Towards Robust Algebraic Multigrid Methods for Nonsymmetric Problems

James William Lottes

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2015

This thesis is dedicated to
my parents Steven and Rosanna
for all of their encouragement and help in pursuing my education,
and to my wife Yin-han,
without whom I would never have embarked on this adventure.

Acknowledgements

I wish to acknowledge Dr. Andy Wathen for his steadfast support and guidance as my supervisor at Oxford University and Dr. Paul Fischer for his support as a student researcher at Argonne National Laboratory. This publication was based on work supported by Award No KUK-C1-013-04, made by King Abdullah University of Science and Technology (KAUST).

Statement of Originality

Chapters 1 and 2 are background material and contain known results. Chapters 3 to 5 are novel. The "form" absolute value investigated in Chapter 3 is new, although it is of course related to many known concepts, as pointed out in that chapter. Chapter 4 develops a new convergence theory for nonsymmetric multigrid methods. Theorem 4.1 is equivalent to previous results, in particular to a result by Notay [39], which generalizes an earlier result for symmetric multigrid by Falgout et al. [19]. However, the equivalence itself is nontrivial (Lemma 4.1). The remaining results and main convergence bounds are new, made possible by the "form" absolute value of Chapter 3. The AMG heuristics developed in Chapter 5 are novel, though inspired by previous methods.

Abstract

When analyzing symmetric problems and the methods for solving them, multigrid and algebraic multigrid in particular, one of the primary tools at the analyst's disposal is the energy norm associated with the problem. The lack of this tool is one of the many reasons analysis of nonsymmetric problems and methods for solving them is substantially more difficult than in the symmetric case. We show that there is an analog to the energy norm for a nonsymmetric matrix A, associated with a new absolute value we term the "form" absolute value. This new absolute value can be described as a symmetric positive definite solution to the matrix equation $A^*|A|^{-1}A = |A|$; it exists and is unique in particular whenever A has positive symmetric part. We then develop a novel convergence theory for a general two-level multigrid iteration for any such A, making use of the form absolute value. In particular, we derive a convergence bound in terms of a smoothing property and separate approximation properties for the interpolation and restriction (a novel feature). Finally, we present new algebraic multigrid heuristics designed specifically targeting this new theory, which we evaluate with numerical tests.

Contents

List of Figures

Chapter 1

Introduction

The solution of large systems of linear equations,

$$Ax = b, \tag{1.1}$$

is a ubiquitous task in scientific computing, forming the core computational kernel of scientific computations arising from a broad range of applications. One remarkable application, and the one for which multigrid (MG) was originally devised, is the numerical solution of some partial differential equation (PDE),

$$Lu = f, \tag{1.2}$$

specified with suitable boundary conditions on some domain Ω. After discretizing the infinite-dimensional problem (1.2) in space and possibly time, and perhaps linearizing, one is led to some finite-dimensional system (1.1) to be solved (at each time step, if time is involved).

Algebraic multigrid (AMG) is a class of iterative methods for solving such linear systems, specifically those that are sparse, as is the case, for example, for typical discretizations of PDEs by grid-based methods. When applicable, AMG is often able to produce a numerical solution using a number of operations scaling linearly or nearly linearly in the number of unknowns. Our specific interest is in AMG applied to "inherently" nonsymmetric problems, for example, discretized convection-diffusion equations with high Péclet number (convection-dominated), as opposed to, say, a nonsymmetric discretization of Poisson's equation. Most of the AMG analysis that has been done, with Notay's work being a notable exception, has been for the symmetric case. Analysis is essential for the development of improved, more robust, and more broadly applicable AMG methods.

We present in Chapter 4 a novel convergence theory of algebraic multigrid (AMG) for nonsymmetric but elliptic operators. This theory makes use of a new norm

associated with such operators, presented in Chapter 3. The background required for these developments is reviewed in Chapter 2, covering Banach spaces, sesquilinear forms, and energy norms. This chapter also establishes the notation used in subsequent chapters. We have chosen to attempt to motivate the material of Chapters 2 to 4, which can be rather abstract, by beginning in this chapter with a focus on specific concrete model problems, which allows many of the abstract notions to come to be introduced first in a concrete setting. This is also background material, and not required for reading later chapters. That is, Chapters 2 to 4 can be read independently of this chapter. After the abstract convergence theory of Chapter 4, Chapter 5 returns to the concrete. Specific heuristics based on the convergence theory of the previous chapter are presented for the various AMG components, and these are tested on the nonsymmetric model problem presented in §1.2.

In this chapter we first look at the simplest of symmetric model problems, Poisson's equation, and briefly cover its discretization by the finite element method (FEM) and its solution by geometric multigrid. The purpose is to introduce in the simplest concrete setting all of the concepts that we will make use of later, including the variatonal setting, linear iterative methods, and multigrid in its general conception. We then look at a nonsymmetric model problem, advection-diffusion, as well as its discretization, which we will use to illustrate and test the abstract theory of later chapters. All of this material is background. Also, for the most part, later chapters can be read independently of the material here.

1.1 A Symmetric Model Problem

This section focuses on Poisson's equation. We begin with a brief presentation of the finite element method (FEM) for this equation, before presenting geometric multigrid applied to the resulting problem. We end with the main ideas of the multigrid convergence theory due to Hackbusch. This is background material, dealing only with a specific, symmetric model problem, and only with geometric multigrid. The intent is to introduce and motivate, in a concrete setting, the ideas we will need in later chapters.

Algebraic multigrid (AMG) methods make as few assumptions about the input matrix as possible, and are certainly not limited to matrices resulting from finite element discretizations. However, one of the key elements of AMG, the Galerkin (or Petrov-Galerkin) coarse operator construction, is directly analogous to the finite element discretization procedure, which comes about from the variatonial setting in

which the FEM is based. The analysis of later chapters will be rooted in just such a variational setting. The presentation of FEM here is meant to provide a concrete introduction to these concepts, not to rule out applicability of the material of later chapters to other discretizations.

1.1.1 Abstract Formulation of Poisson's Equation

Let us consider the problem of finding a solution to Poisson's equation on a bounded open subset Ω of $\mathbb{R}^d$ with Lipschitz continuous boundary $\partial\Omega$:

$$\begin{aligned} -\nabla^2 u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.3}$$

A modern approach is to look for solutions u in a suitable function space. As usual, we will use the Sobolev space $H^1(\Omega)$, the subset of $L^2(\Omega)$ consisting of those functions with weak first derivatives also in $L^2(\Omega)$. The Dirichlet boundary conditions narrow the solution space to $H_0^1(\Omega)$, the subspace of $H^1(\Omega)$ of functions whose trace on $\partial\Omega$ vanishes. The variational form of (1.3) reads: given $f \in H^{-1}(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$\mathfrak{a}[u, v] := \langle \nabla u, \nabla v \rangle = (f, v) \qquad \text{for all } v \in H_0^1(\Omega), \tag{1.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the $L^2(\Omega)$ inner product.

In this and following chapters, the concept of dual space will play a fundamental role. Recall that the continuous dual of a Banach space V, denoted by V^*, consists of all continuous linear functionals on V. Each element $f \in \mathsf{V}^*$ maps each vector $u \in \mathsf{V}$ to the scalar $f[u] =: (f, u)$. The notation (f, u) is called the duality pairing between V^* and V. We will consistently use f and g to denote members of a dual space, and u and v to denote members of the original space. The space V^* is itself a Banach space with norm

$$\|f\| := \sup_{\|u\|=1} |(f, u)|. \tag{1.5}$$

In the case of Sobolev spaces, it is customary to identify $L^2(\Omega)^*$ with $L^2(\Omega)$, so that the duality pairing coincides with the L^2 inner product. Further, we identify $H_0^1(\Omega)^*$ with $H^{-1}(\Omega) = \{f + \sum_{i=1}^d \frac{\partial}{\partial x_i} g_i | f, g_i \in L^2(\Omega)\}$. Then, for example, for $u, v \in H_0^1(\Omega)$, (u, v) has the same interpretation regardless of whether $(\cdot, \cdot)$ is interpreted as the L^2 inner product or the H^{-1}, H_0^1 duality pairing. We have chosen, however, to use angle brackets for inner products and to reserve the notation $(\cdot, \cdot)$ for a duality pairing.

Returning to the variational statement (1.4), the bounded bilinear form $\mathfrak{a} : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ may equivalently be regarded as the bounded linear operator $A : H_0^1(\Omega) \to$

$H_0^1(\Omega)^*$ where $(Au)[v] := \mathfrak{a}[u, v]$ (for details, see §2.1). Under the correspondence, (1.4) becomes

$$(Au, v) = (f, v) \qquad \text{for all } v \in H_0^1(\Omega) \tag{1.6}$$

or simply

$$Au = f. \tag{1.7}$$

The form $\mathfrak{a}$ is coercive as a consequence of the Poincaré-Friedrichs inequality. As it is also bounded, the Lax-Milgram lemma (presented in §2.2) applies, establishing existence and uniqueness of the solution u. Indeed, the lemma asserts more: the solution u depends continuously on f. In this example, $\mathfrak{a}$ is symmetric ($\mathfrak{a}[u, v] = \mathfrak{a}[v, u]$), so that $\mathfrak{a}$ is in fact an inner product. The induced norm is called the energy norm, and we will denote it

$$\|u\|_A := (Au, u)^{1/2}. \tag{1.8}$$

The corresponding dual norm is

$$\|f\|_{A^{-1}} := (f, A^{-1}f)^{1/2} = \sup_{\|u\|_A = 1} |(f, u)|, \tag{1.9}$$

the latter identity being a fact we establish generally below as (2.24).

1.1.2 Discretization by the Finite Element Method (FEM)

The foregoing was only a very brief account, sufficient for our purposes, of the relevant functional analysis and PDE theory. For a detailed treatment of these ideas, we refer the reader to e.g., Evans's monograph on PDE [17]. We will move on to the discretization of (1.4) by the finite element method (FEM). Our presentation is designed to invite the reader to draw an analogy between discretization and multigrid. However, we will be equally brief, as the solution of the discrete system is our main concern, rather than the discretization itself. There are many texts available treating the FEM in detail, e.g. [16,43]. Being self-contained, these texts also serve as additional references for the material of the previous paragraph.

The FEM starts with a triangulation $\mathcal{T}_h$ of Ω into elements, which we shall assume to be d-simplices (e.g., triangles or tetrahedra), each with diameter at most h. Let $\{x_1, \ldots, x_n\}$ be the set of all element vertices not located on the boundary, and $\{x_{n+1}, \ldots, x_{n+n_b}\}$ the set of remaining vertices, all located on the boundary. The nodal basis is defined as the set of functions $\{\phi_1, \ldots, \phi_n\}$ where ϕ_i is the unique continuous function linear on each element satisfying

$$\phi_i(x_j) = \delta_{ij}, \qquad j = 1, \ldots, n + n_b. \tag{1.10}$$

Here δ is the Kronecker delta symbol, equal to 1 when $i = j$ and 0 otherwise. The support of ϕ_i is limited to the set of elements having vertex x_i. We may define a linear operator $P_h : \mathbb{R}^n \to H_0^1(\Omega)$ in matrix form as

$$P_h := \begin{bmatrix} \phi_1 & \cdots & \phi_n \end{bmatrix}. \tag{1.11}$$

P_h maps a vector $\mathbf{u} \in \mathbb{R}^n$ of nodal values to the corresponding piecewise linear interpolant $u = P_h\mathbf{u} \in H_0^1(\Omega)$. Let V_h be the range of P_h, $\mathsf{V}_h := \mathsf{R}(P_h) \subset H_0^1(\Omega)$. Equivalently, V_h is the column space of P_h, i.e., the span of the nodal basis. V_h is the subspace of $H_0^1(\Omega)$ of piecewise linear functions corresponding to the triangulation $\mathcal{T}_h$. Discretization by the Galerkin method proceeds simply by restricting the test and trial space to V_h. The discrete problem reads: find $u_h \in \mathsf{V}_h$ such that

$$(Au_h, v_h) = (f, v_h) \qquad \text{for all } v_h \in \mathsf{V}_h. \tag{1.12}$$

Setting $u_h = P_h\mathbf{u}, v_h = P_h\mathbf{v}$, we have the equivalent problem: find $\mathbf{u} \in \mathbb{R}^n$ such that

$$(AP_h\mathbf{u}, P_h\mathbf{v}) = (f, P_h\mathbf{v}) \qquad \text{for all } \mathbf{v} \in \mathbb{R}^n. \tag{1.13}$$

We may bring P_h across the duality pairing: $(P_h^*AP_h\mathbf{u}, \mathbf{v}) = (P_h^*f, \mathbf{v})$, thus reducing the equation to the matrix system

$$A_h\mathbf{u} = \mathbf{f}_h \tag{1.14}$$

where $A_h := P_h^*AP_h$ and $\mathbf{f}_h := P_h^*f$. Note the entries of the matrix $A_h \in \mathbb{R}^{n\times n}$ and vector $\mathbf{f}_h \in \mathbb{R}^n$ are simply

$$(A_h)_{ij} = \mathfrak{a}[\phi_j, \phi_i], \qquad (\mathbf{f}_h)_i = f[\phi_i]. \tag{1.15}$$

In the case that f is a member of V_h, then $f = P_h\mathbf{v}$ for some $\mathbf{v}$ and $\mathbf{f}_h = Q_h\mathbf{v}$ where $Q_h := P_h^*P_h$ is called the mass matrix. It is a Gram matrix, with entries $(Q_h)_{ij} = \langle\phi_j, \phi_i\rangle$. Note that solution vectors and source vectors have different discrete representations. The symmetry of the matrix A_h is obviously a consequence of the symmetry of $\mathfrak{a}$. Since $\mathfrak{a}$ is coercive, A_h must in fact be symmetric positive definite (SPD). We can also see that A_h is *sparse*: A_{ij} is non-zero only when some element has vertices x_i and x_j. Hence only a few entries in each row of A_h will be nonzero. The mass matrix is similarly SPD and has the same sparsity structure as A_h.

So far we have seen a description of the Galerkin method, without an indication of how good the resulting approximation is. In fact, the Galerkin approximation is optimal in the energy norm. To see this, let $A_h\mathbf{u}_h = \mathbf{f}_h$ so that $P_h\mathbf{u}_h$ is the Galerkin

approximation to the exact solution. By expressing an arbitrary function $u \in \mathsf{V}_h$ as $P_h(\mathbf{u}_h + \delta\mathbf{u})$, we find that the square of the energy norm of the corresponding error is

$$\begin{aligned}
\|A^{-1}f - P_h(\mathbf{u}_h + \delta\mathbf{u})\|_A^2 &= \|f - AP_h(\mathbf{u}_h + \delta\mathbf{u})\|_{A^{-1}}^2 \\
&= \|f - AP_h\mathbf{u}_h\|_{A^{-1}}^2 + 2\,(f - AP_h\mathbf{u}_h, P_h\delta\mathbf{u}) + \|P_h\delta\mathbf{u}\|_A^2 \\
&= \|f - AP_h\mathbf{u}_h\|_{A^{-1}}^2 + 2\,(\mathbf{f}_h - A_h\mathbf{u}_h, \delta\mathbf{u}) + \|P_h\delta\mathbf{u}\|_A^2 \\
&= \|f - AP_h\mathbf{u}_h\|_{A^{-1}}^2 + \|P_h\delta\mathbf{u}\|_A^2,
\end{aligned} \tag{1.16}$$

which is minimized when $\delta\mathbf{u} = 0$, i.e., for the Galerkin approximation. That is, the energy norm of the error of the Galerkin approximation is smaller than that of any other candidate in the solution space V_h.

Now we wish to view the parameter h as varying, so that we may describe what happens as the triangulation $\mathcal{T}_h$ is refined. That is, we are interested in asymptotic behavior in the limit $h \to 0$. This requires additional assumptions. We will assume that $\mathcal{T}_h$ is a *regular* family of triangulations [43, p. 90], meaning that none of the elements may tend toward degeneracy as $h \to 0$. For example, when $d = 2$ the interior angles of all triangles must be bounded away from 0 and π. We will also assume, for simplicity, that the family $\mathcal{T}_h$ is *quasi-uniform* [43, p. 193] [16, p. 57], entailing that the smallest element diameter be within a constant factor of the largest, h.

A central concern for the FEM is error analysis, characterizing the dependence of norms of the error on h. For example, under certain conditions on Ω and f, there exists some C independent of h such that

$$\|u - P_h\mathbf{u}_h\|_A \leq Ch\|f\|_{L^2(\Omega)}, \tag{1.17}$$

where u is the exact solution and $\mathbf{u}_h$ is the discrete solution satisfying $A_h\mathbf{u}_h = \mathbf{f}_h$ (cf. [16, p. 41] [25, p. 139]). Under further conditions, one may show [43, p. 173]

$$\|u - P_h\mathbf{u}_h\|_{L^2(\Omega)} = O(h^2) \qquad \text{as} \quad h \to 0. \tag{1.18}$$

That is, under certain conditions, the method is second-order. Error analysis, however, is not our main concern. The relevance of the limit $h \to 0$ to our purposes is that it corresponds to increasing the size of the matrix A_h, $n = O(h^{-d})$. Approximately solving $A_h\mathbf{u}_h = \mathbf{f}_h$ with work scaling only linearly in n is the ideal—the goal in designing any fast solver such as multigrid.

1.1.3 Matrix Properties

Here we highlight some key properties of the matrices A_h and Q_h. First let us introduce the norm associated with any SPD matrix X,

$$\|\mathbf{u}\|_X := (\mathbf{u}^T X \mathbf{u})^{1/2}. \tag{1.19}$$

Regarding a vector $\mathbf{f}$ as a linear functional via $\mathbf{u} \mapsto \mathbf{f}^T\mathbf{u}$, the corresponding dual norm is $\|\cdot\|_{X^{-1}}$, see (2.24).

The Q_h-norm and A_h-norm are the discrete L^2 and energy norms,

$$\|\mathbf{u}\|_{Q_h} = \|P_h\mathbf{u}\|_{L^2(\Omega)}, \qquad \|\mathbf{u}\|_{A_h} = \|P_h\mathbf{u}\|_A. \tag{1.20}$$

The dual norms enjoy a similar correspondence,

$$\|\mathbf{f}_h\|_{Q_h^{-1}} = \|f\|_{L^2(\Omega)} \quad \text{for } f \in \mathsf{V}_h, \tag{1.21}$$

$$\|\mathbf{f}_h\|_{A_h^{-1}} = \|f\|_{A^{-1}} \quad \text{for } f \in A\mathsf{V}_h, \tag{1.22}$$

where, recall, $\mathbf{f}_h := P_h^* f$. Note that it is important to distinguish between the two types of discrete vectors here, as the discrete L^2-norm for each is different.

While the mass matrix Q_h is not diagonal, it is well approximated by its diagonal part, which we shall call $\tilde{Q}_h$,

$$(\tilde{Q}_h)_{ij} := \delta_{ij}(Q_h)_{ii} = \delta_{ij}\|\phi_i\|^2_{L^2(\Omega)}. \tag{1.23}$$

Using an element by element analysis, one may show [16, p. 295] that

$$\frac{1}{2}\|\mathbf{u}\|^2_{\tilde{Q}_h} \le \|\mathbf{u}\|^2_{Q_h} = \|P_h\mathbf{u}\|^2_{L^2(\Omega)} \le \frac{d+2}{2}\|\mathbf{u}\|^2_{\tilde{Q}_h} \tag{1.24}$$

for any $\mathbf{u}$. Note these equivalance constants are quite mild, do not depend in any way on the geometry of the triangulation $\mathcal{T}_h$, and in particular, are independent of h. Because we have made the quasi-uniform assumption, the discrete L^2 norm is also equivalent to the Euclidean norm $\|\mathbf{u}\|_2 := (\mathbf{u}^T\mathbf{u})^{1/2}$ [16, p. 58] [43, p. 194]: there exist constants c and C independent of h such that

$$ch^d\|\mathbf{u}\|_2^2 \le \|\mathbf{u}\|^2_{Q_h} \le Ch^d\|\mathbf{u}\|_2^2. \tag{1.25}$$

In contrast to the case with $\tilde{Q}_h$, this equivalance exhibits a uniform $h^{d/2}$ scaling factor, as well as constants that depend strongly on the geometric properties of the family of triangulations $\mathcal{T}_h$. It is quite easy to see where the h^d scaling comes from. Consider the constant function 1, whose discretization (ignoring the Dirichlet boundary conditions

for the moment) is the vector of all ones, $\mathbf{1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$. We have $\|1\|^2_{L^2(\Omega)} = |\Omega|$, while $\|\mathbf{1}\|_2^2 = n = O(h^{-d})$.

The inner products and induced norms associated with diagonal SPD matrices are convenient for computational purposes. The 2-norm (associated with the identity matrix) is but one member of this family. We have identified another member, $\tilde{Q}_h$, that enjoys a much more natural correspondence with the underlying PDE problem. We could say that the 2-norm on $\mathbb{R}^n$ corresponds roughly to an L^2 norm on V_h with a weight function depending strongly on the triangulation $\mathcal{T}_h$. This weight and norm, however, are an artifact of the discretization. The 2-norm is certainly useful—it has been put to good use in analyzing the FEM, and in analyzing iterative solution methods and multigrid in particular. Algebraic multigrid, more so than other multigrid methods, is particularly concerned with the actual numerical values of the constants involved in the asymptotic limit $h \to 0$, if the limit is considered at all. From this standpoint, the equivalence between the Q_h and 2-norms is quite unattractive. Indeed, one motivating application for algebraic multigrid is a triangulation with many nearly degenerate elements, in which case the ratio C/c would be quite large. For these reasons, we will almost entirely avoid the 2-norm in later sections.

One may show [16, p. 59] [43, p. 195] that the discrete energy and L^2 norms enjoy the equivalence

$$c\|\mathbf{u}\|^2_{Q_h} \leq \|\mathbf{u}\|^2_{A_h} \leq Ch^{-2}\|\mathbf{u}\|^2_{Q_h} \tag{1.26}$$

for some constants c, C independent of h. The first inequality is an instance of the Poincaré-Friedrichs inequality, applied to $P_h\mathbf{u}$, while the second inequality, called an inverse inequality, is a property of the discretization. Note that the upper bound becomes infinite as $h \to 0$, which corresponds to the fact that the differential operator is unbounded when considered as an operator on $L^2(\Omega)$. Let us also remark that (1.26) is associated with the eigenproblem $A_h\mathbf{u} = \lambda Q_h\mathbf{u}$, the discrete approximation of $-\nabla^2 u = \lambda u$. In particular, the Rayleigh quotient $\|\mathbf{u}\|^2_{A_h}/\|\mathbf{u}\|^2_{Q_h}$ is large for highly oscillatory $\mathbf{u}$.

Combining (1.26) and (1.24) we see that A_h and the diagonal $\tilde{Q}_h$ enjoy the similar equivalence

$$c\|\mathbf{u}\|^2_{\tilde{Q}_h} \leq \|\mathbf{u}\|^2_{A_h} \leq Ch^{-2}\|\mathbf{u}\|^2_{\tilde{Q}_h} \tag{1.27}$$

for different constants. On the other hand, using (1.25), we find the equivalence between the A_h and 2-norms,

$$ch^d\|\mathbf{u}\|_2^2 \leq \|\mathbf{u}\|^2_{A_h} \leq Ch^{d-2}\|\mathbf{u}\|_2^2, \tag{1.28}$$

again for different constants.

The Jacobi iteration, to be considered below, amounts to approximating A_h by its diagonal part, which we shall call D_h,

$$(D_h)_{ij} := \delta_{ij}(A_h)_{ii} = \delta_{ij}\|\phi_i\|_A^2. \tag{1.29}$$

An element by element analysis, similar to that used to establish (1.24), leads to the estimates

$$c_D h^{-2}\|\mathbf{u}\|_{Q_h}^2 \le \|\mathbf{u}\|_{D_h}^2 \le C_D h^{-2}\|\mathbf{u}\|_{Q_h}^2 \tag{1.30}$$

and

$$\frac{c}{C_D}h^2\|\mathbf{u}\|_{D_h}^2 \le \|\mathbf{u}\|_{A_h}^2 \le (d+1)\|\mathbf{u}\|_{D_h}^2. \tag{1.31}$$

The first estimate depends on both the regular and quasi-uniform assumptions, and the constants depend strongly on the geometry of the triangulation. The lower bound of the second estimate follows from combining the Poincaré-Friedrichs inequality $c\|\mathbf{u}\|_{Q_h}^2 \le \|\mathbf{u}\|_{A_h}^2$ from (1.26) with the upper bound of the first estimate. Comparing (1.30) with (1.26), we see that D_h is a poor approximation to A_h for the lower part of the spectrum.

We have considered the equivalence of A_h with the three diagonal matrices $\tilde{Q}_h$, I, and D_h, as well as with the mass matrix Q_h. Note that in each case, the condition number, given by the ratio of the upper to the lower equivalence constants, is proportional to h^{-2}. This generic feature of ill-conditioning of discretized differential operators is, ultimately, the major source of the difficulty in solving the systems they govern.

1.1.4 Iterative Methods

No direct method (one that results in the exact solution in the absence of round-off error) is capable of solving $A_h\mathbf{u} = \mathbf{f}_h$ using a number of operations scaling linearly in the size of the system. The alternative is to use iterative methods, which generate a sequence of iterates $\mathbf{u}_n$ that ideally converge to $A_h^{-1}\mathbf{f}_h$. The error $\mathbf{e}_n := A_h^{-1}\mathbf{f}_h - \mathbf{u}_n$ obeys the "error equation"

$$A_h\mathbf{e}_n = \mathbf{r}_n, \qquad \mathbf{r}_n := \mathbf{f}_h - A_h\mathbf{u}_n. \tag{1.32}$$

The vector $\mathbf{r}_n$ is called the residual. The idea of iterative methods is to approximate $\mathbf{e}_n$ by $B\mathbf{r}_n$ where B is some inexpensive approximation of A_h^{-1}. This leads to the basic form of a linear iteration,

$$\mathbf{u}_{n+1} = \mathbf{u}_n + B\mathbf{r}_n. \tag{1.33}$$

For any B of full rank, the exact solution is the unique fixed point. Because of the sparsity of A_h, evaluating the residual vector requires computational work linear in n. If the application of B also requires linear work, then the overall iterative method will be linear, provided that the number of iterations required for convergence is bounded independent of h. The error evolves according to

$$\mathbf{e}_n = E\mathbf{e}_{n-1} = E^n\mathbf{e}_0, \tag{1.34}$$

where

$$E = I - BA_h \tag{1.35}$$

is called the iteration matrix. The asymptotic convergence rate of the iteration is equal to the spectral radius of E, $\rho(E)$. If B, like A_h, is SPD, then

$$\|E^n\|_{A_h} = \|E\|_{A_h}^n = \rho(E)^n. \tag{1.36}$$

An example of a simple iteration with linear work per iteration is the damped Jacobi iteration, corresponding to the choice $B = \omega D_h^{-1}$, where ω is the damping parameter. The iteration matrix is

$$E_{\text{Jac}} = I - \omega D_h^{-1} A_h. \tag{1.37}$$

From (1.31), we see that the spectrum of $D_h^{-1}A_h$ is bounded by

$$ch^2 \leq \sigma(D_h^{-1}A_h) \leq d + 1 \tag{1.38}$$

with c here corresponding to c/C_D from (1.31). The upper bound of $d+1$ may be pessimistic, but the maximum eigenvalue is at least 1, which implies that the iteration will always diverge whenever $\omega > 2$. In particular, we must have $\omega = O(h^0)$. So long as ω is not too large, the convergence rate of damped Jacobi is bounded by

$$\rho(E_{\text{Jac}}) = \|E_{\text{Jac}}\|_{A_h} \leq 1 - \omega ch^2 = 1 - O(h^2). \tag{1.39}$$

Since $\|E_{\text{Jac}}^k\|_{A_h} \sim 1 - k\omega ch^2$, it follows that $k = O(h^{-2})$ iterations are required to achieve a reduction in the error by a fixed factor. The performance of damped Jacobi deteriorates rapidly as the discretization is refined.

Other standard iterations include Gauss-Seidel and Successive Over Relaxation (SOR). For descriptions of these methods, we refer the reader to standard texts on the subject, e.g., Golub and Van Loan [23, §10.1] or Saad [47]. Like damped Jacobi, they require linear work per iteration and, for our model problem, $O(h^{-2})$, or in the best case, $O(h^{-1})$, iterations to reduce the error by a fixed factor.

1.1.5 Multigrid

The slow convergence of damped Jacobi is directly attributable to the eigenmodes of $D_h^{-1}A_h$ corresponding to the smallest eigenvalues. Because D_h and Q_h are equivalent up to a uniform scaling factor of h^{-2}, these eigenmodes are qualitatively similar to those of $A_h\mathbf{u} = \lambda Q_h\mathbf{u}$, the discrete approximation to $-\nabla^2 u = \lambda u$ with $u = 0$ on $\partial\Omega$. Specifically, the slow-to-converge error components are the smooth, low-frequency ones. Conversely, the damping parameter can be chosen to ensure that all of the highly oscillatory components—the upper half of the spectrum—are reduced by half or more on each iteration. The result is that a few iterations of damped Jacobi, or other similar iterations, results in the error being smooth.

In the early 1960's, Fedorenko [20–22] noticed this behavior and realized that the smooth error would be well represented on a grid of half the resolution. This realization led Fedorenko to implement the first multigrid method, the basic idea being to combine a basic relaxation method such as damped Jacobi, termed the "smoother," with the approximate solution of the error equation on a coarser grid, termed the "coarse-grid correction." The coarse problem involves only a fraction of the number of variables as the original, and is thus already substantially easier to solve than the original. The name multigrid comes from the strategy of handling the coarse problem again by a combination of smoother and correction on an even coarser grid, and so on recursively, leading to a sequence of coarser and coarser grids. Once the problem size has been reduced to a small enough size, sometimes even a single unknown, it can be solved directly.

Here we shall present the basic multigrid iteration for our model problem, followed by the basic ideas of the convergence theory due to Hackbusch [25]. The two components of multigrid, the smoother and the coarse grid correction, are complementary. Neither is effective on its own, but the combination produces a very efficient method. Hackbusch's theory reflects this structure, quantifying the complementary effects of each component, and combining them to yield an overall convergence bound independent of h. While the theory doesn't directly apply to algebraic multigrid, the basic structure and ideas are quite similar.

In the setting of the finite element method, there is a rather natural construction of the coarse grid problem. Let the coarse triangulation be $\mathcal{T}_H$, with $H > h$, and for convenience, assume the fine (original) triangulation $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$ so that $\mathsf{V}_H \subset \mathsf{V}_h$. Then there exists a matrix $P \in \mathbb{R}^{n\times n_c}$ such that

$$P_H = P_h P. \tag{1.40}$$

Here $n_c < n$ is the number of internal vertices of $\mathcal{T}_H$; on a typical sequence of grids $n_c/n \approx 1/2^d$. The function of the prolongation matrix P is to "prolongate" or interpolate a coarse grid vector to a fine grid vector. By definition, the matrices and load vectors at the two levels are related by

$$\begin{aligned} A_H &:= P_H^* A P_H = P^* A_h P, \\ \mathbf{f}_H &:= P_H^* f = P^* \mathbf{f}_h. \end{aligned} \tag{1.41}$$

The discrete solution on the coarse grid, $\mathbf{u}_H = A_H^{-1}\mathbf{f}_H$, corresponds to the function $P_H\mathbf{u}_H = P_h(P\mathbf{u}_H)$. We conclude that the coarse grid approximation to $\mathbf{u}_h$ is

$$\mathbf{u}_h \approx P\mathbf{u}_H = PA_H^{-1}P^*\mathbf{f}_h, \tag{1.42}$$

analogous to the original FEM approximation $u \approx P_h\mathbf{u}_h = P_h A_h^{-1} P_h^* f$. Applying this approximation to the error equation gives the coarse grid correction $\mathbf{e} \approx PA_H^{-1}P^*\mathbf{r}$.

Preceding the coarse grid correction step by k "smoothing" steps of some iterative method like damped Jacobi results in the following concrete "two-grid" iteration,

$$\begin{aligned} \mathbf{v}_{n,0} &= \mathbf{u}_n, \qquad \mathbf{r}_{n,i} := \mathbf{f}_n - A_h\mathbf{v}_{n,i}, \\ \mathbf{v}_{n,i+1} &= \mathbf{v}_{n,i} + B\mathbf{r}_{n,i}, \qquad i = 0, \ldots, k-1, \\ \mathbf{u}_{n+1} &= \mathbf{v}_{n,k} + PA_H^{-1}P^*\mathbf{r}_{n,k}, \end{aligned} \tag{1.43}$$

which is succinctly described by the corresponding iteration matrix,

$$E_{\text{tg}} = (I - PA_H^{-1}P^*A_h)(I - BA_h)^k. \tag{1.44}$$

The symmetric iteration

$$(I - BA_h)^m(I - PA_H^{-1}P^*A_h)(I - BA_h)^m \tag{1.45}$$

is often preferred, but these matrices have identical spectra when $k = 2m$ (see Lemma 4.1 below), and thus it suffices to consider the former for the purposes of analysis. The final element of multigrid is to replace the exact solve A_H^{-1} by another two-grid iteration involving $\mathcal{T}_H$ and an even coarser triangulation. This is repeated until the recursion bottoms out, for example, at a triangulation containing a single node not on the boundary. As Hackbusch [25] notes, the conditions for two-grid convergence are almost sufficient to conclude convergence of the full multigrid iteration. For this reason, we will confine our analysis to the two-grid case.

1.1.5.1 Approximation Property

Here we present the main ideas of the classical multigrid analysis due to Hackbusch [25], confined to our simple example. To begin, we quantify the size of the error of the coarse grid approximation. The easiest way to do this is to appeal to the theory of errors for the underlying FEM. Let us rewrite the error estimate (1.17) as

$$\|(A^{-1} - P_h A_h^{-1} P_h^*) f\|_A \leq Ch\|f\|_{L^2(\Omega)}, \tag{1.46}$$

which holds under certain assumptions on f and Ω. Note that the same estimate holds, under the same assumptions, when h is replaced with H. Combining these using the triangle inequality we find

$$\begin{aligned} &\|(P_h A_h^{-1} P_h^* - P_H A_H^{-1} P_H^*) f\|_A = \\ &\quad \|P_h(A_h^{-1} - P A_H^{-1} P^*)\mathbf{f}_h\|_A \leq C(H+h)\|f\|_{L^2(\Omega)}. \end{aligned} \tag{1.47}$$

We now restrict f to the space V_h (which still leaves $\mathbf{f}_h := P_h^* f$ arbitrary), and use the correspondences between the continuous and discrete norms, (1.20) and (1.21), to obtain

$$\|(A_h^{-1} - P A_H^{-1} P^*)\mathbf{f}_h\|_{A_h} \leq C(\tfrac{H}{h} + 1)h\|\mathbf{f}_h\|_{Q_h^{-1}}. \tag{1.48}$$

This estimate is an instance of what Hackbusch calls the "approximation property."

It is significant that the norm used on the right hand side of (1.48) is the discrete L^2 norm and not the dual of the energy norm. Consider the relationship between these norms, given by

$$c\|\mathbf{f}\|^2_{A_h^{-1}} \leq \|\mathbf{f}\|^2_{Q_h^{-1}} \leq Ch^{-2}\|\mathbf{f}\|^2_{A_h^{-1}}, \tag{1.49}$$

the dual of (1.26). The Rayleigh quotient $\|\mathbf{f}\|^2_{Q_h^{-1}}/\|\mathbf{f}\|^2_{A_h^{-1}}$ is associated with the eigenproblem $Q_h^{-1}\mathbf{f} = \lambda A_h^{-1}\mathbf{f}$, equivalent via $\mathbf{f} = A_h\mathbf{u}$ to $A_h\mathbf{u} = \lambda Q_h\mathbf{u}$, the discrete approximation to $-\nabla^2 u = \lambda u$ with $u = 0$ on $\partial\Omega$. Relative to $\|\mathbf{f}\|_{A_h^{-1}}$, $\|\mathbf{f}\|_{Q_h^{-1}}$ is $O(h^{-1})$ for highly oscillatory $\mathbf{u}$ but only $O(h^0)$ for smooth $\mathbf{u}$. Thus the approximation property (1.48) reflects the fact that the coarse grid correction is more effective for smooth error than for oscillatory error.

1.1.5.2 Smoothing Property

Consider an eigenmode $\mathbf{e}$ of BA_h with corresponding eigenvalue λ. Assume that B is SPD and scaled such that all such eigenvalues lie in the interval $[0, 4/3]$. The vector $\mathbf{e}$

is an eigenmode of the iteration matrix $E = I - BA_h$ with eigenvalue $1 - \lambda$, and is slow to converge if λ is close to 0. Note that

$$\frac{\|A_h \mathbf{e}\|_B^2}{\|\mathbf{e}\|_{A_h}^2} = \lambda. \tag{1.50}$$

So, the B-norm of the residual is small, relative to the energy norm of the error, for the slow-to-converge modes. Now consider the B-norm of the residual after k iterations.

$$\begin{aligned} \sup_{\mathbf{e}\neq\mathbf{0}} \frac{\|A_h(I - BA_h)^k \mathbf{e}\|_B^2}{\|\mathbf{e}\|_{A_h}^2} &= \rho[A_h^{-1}(1 - A_h B)^k A_h B A_h (I - BA_h)^k] \\ &= \rho[BA_h(I - BA_h)^{2k}] \\ &= \sup_{\lambda \in \sigma(BA_h)} \lambda(1 - \lambda)^{2k}. \end{aligned} \tag{1.51}$$

Assuming $k \geq 1$,

$$\sup_{0 \leq \lambda \leq 4/3} \lambda(1 - \lambda)^{2k} = \frac{(2k)^{2k}}{(2k+1)^{2k+1}} \leq \frac{1}{2ek + 1.3}, \tag{1.52}$$

and thus we obtain the estimate

$$\|A_h(I - BA_h)^k \mathbf{e}\|_B \leq (2ek + 1.3)^{-1/2} \|\mathbf{e}\|_{A_h} \tag{1.53}$$

for all $\mathbf{e}$. Note the factor on the right is independent of h and becomes arbitrarily small as k is increased. This is possible because those modes hardly touched by k iterations have correspondingly small B-norm in the residual.

The above result is generic and does not, by itself, indicate whether B is an effective smoother. For that, the coarse grid correction must be effective on modes in the residual with relatively small B-norm, i.e., those not reduced effectively by the smoother. From the approximation property (1.48), we know that the coarse grid correction is effective when the residual has small L^2 norm. For damped Jacobi, $B = \omega D_h^{-1}$, these norms are equivalent up to a uniform scaling in h. We have

$$c_D h^{-2} \|\mathbf{f}\|_{D_h^{-1}}^2 \leq \|\mathbf{f}\|_{Q_h^{-1}}^2 \leq C_D h^{-2} \|\mathbf{f}\|_{D_h^{-1}}^2, \tag{1.54}$$

the dual of (1.30). In particular, $\|\mathbf{f}\|_{Q_h^{-1}} \leq (C_D/\omega)^{1/2} h^{-1} \|\mathbf{f}\|_{\omega D_h^{-1}}$, and thus

$$\|A_h(I - BA_h)^k \mathbf{e}\|_{Q_h^{-1}} \leq \left(\frac{C_D/\omega}{2ek + 1.3} \right)^{1/2} h^{-1} \|\mathbf{e}\|_{A_h}. \tag{1.55}$$

This estimate is an instance of what Hackbusch calls the "smoothing property."

The two-grid iteration matrix factors as

$$E_{\text{tg}} = (A_h^{-1} - PA_H^{-1}P^*)A_h(I - BA_h)^k, \tag{1.56}$$

and so we may combine the approximation property (1.48) and the smoothing property (1.55) to produce the estimate

$$\|E_{\text{tg}}\|_{A_h} \leq C(C_c + 1)\left(\frac{C_D/\omega}{2ek + 1.3}\right)^{1/2} \tag{1.57}$$

where C_c is a bound on the coarsening ratio H/h (the assumption $H = 2h$ is common). The above bound can be made arbitrarily small by increasing the number of smoothing steps k. In particular, there is some minimal k for which the bound is smaller than 1, making (1.57) in fact a convergence bound independent of h. Provided sufficient smoothing steps are taken, the number of two-grid iterations required to reduce the error norm by a fixed amount is thus bounded independent of h.

The bound (1.57) is likely pessimistic, but nonetheless indicates how various factors affect multigrid performance. Performance degrades if the coarsening is more aggressive and the ratio H/h increases, or if the smoother is made less effective by lowering k or ω. Note that k can be increased arbitrarily, at the cost of more work per two-grid iteration, but that the damping parameter ω is limited by the assumption that all eigenvalues of $BA_h = \omega D_h^{-1}A_h$ are less than 4/3. The constant C_D depends on the geometry of the triangulation, and can be quite large, for example, when very thin triangles are present.

For more details on the above theory, we refer the reader to Hackbusch [25], who derives smoothing properties for additional smoothers, including Gauss-Seidel, demonstrates how the assumptions underlying the FEM error analysis used to derive the approximation property may be relaxed, and also derives approximation properties for finite-difference schemes. There were, of course, many others in addition to Hackbusch who worked on the convergence theory of multigrid. Mandel, McCormick, and Ruge [35] showed how to remove the need for sufficiently many smoothing steps and were able to prove convergence for a fixed (small) number of steps. Bramble, Pasciak, Wang, and Xu [3] showed how to remove the regularity assumptions (implicit in the above development in appealing to the underlying theory of errors to derive the approximation property).

1.1.6 Algebraic Multigrid (AMG)

In the presentation of multigrid in the previous section, the sequence of coarse grids were assumed given. This is the distinguishing trait of "geometric" multigrid as opposed to "algebraic" multigrid (AMG), which was introduced by Brandt, McCormick, and Ruge in the early 1980s [4, 6, 7, 46, 48]. In AMG, the coarse "grids" are constructed by some algorithm using heuristics referencing only the available algebraic information, the given matrix A. The coarse "grid" in classical AMG simply refers to a subset of the original unknowns, called "C-variables." The remaining unknowns are termed "F-variables." Note that the unknowns of the coarse problem in the previous section were associated with the vertices of the coarse triangulation, which indeed were a subset of the vertices of the fine triangulation. In the AMG framework, the underlying geometry is forgotten.

The prolongation matrix P, responsible for interpolating the F-variables from the "neighboring" C-variables, must be constructed without reference to the geometry, but instead using the matrix A. In the symmetric case it is natural to take the "restriction" matrix R to be $R = P^*$. Letting B_1 and B_2 be the pre- and post-smoother operators, we then have the two-grid iteration

$$E_{\text{tg}} = (I - B_2A)(I - PA_c^{-1}RA)(I - B_1A), \tag{1.58}$$

where now the coarse matrix A_c is *defined* to be

$$A_c := RAP. \tag{1.59}$$

With the choice $R = P^*$, these definitions are equivalent to those of the previous section. The only difference is that here P is constructed by some procedure, whereas in the previous section P was determined by the given two triangulations. Just as in the previous section, the two-grid iteration becomes a multigrid iteration if the coarse solve A_c^{-1} is approximated by recursively applying the same procedure.

A slogan sharply capturing the difference between geometric and algebraic multigrid, expressed, e.g., in [46], is that geometric approaches employ fixed grids and seek complementary smoothers, while algebraic approaches employ fixed smoothers and seek complementary coarse grids. Actually, both the smoother and the coarse grids are free choices for AMG methods, which was acknowledged from the beginning [6], but the emphasis of most methods has been on the coarse grid. In any case, the freedom to select the coarse grid and adapt it to complement the smoother is a distinguishing feature and major advantage that AMG has over geometric multigrid. Another is that

an AMG method may potentially function as a black box solver, as the procedures for constructing the grids and interpolation operators are automatic, and take as input only the matrix A. A single AMG method, which makes only general assumptions on the input matrix, may work for a broad range of applications, and is not limited to a particular discretization of a particular PDE. Indeed, AMG has been successfully employed in many contexts besides the numerical solution of PDEs. Notice that, for brevity of presentation, we've restricted our attention thus far to the FEM and to the plain Poisson equation. Broadening the analysis of the previous sections to include finite-difference discretizations and non-constant coefficient PDEs would be a substantial task. In contrast, analyses of AMG methods, including our own in Chapter 4, do not reference the PDE or the discretization at all.

There are, of course, also disadvantages to AMG. When a geometric multigrid method works well for a typical application, it is usually more efficient than a more general AMG method. One of the reasons for this is that the coarser operators in AMG tend to lose sparsity, a phenomenon known as "stencil growth."

Early analyses of AMG are given by Brandt [4] and Ruge and Stüben [46]. These focus on the case of symmetric positive definite A, and more specifically on M-matrices—SPD matrices with strictly nonpositive off-diagonal entries. The focal point of the analysis is to define the error to be smooth if it is slow to converge with respect to the smoother, regardless of whether this is correlated with some geometrical smoothness or not. The problem of finding a coarse grid correction that complements the smoother is then the same as finding a matrix P whose columns approximately span the space of smooth vectors as so defined. Quantitatively, the authors say that an error vector $\mathbf{e}$ is "algebraically smooth" if

$$\|A\mathbf{e}\|_{D^{-1}} \ll \|\mathbf{e}\|_A \tag{1.60}$$

where D is the diagonal part of A. Indeed, as we established in §1.1.5.2, the ratio $\|A\mathbf{e}\|_{D^{-1}}^2/\|\mathbf{e}\|_A^2$ is the Rayleigh quotient precisely related to the speed of convergence of the Jacobi iteration, and is small for those error components slow to converge. It is then noticed that smooth error "varies slowly in the direction of strong connections" [46]. That is, for smooth error $\mathbf{e}$, $|e_i - e_j|$ is small when $|a_{ij}|/a_{ii}$ is large. This basic intuition forms the basis of the classic Ruge-Stüben coarsening procedure. "Connections" between variables are classified as being weak or strong, according to some given threshold parameter, and the variables with the most strong connections are made C-variables until all remaining F-variables are strongly connected to one or more C-variables.

Since the introduction of AMG in the 1980's, there has been a great deal of research to refine and improve its various aspects and make it suitable for more applications. For example, a recent effort to refine the "strength of connection" heuristic is the notion of energy-based strength of connection [8]. An entirely alternative approach to coarsening is "compatible relaxation," due to Brandt [5], in which a given coarsening is considered good when relaxation restricted to the F-variables is fast to converge.

Another interesting approach is element-based AMG [30], in which the structure of the matrix as a sum of many individual element stiffness matrices is maintained, and coarsening is done by agglomerating elements. This approach, when applicable, enjoys many theoretical advantages, including robust interpolation and the lack of stencil growth. Unfortunately, there is no "strength of connection" analogue for element-based AMG, and no good simple and robust coarsening procedure is known.

Wan, Chan, and Smith [52] proposed solving a global optimization problem to find energy-minimizing interpolation weights. Similarly, "smoothed aggregation" methods, introduced earlier by Vaněk [50], reduce the energy norms of the coarse basis vectors (the columns of P) by relaxing them with an iteration like Jacobi.

Falgout, Vassilevski, and Zikatanov [18, 19] developed a two-level/two-grid convergence theory for AMG featuring an approximation property that involves the smoother. This is very much in keeping with the spirit of AMG, that the coarse grids be adapted to complement the smoother. However, not having a separate smoothing property complicates matters, and in subsequent papers where the framework is used, a simple diagonal smoother (Jacobi) is often assumed, with the suggestion that it should be possible to extend the ideas to more general smoothers. The book by Vassilevski [51] is also an excellent reference for this theory, containing also some extensions to multigrid/multilevel methods. One key tool in the multilevel analysis is an identity due to Xu and Zikatanov [53] that replaced earlier bounds.

Notay [39] has recently made progress in extending the convergence theory to the nonsymmetric case. Theorem 4.1 of the convergence theory presented in Chapter 4, describing the spectrum of the two-grid iteration operator is in agreement with the main theorem of Notay's paper. An application of some of these ideas to convection-diffusion in particular, using piecewise-constant interpolation and $R = P^*$, can be found in [40]. This is similar to earlier work by Kim, Xu, and Zikatanov [32] on AMG for convection-diffusion, with interpolation based on graph-matching, effectively also giving piece-wise constant interpolation.

1.2 A Nonsymmetric Model Problem

The example PDE we choose for a concrete presentation of a nonsymmetric problem is the advection-diffusion equation $-p\nabla^2 u + \mathbf{c}\cdot\nabla u = f_S$ governing the diffusive transport of some quantity u that is being advected by the flow field $\mathbf{c}$ with source term f_S. We assume the reader has some familiarity with such equations. When advection is dominant, as occurs frequently in practice, the equation is highly nonsymmetric, making it a suitable model equation. We will be using this model problem and its discretization for illustration purposes throughout.

1.2.1 Model Problem with Adjoint

Our presentation will closely follow that of Quarteroni and Valli [43, §6.1]. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega$, and let $\partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$ be a partition of the boundary into two disjoint open sets, each Lipschitz continuous, and with Γ_D nonempty. Consider the pair of adjoint boundary value problems (BVPs),

$$\boxed{\begin{aligned} \mathcal{L}u &= f_S \quad \text{on } \Omega, \\ u &= f_D \quad \text{on } \Gamma_D, \\ [p\nabla u - \mathbf{b}u]\cdot\mathbf{n} &= f_N \quad \text{on } \Gamma_N, \end{aligned}} \qquad \boxed{\begin{aligned} \mathcal{L}^\dagger u &= g_S \quad \text{on } \Omega, \\ u &= g_D \quad \text{on } \Gamma_D, \\ [p\nabla u + \mathbf{c}u]\cdot\mathbf{n} &= g_N \quad \text{on } \Gamma_N, \end{aligned}} \tag{1.61}$$

involving the operators

$$\begin{aligned} \mathcal{L}u &:= -\nabla\cdot(p\nabla u) + \nabla\cdot(\mathbf{b}u) + \mathbf{c}\cdot\nabla u + qu, \\ \mathcal{L}^\dagger u &= -\nabla\cdot(p\nabla u) - \nabla\cdot(\mathbf{c}u) - \mathbf{b}\cdot\nabla u + qu, \end{aligned} \tag{1.62}$$

where $p, q : \Omega \to \mathbb{R}$, $\mathbf{b}, \mathbf{c} : \Omega \to \mathbb{R}^d$ are real functions on Ω, and $\mathbf{n}$ is the outward-pointing unit normal on the boundary. The dagger superscript indicates that $\mathcal{L}^\dagger$ is the formal adjoint of $\mathcal{L}$; we reserve the superscript star for the "true" adjoint. To ensure that the BVPs (1.61) are well-posed, we will make further assumptions on the coefficients and problem data, which we discuss in the next section. In particular, we will assume that $p \geq p_0$ for some $p_0 > 0$, $\nabla\cdot(\mathbf{b}-\mathbf{c}) \geq 0$, and $q \geq 0$ almost everywhere.

Before demonstrating the adjointness of these problems, let us make some quick comments about them. For both problems, there are two first-order terms, which is perhaps redundant as, e.g., $\nabla\cdot(\mathbf{b}u) = \mathbf{b}\cdot\nabla u + (\nabla\cdot\mathbf{b})u$, so that this term can be absorbed into the others. We include both first-order terms so that the adjoint problem takes the same form. Notice that the adjoint problem can be obtained by systematically making the replacements $\mathbf{b} \leftarrow -\mathbf{c}$, $\mathbf{c} \leftarrow -\mathbf{b}$. In words, the direction of the flow field is reversed for the adjoint. A problem is symmetric when it is the same as

its adjoint. The nonsymmetry of the above problems arises from the first-order terms, i.e., the advection. Nonsymmetry is our central theme; this is why we have introduced the adjoint at the start. We have also done this to highlight the symmetrical structure of the adjoint problem. The two problems are not the same, yet they have an identical structure. This is an idea that will underly our entire approach to nonsymmetry: we will exploit the the symmetry of the structure underlying the nonsymmetric problem.

What it means for the above BVPs to be adjoint is that, for all u and v in the domain of their respective operators satisfying the *homogeneous* versions of the boundary conditions ($f_D = g_D = 0$, $f_N = g_N = 0$), $\langle \mathcal{L}u, v\rangle = \langle u, \mathcal{L}^\dagger v\rangle$, where the inner products are the L^2 inner product on Ω. This is easiest to demonstrate if we first define the form

$$\mathfrak{a}[u, v] := \langle p\nabla u, \nabla v\rangle - \langle \mathbf{b}u, \nabla v\rangle + \langle \nabla u, \mathbf{c}v\rangle + \langle qu, v\rangle, \tag{1.63}$$

which will be the basis of the weak formulations of the BVPs. Applying integration by parts to the expression defining $\mathfrak{a}[u, v]$, in either direction, and applying the divergence theorem, we see that

$$\begin{aligned} \mathfrak{a}[u, v] &= \langle \mathcal{L}u, v\rangle + \langle [p\nabla u - \mathbf{b}u] \cdot \mathbf{n}, v\rangle_{\partial\Omega} \\ &= \langle u, \mathcal{L}^\dagger v\rangle + \langle u, [p\nabla v + \mathbf{c}v] \cdot \mathbf{n}\rangle_{\partial\Omega}. \end{aligned} \tag{1.64}$$

Here the subscripted inner products are the L^2 inner product on the boundary. We see now why the particular boundary conditions above were chosen. When u and v satisfy the homogeneous boundary conditions, the boundary integrals vanish and $\langle \mathcal{L}u, v\rangle = \langle u, \mathcal{L}^\dagger v\rangle$ as claimed.

1.2.2 Weak Form

Let us suppose $p, q \in L^\infty(\Omega)$, $\mathbf{b}, \mathbf{c} \in L^\infty(\Omega)^d$. These constraints on the coefficients are more than sufficient to ensure that $\mathfrak{a}[u, v]$ is well-defined and indeed *bounded* on all of $H^1(\Omega) \times H^1(\Omega)$ [43, p. 164]:

$$|\mathfrak{a}[u, v]| \leq C\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}, \tag{1.65}$$

for some constant C. A few more assumptions will ensure a kind of lower bound on $\mathfrak{a}$. We will assume that the BVPs are elliptic: there is some $p_0 > 0$ such that $p \geq p_0$ almost everywhere (a.e.) in Ω. Let us assume, in addition, that q is nonnegative a.e., that $\nabla \cdot (\mathbf{b} - \mathbf{c})$ is well defined and nonnegative a.e., which holds in particular for solenoidal flow, and also that $(\mathbf{b} - \mathbf{c}) \cdot \mathbf{n} \leq 0$ a.e. on Γ_N. This excludes, for example,

the possibility of prescribing the normal derivative of u on an inflow boundary in either BVP. Then it follows that $\mathfrak{a}$ has *coercive* real part: for all $u \in H^1(\Omega)$,

$$\operatorname{Re} \mathfrak{a}[u,u] \geq c\|u\|^2_{H^1(\Omega)} \tag{1.66}$$

for some $c \geq 0$. This statement holds for the complex Sobolev space: in the real setting, taking the real part is superfluous and one just says that $\mathfrak{a}$ is coercive. For the proof, and for more general conditions under which such a statement holds, we refer the reader to Quarteroni and Valli [43, p. 165]. Later we shall see that coercivity and boundedness together guarantee the existence of unique solutions to the weak forms of the BVPs.

To state the weak form of the adjoint BVP, we will make use of the adjoint of $\mathfrak{a}$, defined by

$$\mathfrak{a}^*[u,v] := \overline{\mathfrak{a}[v,u]}. \tag{1.67}$$

Notice that this is perhaps a simpler notion than that of the adjoint of an operator: one merely exchanges the order of the arguments, and takes the complex conjugate. For our form $\mathfrak{a}[\cdot,\cdot]$, the adjoint may be obtained by making the replacements $\mathbf{b} \leftarrow -\mathbf{c}$, $\mathbf{c} \leftarrow -\mathbf{b}$, just as was the case for the original statement of the adjoint BVP. The structure of $\mathfrak{a}[\cdot,\cdot]$ also makes it obvious that the second- and zeroth-order terms are symmetric, whereas the first-order terms are not.

The trace theorem (see, e.g., [43, p. 10]) establishes the existence of a unique continuous linear map, called a trace operator, from $H^1(\Omega)$ to $H^{1/2}(\partial\Omega)$ mapping functions to their trace on the boundary, which enables us to talk about the "values" a Sobolev space function takes on the boundary. This is needed to enforce Dirichlet boundary conditions. The trace theorem holds also when $\partial\Omega$ is replaced with the subset Γ_D. We define $H^1_{\Gamma_D}(\Omega) \subset H^1(\Omega)$ as the closed subspace of functions whose trace on Γ_D vanishes. We will also need to make use of the inverse trace theorem, which says that the trace operator has a (non-unique) continuous right inverse. That is, arbitrary boundary data in $H^{1/2}(\partial\Omega)$, or $H^{1/2}(\Gamma_D)$, can be extended to the interior to obtain a function in $H^1(\Omega)$, obviously in a non-unique way.

With these definitions in place, we can state the precise weak forms of the BVPs presented earlier,

$$\boxed{\begin{aligned} \mathcal{L}u &= f_S \quad \text{on } \Omega, \\ u &= f_D \quad \text{on } \Gamma_D, \\ [p\nabla u - \mathbf{b}u]\cdot\mathbf{n} &= f_N \quad \text{on } \Gamma_N, \end{aligned}} \qquad \boxed{\begin{aligned} \mathcal{L}^\dagger v &= g_S \quad \text{on } \Omega, \\ v &= g_D \quad \text{on } \Gamma_D, \\ [p\nabla v + \mathbf{c}v]\cdot\mathbf{n} &= g_N \quad \text{on } \Gamma_N. \end{aligned}} \tag{1.68}$$

We will assume that the boundary data satisfy $f_S, g_S \in L^2(\Omega)$, $f_D, g_D \in H^{1/2}(\Gamma_D)$, $f_N, g_N \in L^2(\Gamma_N)$. The constraint on the Dirichlet data ensures that f_D, g_D can be extended to some $\tilde{f}_D, \tilde{g}_D \in H^1(\Omega)$, such that, e.g., the trace of $\tilde{f}_D$ on Γ_D is f_D. Letting $\mathsf{V} = H^1_{\Gamma_D}(\Omega)$, the weak form of the first problem reads: find $u_0 \in \mathsf{V}$ such that

$$\mathfrak{a}[u_0, v] = (f_S, v) - \mathfrak{a}[\tilde{f}_D, v] + (f_N, v)_{\Gamma_N} =: f[v] \qquad \text{for all } v \in \mathsf{V}. \tag{1.69}$$

The actual solution sought is $u = \tilde{f}_D + u_0$. Similarly, the weak form of the adjoint reads: find $u_0 \in \mathsf{V}$ such that

$$\mathfrak{a}^*[u_0, v] = (g_S, v) - \mathfrak{a}^*[\tilde{g}_D, v] + (g_N, v)_{\Gamma_N} =: g[v] \qquad \text{for all } v \in \mathsf{V}, \tag{1.70}$$

where the actual solution sought is $u = \tilde{g}_D + u_0$. Here we have introduced the semilinear forms f and g, which collect together the terms involving all of the problem data. The constraints on the problem data ensure that f and g are continuous. That is, $f, g \in \mathsf{V}^*$. Defining the operator $A : \mathsf{V} \to \mathsf{V}^*$ by $(Au)[v] = (Au, v) := \mathfrak{a}[u, v]$, we may then write the weak forms of the adjoint problems simply as

$$Au_0 = f, \qquad A^* v_0 = g. \tag{1.71}$$

Making use of the integration by parts calculation from earlier,

$$\begin{aligned} (Au, v) &= \langle \mathcal{L}u, v\rangle + \langle [p\nabla u - \mathbf{b}u] \cdot \mathbf{n}, v\rangle_{\partial\Omega}, \\ (A^* u, v) &= \langle \mathcal{L}^\dagger u, v\rangle + \langle [p\nabla u + \mathbf{c}u] \cdot \mathbf{n}, v\rangle_{\partial\Omega}, \end{aligned} \tag{1.72}$$

one readily verifies that a classical (strong) solution to one of the original BVPs is necessarily a solution to the corresponding weak version. On the other hand, because A is bounded with coercive real part, the Lax-Milgram lemma (discussed in §2.2), ensures that the problems (1.71) have unique solutions that depend continuously on the problem data.

1.2.3 Discretization

We need to discretize (1.71) if we want a model problem to which we may apply AMG. The simple Galerkin approach, a symmetric construction which works very well in the symmetric case, is especially ill-suited to highly nonsymmetric problems. For the advection-diffusion problem, the poor accuracy of the Galerkin scheme can manifest in practice as, for example, spurious oscillations propagating from an underresolved boundary or interior layer into regions where the solution could be well-resolved. Since the highly nonsymmetric case (advection-dominated) is the one we are interested

in considering for AMG, we must consider a more suitable discretization, unless we are prepared for our model problem to be unrealistic. A full discussion of the variety of stabilization schemes used to recover a more suitable discretization for the advection-diffusion problem would take us too far afield. There are many treatments in the literature, including Quarteroni and Valli [43, §8.3]. Instead we shall simply describe the discretization we will use in our examples, the streamline diffusion scheme described in [16, §3.3.2]. The scheme is due originally to Hughes and Brookes [11] under the name streamline/upwind Petrov-Galerkin (SUPG), but we will make the parameter choices suggested in the first reference.

For simplicity, let us restrict our attention to the case $\mathbf{b} = 0$ and $q = 0$, so that the weak form of the BVP (omitting the adjoint) reads: find $u \in \mathsf{V}$ such that

$$\mathfrak{a}[u, v] = \langle p\nabla u, \nabla v\rangle + \langle \nabla u, \mathbf{c}v\rangle = \langle f_S, v\rangle + \langle f_N, v\rangle_{\Gamma_N} = f[v] \tag{1.73}$$

for all $v \in \mathsf{V}$. For the streamline diffusion scheme, one considers the modified forms

$$\mathfrak{a}_{sd}[u, v] = \mathfrak{a}[u, v] + \sum_k \delta_k \langle -\nabla \cdot (p\nabla u) + \mathbf{c} \cdot \nabla u, \mathbf{c} \cdot \nabla v\rangle_k, \tag{1.74}$$

$$f_{sd}[v] = f[v] + \sum_k \delta_k \langle f_S, \mathbf{c} \cdot \nabla v\rangle_k, \tag{1.75}$$

where the sums are over all elements of the triangulation of the domain, the subscripted inner product is the L^2 inner product on the indicated element, and where the parameter δ_k for element k is chosen based on an element Péclet number P_k,

$$P_k = \frac{\|\mathbf{c}_k\|(h_k/2)}{p_k}, \qquad \delta_k = \begin{cases} \frac{h_k/2}{\|\mathbf{c}_k\|}\left(1 - \frac{1}{P_k}\right) & P_k > 1 \\ 0 & P_k \leq 1 \end{cases}. \tag{1.76}$$

Here $\mathbf{c}_k$ and p_k are the velocity and diffusion evaluated at the centroid of element k, and h_k is the length of the element in the direction of $\mathbf{c}_k$. The streamline diffusion discretization is simply the Galerkin scheme applied to the modified forms.

Due to the presence of second derivatives of u, the modified form is no longer well-defined (i.e., continuous) on all of $\mathsf{V} \times \mathsf{V}$. However, provided p is smooth enough in each element, the form is well-defined on the discretization space, since functions in this space are C^∞ within each element.

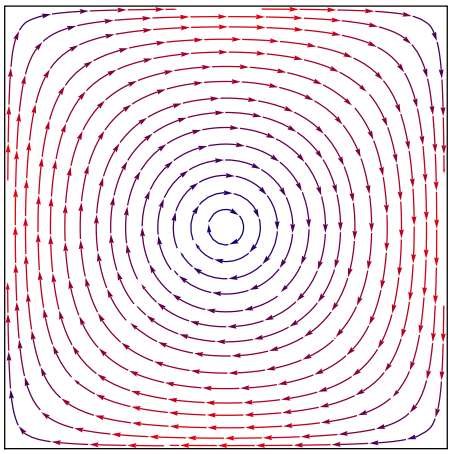
Note that for a classical solution u, the terms added to the forms on the left- and right-hand-sides of the weak problem are equal, so that a classical solution remains a solution to the modified problem. In this sense the scheme is a strongly consistent stabilization. Also note that the parameter δ_k is 0 wherever the mesh is well-resolved so that the element Péclet number is smaller than 1. Hence, locally in well-resolved

areas, the discretization reduces to the Galerkin discretization. This implies, of course, that as the mesh is resolved globally ($h \to 0$), the discretization also reduces to the Galerkin one globally.

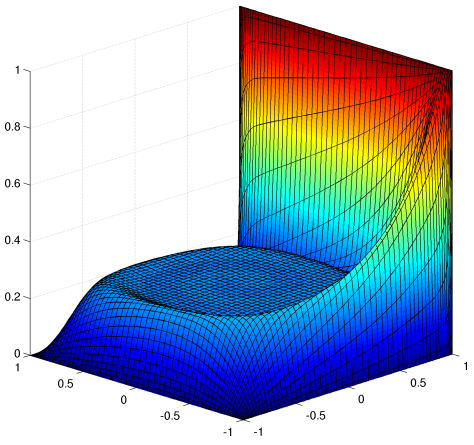
For a full derivation and discussion of the scheme, including the rationale behind the choice for the parameter δ_k, we refer the reader to [16, §3.3.2]. The original SUPG derivation was motivated as a Petrov-Galerkin scheme: one keeps the original forms and finite-dimensional trial space V_h, but replaces the test space with $\mathsf{W}_h = \{v + \delta \mathbf{c} \cdot \nabla v \mid v \in \mathsf{V}_h\}$. This description is incomplete by itself, as the original form $\mathfrak{a}$ is not well-defined on this test space, as it consists of functions which are generally not continuous across element boundaries, and hence not in V. The streamline-diffusion scheme described above can be seen as the result of applying remedies to this Petrov-Galerkin idea. In particular, we can see that the problematic higher-order term has been handled simply by splitting the integral into a sum of integrals over each element and ignoring the problematic discontinuities between elements.

Figure 1.1 shows the discrete solution for the "double glazing problem," taken from [16, §3.1], for the domain $[-1, 1]^2$ discretized by a 63 by 63 Chebyshev grid. The advection velocity $\mathbf{c}$ is as stated and illustrated in the figure, the diffusion coefficient p is the constant $1/200$, and the boundary conditions are Dirichlet with $u = 0$ on three sides and $u = 1$ on the fourth. The Chebyshev spacing was chosen to give clustering of nodes near the boundaries to mimic what would be produced by an adaptive method or by hand in order to capture the boundary layers of the solution. This double glazing problem will be one of the test problems for the numerical tests in Chapter 5, and we will also have occassion to refer to it for illustration purposes, for example, to illustrate the "form" absolute value of Chapter 3.

The next chapter switches gears from the concrete to the abstract, covering in more depth the abstract framework upon which the theory of subsequent chapters is based, and which, in particular, encompasses the two model problems of this chapter.



$$c_x = \quad 2y(1-x^2)$$
$$c_y = -2x(1-y^2)$$



$p = 0.005$

Figure 1.1: Double glazing problem

Chapter 2

Theoretical Foundations

In the following two chapters, we will first investigate a new "absolute value" norm, which we will then use to develop a general two-level multigrid convergence theory applicable to nonsymmetric problems. Both of the subsequent chapters will be very general, assuming only that we are interested in solving problems of the form

$$Au = f \tag{2.1}$$

with $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ being a bounded linear operator from a complex reflexive Banach space to the adjoint space. Additionally, the real part $H = \frac{1}{2}(A + A^*)$ of A will be assumed to be positive, which suffices, by the Lax-Milgram lemma, Theorem 2.1, to ensure the existence of $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$, so that there is a unique solution u to (2.1) depending continuously on f.

Taking $\mathsf{V} = \mathbb{C}^n$ reduces this framework to the familiar matrix one, with (2.1) reducing to a system of linear equations governed by the matrix A. This and following chapters may be read with the assumption $\mathsf{V} = \mathbb{C}^n$ in mind, in which case one may ignore the functional analysis concepts that come up. As with (2.1), the equations should be readily interpretable in the matrix setting, as the notation matches standard matrix notation.

On the other hand, the above framework is general enough to encompass elliptic PDEs such as those encountered in the previous chapter. Hence, all of the concepts we consider will be applicable not just to discretized systems, but also to the underlying PDE as well. It is a feature of the finite element discretization that concepts will tend to have a compatible meaning whether we instantiate our framework to apply to the PDE or to its discretization. For example, $\|u\|_A^2$ may be, for a symmetric second-order elliptic problem, the energy of an arbitrary H^1 function in the one case, or the energy of some member of the finite-dimensional subspace of discretization functions in the

other. We may thus draw insight as to the meaning of concepts in the discrete case from the infinite-dimensional one. This applies, for example, to the new norm we investigate in the next chapter.

The spaces V we have in mind (H^1 and $\mathbb{C}^n$) are in fact Hilbert spaces. Indeed, as we shall see, our premises on the operator $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ in (2.1), that it have positive real part, ensures that V can always be equipped with a Hilbert space structure, with a norm equivalent to the Banach norm. In other words, our assumptions imply that V is topologically isomorphic to a Hilbert space. The reason we do not make the Hilbert space structure an assumption is that we regard the particular inner products (the Sobolev and Euclidean) and associated norms of these spaces as not being particularly fundamental. By assuming only a Banach space structure, we essentially make these inner products and their associated geometry unavailable to our developments. We will instead investigate a number of alternative Hilbert space structures, such as that given by the energy inner product, that are naturally associated with the given problem. The only aspect of the given spaces that we consider important is the topology, and this is captured by the Banach metric.

The reason for assuming a complex Banach space is that we will need to make use of the complex numbers almost immediately (so that the quadratic form is sensitive to the nonsymmetry of the operator), and it seems simplest to just treat the complex case in full at the outset. Instantiating results to the real case is usually trivial (e.g., ignoring complex conjugation), whereas generalizing in the other direction is less straight-forward.

In this chapter we will review the basic concepts of our abstract framework. For a functional analyst, all of the material should be review or at least very straight-forward, albeit, rendered in a matrix-like notation: favoring operators over forms, and denoting Hilbert spaces by the operator associated with the inner product. For a numerical analyst used to the matrix setting, the emphasis will be in focusing on Hilbert spaces other than the Euclidean one, and the notation we use should be fairly standard in this setting.

2.1 Sesquilinear Forms

The following definitions are standard and follow Kato [31]. Let V be a reflexive complex Banach space. The adjoint space V^* is defined to consist of bounded semilinear (alternatively conjugate- or antilinear) forms on V, as opposed to linear ones. We will

reserve the notation $(\cdot,\cdot)$ for the duality pairing of V^* and V, so that, for any bounded semilinear form $f \in \mathsf{V}^*$ and vector $u \in \mathsf{V}$,

$$(f, v) := f[v]. \tag{2.2}$$

Note that the pairing $(\cdot,\cdot)$ is linear in the first argument and semilinear in the second. As above, we will consistently use the letters f and g to denote elements of an adjoint space, and the letters u, v, and w for elements of the main space. In the matrix setting, $\mathsf{V} = \mathbb{C}^n$, we identify V^* with $\mathbb{C}^n$ by letting $(\cdot,\cdot)$ be the usual Euclidean inner product.

Let $\mathfrak{a}$ be a bounded sesquilinear form on V. A sesquilinear form is defined to be linear in its first argument and semilinear in its second, matching the convention for the duality pairing. "Sesqui-" means "one and a half" and serves as a mnemonic for this convention. Taken together, these conventions allow the form $\mathfrak{a}$ to equivalently be considered as a bounded linear operator $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$, defined by

$$(Au, v) = (Au)[v] = \mathfrak{a}[u, v], \tag{2.3}$$

where $\mathscr{B}(\mathsf{V}, \mathsf{W})$ is the space of bounded linear operators from V to W. In the matrix setting $\mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ is simply $\mathbb{C}^{n\times n}$, the space of all $n \times n$ matrices.

That boundedness of the form $\mathfrak{a}$ implies boundedness of the operator A is straightforward. If $\mathfrak{a}[u, v]$ has bound C, i.e., $|\mathfrak{a}[u, v]| \le C\|u\|_{\mathsf{V}}\|v\|_{\mathsf{V}}$, then so does A, since

$$\|Au\|_{\mathsf{V}^*} = \sup_{v\in\mathsf{V}\setminus\{0\}} \frac{|(Au, v)|}{\|v\|_{\mathsf{V}}} \le \sup_{v\in\mathsf{V}\setminus\{0\}} \frac{C\|u\|_{\mathsf{V}}\|v\|_{\mathsf{V}}}{\|v\|_{\mathsf{V}}} = C\|u\|_{\mathsf{V}}. \tag{2.4}$$

In the other direction, if $\|A\| \le C$, then

$$|\mathfrak{a}[u, v]| = |(Au, v)| \le \|Au\|_{\mathsf{V}^*}\|v\|_{\mathsf{V}} \le C\|u\|_{\mathsf{V}}\|v\|_{\mathsf{V}}. \tag{2.5}$$

That is, not only is A bounded, but it has the same bounding constant as $\mathfrak{a}$.

Conversely, if we are given an operator $A \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$, then (2.3) can be taken to define the sesquilinear form $\mathfrak{a}$, which by the above arguments is bounded with the same bounding constant as A. That is, we have established an isometric isomorphism between the space $\mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ and the space of bounded sesquilinear forms on V. In light of this correspondence, we shall sometimes refer to an operator $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ as a bounded sesquilinear form, leaving the isomorphic mapping implied.

The sesquilinear form $\mathfrak{a}$ is associated with the quadratic form

$$\mathfrak{a}[u] := \mathfrak{a}[u, u] = (Au, u). \tag{2.6}$$

In fact, the sesquilinear form is determined by the quadratic form via the polarization identity

$$\mathfrak{a}[u,v] = \frac{1}{4}(\mathfrak{a}[u+v] - \mathfrak{a}[u-v] + i\mathfrak{a}[u+iv] - i\mathfrak{a}[u-iv]). \tag{2.7}$$

This one-to-one correspondence between sesquilinear and quadratic forms relies on the existence of the scalar i. That is, the correspondence would not hold if V were a real instead of complex Banach space. The adjoint form $\mathfrak{a}^*$ is defined by

$$\mathfrak{a}^*[u,v] := \overline{\mathfrak{a}[v,u]} = \overline{(Av,u)} = (A^*u,v). \tag{2.8}$$

Note $A^* \in \mathscr{B}(\mathsf{V}^{**},\mathsf{V}^*) = \mathscr{B}(\mathsf{V},\mathsf{V}^*)$ as V was assumed to be reflexive. Thus, A^* is the operator associated with $\mathfrak{a}^*$ in exactly the same way as A is associated with $\mathfrak{a}$. In the matrix case, A^* is the conjugate transpose of A, as the notation suggests.

The form $\mathfrak{a}$ is *symmetric* when $\mathfrak{a}^* = \mathfrak{a}$, which is equivalent to A being equal to its adjoint ($A^* = A$). In this case the quadratic form $\mathfrak{a}[u]$ is real. In light of the correspondence (2.3), where we think of A as a sesquilinear form, we will call A symmetric when it is equal to its adjoint. There is a slight danger of confusion in the matrix setting, as the term "Hermitian" is more common, "symmetric" possibly referring to a matrix being equal to its regular (not conjugate) transpose. For a real matrix there is no distinction, and the condition $A^* = A$ (the "Hermitian" one) is the generalization of symmetry to the complex case more appropriate for our uses. Here, symmetric always means Hermitian.

The Cartesian decomposition of the quadratic form $\mathfrak{a}[u]$ is

$$\mathfrak{a}[u] = \mathfrak{h}[u] + i\mathfrak{k}[u], \qquad \mathfrak{h} := \frac{1}{2}(\mathfrak{a}+\mathfrak{a}^*), \qquad \mathfrak{k} := \frac{1}{2i}(\mathfrak{a}-\mathfrak{a}^*). \tag{2.9}$$

Note that $\mathfrak{h}$ and $\mathfrak{k}$ are both symmetric, giving rise to real quadratic forms. The associated symmetric operators are

$$H_A := \frac{1}{2}(A+A^*), \quad \text{and} \quad K_A := \frac{1}{2i}(A-A^*). \tag{2.10}$$

We will refer to these as the *real* and *imaginary* parts of A. Subsequently, we will usually shorten H_A and K_A to H and K. We will occasionally wish to refer to the real and imaginary parts of other operators. For example, note that

$$H_{A^*} = H_A, \qquad K_{A^*} = -K_A. \tag{2.11}$$

For the case of A being a real matrix, H is the symmetric part of A, while iK is the anti-symmetric part.

$\mathbb{C}$	$\mathscr{B}(\mathsf{V}, \mathsf{V}^*)$
$\overline{z}$	A^*
$z \in \mathbb{R}$ if $z = \overline{z}$	$(Au, u) \in \mathbb{R}$ if $A = A^*$
$z = a + ib$	$A = H + iK$

Table 2.1: Analogy between complex numbers and sesquilinear forms

Table 2.1 summarizes the beginning of an analogy that arises between sesquilinear forms and the complex numbers, including the concepts introduced so far. In the analogy, the adjoint corresponds to complex conjugation, and "symmetric" corresponds to "real." Note that the second column reduces to the first in the scalar case $\mathsf{V} = \mathbb{C}$.

2.2 Lax-Milgram Lemma

A sufficient condition for A^{-1} to exist, be continuous, and be defined on all of V^* is for $\mathfrak{a}$ to be coercive. This statement is a variant of the Lax-Milgram lemma, usually stated for a bilinear form on a real Hilbert space. The lemma will play a central role in our development, as its premises serve to define (nearly, see the corollary below) the class of problems we are interested in solving. It is easy to find the complex version of the lemma, for which the proof is essentially the same. Roşca appears to have been the first to consider its generalization to Banach spaces [45]. As it is difficult to find a version of the proof with both generalizations, and as it is relatively short, a proof is included here, for completeness. But note that considering a complex instead of a real space adds essentially no difficulty to the proof. This is the only result we will consider that involves topological notions, where the possibility of infinite dimensions introduces subtleties not present in the finite-dimensional case. As the result itself is (nearly) standard, the proof may be skipped, as none of the following material will require the concepts it uses.

Theorem 2.1 (Lax-Milgram)**.** *Given $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ with V a reflexive complex Banach space, suppose*

$$|(Au, u)| \geq c\|u\|_{\mathsf{V}}^2 \tag{2.12}$$

for all $u \in \mathsf{V}$ and some $c > 0$. Then A is invertible and $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$ with $\|A^{-1}\| \leq c^{-1}$.

Proof. From the coercivity assumption and the generalized Cauchy-Schwarz inequality, we find that for all $u \in \mathsf{V}$,

$$c\|u\|_{\mathsf{V}}^2 \leq |(Au, u)| \leq \|Au\|_{\mathsf{V}^*}\|u\|_{\mathsf{V}} \tag{2.13}$$

so that

$$\|Au\|_{\mathsf{V}^*} \geq c\|u\|_{\mathsf{V}}. \tag{2.14}$$

We conclude [31, p. 146] that A has bounded inverse A^{-1} with domain $\mathsf{D}(A^{-1}) = \mathsf{R}(A)$, the range of A, and that $\|A^{-1}\| \leq c^{-1}$. We must yet show $\mathsf{R}(A) = \mathsf{V}^*$.

Because $|(A^*u, u)| = |\overline{(Au, u)}| = |(Au, u)|$, we see that A^* is also coercive on V, which by reflexivity is the whole domain of A^* ($\mathsf{D}(A^*) = \mathsf{V}^{**} = \mathsf{V}$). The above argument thus applies equally well to A^*, and we conclude in particular that the null space of A^* is trivial, $\mathsf{N}(A^*) = \{0\}$. Recall that the annihilator $\mathsf{S}^\perp$ of a set $\mathsf{S} \subseteq \mathsf{V}$ is the closed subspace of V^* consisting of those forms f for which $(f, u) = 0$ for all $u \in \mathsf{S}$, and that $\mathsf{S}^{\perp\perp}$ is the closure of the span of S [31, p. 136], provided V is reflexive as we have assumed. Since $\mathsf{N}(A^*) = \mathsf{R}(A)^\perp$ [31, p. 168], it follows that the closure of $\mathsf{R}(A)$ is

$$\mathsf{R}(A)^{\perp\perp} = \mathsf{N}(A^*)^\perp = \{0\}^\perp = \mathsf{V}^*. \tag{2.15}$$

If we can show that $\mathsf{R}(A)$ is closed, then it would follow that $\mathsf{R}(A) = \mathsf{V}^*$.

But it is a standard result that a bounded operator has closed range when it is bounded below. Here, the argument runs as follows. Let Au_n be an arbitrary Cauchy sequence in $\mathsf{R}(A)$ so that $\|Au_n - Au_m\|_{\mathsf{V}^*} \to 0$ as $n, m \to \infty$. From (2.14) we see that $\|u_n - u_m\|_{\mathsf{V}} \leq c^{-1}\|Au_n - Au_m\|_{\mathsf{V}^*} \to 0$ so that u_n is a Cauchy sequence in V. By completeness, the sequence u_n has some limit u, and by continuity (boundedness of A) $Au_n \to Au$ in V^*. Thus $\mathsf{R}(A)$ is closed, and it follows that $\mathsf{R}(A) = \mathsf{V}^*$ and $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$. □

We will actually be concerned with a slightly more restricted class of operators A, namely those with positive real part H.

Corollary. *If $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ has positive real part H, i.e.,*

$$\mathrm{Re}\,(Au, u) = (Hu, u) \geq c\|u\|_{\mathsf{V}}^2 \tag{2.16}$$

for all $u \in \mathsf{V}$ and some $c > 0$, then $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$.

Proof. $|(Au, u)| \geq \mathrm{Re}\,(Au, u)$. □

2.3 The "Energy" and Dual Hilbert Space

A symmetric form $\mathfrak{h}$ has lower bound c, written $c \leq \mathfrak{h}$, if

$$\mathfrak{h}[u] \geq c\|u\|_{\mathsf{V}}^2 \tag{2.17}$$

for all $u \in \mathsf{V}$. For the corresponding linear operator $H : \mathsf{V} \to \mathsf{V}^*$, defined by $(Hu, v) = \mathfrak{h}[u, v]$, we write $c \leq H$. Assume $0 < c \leq \mathfrak{h}$. Then the sesquilinear form $\mathfrak{h}$ is an inner product on V. Assuming $\mathfrak{h}$ (and thus H) is bounded, the induced norm

$$\|u\|_H^2 := \mathfrak{h}[u] = (Hu, u) \tag{2.18}$$

is equivalent to the Banach norm on V, as for all $u \in \mathsf{V}$,

$$c^{1/2}\|u\|_{\mathsf{V}} \leq \|u\|_H \leq \|H\|^{1/2}\|u\|_{\mathsf{V}}. \tag{2.19}$$

In particular, V is complete under the new metric and thus a Hilbert space. The Lax-Milgram lemma implies that H has inverse $H^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$ with $\|H^{-1}\| \leq c^{-1}$. The form defined by H^{-1} is also an inner product, as it is symmetric with positive lower bound, since, for all $f \in \mathsf{V}^*$,

$$(H^{-1}f, f) = \|H^{-1}f\|_H^2 \geq c\|H^{-1}f\|_{\mathsf{V}}^2 \geq \frac{c}{\|H\|^2}\|f\|_{\mathsf{V}^*}^2, \tag{2.20}$$

i.e., $c\|H\|^{-2} \leq H^{-1}$. Note here that the duality pairing is of $\mathsf{V}^{**} = \mathsf{V}$ and V^*. The arguments above apply also to H^{-1}, and V^* is a Hilbert space under the inner product defined by H^{-1}. In this case the norm equivalence is given by

$$\frac{c^{1/2}}{\|H\|}\|f\|_{\mathsf{V}^*} \leq \|f\|_{H^{-1}} \leq c^{-1/2}\|f\|_{\mathsf{V}^*} \tag{2.21}$$

for all $f \in \mathsf{V}^*$. By construction, H is the isometric Riesz representation mapping between these dual Hilbert spaces.

$$\|Hu\|_{H^{-1}} = \|u\|_H, \qquad \|H^{-1}f\|_H = \|f\|_{H^{-1}}. \tag{2.22}$$

Versions of many familiar results involving the Banach norms arise with the Hilbert space norms replacing the Banach norms. We first consider the generalized Cauchy-Schwarz inequality $|(f, u)| \leq \|f\|_{\mathsf{V}^*}\|u\|_{\mathsf{V}}$, which is little more than the definition of $\|f\|_{\mathsf{V}^*}$. We see that

$$|(f, u)| = |\mathfrak{h}[H^{-1}f, u]| \leq \|H^{-1}f\|_H\|u\|_H = \|f\|_{H^{-1}}\|u\|_H \tag{2.23}$$

for all $u \in \mathsf{V}, f \in \mathsf{V}^*$, where we have used the Cauchy-Schwarz inequality for the inner product $\mathfrak{h}$. The Hilbert space norms have the following alternative characterizations.

$$\|f\|_{H^{-1}} = \sup_{\|u\|_H=1} |(f, u)|, \qquad \|u\|_H = \sup_{\|f\|_{H^{-1}}=1} |(f, u)|. \tag{2.24}$$

By the generalized Cauchy-Schwarz inequality (2.23), $\|f\|_{H^{-1}} \geq |(f,u)|$ for $\|u\|_H = 1$, with equality holding for $u = H^{-1}f/\|H^{-1}f\|_H$. This establishes the first identity; the second may be established by similar reasoning. The characterization of $\|f\|_{H^{-1}}$ in (2.24) is the definition of the norm of the adjoint space, and so we have just proved that these Hilbert spaces are dual, as previously asserted.

Suppose we have two reflexive complex Banach spaces, V_1 and V_2, and let $H_1 \in \mathscr{B}(\mathsf{V}_1, \mathsf{V}_1^*)$ and $H_2 \in \mathscr{B}(\mathsf{V}_2, \mathsf{V}_2^*)$ be positive symmetric operators. To refer explicitly to the norm of T as an operator between the Hilbert spaces associated with H_1 and H_2 we will use the notation

$$\|T\|_{H_2,H_1} := \sup_{u \in \mathsf{V}_1 \setminus \{0\}} \frac{\|Tu\|_{H_2}}{\|u\|_{H_1}}, \tag{2.25}$$

which is the standard definition of the norm of an operator between two Banach spaces (here with the Hilbert space norms). When $H_1 = H_2 = H$, we will write simply $\|T\|_H$.

The norm of the adjoint of an operator is equal to the norm of the operator,

$$\|T^*\|_{H_1^{-1},H_2^{-1}} = \|T\|_{H_2,H_1}. \tag{2.26}$$

This is just the familiar identity $\|T^*\| = \|T\|$ [31, p. 154], but applied to T considered as an operator between the two Hilbert spaces, and taking account that T^* is an operator between the two dual spaces.

2.4 Sectorial Forms

Let us return our attention now to the nonsymmetric operator A, which we shall assume to have positive real part $H > 0$. If the imaginary part is bounded by

$$|\mathfrak{k}[u]| \leq (\tan\theta)\mathfrak{h}[u], \tag{2.27}$$

for all $u \in \mathsf{V}$, then $\mathfrak{a}$ is said to be sectorial with vertex 0 and corresponding semi-angle θ (see Figure 2.1). If A is bounded in addition to having positive real part, then $\mathfrak{a}$ is always sectorial, as we may take

$$\tan\theta = \|K\|_{H^{-1},H}. \tag{2.28}$$

This follows from the generalized Cauchy-Schwarz inequality,

$$|\mathfrak{k}[u]| = |(Ku,u)| \leq \|Ku\|_{H^{-1}}\|u\|_H \leq \|K\|_{H^{-1},H}\|u\|_H^2 = \|K\|_{H^{-1},H}\mathfrak{h}[u]. \tag{2.29}$$

We see that the angle θ, which falls in the range $0 \leq \theta < \pi/2$, is a direct measure of the "nonsymmetry" of A, being a measure of the size of the imaginary part relative to

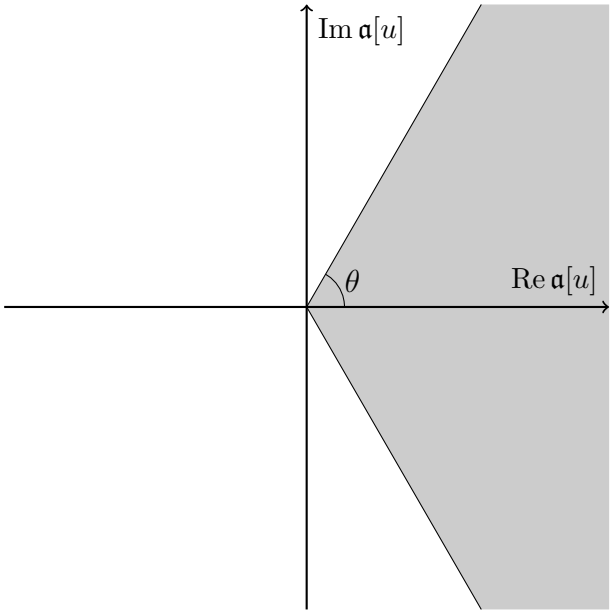


Figure 2.1: Sectorial form with semi-angle θ. The quadratic form $\mathfrak{a}[u]$ takes on values in the shaded sector of the complex plane.

the real part. Note that $\theta = 0$ for positive symmetric A, as was the case for Poisson's equation as examined in §1.1.1, while for the advection-diffusion example of §1.2.1, we have on the other hand $\theta \to \pi/2$ as the diffusion coefficient $p \to 0$.

It is easy to construct new Hilbert spaces that "pair" with those associated with H and H^{-1} to make A an isometry. Consider $\|Au\|_{H^{-1}}$.

$$\|Au\|^2_{H^{-1}} = (H^{-1}Au, Au) = (A^*H^{-1}Au, u). \tag{2.30}$$

Let us construct the norm $\|u\|_M = \|Au\|_{H^{-1}}$ by defining

$$M := A^*H^{-1}A = AH^{-1}A^* = H + KH^{-1}K, \tag{2.31}$$

which follows since $A = H + iK$ while $A^* = H - iK$. Note that $0 \leq KH^{-1}K = K^*H^{-1}K$ so that $H \leq M$, i.e., $(Hu, u) \leq (Mu, u)$ for all $u \in \mathsf{V}$. As for H, we will occasionally use the notation M_A when we wish to make the dependence on A explicit. Here we remark that $M_A = M_{A^*}$. The Lax-Milgram lemma applies to M, so that $M^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$. Indeed, M^{-1} is the real part of A^{-1}.

$$M^{-1} = A^{-1}HA^{-*} = \tfrac{1}{2}A^{-1}(A + A^*)A^{-*} = \tfrac{1}{2}(A^{-1} + A^{-*}) \tag{2.32}$$

In other words,

$$H_{A^{-1}} = M_A^{-1}, \quad \text{and hence} \quad M_{A^{-1}} = H_A^{-1}. \tag{2.33}$$

By construction, we have the following identities, analogous to (2.22), which state that A is an isometry between the indicated Hilbert spaces.

$$\begin{aligned} \|Au\|_{H^{-1}} &= \|u\|_M, \qquad \|A^{-1}f\|_M = \|f\|_{H^{-1}}, \\ \|Au\|_{M^{-1}} &= \|u\|_H, \qquad \|A^{-1}f\|_H = \|f\|_{M^{-1}}. \end{aligned} \tag{2.34}$$

These identities remain valid when A is replaced by A^*. In any of these identities, note that H and M may be exchanged—they enjoy a kind of "duality," though not in the sense of the dual space. The sectorial semi-angle provides an equivalence between the two norms. We have

$$\|u\|_H \le \|u\|_M \le (\sec\theta)\|u\|_H \tag{2.35}$$

for any u. This follows from (2.28) and (2.31), noting that

$$\|u\|_M^2 - \|u\|_H^2 = \|Ku\|_{H^{-1}}^2 \le (\tan\theta)^2\|u\|_H^2 \tag{2.36}$$

and $\sec^2\theta = 1 + \tan^2\theta$. Dually, by taking $u = A^{-1}f$ we see that

$$\|f\|_{M^{-1}} \le \|f\|_{H^{-1}} \le (\sec\theta)\|f\|_{M^{-1}} \tag{2.37}$$

for all $f \in \mathsf{V}^*$.

By making use of the isometry identities (2.34) and the generalized Cauchy-Schwarz inequality (2.23), as well as the corresponding one for M, we see that

$$\begin{aligned} |(Au, v)| &\le \|Au\|_{H^{-1}}\|v\|_H = \|u\|_M\|v\|_H, \\ |(Au, v)| &\le \|Au\|_{M^{-1}}\|v\|_M = \|u\|_H\|v\|_M. \end{aligned} \tag{2.38}$$

Each member of this symmetric pair of inequalities resembles a Cauchy-Schwarz inequality, though two norms are involved and the form $\mathfrak{a}$ is not symmetric. The weaker result

$$|(Au, v)| \le (\sec\theta)\|u\|_H\|v\|_H \tag{2.39}$$

then follows from (2.35). For the quadratic form, we have

$$\|u\|_H^2 \le |(Au, u)| \le \|u\|_H\|u\|_M \le (\sec\theta)\|u\|_H^2. \tag{2.40}$$

These inequalities bounding the quadratic form, apart from the intermediate bound involving M, can be found in Kato [31, p. 311], and are evident from studying Figure 2.2. The dual versions of these inequalities are

$$|(f, A^{-1}g)| \le \begin{matrix} \|f\|_{H^{-1}}\|g\|_{M^{-1}} \\ \|f\|_{M^{-1}}\|g\|_{H^{-1}} \end{matrix} \le (\sec\theta)\|f\|_{M^{-1}}\|g\|_{M^{-1}}, \tag{2.41}$$

$$\|f\|_{M^{-1}}^2 \le |(f, A^{-1}f)| \le \|f\|_{M^{-1}}\|f\|_{H^{-1}} \le (\sec\theta)\|f\|_{M^{-1}}^2. \tag{2.42}$$

These follow from the former by taking $u = A^{-1}f$, $v = A^{-1}g$ and making use of the isometry identities (2.34).

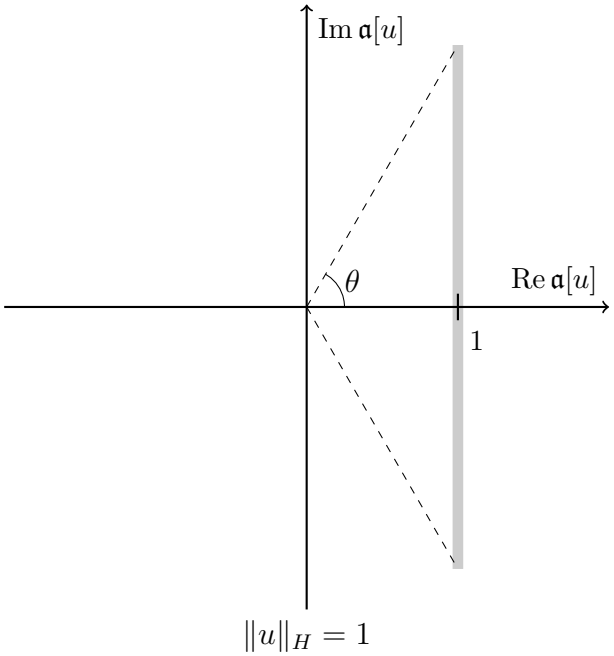


Figure 2.2: When u is normalized such that $\|u\|_H = 1$, the quadratic form $\mathfrak{a}[u]$ takes on values on the shaded line segment in the complex plane.

2.5 On Energy

It is perhaps worth pausing to consider the physical significance of the abstract Hilbert spaces we have been considering. The norm $\|u\|_H$ is often called an "energy" norm. It is indeed sometimes the case that $\|u\|_H^2$ corresponds to a physical energy. For example, the Poisson equation (see §1.1.1) can be used to model the condition of equilibrium for a membrane subjected to some load. In this case $\|u\|_H^2$ is the potential energy of the membrane.

More generally, consider a time-dependent version of $Au = f$,

$$Q\dot{u} + Au = f, \tag{2.43}$$

where $Q \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ is nonnegative symmetric and $\dot{u} := du/dt$. For example, if we take Q to be the operator corresponding to the L^2 inner product, then for our Poisson equation example (§1.1.1) we recover the heat equation, while our steady-state advection-diffusion example (§1.2.1) becomes time-dependent. In either case, if we test (2.43) on u and take the real part, we find

$$\frac{d}{dt}\left(\tfrac{1}{2}\|u\|_Q^2\right) = -\|u\|_H^2 + \operatorname{Re}(f, u) \tag{2.44}$$

since $\frac{d}{dt}\|u\|_Q^2 = (Qu, \dot{u}) + (Q\dot{u}, u) = 2\operatorname{Re}(Q\dot{u}, u)$. As we have assumed $H > 0$, we can infer that the "energy" $\|u\|_Q^2$ cannot increase in the absence of a source (i.e.,

$f = 0$). We can also see that the imaginary part K of A leaves $\|u\|_Q^2$ invariant. This is an example of an "energy" method, a technique for analyzing PDEs (see e.g., Evans [17, pp. 42, 65, 86]). Despite the fact that the Navier-Stokes equations are a nonlinear and a vector version of the advection-diffusion example, an energy *inequality* similar to the above equality nevertheless holds. For that case, u is the fluid velocity vector, $\frac{1}{2}\|u\|_Q^2$ is the kinetic energy of the fluid, and $-\|u\|_H^2$ is the viscous dissipation of kinetic into thermal energy. That the energy norm does not depend on the imaginary part K physically corresponds to the fact that the convection transports kinetic energy without diminishing it.

The M norm does not appear to have any common name. It is the dual to the "energy" norm associated with A^{-1} (recall $M_A^{-1} = H_{A^{-1}}$), but there is no similar physical significance to attach to it, as, e.g., there is no physically significant analog of the time-dependent equation (2.43) for A^{-1}.

It is apparent that the energy norm $\|u\|_H^2$ is useful and has importance in certain contexts even for nonsymmetric problems. However, for other purposes it is inadequate. Returning to the problem we are actually concerned with, $Au = f$, one might like to use the Cauchy-Schwarz type inequality (2.39) or the bound on the quadratic form (2.40) in a convergence analysis of the multigrid method. Unfortunately, these bounds involve $\sec\theta$, and as $\sec\theta \to \infty$ as $\theta \to \pi/2$, they become useless for the highly nonsymmetric problems that interest us. We can see that this difficulty stems precisely from the fact that $\|u\|_H^2$ is completely independent of the imaginary (nonsymmetric) part of A, which we have just seen to be a useful and signficant property in a different context.

In this chapter, we have reviewed the basic concepts of the abstract setting we have chosen for the analysis of multigrid methods applied to a nonsymmetric form defined by an operator A with positive real part H. We have also looked at some Hilbert spaces associated with A, such as that associated to H and the "energy" norm. One more such Hilbert space and norm will be investigated in the next chapter, one associated with a new notion of the absolute value $|A|$ of A, which, unlike H, does take account of the imaginary part K of A, and which will prove ideal for the multigrid analysis in the subsequent chapter.

Chapter 3

Form Absolute Value

In the last chapter we considered a pair of Hilbert spaces that arose from considering the Cartesian decomposition of a sesquilinear form on a complex reflexive Banach space V, defined by an operator $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$, including the "energy" Hilbert space defined by taking the inner product to be the real part H of A. In this chapter we investigate a new absolute value $|A|$ of A that arises as the symmetric factor of a special polar decomposition of A. Properties hold for the associated norm $\|\cdot\|_{|A|}$ that generalize useful properties of the energy norm (to which it reduces) in the positive symmetric case, without degrading when A is highly nonsymmetric.

Consider the case when $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ defines a symmetric and coercive form (i.e., when A is symmetric positive definite in the matrix case). The boundedness and positivity of A can be written in the form

$$\begin{aligned} |(Au, v)| &\le C\|u\|_{\mathsf{V}}\|v\|_{\mathsf{V}} \quad \text{for all } u, v \in \mathsf{V}, \quad \text{and} \\ (Au, u) &\ge c\|u\|_{\mathsf{V}}^2 \qquad\quad \text{for all } u \in \mathsf{V}, \end{aligned} \tag{3.1}$$

for some $C, c > 0$. Note, as in the previous chapter, we will reserve the notation $(\cdot, \cdot)$ for the duality pairing between a space and its adjoint. The same bounds hold, with different constants, for any norm equivalent to the Banach norm (any norm in the matrix case). It is easy to see that the "energy" norm $\|\cdot\|_A$ defined by $\|u\|_A^2 := (Au, u)$ is the unique norm for which one can take $C = c = 1$. It is this property that causes the energy norm to find so many applications.

In the more general case, letting $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ be nonsymmetric, suppose that A has a bounded inverse $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$ (in the matrix case this just means A is nonsingular). Then A is bounded and *weakly* coercive,

$$|(Au, v)| \le C\|u\|_{\mathsf{V}}\|v\|_{\mathsf{V}} \qquad \text{for all } u, v \in \mathsf{V}, \quad \text{and} \tag{3.2}$$

$$\sup_{v \in \mathsf{V}\setminus\{0\}} \frac{|(Au, v)|}{\|v\|_{\mathsf{V}}} \ge c\|u\|_{\mathsf{V}} \qquad \text{for all } u \in \mathsf{V}. \tag{3.3}$$

This statement is the converse of the Babuška-Lax-Milgram theorem generalized to reflexive Banach spaces, and was proven by Roşca [44]. Again, the same inequalities hold with different constants for any equivalent norm. A norm for which one can take $C = c = 1$ is a kind of generalization of the energy norm to the nonsymmetric case. When A is symmetric, we recover the usual energy norm, as in this case the supremum in the second inequality is attained for $v = u$, so that weak and normal coercivity are equivalent. We shall see that this generalization of the energy norm exists under certain conditions and can be expressed as $\|\cdot\|_X$ for some positive symmetric operator $X \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$, which can naturally be regarded as an absolute value of A, occurring as the positive factor of a special polar decomposition of A.

The polar decomposition of a sesquilinear form on a Hilbert space does not appear to be a standard concept. There is an obvious way to define it, however, that, under the standard identification between sesquilinear forms and operators on a Hilbert space, agrees with the usual polar decomposition of an operator. The absolute value of a Hilbert space operator is the positive factor of its polar decomposition. The corresponding sesquilinear form is the obvious choice for the definition of absolute value for a sesquilinear form. We will review the construction in the first section below.

Note, however, that we have not assumed a Hilbert space structure on V. Which Hilbert space should be used for the polar decomposition? For our PDE examples, do we use the Sobolev space H^1, or perhaps L^2 (on which our forms are unbounded)? Should we use the Euclidean inner product in the matrix case? None of these choices (which would have been "canonical" if V had been assumed to be a Hilbert space) in general lead to $C = c = 1$ in (3.2) and (3.3). A more interesting choice is to use the inner product associated with the positive factor of the polar decomposition itself. As will be shown, this apparently circular choice of inner product is equivalent to the induced norm satisfying (3.2) and (3.3) with $C = c = 1$, i.e., being the generalization of the energy norm described above. The "new" absolute value $|A|$ we define as the positive factor of this special polar decomposition, and we will refer to it as the *form* absolute value. The special choice of inner product that defines the form absolute value does not appear to have been investigated before. It is in this sense that it is new. But note that it remains an instance, albeit a remarkable one, of the standard Hilbert space concept. Other authors have considered, on the other hand, *generalizations* of the polar decomposition to spaces in which the scalar product is indefinite (e.g., [2]) or even non-Hermitian (e.g., [27]).

The apparent circularity in the definition is also what makes this absolute value interesting to investigate. For example, once a Hilbert space has been given, the polar decomposition always exists and is unique. It is far from obvious, however, whether the special choice of inner product of the form absolute value can always be made, or that it can be made in only one way. Indeed, we demonstrate that the form absolute value of a general form may not exist, and may not be unique if it does.

This chapter is organized as follows. We start in §3.1 by reviewing the polar decomposition in a given Hilbert space. We single out the remarkable polar decomposition just described to define the form absolute value $|A|$, before considering alternate characterizations in §3.2. In §3.3, we focus on the case when the real part of A is positive, and we show that $|A|$ exists and is unique in this case. We go on to characterize existence and uniqueness for general matrices in §3.4. Illustrative examples are examined in §3.5, and computational issues are investigated in §3.6.

3.1 Polar Decomposition of a Form

Let V be a complex reflexive Banach space. Recall from §2.1 that every bounded sesquilinear form corresponds uniquely to an operator $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$. Assume such an A is given, and assume A has bounded inverse $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$. Further, suppose we have symmetric positive $X \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$, so that it defines an inner product making V into a Hilbert space, as per §2.3. Let us denote this Hilbert space by H_X and its inner product by $\langle u, v\rangle_X := (Xu, v)$. Since

$$(Au, v) = \langle X^{-1}Au, v\rangle_X = \langle u, X^{-1}A^*v\rangle_X \tag{3.4}$$

for all $u, v \in \mathsf{V}$, we see that the form given by A corresponds to the operator $S = X^{-1}A$ on H_X, and that the adjoint of this operator is $T = X^{-1}A^*$. Note that $T \in \mathscr{B}(\mathsf{V})$ is not given by $S^* = A^*X^{-1} \in \mathscr{B}(\mathsf{V}^*)$, but is related to it by the Riesz mapping X, $T = X^{-1}S^*X$. This apparent discrepancy, which makes the notation somewhat awkward, is due to not having identified the Hilbert space with its dual.

The polar decomposition of the operator S on H_X is given by

$$S = U|S| = |T|U, \quad |S| = (TS)^{1/2}, \quad |T| = (ST)^{1/2} \tag{3.5}$$

where the square roots are the unique positive selfadjoint ones, so, e.g., for all $u, v \in \mathsf{V}$,

$$\langle |S|u, v\rangle_X = \langle u, |S|v\rangle_X, \qquad \langle |S|u, u\rangle_X \geq c\|u\|^2 \text{ for some } c > 0. \tag{3.6}$$

The factor U is a unitary operator on H_X, so that

$$U^*XU = X, \tag{3.7}$$

i.e., U preserves the inner product defined by X. Note, had we not assumed A had a bounded inverse, U would in general just be a partial isometry. For details of this standard polar decomposition construction, see e.g., Kato [31, p. 334].

It seems natural to define

$$|A|_X := X(X^{-1}A^*X^{-1}A)^{1/2} = X(TS)^{1/2} = X|S| \tag{3.8}$$

so that $(|A|_X u, v) = \langle |S|u, v\rangle_X$. Since $|S|$ is a positive selfadjoint operator on H_X, it follows that $|A|_X$ is positive symmetric, $0 < |A|_X = |A|_X^*$. We have $A = XU|S| = U^{-*}X|S|$ and $A = X|T|U$. Thus we have the polar decomposition of A

$$A = U_X^{-*}|A|_X = |A^*|_X U_X, \tag{3.9}$$

where we have added a subscript on U to make its dependence on X, which determines the Hilbert space, explicit.

As noted in the introduction, the polar decomposition of a sesquilinear form on a Hilbert space does not appear to be a standard concept, but equations (3.8) and (3.9) give the obvious definition, linked to the standard polar decomposition of a Hilbert space operator via the standard identification between bounded operators and bounded sesquilinear forms on a Hilbert space.

3.2 The Form Absolute Value

Whenever V can be given a Hilbert space structure, (3.9) defines an infinite family of polar decompositions of A, indexed by the choice of inner product specified by X. In the matrix setting, for example, we recover among these the "plain" polar decomposition when X is the identity matrix.

It is sometimes possible to find X such that

$$|A|_X = X. \tag{3.10}$$

For example, we shall see in the next section that such an X exists (and is unique) if the real part H of A is positive. Whenever such X can be found, the unitary factor U_X preserves the inner product defined by the positive factor $|A|_X$. This is in sharp contrast with the usual case, in which the inner product is specified up front. As

$|A|_X = X$ is the choice that most interests us, since it leads to a number of remarkable properties, and since we will be using it exclusively in subsequent chapters, it is convenient to reserve the unsubscripted notation $|A|$ for it, and to call it simply an absolute value of A. This leads to the following definition.

Definition. *Given $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ with bounded inverse $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$, we call $|A|$ an absolute value of A whenever $0 < |A| = |A|^*$, i.e., $|A|$ defines a coercive symmetric form, and when*

$$A = |A|U \quad \textit{where} \quad U^*|A|U = |A| \quad \textit{for some } U \in \mathscr{B}(\mathsf{V}). \tag{3.11}$$

This is a nonstandard definition of absolute value in that it singles out a rather particular set of Hilbert spaces. Again, let us emphasize that this definition is being taken for convenience, as we will be referring to it with some frequency. On the other hand, given the "naturalness" that shall be demonstrated for this choice of Hilbert space, there is a case to be made that this is a natural definition for the polar decomposition of a sesquilinear form on a *Banach* space, i.e., when no Hilbert space structure has been assumed a priori. In light of this, when we need to distinguish the above definition from other absolute values, we shall refer to it as the *form* absolute value.

The following list gives equivalent characterizations of $|A|$, any of which could have been taken as a definition.

Theorem 3.1. *$|A| \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ is an absolute value of A whenever $|A|$ is positive symmetric and any of the following equivalent conditions hold.*

(*i*) $A = |A|U$ *or* $A = U^{-*}|A|$ *where* $U^*|A|U = |A|$ *for some* $U \in \mathscr{B}(\mathsf{V})$

(*ii*) $A^*|A|^{-1}A = |A|$ *or* $A|A|^{-1}A^* = |A|$

(*iii*)

$$\begin{aligned} |(Au, v)| &\le \|u\|_{|A|}\|v\|_{|A|} && \textit{for all } u, v \in \mathsf{V}, \quad \textit{and} \\ \sup_{v \in \mathsf{V}\setminus\{0\}} \frac{|(Au, v)|}{\|v\|_{|A|}} &\ge \|u\|_{|A|} && \textit{for all } u \in \mathsf{V}. \end{aligned}$$

Before demonstrating the asserted equivalences, let us make some remarks. In the matrix setting ($\mathsf{V} = \mathbb{C}^n$), the definition above conflicts with standard usage, as $|A|$ usually refers to the absolute value associated with the Euclidean inner product, given by $|A| = (A^*A)^{1/2}$. Note that, like $(A^*A)^{1/2}$, the characterizations (i) and (ii) above

Standard $\mathscr{B}(\mathsf{H})$	Scalar $\mathbb{C}$	Form $\mathscr{B}(\mathsf{V}, \mathsf{V}^*)$
$A = U\|A\| = \|A^*\|U$	$z = e^{i\theta}\|z\| = \|z\|e^{i\theta}$	$A = U^{-*}\|A\| = \|A\|U$
$0 \leq \|A\| = \|A\|^*$	$0 \leq \|z\| = \overline{\|z\|}$	$0 \leq \|A\| = \|A\|^*$
$\|Uu\|_{\mathsf{H}} = \|u\|_{\mathsf{H}}$	$\|e^{i\theta}u\| = \|u\|$	$\|Uu\|_{\|A\|} = \|u\|_{\|A\|}$
$\|A\| := (A^*A)^{1/2}$	$\|z\| = (\overline{z}z)^{1/2}$	—
—	$\overline{z}\|z\|^{-1}z = \|z\|$	$A^*\|A\|^{-1}A = \|A\|$
—	$\|\overline{z}\| = \|z\|$	$\|A^*\| = \|A\|$
$\|A^{-1}\| = \|A^*\|^{-1}$	$\|z^{-1}\| = \|z\|^{-1}$	$\|A^{-1}\| = \|A\|^{-1}$

Table 3.1: Comparisons of the Hilbert space, scalar, and form absolute values

both reduce to the familiar scalar absolute value when A is a scalar ($\mathsf{V} = \mathbb{C}$). That is, we have two different generalizations of the scalar absolute value, as illustrated in Table 3.1. In the more general abstract setting we have been considering, A^*A is not even a well-formed expression: the codomain of A is V^* while the domain of A^* is V.

From (i), we see that $A^{-1}A^* = U^{-2}$ and so

$$|A| = AU^{-1} = A(A^{-1}A^*)^{1/2} \tag{3.12}$$

for *some* square root of $A^{-1}A^*$, if $|A|$ is an absolute value of A. Under certain conditions in the matrix setting, this expression solves the Riccati equation to which (ii) is equivalent [26, p. 44]. However, it is far from clear which square root(s) would make $|A|$ positive symmetric, or if any such square root exists at all. As a trivial example, when A is the scalar -1, we have $A^{-1}A^* = 1$, and the "correct" square root to take is -1, not the principal one. Perhaps surprisingly, the constraint that $|A|$ be positive does not in general single out a particular square root, and in general $|A|$ is not unique. In the next section, we shall show that $|A|$ exists and is unique whenever A has positive real part. In this case, in fact, the above formula holds with the square root taken to be the principal one. We give a full characterization of existence and uniqueness in the matrix case in §3.4. For the remainder of this section, we will fix $|A|$ as one particular absolute value, assumed to exist.

Let us now show that conditions (i)–(iii) are indeed equivalent. First, note that both $|A|$ and U have bounded inverses: the Lax-Milgram lemma (Theorem 2.1) applies to $|A|$ while $U^{-1} = A^{-1}|A|$. Assuming $U^*|A|U = |A|$, then

$$A = |A|U \Leftrightarrow A = U^{-*}(U^*|A|U) \Leftrightarrow A = U^{-*}|A|, \tag{3.13}$$

and either equation in (i) implies the other. For (ii), we have

$$A^*|A|^{-1}A = |A| \Leftrightarrow A^{-1}|A|A^{-*} = |A|^{-1} \Leftrightarrow |A| = A|A|^{-1}A^*, \tag{3.14}$$

and again either equation implies the other.

Now, assuming (i), we have

$$A^*|A|^{-1}A = A^*U = U^*|A|U = |A| \tag{3.15}$$

so that (i) $\Rightarrow$ (ii). Assuming (ii), we can choose $U := |A|^{-1}A$. Then,

$$U^*|A|U = A^*|A|^{-1}|A||A|^{-1}A = A^*|A|^{-1}A = |A|, \tag{3.16}$$

and thus (ii) $\Rightarrow$ (i).

We have already discussed condition (i), the first half of which is the definition we took for the form absolute value $|A|$, in which $|A|$ appears as a factor in a special polar decomposition. Condition (ii) is equivalent to the statement that

$$\|Au\|_{|A|^{-1}} = \|A^*u\|_{|A|^{-1}} = \|u\|_{|A|} \tag{3.17}$$

for all $u \in \mathsf{V}$. Recall from §2.3 that $\|\cdot\|_{|A|^{-1}}$ is the norm dual to $\|\cdot\|_{|A|}$:

$$\|f\|_{|A|^{-1}} = \sup_{u \neq 0} \frac{|(f,u)|}{\|u\|_{|A|}}. \tag{3.18}$$

So, (3.17) states that A and A^* are isometric mappings from the Hilbert space associated with $|A|$ to the dual Hilbert space. If we apply (3.18) to (3.17), we see that (3.17) is equivalent to

$$\sup_{v \neq 0} \frac{|(Au,v)|}{\|v\|_{|A|}} = \|u\|_{|A|}, \tag{3.19}$$

which is equivalent to the two inequalities of (iii), proving (ii) $\Leftrightarrow$ (iii). The first of these inequalities resembles a Cauchy-Schwarz inequality,

$$|(Au,v)| \leq \|u\|_{|A|}\|v\|_{|A|}, \tag{3.20}$$

except that the form defined by A is not symmetric. This inequality can also be worked out directly from the polar decomposition of (i). From this exercise, one sees that equality holds in (3.20) precisely when v is a scalar multiple of Uu.

When A is positive symmetric, it is obvious from (ii) that $|A| = A$ is one absolute value, and later we shall prove that it is the only one. In this case, (3.17) and (iii) reduce to familiar properties of the energy norm $\|\cdot\|_A$ and its dual that characterize their natural association with A. In the general case, the norm $\|\cdot\|_{|A|}$ and its dual $\|\cdot\|_{|A|^{-1}}$ are naturally associated with the nonsymmetric form A in the same way. In this sense $\|\cdot\|_{|A|}$ generalizes the energy norm.

Let us close this section by considering absolute values of some forms related to A.

Theorem 3.2. *If $|A|$ is an absolute value of $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$, then particular absolute values of A^*, A^{-1}, αA, and Q^*AQ are given by*

$$\begin{aligned} |A^*| &= |A|, \\ |A^{-1}| &= |A|^{-1}, \\ |\alpha A| &= |\alpha||A|, \\ |Q^*AQ| &= Q^*|A|Q, \end{aligned} \tag{3.21}$$

where α is an arbitrary non-zero (complex) scalar, and $Q \in \mathscr{B}(\mathsf{W}, \mathsf{V})$ is an arbitrary bounded operator from a complex reflexive Banach space W to V with a bounded inverse $Q^{-1} \in \mathscr{B}(\mathsf{V}, \mathsf{W})$.

The theorem is easily verified in each case by checking condition (ii) of Theorem 3.1. Note that, as indicated in Table 3.1, $|A^*| = |A|$ generalizes a property that holds in the scalar case but fails to hold in general for the standard absolute value of a Hilbert space operator, in which case $|A^*| = |A|$ holds if and only if A is normal. One nice implication of $|A^*| = |A|$ is that the generalized energy norm $\|\cdot\|_{|A|}$ is equally suited to the problem governed by A as to the adjoint problem governed by A^*.

Congruence transformations are the natural change of basis transformations for sesquilinear forms,

$$(AQu, Qv) = (Q^*AQu, v), \tag{3.22}$$

and so it is quite natural that the form absolute value be invariant under them.

3.3 Positive Case

In this section, we focus on the case where the real part H of $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ is positive. In this case there is a unique absolute value, given by the geometric mean of H and $M = A^*H^{-1}A$. Recall that H and M are defined by (2.10) and (2.31) and were discussed in the previous chapter.

Theorem 3.3. *If the real part H of $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ is positive, then $|A|$ exists, is unique, and is equal to the geometric mean of H and M, given by*

$$|A| = H(H^{-1}M)^{1/2}, \tag{3.23}$$

where $(H^{-1}M)^{1/2}$ denotes the unique positive square root.

The geometric mean of two positive sesquilinear forms was introduced by Pusz and Woronowicz [42]. Perhaps more widely known is the related concept of the geometric mean of two positive operators on a Hilbert space, studied by, for example, by Kubo and Ando [33]. The two definitions are equivalent when applied to two positive Hermitian matrices. An accessible introduction to the matrix sub-case is the article by Lawson and Lim [34]. It is also discussed in Higham's book on functions of matrices [26, p. 46]. In the scalar case (3.23) reduces to the familiar geometric mean of two scalars, and the theorem is easily verified. To prove the general case, we make use of Hilbert space theory.

Proof. Define $T := H^{-1}M$. T is positive selfadjoint as an operator on the Hilbert space H_H formed from the inner product $\langle u, v\rangle_H := (Hu, v)$ for $u, v \in \mathsf{V}$;

$$\langle Tu, v\rangle_H = (Mu, v) = \langle u, Tv\rangle_H, \qquad \langle Tu, u\rangle_H = \|u\|_M^2 \geq \|u\|_H^2. \tag{3.24}$$

There is precisely one nonnegative selfadjoint operator $T^{1/2}$ such that $(T^{1/2})^2 = T$ [31, p. 281], and further [31, p. 282]

$$\langle T^{1/2}u, u\rangle_H \geq \|u\|_H^2. \tag{3.25}$$

To prove existence, we shall show that $|A| = HT^{1/2}$ satisfies condition (ii) of Theorem 3.1, $A^*|A|^{-1}A = |A|$. That $|A| = |A|^*$ follows from the selfadjointness of $T^{1/2}$.

$$(|A|u, v) = \langle T^{1/2}u, v\rangle_H = \langle u, T^{1/2}v\rangle_H = (u, |A|v). \tag{3.26}$$

The positivity of $|A|$ follows from (3.25), which becomes

$$\|u\|_H^2 \leq (|A|u, u) = \|u\|_{|A|}^2. \tag{3.27}$$

The operator $H^{-1}A^*|A|^{-1}A$ is also positive selfadjoint on H_H.

$$\langle H^{-1}A^*|A|^{-1}Au, v\rangle_H = (|A|^{-1}Au, Av) = \langle u, H^{-1}A^*|A|^{-1}Av\rangle_H. \tag{3.28}$$

Further, its square satisfies

$$\begin{aligned}(H^{-1}A^*|A|^{-1}A)^2 &= (H^{-1}A^*T^{-1/2}H^{-1}A)^2 \\ &= H^{-1}A^*T^{-1/2}(H^{-1}M)T^{-1/2}H^{-1}A \\ &= H^{-1}A^*H^{-1}A = H^{-1}M = T.\end{aligned} \tag{3.29}$$

As T has exactly one positive selfadjoint square root, we conclude that $H^{-1}A^*|A|^{-1}A = T^{1/2}$, and thus $A^*|A|^{-1}A = |A|$.

To prove uniqueness, let $|A|$ be any absolute value of A and define $G := H^{-1}|A|$. Let us first show that $A^*|A|^{-1}A = |A|$ implies $G^2 = T$. We have

$$\begin{aligned} |A|^{-1}A|A|^{-1} &= A^{-*}, \quad \text{and} \\ |A|^{-1}A^*|A|^{-1} &= A^{-1}. \end{aligned} \tag{3.30}$$

Averaging these equations, we see that

$$|A|^{-1}H|A|^{-1} = M^{-1}, \quad \text{so} \quad |A|H^{-1}|A| = M, \tag{3.31}$$

and thus $G^2 = T$. Now G is a positive selfadjoint operator on H_H,

$$\langle Gu, v\rangle_H = (|A|u, v) = \langle u, Gv\rangle_H, \qquad \langle Gu, u\rangle_H = \|u\|^2_{|A|}, \tag{3.32}$$

and since T has exactly one positive selfadjoint square root, it follows that $G = T^{1/2}$, proving uniqueness. □

The observant reader may have noticed that equation (3.23) expressing $|A|$ as the geometric mean of H and M also identifies it as $|A|_H$, the positive factor of the polar decomposition of A with respect to the H inner product, defined in (3.8). It also holds that $|A| = |A|_M$, though we omit the proof. Accordingly, $U = |A|^{-1}A$ is unitary in each of the corresponding Hilbert spaces. To see this more directly, recall condition (i) of Theorem 3.1,

$$A = |A|U = U^{-*}|A|, \qquad U^*|A|U = |A|. \tag{3.33}$$

Multiplying the second equation on the right by U, we see that

$$U^*AU = A, \quad \text{and so also} \quad UA^{-1}U^* = A^{-1}. \tag{3.34}$$

Averaging each equation with its adjoint, and then inverting the second, produces

$$U^*HU = H \quad \text{and} \quad U^*MU = M. \tag{3.35}$$

Interestingly, the $|A|$ norm is bounded by the geometric mean of the H and M norms. Recall from equation (2.38) that

$$|(Au, v)| \le \|u\|_M \|v\|_H \tag{3.36}$$

for all $u, v \in \mathsf{V}$. Now take $v = Uu$. Then, since $U^*A = |A|$,

$$|(Au, Uu)| = \|u\|^2_{|A|} \le \|u\|_M \|Uu\|_H = \|u\|_M \|u\|_H \tag{3.37}$$

for all $u \in \mathsf{V}$. Combining this result with (3.27) and (2.35), we see that

$$\|u\|_H \le \|u\|_{|A|} \le \|u\|_H^{1/2} \|u\|_M^{1/2} \le (\sec\theta)^{1/2} \|u\|_H, \tag{3.38}$$

where θ is the sectorial half-angle of A (see §2.4).

In fact, a similar geometric mean bound holds for operators.

Theorem 3.4. *Given* $L \in \mathscr{B}(\mathsf{V}_1, \mathsf{V}_2), A_1 \in \mathscr{B}(\mathsf{V}_1, \mathsf{V}_1^*), A_2 \in \mathscr{B}(\mathsf{V}_2, \mathsf{V}_2^*)$, *let* $H_i := \frac{1}{2}(A_i + A_i^*), M_i := A_i^* H_i^{-1} A_i, i = 1, 2$ *and suppose* H_1 *and* H_2 *are positive. Then,*

$$\|L\|_{|A_2|,|A_1|} \leq \|L\|_{H_2,H_1}^{1/2} \|L\|_{M_2,M_1}^{1/2}. \tag{3.39}$$

Before we give the proof, let us remark that this result identifies the Hilbert space associated with $|A|$ as an exact interpolation space, of exponent $\frac{1}{2}$, between the spaces associated with H and M. For a discussion of interpolation space theory applied to Hilbert spaces, see Donoghue [14] and McCarthy [37]. The use of interpolation space theory is unusual here in that the interpolating spaces have the same topology, whereas the usual motivation is to find a space with a topology intermediate between two different topologies (e.g., between two Sobolev spaces). The proof below is essentially an application of McCarthy's general existence proof for such a space [37].

Proof. Let $\langle u, v\rangle_{|A_1|} := (|A_1|u, v)$.

$$\|L\|_{|A_2|,|A_1|}^2 = \sup_{u \in \mathsf{V}_1 \setminus \{0\}} \frac{(|A_2|Lu, Lu)}{(|A_1|u, u)} = \sup_{u \in \mathsf{V}_1 \setminus \{0\}} \frac{\langle |A_1|^{-1} L^* |A_2| Lu, u\rangle_{|A_1|}}{\langle u, u\rangle_{|A_1|}} \tag{3.40}$$

The convex hull of the numerical range of a selfadjoint operator on a Hilbert space is the closure of the convex hull of its spectrum, and $|A_1|^{-1}L^*|A_2|L$ is nonnegative selfadjoint as an operator on the Hilbert space associated with $|A_1|$. So, letting $\rho(\cdot)$ denote the spectral radius,

$$\begin{aligned}
\|L\|_{|A_2|,|A_1|}^2 &= \rho(|A_1|^{-1} L^* |A_2| L) \\
&\leq \||A_1|^{-1} L^* |A_2| L\|_{H_1} \\
&\leq \||A_1|^{-1} L^* |A_2|\|_{H_1,H_2} \|L\|_{H_2,H_1} \\
&= \||A_2| L |A_1|^{-1}\|_{H_2^{-1},H_1^{-1}} \|L\|_{H_2,H_1} \\
&= \|L\|_{M_2,M_1} \|L\|_{H_2,H_1},
\end{aligned} \tag{3.41}$$

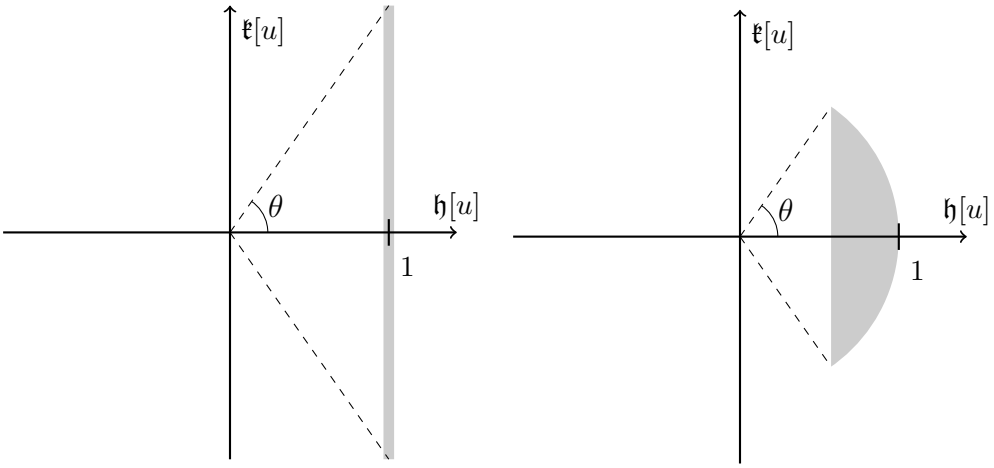
where for the last two steps we have used (2.26) and the fact that $\||A_i|u\|_{H_i^{-1}} = \|u\|_{M_i}$, which follows directly from (3.31). □

A visual comparison of the H and $|A|$ norms is made in Figure 3.1. In it, we see shaded regions indicating the values that the quadratic form $\mathfrak{a}[u] = (Au, u)$ can possibly take in the complex plane, when u is normalized to have unit H or $|A|$ norm. The numerical range of the form $\mathfrak{a}$ defined by A, considered as a sesquilinear form on the Hilbert spaces associated with H and $|A|$, is necessarily a convex subset of the corresponding shaded regions. Note that these numerical ranges are the same as those of the operators $H^{-1}A = 1 + iH^{-1}K$ and $|A|^{-1}A = U$ on the respective Hilbert spaces,

and these numerical ranges must in turn contain the spectrum of these operators. That the two operators are quite different (and operate on different Hilbert spaces) explains why the shapes of the shaded regions are so different (containing only the point 1 in common). The first figure is repeated from §2.4, Figure 2.2. Figure 3.1b illustrates both that

$$\text{Re } \mathfrak{a}[u] = \mathfrak{h}[u] = \|u\|_H^2 \geq (\cos\theta)\|u\|_{|A|}^2, \tag{3.42}$$

which follows from (3.38), and that $|(Au, u)| \leq \|u\|_{|A|}^2$, which follows from the Cauchy-Schwarz type inequality (3.20).



(a) $\|u\|_H = 1$ (b) $\|u\|_{|A|} = 1$

Figure 3.1: Regions of the complex plane (shaded) in which the quadratic form $\mathfrak{a}[u] = (Au, u)$ can take values, for two different normalizations of u.

We close this section by considering whether the absolute value of a restriction of the form defined by A bears any relation to the corresponding restriction of $|A|$. In other words, is there any relationship between $|P^*AP|$ and $P^*|A|P$, where P is the injection from a closed subspace of V? Recall that, when P is not merely an injection, but rather an operator with bounded inverse (and hence surjective), these two expressions define identical forms, by Theorem 3.2. For general P, there is no simple relationship between these operators. However, the next proposition shows, at least, that the same bounds in terms of the H and M norms govern both equally.

Proposition 3.1. *Let $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ have coercive real part and let $P \in \mathscr{B}(\mathsf{W}, \mathsf{V})$ be a bounded operator from the complex reflexive Banach space W such that $\|Pu\| \geq c\|u\|$ holds for some $c > 0$ and all $u \in \mathsf{W}$. Then*

$$\|Pu\|_H \leq \frac{\|u\|_{P^*|A|P}}{\|u\|_{|P^*AP|}} \leq \|Pu\|_H^{1/2}\|Pu\|_M^{1/2} \tag{3.43}$$

for all $u \in \mathsf{W}$. That is, the indicated lower and upper bounds hold each for $\|u\|_{P^*|A|P}$ and $\|u\|_{|P^*AP|}$.

Proof. The bounds for $\|u\|_{P^*|A|P}$ result directly by applying (3.38) to Pu. The lower bound on P implies that the real part P^*HP of P^*AP is positive, so that we may apply (3.38) to P^*AP, producing

$$\|Pu\|_H \leq \|u\|_{|P^*AP|} \leq \|Pu\|_H^{1/2} \|P^*APu\|_{(P^*HP)^{-1}}^{1/2}, \tag{3.44}$$

where we have also expanded the analog of $M = A^*H^{-1}A$ for P^*AP. The result follows by noting that $\|P^*APu\|_{(P^*HP)^{-1}} \leq \|APu\|_{H^{-1}} = \|Pu\|_M$, which holds because

$$0 \leq \|H^{-1}f - P(P^*HP)^{-1}P^*f\|_H^2 = \|f\|_{H^{-1}}^2 - \|P^*f\|_{(P^*HP)^{-1}}^2 \tag{3.45}$$

for all $f \in \mathsf{V}^*$, and in particular for $f = APu$. □

3.4 Existence and Uniqueness

Given the satisfactory properties of the absolute value concept we have introduced in the case of positive real part, when uniqueness and existence are guaranteed, it is natural to wonder if the concept still applies in a more general setting. In general, the answer is no: as we shall demonstrate, the absolute value may not exist, and when it does, it is not always unique.

For convenience, we shall focus on the matrix case ($\mathsf{V} = \mathbb{C}^n$), and assume only that $A \in \mathbb{C}^{n\times n}$ is nonsingular. This suffices to establish the negative answers to the questions of existence and uniqueness.

The material in this section is somewhat technical and may skipped, as it is not needed in the remaining chapters, in which the absolute value will only be used in the case of positive real part. We include the material to complete the basic picture of the new form absolute value by settling the fundamental questions of existence and uniqueness. The precise conditions of when the form absolute value of a nonsingular matrix exists and when it is unique are given in the statements of Theorems 3.5 and 3.6.

Recall from Theorem 3.2 that the absolute value is invariant under congruence transforms: given a nonsingular matrix $Q \in \mathbb{C}^{n\times n}$, one absolute value of Q^*AQ is $Q^*|A|Q$. It follows that a matrix has an absolute value precisely when any member of its congruency class has one. Similarly, the absolute value of a matrix is unique precisely when any member of the congruency class of the matrix has a unique absolute value. The approach of this section will be to resolve the questions of existence and uniqueness by examining the canonical forms that generate the congruency classes.

3.4.1 Existence

It is sometimes possible to diagonalize a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ by congruence. Whenever this is possible, the congruence transformation can always be scaled so that the resulting diagonal entries have unit modulus,

$$Q^* A Q = D, \qquad D = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n}). \tag{3.46}$$

In the singular case, D may have zeros on the diagonal as well. The fact that the angles θ_i and number of zeros are canonical (uniquely determined by A up to ordering) was shown by Johnson and Furtado [29]. When $A = A^*$ this fact reduces to Sylvester's law of inertia. Johnson and Furtado call matrices that are diagonalizable by congruence unitoid. In the general case, there exists a canonical form analogous to the Jordan form, which has been worked out by Horn and Sergeichuck [28]. The class of unitoid matrices is precisely the class of matrices that have absolute values.

Theorem 3.5. *An absolute value of nonsingular $A \in \mathbb{C}^{n\times n}$ exists if and only if A is diagonalizable by congruence.*

Proof. Suppose that A is diagonalizable by congruence, so that (3.46) holds. One absolute value of D is given by the identity matrix, since $D^* I^{-1} D = I$. Applying Theorem 3.2, it follows that one absolute value of $A = Q^{-*} D Q^{-1}$ is $|A| = Q^{-*} Q^{-1}$. The corresponding polar decomposition for A is

$$A = Q^{-*} D Q^{-1} = \underbrace{(Q^{-*} Q^{-1})}_{|A|} \underbrace{(Q D Q^{-1})}_{U} = \underbrace{(Q^{-*} D Q^*)}_{U^{-*}} \underbrace{(Q^{-*} Q^{-1})}_{|A|}. \tag{3.47}$$

Conversely, suppose that nonsingular A has an absolute value $|A|$,

$$A = |A| U = U^{-*} |A|, \qquad U^* |A| U = |A|. \tag{3.48}$$

The matrix U is a unitary operator on the Hilbert space associated with the $|A|$ inner product. It follows that U must be diagonalizable with eigenvalues on the unit circle, and moreover that it is always possible to find an orthonormal set of eigenvectors,

$$UQ = QD, \qquad Q^* |A| Q = I. \tag{3.49}$$

Here the columns of Q are the orthonormal eigenvectors (orthonormal with respect to the $|A|$ inner product), and D is the diagonal matrix of eigenvalues, each having unit modulus. We have

$$Q^* A Q = Q^* |A| U Q = Q^* |A| Q D = D, \tag{3.50}$$

and so A is diagonalizable by congruence. □

In their paper, Johnson and Furtado [29] remark that the canonical angles of a matrix appear to be unrelated to the eigenvalues of the unitary factor in the polar decomposition (with respect to the Euclidean inner product). On the other hand, we see from (3.47) that the canonical angles are precisely related to the eigenvalues of the unitary factor in the special polar decomposition that defines $|A|$.

While every square matrix is arbitrarily close to one that is diagonalizable by similarity transformations, it would be an error to transfer this intuition over to congruence transformations. An example of a matrix that is not diagonalizable by congruence, and thus has no absolute value, is

$$\begin{bmatrix} 0 & 1 \\ \mu & 0 \end{bmatrix} \tag{3.51}$$

with $|\mu| > 1$; this is the simplest instance of one of the canonical block types proposed by Horn and Sergeichuck. In any metric, the nearest diagonalizable matrix is a positive distance away. The class of matrices for which an absolute value exists is thus fairly restrictive: it excludes all matrices whose canonical forms contain a block of this type.

The diagonal canonical form also allows us to prove the following result about the positive case investigated in §3.3.

Proposition 3.2. *If $A \in \mathbb{C}^{n\times n}$ has positive Hermitian part $0 < H = \frac{1}{2}(A + A^*)$, then the unique absolute value of A is*

$$|A| = A(A^{-1}A^*)^{1/2} = A^*(A^{-*}A)^{1/2}, \tag{3.52}$$

where the square roots taken are the principal ones.

Proof. Existence and uniqueness of $|A|$ was established in Theorem 3.3. We need to show that $U = |A|^{-1}A$ is the principal square root of $A^{-*}A$ and that U^{-1} is the principal square root of $A^{-1}A^*$.

By the previous theorem, (3.46) holds for some nonsingular Q. Since $Q^*HQ = \frac{1}{2}(D + D^*) = \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_n)$ is positive, it follows that $-\pi/2 < \theta_i < \pi/2$ holds for each canonical angle θ_i. Now

$$A^{-*}A = QD^2Q^{-1} \quad \text{and} \quad A^{-1}A^* = QD^{-2}Q^{-1} \tag{3.53}$$

have square roots

$$U = QDQ^{-1} \quad \text{and} \quad U^{-1} = QD^{-1}Q^{-1}. \tag{3.54}$$

Because the eigenvalues of U and U^{-1} all lie in the right half plane, these must be the principal square roots [26, p. 20]. □

The formulae of this last theorem closely resemble the definition of the geometric mean of two positive Hermitian matrices, except, of course, that A and A^* are Hermitian only in the trivial case. Intuitively, however, they suggest that $|A|$ lies halfway between A and A^*.

3.4.2 Uniqueness

Clearly, one absolute value of

$$E_{m,n} = \begin{bmatrix} -I_m & \\ & I_n \end{bmatrix} \tag{3.55}$$

is the identity matrix I_{m+n}. Recall from Theorem 3.2 that, provided Q is nonsingular, $|Q^*E_{m,n}Q| = Q^*|E_{m,n}|Q = Q^*Q$ is an absolute value of $Q^*E_{m,n}Q$. There is a set of matrices Q—namely the pseudo-unitary group $U(n,m)$—that preserve the indefinite form $E_{m,n}$, $Q^*E_{m,n}Q = E_{m,n}$. This implies that, for any such Q, Q^*Q is an absolute value of $E_{m,n}$. Now, since the unitary group $U(n+m)$ of matrices for which $Q^*Q = I_{m+n}$ is quite different in general from the pseudo-unitary group, we see that there are many absolute values of $E_{m,n}$, unless $m = 0$ or $n = 0$, i.e., in the degenerate cases of $E_{m,n}$ being the identity matrix or its additive inverse.

We can explicitly describe all absolute values of $E_{m,n}$ as follows. Without loss of generality, assume $n \geq m$ (otherwise one may consider a permutation of $-E_{m,n}$). Let $B \in \mathbb{C}^{m\times n}$ be arbitrary. We shall see that each choice of B leads to a different absolute value of $E_{m,n}$, and conversely that each absolute value corresponds to some choice of B. Let the singular value decomposition of B be

$$B = U\Sigma V^* = U \begin{bmatrix} \sinh 2\Theta & O \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^* \tag{3.56}$$

where $U^*U = I_m$, $V^*V = I_n$, Σ is rectangular diagonal with nonnegative real entries, and Θ is square diagonal. Note that, because sinh is a one-to-one mapping from the set of nonnegative reals onto itself, Θ also has real nonnegative entries. From the factors of its decomposition we can construct the hyperbolic rotation

$$Q = \begin{bmatrix} U & & \\ & V_1 & V_2 \end{bmatrix} \begin{bmatrix} \cosh\Theta & \sinh\Theta & \\ \sinh\Theta & \cosh\Theta & \\ & & I_{n-m} \end{bmatrix} \begin{bmatrix} U & & \\ & V_1 & V_2 \end{bmatrix}^*. \tag{3.57}$$

It is easily verified that $Q^*E_{m,n}Q = E_{m,n}$, i.e., that Q is a member of the pseudo-unitary group (and hence nonsingular). Therefore, one absolute value of $E_{m,n}$ is Q^*Q, which is given in terms of B by

$$X_B = Q^*Q = \begin{bmatrix} (I_m + BB^*)^{1/2} & B \\ B^* & (I_n + B^*B)^{1/2} \end{bmatrix}, \tag{3.58}$$

as is readily verified. One may also check condition (ii) of Theorem 3.1 directly: one finds that $X_B^{-1} = X_{-B}$ and $E_{m,n}^* X_{-B} E_{m,n} = X_B$. Conversely, it is straightforward to verify that *all* absolute values of $E = E_{m,n}$ take the form X_B for some B. This follows from expanding the equation $E^{-*}|E|E^{-1}|E| = I$, implied by condition (ii) of Theorem 3.1, in block form.

Now let us consider the general case. Let A be nonsingular and diagonalizable by congruence, so that an absolute value of A exists. Further, let

$$Q^*AQ = D, \qquad D = \begin{bmatrix} e^{i\theta_1} E_{m_1,n_1} & & \\ & \ddots & \\ & & e^{i\theta_s} E_{m_s,n_s} \end{bmatrix} \tag{3.59}$$

where $-\pi/2 < \theta_k \leq \pi/2$ and each θ_k is distinct. This can always be arranged by a suitable reordering of the columns of Q. Let $Q^{-*}XQ^{-1}$ be any absolute value of A. Then $A^*QX^{-1}Q^*A = Q^{-*}XQ^{-1}$ so that

$$D^*X^{-1}D = X, \tag{3.60}$$

i.e., X is an absolute value of D. Since $D^*D = I$, we have $D = XDX$ and thus $XD^2 = XDXDX = D^2X$. The square of D is diagonal with block structure

$$D^2 = \begin{bmatrix} e^{2i\theta_1} I_{m_1+n_1} & & \\ & \ddots & \\ & & e^{2i\theta_s} I_{m_s+n_s} \end{bmatrix}. \tag{3.61}$$

Note that the restrictions on the angles θ_k ensures that each $e^{2i\theta_k}$ is distinct. Suppose the jth coordinate is in block k so that, letting q_j be the jth column of the identity matrix,

$$D^2 q_j = e^{2i\theta_k} q_j. \tag{3.62}$$

Multiplying on the left by X, and using the fact that X and D^2 commute, we see that

$$D^2(Xq_j) = e^{2i\theta_k}(Xq_j). \tag{3.63}$$

That is, the jth column of X is an eigenvector of D^2 with eigenvalue $e^{2i\theta_k}$. Consequently, the coordinates of Xq_j outside of block k must be zero, and so X must take on the block diagonal structure

$$X = \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_s \end{bmatrix}, \tag{3.64}$$

where X_k has size $m_k + n_k$ to match the block structures of D^2 and D. From here, we can conclude that X_k must be an absolute value of E_{m_k,n_k}, which we already know is unique if and only if either $m_k = 0$ or $n_k = 0$ (so that $E_{m_k,n_k} = I_{n_k}$ or $-I_{m_k}$). This proves the following theorem.

Theorem 3.6. *Nonsingular $A \in \mathbb{C}^{n\times n}$ has a unique absolute value $|A|$ if and only if*

- *A is diagonalizable by congruence, and*
- *A and $-A$ have no canonical angles in common.*

3.5 Examples

3.5.1 A 2 × 2 Matrix

For a concrete illustration of the concepts of this chapter consider the matrix

$$A = \begin{bmatrix} 15 & 24 \\ 0 & 15 \end{bmatrix}, \tag{3.65}$$

which has been artificially chosen so that the numbers work out nicely. Note that A is not diagonalizable—at least, not by similarity transforms (which is often the default sense of the term). This example nicely illustrates that this sense of diagonalizability is not the important one when A is regarded as defining a sesquilinear form (as opposed to an endomorphic operator). The Cartesian decomposition $A = H + iK$ and the matrix $M = A^*H^{-1}A$ are given in this case by

$$H = \begin{bmatrix} 15 & 12 \\ 12 & 15 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & -12i \\ 12i & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 41\frac{2}{3} & 33\frac{1}{3} \\ 33\frac{1}{3} & 41\frac{2}{3} \end{bmatrix} = \tfrac{25}{9}H. \tag{3.66}$$

Because H is symmetric positive definite, we know that $|A|$ exists and is unique. The geometric mean of H and M is particularly simple to work out in this case, and we find that the absolute value of A is

$$|A| = H(H^{-1}M)^{1/2} = \tfrac{5}{3}H = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}. \tag{3.67}$$

The factor U is then given by $U = |A|^{-1}A$. Alternatively, one may directly calculate

$$U^2 = A^{-*}A = \begin{bmatrix} 1 & \frac{8}{5} \\ -\frac{8}{5} & \frac{39}{25} \end{bmatrix}, \quad U = (A^{-*}A)^{1/2} = \begin{bmatrix} \frac{5}{3} & \frac{4}{3} \\ -\frac{4}{3} & -\frac{7}{15} \end{bmatrix}, \tag{3.68}$$

without first determining $|A|$. This then gives an alternate way to calculate $|A| = U^*A$, agreeing with what we had before. That U is unitary on $\mathsf{H}_{|A|}$, i.e., $U^*|A|U = |A|$,

is easily verified. Indeed, we have the eigendecomposition of U given by $UQ = QD$ where

$$Q = \frac{1}{15\sqrt{2}} \begin{bmatrix} 4+3i & 4-3i \\ -5 & -5 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \frac{3+4i}{5} & 0 \\ 0 & \frac{3-4i}{5} \end{bmatrix}. \tag{3.69}$$

Note that the eigenvalues of U (the diagonal elements of D) are of unit modulus, as must be the case for a unitary operator. We have also chosen normalized eigenvectors (the columns of Q) so that they form an orthonormal basis of $\mathsf{H}_{|A|}$: $Q^*|A|Q = I$. It follows that Q diagonalizes A by congruence, $Q^*AQ = D$, and we see that the canonical angles of A are $\tan^{-1} \pm\frac{4}{3} \approx \pm 53.1°$. The sectorial half-angle θ of A is the magnitude of the largest canonical angle, and this agrees with (2.28), $\tan\theta = \|K\|_{H^{-1},H}$.

3.5.2 1D Advection Diffusion

For an illustrative infinite-dimensional example, consider the one-dimensional advection diffusion operator A defined on the function space $\mathsf{V} = \{u \in H^1_{\mathrm{p}}[-\pi,\pi] : \langle 1, u\rangle = 0\}$, consisting of periodic functions on $[-\pi,\pi]$ with one weak derivative and zero mean. Let A be given by

$$(Au, v) = p\langle u', v'\rangle + c\langle u', v\rangle, \tag{3.70}$$

where $p > 0$, $c \in \mathbb{R}$, and where $\langle\cdot,\cdot\rangle$ on the right of the equation denotes the L^2 inner product on $[-\pi,\pi]$. Note that the mean constraint $\langle 1, u\rangle = 0$ removes what would otherwise be a one-dimensional null space of constant functions.

The adjoint and real part are given by

$$(A^*u, v) = p\langle u', v'\rangle - c\langle u', v\rangle, \quad (Hu, v) = p\langle u', v'\rangle. \tag{3.71}$$

Because $p > 0$, the form defined by H is coercive as a consequence of the Poincaré inequality (and using the zero mean constraint). Consequently, we know that $|A|$ exists and is unique. One approach to analyzing the absolute value is to look for eigenfunctions of U using the correspondence

$$U^2u = \lambda^2 u \quad \Leftrightarrow \quad |A|Uu = \lambda^2|A|U^{-1}u \quad \Leftrightarrow \quad Au = \lambda^2 A^*u. \tag{3.72}$$

Since, for sufficiently smooth u, $(Au, v) = \langle -pu'' + cu', v\rangle$ and $(A^*u, v) = \langle -pu'' - cu', v\rangle$, the infinitely smooth ansatz $u = e^{ikx}$ leads to

$$(pk^2 + ick)\langle e^{ikx}, v\rangle = \lambda^2(pk^2 - ick)\langle e^{ikx}, v\rangle \tag{3.73}$$

and we find

$$\lambda = \frac{pk^2 + ick}{|pk^2 + ick|} \tag{3.74}$$

where the square root was chosen to lie in the right half of the complex plane. As expected, the eigenvalues λ of U have unit modulus. In summary, the eigenfunctions and eigenvalues of U are given by

$$Ue^{ikx} = e^{i\theta_k}e^{ikx}, \quad \tan\theta_k = \frac{c}{pk} \tag{3.75}$$

for $k \in \mathbb{Z} \setminus \{0\}$. The tangent of each canonical angle is a Péclet number with the characteristic length being the inverse of the corresponding wave number. For the case $c = 1$ and $p = 1/4$, these eigenvalues are plotted in Figure 3.2. Also shown is the field of values $\{(Au, u) : \|u\|_{|A|} = 1\}$, which is precisely the convex hull of these eigenvalues. Note that the sectorial half-angle $\theta = |\theta_1|$ corresponds to the smallest wave number $k = 1$.

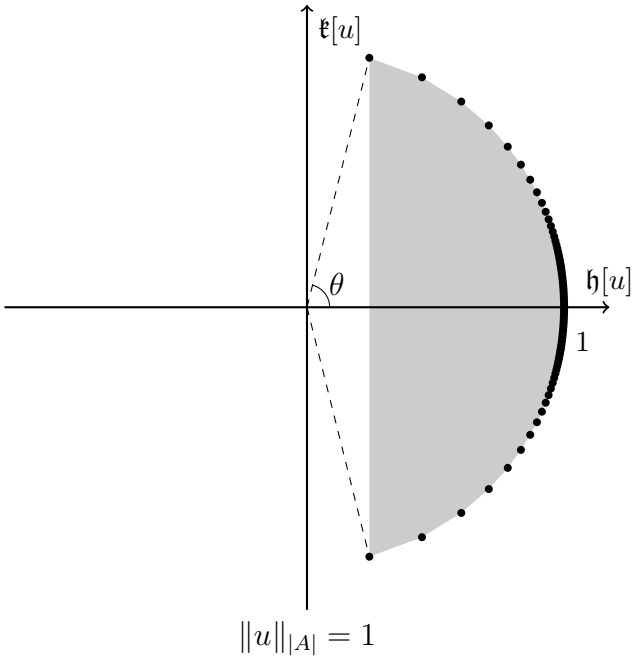


Figure 3.2: The field of values (shaded) of the form A on the Hilbert space $\mathsf{H}_{|A|}$ for the 1D convection diffusion example with $c = 1$ and $p = 1/4$, shown with the eigenvalues $e^{i\theta_k}$ of U.

The absolute value $|A|$ applied to one of the eigenfunctions of U is

$$\begin{aligned} |A|e^{ikx} &= AU^{-1}e^{ikx} \\ &= e^{-i\theta_k}Ae^{ikx} \\ &= e^{-i\theta_k}(pk^2 + ick)\langle e^{ikx}, \cdot\rangle \\ &= |pk^2 + ick|\langle e^{ikx}, \cdot\rangle. \end{aligned} \tag{3.76}$$

Here we are being careful to emphasize that $|A|e^{ikx}$ is a functional, an element of V^*. Since $\langle e^{imx}, e^{inx}\rangle = 2\pi\delta_{mn}$, it is clear that the eigenfunctions of U form an orthogonal basis of $\mathsf{H}_{|A|}$, as expected. Since these eigenfunctions are, in this case, the Fourier basis, (3.76) reveals $|A|$ to define the form corresponding to a pseudo-differential operator of order 2, with symbol $P(k) = |pk^2 + ick|$. Since for any $u \in \mathsf{V}$ we may write

$$u = \sum_{k\in\mathbb{Z}} u_k e^{ikx}, \qquad u_k := \tfrac{1}{2\pi}\langle u, e^{ikx}\rangle \tag{3.77}$$

it follows that

$$\|u\|^2_{|A|} = 2\pi \sum_{k\in\mathbb{Z}} |pk^2 + ick||u_k|^2, \tag{3.78}$$

which may be compared with

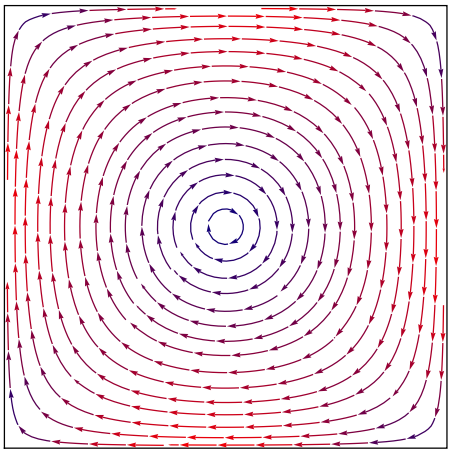
$$\|u\|^2_H = 2\pi \sum_{k\in\mathbb{Z}} pk^2|u_k|^2, \tag{3.79}$$

$$\|u\|^2_M = 2\pi \sum_{k\in\mathbb{Z}} (pk^2 + \tfrac{c^2}{p})|u_k|^2, \quad \text{and} \tag{3.80}$$

$$\|u\|^2_{\mathsf{V}} = 2\pi \sum_{k\in\mathbb{Z}} (k^2+1)|u_k|^2, \tag{3.81}$$

the last being the standard Sobolev H^1 norm.

3.5.3 2D Advection Diffusion



$$c_x = \quad 2y(1-x^2)$$
$$c_y = -2x(1-y^2)$$

Figure 3.3: Advection velocity for 2D example

Recall the discretized 2D advection diffusion example presented in §1.2.3, posed on $\Omega = [-1,1]^2$ with Dirichlet boundary conditions, the advection velocity specified

as shown in Figure 3.3, and diffusion coefficient $p = 1/200$. Three of the first few eigenfunctions of $U = |A|^{-1}A$ are shown in Figure 3.4, where A is the *discretized* operator on a 63 by 63 Chebyshev grid, using the streamline diffusion scheme discussed in §1.2.3. The third column in this figure consists of phase portraits of the eigenfunctions, in which the complex argument is indicated by color, as on a color wheel, while magnitude is indicated by brightness (so that, in particular, zero is black). Note that the primary direction of phase change is in the direction of the advection, the source of the nonsymmetry: these 2D eigenfunctions appear to be, roughly, waves along streamlines, generalizing the 1D example. Also note that, as in the 1D example, the larger canonical angles correspond to the lower wave numbers.

While the shapes of the displayed eigenfunctions should be close to those that would come from the underlying infinite-dimensional operator, the corresponding canonical angles certainly are not. This is due to the fact that the streamline diffusion scheme is closer to being a Petrov-Galerkin scheme than a strictly Galerkin one, which would approximately preserve canonical angles. The added stabilizing terms of the streamline diffusion scheme have a "symmetrizing" effect, reducing the maximum canonical angle. The gap between θ_1 and $\pi/2$ approximately halves when the mesh is refined from a 31 by 31 to the 63 by 63 Chebyshev grid used here, and this behavior is expected to continue as the mesh is refined, until the element Péclet numbers become less than one and the streamline diffusion scheme reduces to a standard Galerkin discretization.

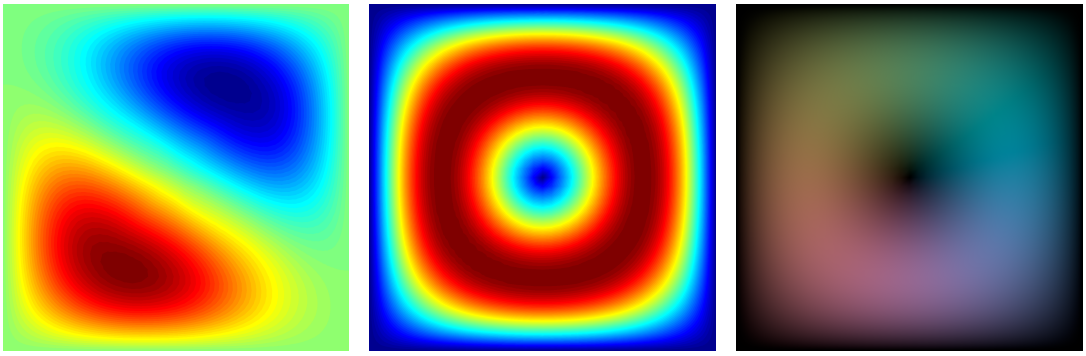
3.6 Computation

Here we consider methods for computing the absolute value of a given matrix $A \in \mathbb{C}^{n\times n}$ or $\mathbb{R}^{n\times n}$ requiring $O(n^3)$ operations. Of course, these methods are only practical for small to moderately sized matrices. For a large sparse matrix, such as those to which we wish to apply AMG, the absolute value would be expected to be dense anyway, such that merely storing it would be prohibitive, much less computing it. Nevertheless, we shall look in §5.3.2 at an interpolation strategy that involves computing the absolute values of small submatrices (the sizes of which correspond to the interpolation stencil width), making what follows of some practical interest.

3.6.1 General Case

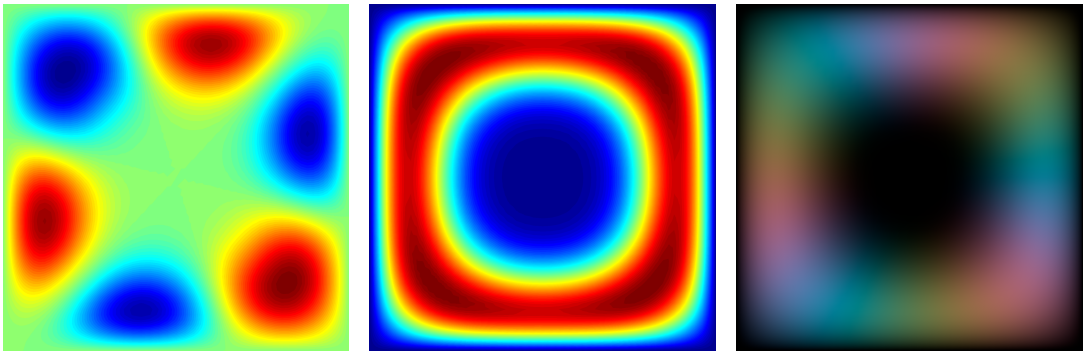
Let us first consider the general case, in which $H = \frac{1}{2}(A+A^*)$ is not necessarily positive. The following construction follows that for computing the canonical congruence form

$\theta_1 \approx 85.8°$



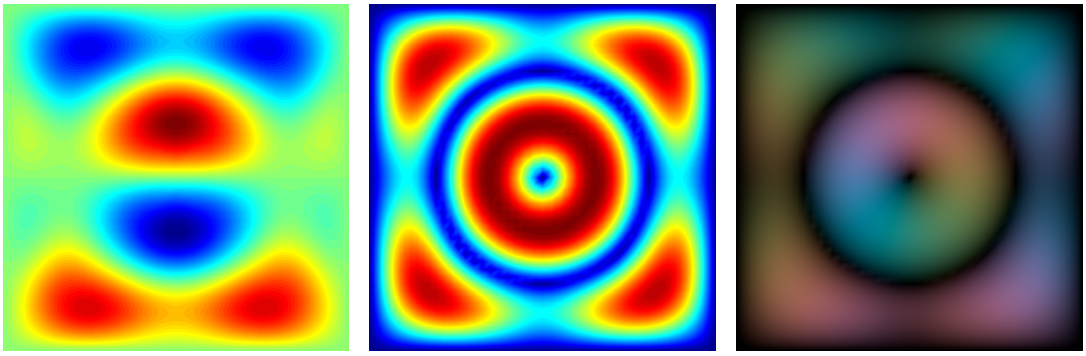
(a) $\operatorname{Re} u_1$ (b) $|u_1|$ (c) $\arg u_1$

$\theta_5 \approx 83.1°$



(d) $\operatorname{Re} u_5$ (e) $|u_5|$ (f) $\arg u_5$

$\theta_9 \approx 80.8°$



(g) $\operatorname{Re} u_9$ (h) $|u_9|$ (i) $\arg u_9$

Figure 3.4: Real part, magnitude, and phase portrait of some of the first few eigenfunctions of U for the 2D advection diffusion example.

of Horn and Sergeichuk [28]. The construction starts with the computation of the eigendecomposition

$$AV = A^*VD, \tag{3.82}$$

where D is the diagonal matrix of eigenvalues, using a general purpose generalized eigenproblem solver. It may happen that a full set of eigenvectors do not exist, which would show up in practice as a singular or nearly singular computed V. In this case no absolute value exists. It may also happen that two or more of the eigenvalues have modulus different than 1 (such eigenvalues come in pairs). In this case also, no absolute value exists. Otherwise, A is diagonalizable by congruence, so that an absolute value exists, and, provided the columns of V corresponding to equal eigenvalues are grouped together, V^*AV is necessarily block diagonal [28, §5],

$$V^*AV = \begin{bmatrix} e^{i\theta_1}X_1 & & \\ & \ddots & \\ & & e^{i\theta_1}X_s \end{bmatrix}, \qquad D = \begin{bmatrix} e^{2i\theta_1} & & \\ & \ddots & \\ & & e^{2i\theta_s} \end{bmatrix}, \tag{3.83}$$

with blocks for each distinct eigenvalue. Taking each angle θ_j to be in the interval $0 \le \theta_j \le \pi$, each X_j is Hermitian [28, §5]. Since

$$|A| = V^{-*}|V^*AV|V^{-1} = V^{-*}\begin{bmatrix} |X_1| & & \\ & \ddots & \\ & & |X_s| \end{bmatrix}V^{-1}, \tag{3.84}$$

the problem has been reduced to computing an absolute value of a Hermitian matrix. One such is $|X_j| = (X_j^2)^{1/2}$, the principal square root of X_j^2, so that

$$|A| = V^{-*}\begin{bmatrix} (X_1^2)^{1/2} & & \\ & \ddots & \\ & & (X_s^2)^{1/2} \end{bmatrix}V^{-1} \tag{3.85}$$

is one absolute value of A. Whenever any of the X_j are indefinite, then, as we have seen, there are infinitely many other absolute values.

3.6.2 Schur Decomposition

If the real part of A, $H = \frac{1}{2}(A + A^*)$, is positive, then more direct methods are available for computing $|A|$. In particular, it is possible to avoid computing the eigenvectors of U and to avoid complex arithmetic when A is real. If a procedure for computing the principal matrix square root is available, then one may use the

formula $|A| = A(A^{-1}A^*)^{1/2}$ (Proposition 3.2). In Matlab, for example, the syntax is `A*sqrtm(A\A')`.

One implementation strategy for computing $(A^{-1}A^*)^{1/2}$ is to compute the Schur decomposition of $A^{-1}A^*$. In this case, it is probably better to compute a generalized Schur decomposition, or QZ decomposition,

$$A = QSZ^*, \qquad A^* = QTZ^* \tag{3.86}$$

with Q, Z unitary and S, T upper triangular. When A is real, a real decomposition exists with S quasi-triangular instead of triangular, with 1×1 or 2×2 blocks along the diagonal. In either case we have

$$|A| = QT(T^{-1}S)^{1/2}Z^*, \qquad U = Z(T^{-1}S)^{1/2}Z^*, \tag{3.87}$$

with the square roots being principal. Hence the problem is reduced to a triangular solve and taking the square root of a triangular or quasi-triangular matrix, which are much easier problems. Explicit algorithms for computing the square roots are given in Higham [26, §6.2].

The reason the above method fails when H is not positive is that, in this case, the principal square root of $T^{-1}S$ is likely the wrong one so that $QT(T^{-1}S)^{1/2}Z^*$ fails to be positive.

3.6.3 Newton Iteration

Let us continue to assume that the real part H of A is positive. Consider the fixed-point iteration given by

$$\begin{aligned} X_1 = H, \qquad X_{k+1} &= \tfrac{1}{2}(X_k + A^* X_k^{-1} A) \\ &= \tfrac{1}{2}(X_k + A X_k^{-1} A^*). \end{aligned} \tag{3.88}$$

The final equality may be established by induction (these are two different iterations for other choices of X_1). The equation for the fixed point of (3.88) is equivalent to the defining equation of the absolute value, and so $|A|$ is the unique fixed point. This iteration can be implemented using a Cholesky factorization and structured in a way such that the computed iterates remain symmetric in the presence of round-off error:

$$L_k L_k^* = X_k, \qquad X_{k+1} = \tfrac{1}{2}[X_k + (L_k^{-1}A)^*(L_k^{-1}A)]. \tag{3.89}$$

Also note that the iteration can be implemented in real arithmetic when A is real.

Taking $U_k = A^{-*}X_k$, we see that (3.88) corresponds implicitly to the iteration

$$U_0 = I, \qquad U_{k+1} = \tfrac{1}{2}(U_k + U_k^{-1}A^{-*}A), \tag{3.90}$$

which is Newton's method for computing $(A^{-*}A)^{1/2}$. Newton's method for the matrix square root is discussed in detail in Higham's book [26, §6.3]. The eigenvalues of $A^{-*}A$ are $\lambda_j = e^{2i\theta_j}$ where $-\frac{\pi}{2} < \theta_j < \frac{\pi}{2}$ are the canonical angles of A. It follows that

$$\max_{jk} \tfrac{1}{2}|1 - \lambda_j^{1/2}\lambda_k^{-1/2}| = \max_{jk} \tfrac{1}{2}|1 - e^{i(\theta_j - \theta_k)}| \le \tfrac{1}{2}|1 - e^{2i\theta}| < 1 \tag{3.91}$$

where $\theta = \max_j |\theta_j| < \frac{\pi}{2}$ is the sectorial half-angle of A. Thus, according to the stability analysis of Higham [26, §6.4], the iteration (3.90) is stable, and this implies that (3.88) is stable as well.

Let Q be a nonsingular matrix that diagonalizes A by congruence, $Q^*AQ = e^{i\Theta}$, where $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_n)$ is the diagonal matrix of the canonical angles of A. Let D_k be defined by $Q^*X_kQ = D_k$. Then,

$$D_1 = \cos\Theta, \qquad D_{k+1} = \tfrac{1}{2}(D_k + e^{-i\Theta}D_k^{-1}e^{i\Theta}) = \tfrac{1}{2}(D_k + D_k^{-1}). \tag{3.92}$$

The last equality holds because D_k is diagonal, which follows by induction. This leads to the following result characterizing the accuracy of the iterates.

Proposition 3.3. *For all $u \in \mathbb{C}^n$ and $k > 1$,*

$$1 \le \frac{\|u\|_{X_k}^2}{\|u\|_{|A|}^2} \le \gamma_k \tag{3.93}$$

holds for the iterates X_k of (3.88) *where*

$$\gamma_1 = \sec\theta, \qquad \gamma_{k+1} = \tfrac{1}{2}(\gamma_k + \gamma_k^{-1}) = \coth(2^k \operatorname{arcoth}\sec\theta), \tag{3.94}$$

with θ being the sectorial half-angle of A.

The proof is that the Rayleigh quotient

$$\frac{\|u\|_{X_k}^2}{\|u\|_{|A|}^2} = \frac{\|Q^{-1}u\|_{D_k}^2}{\|Q^{-1}u\|_I^2} \tag{3.95}$$

is bounded by the extremal eigenvalues of D_k, which of course are simply its smallest and largest diagonal entries. All diagonal entries of D_1 are contained in the interval $[\cos\theta, 1]$, where θ is the sectorial half-angle of A, while the entries of D_k for $k > 1$ are contained in the interval $[1, \gamma_k]$. The closed form given for γ_k is easily verified to satisfy the recursion.

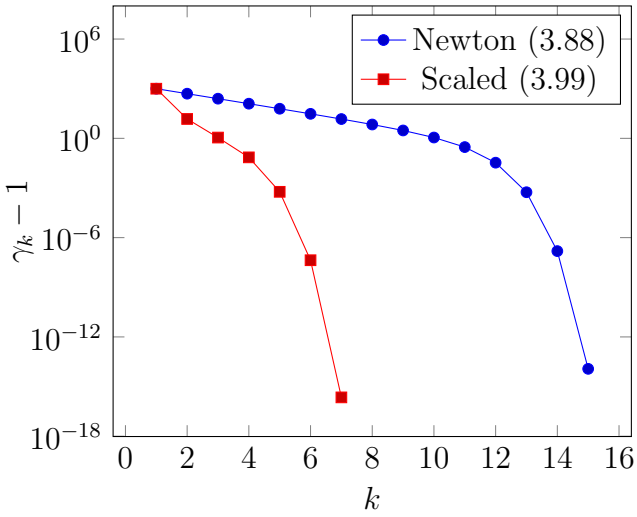


Figure 3.5: Convergence of the iterations (3.88) and (3.99) for $\gamma_1 = \sec\theta = 10^3$.

The proposition applies also at $k = 1$ for the alternate iteration with $X_1 = M = A^*H^{-1}A$. Either choice of X_1 leads to $X_2 = \frac{1}{2}(H + M)$. While there is no computational advantage to choosing $X_1 = M$, it makes the description of the iteration more uniform: in this case the iterates approximate $|A|$ monotonically from above.

Figure 3.5, which plots $\gamma_k - 1$, shows the convergence behavior in the case $\sec\theta = 10^3$ ($\theta = 89.94°$). At first, the convergence is approximately linear, with the error approximately halving in each iteration. Once γ_k falls below 2, the quadratic convergence rate of Newton's method becomes evident.

The preceding analysis suggests how the iteration can be accelerated. If a scaling parameter $\alpha > 0$ is introduced into the iteration,

$$X_{k+1} = \tfrac{1}{2}(\alpha^{-1}X_k + \alpha A^* X_k^{-1} A), \tag{3.96}$$

then in the diagonalizing basis we have

$$D_{k+1} = \tfrac{1}{2}(\alpha^{-1}D_k + \alpha D_k^{-1}). \tag{3.97}$$

This mapping sends diagonal entries from the interval $[1, \gamma]$ to the interval

$$[1, \max\{\tfrac{1}{2}(\tfrac{1}{\alpha} + \alpha), \tfrac{1}{2}(\tfrac{\gamma}{\alpha} + \tfrac{\alpha}{\gamma})\}]. \tag{3.98}$$

The optimal value of α is $\gamma^{1/2}$.

Putting the pieces together leads to the following iteration, starting with γ_1 some

given estimate, or possibly a bound, of $\sec\theta$:

$$\begin{aligned} X_1 &= H, \\ X_2 &= \tfrac{1}{2}(\gamma_1^{1/2} X_1 + \gamma_1^{-1/2} A^* X_1^{-1} A), \\ \gamma_{k+1} &= \tfrac{1}{2}(\gamma_k^{1/2} + \gamma_k^{-1/2}), \\ X_{k+1} &= \tfrac{1}{2}(\gamma_k^{-1/2} X_k + \gamma_k^{1/2} A^* X_k^{-1} A), \qquad k > 1. \end{aligned} \tag{3.99}$$

The first iteration is different as X_1 is H and not M; essentially the two terms are swapped. The following result about convergence follows straight-forwardly from the analysis above.

Proposition 3.4. *For the iterates X_k of (3.99), all $u \in \mathbb{C}^n$, and $k > 1$,*

$$1 \le \frac{\|u\|_{X_k}^2}{\|u\|_{|A|}^2} \le \gamma_k \tag{3.100}$$

provided that

$$\sec\theta \le \gamma_1 \tag{3.101}$$

where θ is the sectorial half-angle of A.

If $L_1 L_1^* = X_1 = H$ is the Cholesky decomposition of X_1, then one computable bound of $\sec\theta$ that can be taken for γ_1 is given by

$$\sec\theta = \|L_1^{-1} A L_1^{-*}\|_2 \le \|L_1^{-1} A L_1^{-*}\|_1^{1/2} \|L_1^{-1} A L_1^{-*}\|_\infty^{1/2}, \tag{3.102}$$

requiring one additional triangular solve during the first iteration ($L_1^{-1}A$ is required anyway). A Matlab implementation of the scaled iteration using this bound is shown in Figure 3.6.

Figure 3.5 compares the convergence speed of the original and scaled Newton iteration, showing the much faster speed of the scaled version, especially during the initial stage when the convergence of the unscaled iteration is approximately linear. The comparison is slightly unfair in that it is assumed that γ_1 is taken to be $\sec\theta$ exactly. However, as illustrated in Table 3.2, the number of iterations required to achieve a fixed accuracy is a remarkably flat function of $\sec\theta$, especially for θ close to $\pi/2$. Consequently, it does not make much difference if the estimate of $\sec\theta$ taken for γ_1 is off by a even a moderately large multiplicative factor. Also note that, if $\gamma_1 > 1$ is an underestimate of $\sec\theta$ rather than an upper bound, then (3.99) will always converge faster than (3.88). Indeed, (3.88) is an instance of (3.99) with the trivial estimate $\gamma_1 = 1$. The advantage of using a bound is that then the accuracy of the iteration is known a priori; otherwise some other convergence criteria needs to be monitored.

```
function X = absf( A )
%ABSF The (form) absolute value of a matrix
%   Given A such that (A+A') is SPD,
%   returns SPD X with A'*X\A = X .

X = full(A+A')/2;
L = chol(X,'lower');
LA = L\A;
LAL = LA/L';
gamma = sqrt(norm(LAL,1)*norm(LAL,inf));
alpha = 1/sqrt(gamma);
for i=1:20
  X = (X/alpha + alpha*(LA'*LA))/2;
  if gamma-1 < 1e-8; break; end
  gamma = (alpha+1/alpha)/2;
  if gamma <= 1; break; end
  L = chol(X,'lower');
  LA = L\A;
  alpha = sqrt(gamma);
end

end
```

Figure 3.6: A Matlab implementation of iteration (3.99)

n	$\sec\theta$	$\theta\,(^\circ)$
2	1.00000003	0.014
3	1.00048	1.8
4	1.064	19.9
5	2.03	60.5
6	14.5	86.0
7	835.7	89.93
8	2.8×10^{6}	$90 - 2.1 \times 10^{-5}$
9	3.1×10^{13}	$90 - 1.8 \times 10^{-12}$

Table 3.2: Maximum $\gamma_1 = \sec\theta$ such that $\gamma_n \leq 1 + 10^{-16}$ in the scaled iteration (3.99)

Chapter 4

Convergence Theory

We will now present a convergence theory for a fairly general two-level algorithm applied to an elliptic nonsymmetric operator A. While generally applicable, the theory is especially well-suited for algebraic multigrid (AMG). The specific assumptions we will make on A are the following.

Let $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ define a bounded sesquilinear form on the complex, reflexive Banach space V, and let the real part H of A be positive, $0 < c \leq H$.

We have mentioned these assumptions in previous chapters. They allow us to conclude, by the Lax-Milgram lemma (Theorem 2.1), that A has bounded inverse $A^{-1} \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$, and that A has a unique form absolute value $|A|$.

Our main application will of course be the matrix setting $\mathsf{V} = \mathbb{C}^n$. There are several reasons the more abstract setting we have chosen for the theory is useful, and not simply a distraction. The distinction between the solution space V and its adjoint V^* is a useful constraint that guides the development of the theory. For example, the sometimes seen multigrid constraint $RP = I$, that the restriction of the prolongation of a vector should be the original vector, is not well-formed in the abstract context, as P operates between solution spaces, whereas R operates between their adjoints. This is not to say such a constraint is never appropriate for multigrid—just that it does not fit the conceptual framework used here, a framework which is natural for AMG given the Petrov-Galerkin coarse operator construction. Another useful constraint is that we have not assumed a Hilbert space structure on V, which makes the Euclidean inner product on $\mathbb{C}^n$ and the Sobolev inner product on $H^1(\Omega)$ unavailable. Use of the associated Hilbert spaces would lead to inequalities featuring problem-dependent constants, which could be avoided by using a space adapted to the problem, such as that associated with the absolute value $|A|$ discussed in the last chapter. For AMG, such constants are particularly undesirable, as target applications are typically those

for which the constants involved are quite poor (e.g., meshes with very high aspect ratio cells).

It is also a strength of the theory that it can, at least in principle, be applied to a continuous multigrid method, applied to an infinite-dimensional PDE problem. This is a point of view that is common, for example, with Schwarz (domain decomposition) methods, but is perhaps especially unusual to find for *algebraic* multigrid methods. Nonetheless, the only additional subtlety that arises in treating the possibly infinite-dimensional case is the occasional need to demonstrate boundedness of an operator, which will usually be fairly trivial. A reader interested only in the matrix case can ignore these demonstrations, as all matrices are of course bounded linear operators.

Our analysis will be primarily concerned with the two-level iteration operator $E \in \mathscr{B}(\mathsf{V})$ defined as

$$E := (1 - BA)(1 - P_r \hat{A}_c^{-1} P_s^* A) \tag{4.1}$$

where $B \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$ is the smoother operator, $P_r, P_s \in \mathscr{B}(\mathsf{V}_c, \mathsf{V})$ are prolongation operators defined on a closed subspace V_c of V, and $\hat{A}_c := P_s^* A P_r$ is the coarse operator. We further require the projections of the prolongation operators onto V_c along some closed complement V_f of V_c to be the identity. We will elaborate more on these components and constraints as we proceed with our analysis.

The notation used here is somewhat nonstandard. One usually sees P for P_r and R for P_s^*. The notation has been chosen to make evident the fact that P_r and P_s play symmetric roles, as discussed in the next section. The subscripts r and s are mnemonics for "trial" and "test": the coarse trial and test spaces are the ranges of P_r and P_s. The hat on the coarse operator is also unusual; its presence will be explained in the section on the hierarchical decomposition below. The form of the coarse grid correction factor $(1 - P_r \hat{A}_c^{-1} P_s^* A)$ is otherwise standard.

The smoothing iteration factor $(1 - BA)$ has been chosen to be quite general. It encompasses, for example, k pre- or post-smoothing steps of the iteration $1 - M^{-1}A$ determined by a some preconditioner M via $1 - BA = (1 - M^{-1}A)^k$, or a combination of different pre- and post-smoothers via

$$1 - BA = (1 - M_1^{-1}A)^{k_1}(1 - M_2^{-1}A)^{k_2}, \tag{4.2}$$

and even Chebyshev-accelerated versions of the above. The reason the pre- and post-smoothing iterations can be grouped together is that we will mostly be concerned with the spectrum of E, which is not affected when the pre-smoothing iteration factor is moved to the front of E, as a consequence of Proposition 4.1 below. Additionally, we shall see that it is not necessary for convergence that B be invertible, which is why

we do not assume, say, $B = M^{-1}$. This is also the reason for the choice of symbol B, aside from preventing confusion with the operator $M = A^*H^{-1}A$ of Chapter 2.

A brief outline of our analysis is as follows. The Petrov-Galerkin form assumed for the coarse operator, $\hat{A}_c = P_s^* A P_r$, ensures that we can find a block representation of A in which $\hat{A}_c$ appears as a sub-block. In a matrix setting, this would involve a change of coordinates. We call this block form the hierarchical representation, though we only have two levels. The full multilevel version of this representation features prominently in the hierarchical basis method, demonstrated by Griebel [24] to be mathematically equivalent to multigrid. With the coarse operator appearing as a block, a further Schur complement decomposition leads to the representation of A as

$$\begin{bmatrix} S_f & \\ & \hat{A}_c \end{bmatrix} \tag{4.3}$$

in which the coarse operator is decoupled from the rest of A, its Schur complement. Theorem 4.1, which reproduces part of the main result from Yvan Notay's analysis [39], demonstrates that the asymptotic convergence rate of the two-level iteration depends on the action of a projection of the smoother to an "F-relaxation" on S_f. We then proceed by providing a convergence bound isolating the effect of the interpolation and restriction from each other, making key use of the absolute value norm from the previous chapter.

A point worth emphasizing is that the algorithm is fixed by (4.1), and that the constructions below are only for the purposes of analyzing and understanding the given algorithm, and are not intended to be used in implementation. For example, while the hierarchical basis method and multigrid are mathematically equivalent, they are distinguished by implementation choices (specifically, in the former coarse basis vectors with global support are explicitly present, while in multigrid they are implicitly stored in "factored form" in the prolongation operators). Although the ideas of the hierarchical basis method feature below, this does not imply an endorsement of that specific implementation.

4.1 Symmetry

If we replace A by A^*, B by B^*, and swap P_r and P_s, we obtain

$$E_* := (1 - B^*A^*)(1 - P_s \hat{A}_c^{-*} P_r^* A^*), \tag{4.4}$$

a two-level iteration operator for the adjoint operator A^*. Momentarily, we will demonstrate that the spectra of E and E_* are essentially mirror images of each other

about the real axis. Suppose we develop a theory containing results about $\sigma(E)$. We may consider the "dual" theory which results by making the above replacements, and translating the conclusions about $\sigma(E_*)$ into conclusions about $\sigma(E)$. A major goal is to develop a symmetric theory, such that the conclusions of the dual theory are identical to the original. We have taken care with notation to make the symmetry obvious; generally symbols with r and s subscripts swap for the dual theory, while operators with f and c subscripts are replaced with their adjoints.

That $\sigma(E) \cup \{0\} = \overline{\sigma(E_*)} \cup \{0\}$ is a consequence of

$$E_*^* = A(1 - P_r \hat{A}_c^{-1} P_s^* A)(1 - BA)A^{-1}, \tag{4.5}$$

by the following (basic and well-known) proposition with $A_1 = A(1 - P_r \hat{A}_c^{-1} P_s^* A)$ and $A_2 = (1 - BA)A^{-1}$.

Proposition 4.1. *Consider bounded operators $A_1 \in \mathscr{B}(\mathsf{V}_1, \mathsf{V}_2)$ and $A_2 \in \mathscr{B}(\mathsf{V}_2, \mathsf{V}_1)$. Possibly apart from 0, the spectra of $A_1 A_2$ and $A_2 A_1$ are identical,*

$$\sigma(A_1 A_2) \cup \{0\} = \sigma(A_2 A_1) \cup \{0\}. \tag{4.6}$$

Proof. If $\zeta \neq 0$ is in the resolvent set of $A_1 A_2$, then $A_1 A_2 - \zeta$ has bounded inverse $(A_1 A_2 - \zeta)^{-1}$. But then $A_2 A_1 - \zeta$ has bounded inverse

$$\zeta^{-1}[A_2(A_1 A_2 - \zeta)^{-1} A_1 - 1], \tag{4.7}$$

so that ζ is also in the resolvent set of $A_2 A_1$. □

4.2 Coarsening

In the classic AMG method, the degrees of freedom are partitioned by some automatic procedure into a set of "C-variables" and "F-variables." The C-variables become the coarse degrees of freedom; "C" is a mnemonic for coarse, and "F" for fine. Note, while the "fine" degrees of freedom might refer to all variables, the term F-variables denotes a set that is disjoint from the C-variables.

We abstract this algebraic notion of coarsening as follows. Suppose we are given complementary closed subspaces V_f, V_c of V.

$$\mathsf{V} = \mathsf{V}_f \oplus \mathsf{V}_c \tag{4.8}$$

In the classic AMG case, the intent is that V_c would be the set of vectors with nonzero components only at the C-variables, and similarly for V_f. Let us denote by π_c the

operator that gives the components of a vector with respect to this decomposition. Thus, for $u_f \in \mathsf{V}_f$ and $u_c \in \mathsf{V}_c$, we have

$$\pi_c(u_f + u_c) = \begin{bmatrix} u_f \\ u_c \end{bmatrix}, \qquad \pi_c^{-1} \begin{bmatrix} u_f \\ u_c \end{bmatrix} = u_f + u_c. \tag{4.9}$$

The inverse, of course, consists of the pair of trivial injections from V_f and V_c into V. In particular, π_c^{-1} is bounded. The boundedness of the pair of projections that make up π_c is a consequence of the closedness of V_f and V_c [31, p. 167]. Thus $\pi_c \in \mathscr{B}(\mathsf{V}, \mathsf{V}_f \times \mathsf{V}_c)$. Following Kato [31, p. 164], we take the norm on $\mathsf{V}_f \times \mathsf{V}_c$ to be

$$\left\| \begin{bmatrix} u_f \\ u_c \end{bmatrix} \right\| := (\|u_f\|^2 + \|u_c\|^2)^{1/2} \tag{4.10}$$

to ensure that $(\mathsf{V}_f \times \mathsf{V}_c)^* = \mathsf{V}_f^* \times \mathsf{V}_c^*$.

Our decomposition induces a block structure on A, for which we introduce the notation

$$[A]_c := \pi_c^{-*} A \pi_c^{-1} =: \begin{bmatrix} A_f & A_r \\ A_s^* & A_c \end{bmatrix}. \tag{4.11}$$

In the classic AMG case, this is simply the block structure that A takes on when the F-variables are ordered first, and π_c could be taken to be the permutation that accomplishes this re-ordering. Because the components of π_c^{-1} are just the trivial injections from V_f and V_c into V, the blocks of $[A]_c$ define forms that are simply restrictions of the form defined by A. For example, the sesquilinear form defined by A_f is just the restriction of the form defined by A to $\mathsf{V}_f \times \mathsf{V}_f$. The motivation for the adjoint on A_s^* is so that A_r and A_s, which play symmetric roles, have the same form, namely as members of $\mathscr{B}(\mathsf{V}_c, \mathsf{V}_f^*)$. The subscripts r and s on these blocks were chosen because they will play an important role in the determination of the coarse trial and test spaces. The other operators have block representations also. We define

$$[B]_c := \pi_c B \pi_c^*, \tag{4.12}$$

and introduce $[P_r]_c, [P_s]_c$ in the next section.

Recall that H, the real part of A, was assumed to be positive. Because the real part H_f of A_f is a restriction of H, it too is positive. Similar remarks apply to A_c and its real part. The Lax-Milgram lemma (Theorem 2.1) thus applies to both operators so that

$$A_f^{-1} \in \mathscr{B}(\mathsf{V}_f, \mathsf{V}_f^*), \qquad A_c^{-1} \in \mathscr{B}(\mathsf{V}_c, \mathsf{V}_c^*). \tag{4.13}$$

Note that these results make crucial use of the assumption that the real part H of A is positive. Saddle-point systems provide examples of operators that possess bounded inverses, but which violate the assumption of positive real part, and for which (4.13) generally fails. Indeed, we can find easy examples where, e.g., $A_c = 0$.

4.3 Hierarchical Decomposition

In classical AMG, the "prolongation" operators $P_r, P_s \in \mathscr{B}(\mathsf{V}_c, \mathsf{V})$ are constrained structurally as

$$[P_j]_c := \pi_c P_j =: \begin{bmatrix} W_j \\ 1 \end{bmatrix}, \tag{4.14}$$

where $W_j \in \mathscr{B}(\mathsf{V}_c, \mathsf{V}_f)$ and where the subscript j in this section ranges over r and s. In the discrete case, the constraint captures the property that the coarse variables are a subset of the original variables, and the "weights" W_j determine how the F-variables are interpolated from the C-variables.

A technical detail needed later is that the form assumed for the prolongation operators ensure them to be bounded not only from above, but also from below. We can write the structural constraint as $\begin{bmatrix} 0 & 1 \end{bmatrix} \pi_c P_j u_c = u_c$, or simply as $\pi_{cf} P_j u_c = u_c$, where π_{cf} denotes the second component of π_c, the projection onto V_c along V_f. Taking norms, we find

$$\frac{1}{\|\pi_{cf}\|} \|u_c\|_{\mathsf{V}_c} \leq \|P_j u_c\|_{\mathsf{V}} \leq \|P_j\| \|u_c\|_{\mathsf{V}_c}. \tag{4.15}$$

The prolongation operators define coarse trial and test subspaces via their range.

$$\mathsf{W}_j := \mathsf{R}(P_j). \tag{4.16}$$

We have two corresponding "hierarchical" decompositions.

$$\mathsf{V} = \mathsf{V}_f \oplus \mathsf{W}_j. \tag{4.17}$$

Let us define the operators $\pi_{h,j}^{-1} \in \mathscr{B}(\mathsf{V}_f \times \mathsf{V}_c, \mathsf{V})$ to consist of pairs of injections corresponding to each of these decompositions by

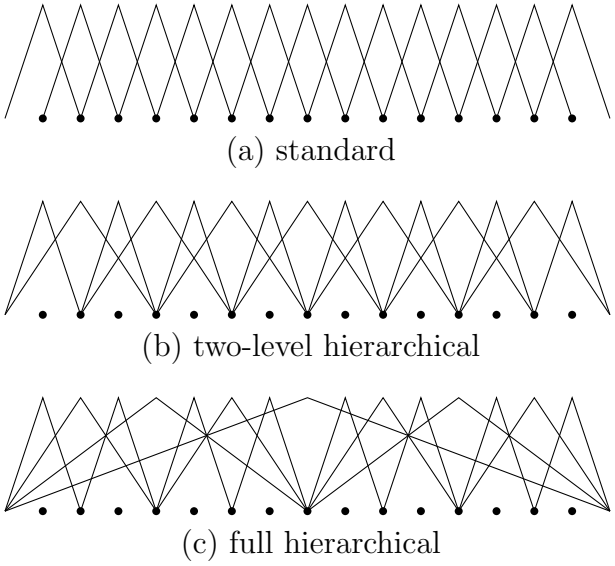
$$\pi_{h,j}^{-1} \begin{bmatrix} u_f \\ u_c \end{bmatrix} := u_f + P_j u_c. \tag{4.18}$$

In contrast to the previous section, here we are representing each $u_j \in \mathsf{W}_j$ indirectly by its preimage $P_j^{-1} u_j \in \mathsf{V}_c$. In components, we have

$$T_j := \pi_c \pi_{h,j}^{-1} = \begin{bmatrix} 1 & W_j \\ & 1 \end{bmatrix}, \qquad T_j^{-1} = \pi_{h,j} \pi_c^{-1} = \begin{bmatrix} 1 & -W_j \\ & 1 \end{bmatrix}. \tag{4.19}$$

Here and throughout, we adopt the convention that any missing blocks in a matrix are 0. We see that $\pi_{h,j}$ exists and is bounded, justifying the notation $\pi_{h,j}^{-1}$. In the matrix setting, the operators T_j are change of basis matrices that transform to hierarchical coordinates.

An example hierarchical decomposition is illustrated in Figure 4.1. Figure 4.1a shows the standard hat basis functions of a row of sixteen linear finite elements. Figure 4.1b illustrates a possible two-level hierarchical basis, where each basis function corresponds to a column of T_j. Half of the basis functions, corresponding to the F-variables and the first block column of T_j, are identical to the original basis functions, while the remaining basis functions, corresponding to the C-variables and the second block column of T_j, have wider support (coming from nonzeros in W_j).



(a) standard

(b) two-level hierarchical

(c) full hierarchical

Figure 4.1: Example hierarchical basis

Define hierarchical representations of A and B as

$$[A]_h := \pi_{h,s}^{-*} A \pi_{h,r}^{-1} = T_s^* [A]_c T_r = \begin{bmatrix} A_f & \hat{A}_r \\ \hat{A}_s^* & \hat{A}_c \end{bmatrix}, \tag{4.20}$$

$$[B]_h := \pi_{h,r} B \pi_{h,s}^* = T_r^{-1} [B]_c T_s^{-*} =: \begin{bmatrix} \hat{B}_f & \hat{B}_r \\ \hat{B}_s^* & \hat{B}_c \end{bmatrix} \tag{4.21}$$

where

$$\hat{A}_r := A_f W_r + A_r, \quad \hat{A}_s := A_f^* W_s + A_s, \quad \hat{A}_c := P_s^* A P_r. \tag{4.22}$$

The significance of the hierarchical representation is that the coarse operator $\hat{A}_c := P_s^* A P_r$ appears as a sub-block. The choice of the hat symbol to mark the blocks of the hierarchical representation was chosen to be reminiscent of Figure 4.1. Note that, formally, $[A]_c$ and $[A]_h$ are in the same space of bounded operators, both defining sesquilinear forms on $\mathsf{V}_f \times \mathsf{V}_c$, and are different. We have set up our notation to emphasize that they are representations, via different topological isomorphisms, of the same operator A. We delay discussion of $[B]_h$ to §4.5.

The particular choice of weights

$$W_r = -A_f^{-1} A_r, \qquad W_s = -A_f^{-*} A_s, \tag{4.23}$$

causes the coupling operators $\hat{A}_r$ and $\hat{A}_s$ to vanish. For this reason, these are termed the ideal weights. Note that the existence and boundedness of the ideal weights depends on the bounded invertibility of A_f, which we demonstrated in (4.13) using the assumption that A had positive real part. In the matrix setting, the ideal weights are generally full owing to the inverses appearing in their definition, whereas for efficiency reasons, it is important that the chosen weights be sparse, with most entries zero. This motivates a definition of the approximation error, the difference between the actual and ideal weights.

$$\begin{aligned} F_r &:= -A_f^{-1} A_r - W_r = -A_f^{-1} \hat{A}_r, \\ F_s &:= -A_f^{-*} A_s - W_s = -A_f^{-*} \hat{A}_s. \end{aligned} \tag{4.24}$$

The ideal weights give rise to a Schur complement decomposition involving

$$S_c := A_c - A_s^* A_f^{-1} A_r, \tag{4.25}$$

the Schur complement of A_f. In fact it is easy to check that

$$\hat{A}_c = S_c + \hat{A}_s^* A_f^{-1} \hat{A}_r = S_c + F_s^* A_f F_r. \tag{4.26}$$

As we might expect, $\hat{A}_c$ reduces to S_c when either the trial or test weights is ideal. In this case, at least, the coarse operator has positive real part.

Proposition 4.2. *The real part of S_c is positive.*

Proof. For the purposes of this proof, fix $W_s = W_r = -A_f^{-1} A_r$ so that $F_r = 0$. While F_s may not be small, we still have $\hat{A}_c = P_r^* A P_r = S_c + F_s^* A_f F_r = S_c$. Consequently,

$$\operatorname{Re}(S_c u_c, u_c) = \operatorname{Re}(A P_r u_c, P_r u_c) = \|P_r u_c\|_H^2 \geq c \|P_r u_c\|_{\mathsf{V}}^2 \geq \frac{c}{\|\pi_{cf}\|^2} \|u_c\|_{\mathsf{V}_c}^2, \tag{4.27}$$

where we have used the lower bound (4.15) on P_r. Note that fixing $W_s = W_r = -A_f^{-*} A_s$ would have led to the same conclusion. □

If we wish to apply our theory recursively, we must ensure that the coarse operator $\hat{A}_c$ satisfies the assumptions we made on A. In particular, because in general $P_s \neq P_r$, it is not automatic that $\hat{A}_c = P_s^* A P_r$ has positive real part. We have demonstrated positivity for the case of ideal weights. From (4.26) we see that $\hat{A}_c$ will be positive in the general case if F_r and F_s are small enough.

Proposition 4.3. *Let $\tilde{H}_c := H_{S_c} = \frac{1}{2}(S_c + S_c^*)$ be the real part of S_c. If*

$$\|F_r\|_{|A_f|,\tilde{H}_c}\|F_s\|_{|A_f|,\tilde{H}_c} < 1 \tag{4.28}$$

then $\hat{A}_c$ has positive real part.

Proof. We have

$$\begin{aligned} \operatorname{Re}(\hat{A}_c u_c, u_c) &= \operatorname{Re}(S_c u_c, u_c) + \operatorname{Re}(A_f F_r u_c, F_s u_c) \\ &\geq \|u_c\|^2_{\tilde{H}_c} - |(A_f F_r u_c, F_s u_c)|, \end{aligned} \tag{4.29}$$

while by the Cauchy-Schwarz type inequality (3.20),

$$\begin{aligned} |(A_f F_r u_c, F_s u_c)| &\leq \|F_r u_c\|_{|A_f|}\|F_s u_c\|_{|A_f|} \\ &\leq \|F_r\|_{|A_f|,\tilde{H}_c}\|F_s\|_{|A_f|,\tilde{H}_c}\|u_c\|^2_{\tilde{H}_c}. \end{aligned} \tag{4.30}$$

Hence,

$$\begin{aligned} \operatorname{Re}(\hat{A}_c u_c, u_c) &\geq (1 - \|F_r\|_{|A_f|,\tilde{H}_c}\|F_s\|_{|A_f|,\tilde{H}_c})\|u_c\|^2_{\tilde{H}_c} \\ &\geq c\|\pi_{cf}\|^{-2}(1 - \|F_r\|_{|A_f|,\tilde{H}_c}\|F_s\|_{|A_f|,\tilde{H}_c})\|u_c\|^2_{\mathsf{V}_c}. \end{aligned} \tag{4.31}$$

□

The premise of this proposition is fairly stringent. In particular, for a symmetric problem with $W_r = W_s$, $\hat{A}_c$ is of course always positive, and yet the bound in the premise may well fail to hold. In the language of Proposition 4.6 below, the bound in the symmetric case can be read as requiring that the abstract angle ϕ between the subspaces V_f and $\mathsf{W}_r = \mathsf{W}_s$, which is always in the range $0 < \phi \leq \pi/2$, be at least $\pi/4$. The following alternate bound, which is always satisfied in the symmetric case, explains this apparent gap. As the proof is technical and not particularly illuminating, it may be skipped.

Proposition 4.4. *$[A]_h$ (and thus $\hat{A}_c$) has coercive real part provided*

$$\|F_r - F_s\|_{M_f,\tilde{H}_c} < 2, \tag{4.32}$$

where $M_f := M_{A_f} = A_f^* H_f^{-1} A_f$.

Proof. Let

$$\hat{T}_s = \begin{bmatrix} 1 & -A_f^{-*}\hat{A}_s \\ & 1 \end{bmatrix} = \begin{bmatrix} 1 & F_s \\ & 1 \end{bmatrix}. \tag{4.33}$$

Because $\hat{T}_s$ is bounded with bounded inverse, it suffices to show that $c' \leq \hat{T}_s^*[A]_h\hat{T}_s$ for some $c' > 0$, for then

$$\mathrm{Re}\,([A]_h u, u) = \mathrm{Re}\,(\hat{T}_s^*[A]_h\hat{T}_s\hat{T}_s^{-1}u, \hat{T}_s^{-1}u) \geq c'\|\hat{T}_s^{-1}u\|_{\mathsf{V}}^2 \geq c'\|\hat{T}_s\|^{-2}\|u\|_{\mathsf{V}}^2. \tag{4.34}$$

Now,

$$\hat{T}_s^*[A]_h\hat{T}_s = \begin{bmatrix} A_f & -A_f(F_r - F_s) \\ & S_c \end{bmatrix} \tag{4.35}$$

so that

$$\begin{aligned} \mathrm{Re}\left([A]_h\hat{T}_s\begin{bmatrix}u_f\\u_c\end{bmatrix}, \hat{T}_s\begin{bmatrix}u_f\\u_c\end{bmatrix}\right) &= \|u_f\|_{H_f}^2 - \mathrm{Re}\,(A_f(F_r - F_s)u_c, u_f) + \|u_c\|_{\tilde{H}_c}^2 \\ &\geq \|u_f\|_{H_f}^2 - \|u_f\|_{H_f}\|(F_r - F_s)u_c\|_{M_f} + \|u_c\|_{\tilde{H}_c}^2 \\ &\geq \|u_f\|_{H_f}^2 - \|F_r - F_s\|_{M_f,\tilde{H}_c}\|u_f\|_{H_f}\|u_c\|_{\tilde{H}_c} + \|u_c\|_{\tilde{H}_c}^2 \\ &\geq \frac{1 - (\frac{1}{2}\|F_r - F_s\|_{M_f,\tilde{H}_c})^2}{2}(\|u_f\|_{H_f}^2 + \|u_c\|_{\tilde{H}_c}^2). \end{aligned} \tag{4.36}$$

Note the use of the bound (2.38) in the second line, while the final line is an instance of the inequality

$$\begin{aligned} a^2 - 2\gamma ab + b^2 &= \tfrac{1}{2}(a - \gamma b)^2 + \tfrac{1}{2}(b - \gamma a)^2 + \frac{1 - \gamma^2}{2}(a^2 + b^2) \\ &\geq \frac{1 - \gamma^2}{2}(a^2 + b^2), \end{aligned} \tag{4.37}$$

which holds for all real a, b, γ. We complete the proof by noting that $\|u_f\|_{H_f}^2 + \|u_c\|_{\tilde{H}_c}^2$ defines the square of a norm equivalent to the Banach one. Concretely, for $u \in \mathsf{V}_f \times \mathsf{V}_c$ with components u_f and u_c,

$$\|u_f\|_{H_f}^2 + \|u_c\|_{\tilde{H}_c}^2 \geq c\|u_f\|_{\mathsf{V}}^2 + \frac{c}{\|\pi_{cf}\|^2}\|u_c\|_{\mathsf{V}}^2 \geq c\min(1, \|\pi_{cf}\|^{-2})\|u\|_{\mathsf{V}_f\times\mathsf{V}_c}^2. \tag{4.38}$$

□

Note that the premise of this lemma is always satisfied in the symmetric case, for then $F_r - F_s = 0$. This is a rather fortuitous cancellation, however, which we have no reason to expect in the general nonsymmetric case. As soon as we take the obvious step of using the triangle inequality to bound the norm of $F_r - F_s$ by a sum of norms, the premise becomes strictly stronger than that of Proposition 4.3.

4.4 Biorthogonal Decomposition

Let us assume that $\hat{H}_c := H_{\hat{A}_c} = \frac{1}{2}(\hat{A}_c + \hat{A}_c^*)$ is positive so that $\hat{A}_c^{-1} \in \mathscr{B}(\mathsf{V}_c^*, \mathsf{V}_c)$ by the Lax-Milgram lemma (Theorem 2.1). Our goal is to construct decompositions for the trial and test spaces

$$\mathsf{V} = \tilde{\mathsf{V}}_{f,r} \oplus \mathsf{W}_r, \qquad \mathsf{V} = \tilde{\mathsf{V}}_{f,s} \oplus \mathsf{W}_s \tag{4.39}$$

that are biorthogonal with respect to the form defined by A. We can accomplish this by arranging things so as to end up with a Schur complement decomposition of $[A]_h$, $Q_s^*[A]_h Q_r = [A]_o$. As before, let us define our decompositions via operators $\pi_{o,s}^{-1}, \pi_{o,r}^{-1} \in \mathscr{B}(\mathsf{V}_f \times \mathsf{V}_c, \mathsf{V})$ consisting of pairs of injections. Let

$$\pi_{o,r}^{-1} \begin{bmatrix} u_f \\ u_c \end{bmatrix} := (1 - P_r \hat{A}_c^{-1} \hat{A}_s^*) u_f + P_r u_c, \tag{4.40}$$

$$\pi_{o,s}^{-1} \begin{bmatrix} u_f \\ u_c \end{bmatrix} := (1 - P_s \hat{A}_c^{-*} \hat{A}_r^*) u_f + P_s u_c, \tag{4.41}$$

so that, e.g., $\tilde{\mathsf{V}}_{f,r} = \mathsf{R}(1_{\mathsf{V}_f} - P_r \hat{A}_c^{-1} \hat{A}_s^*)$. Clearly these injections are bounded. In terms of the hierarchical decompositions, we have

$$Q_r := \pi_{h,r} \pi_{o,r}^{-1} = \begin{bmatrix} 1 & \\ -\hat{A}_c^{-1} \hat{A}_s^* & 1 \end{bmatrix}, \qquad Q_s := \pi_{h,s} \pi_{o,s}^{-1} = \begin{bmatrix} 1 & \\ -\hat{A}_c^{-*} \hat{A}_r^* & 1 \end{bmatrix}, \tag{4.42}$$

which are the desired factors of the Schur complement decomposition. As was the case with the operators T_j, these operators clearly have bounded inverses, so that $\pi_{o,r}$, $\pi_{o,s}$ and their component projections are bounded. Letting j stand for r or s, note that $\begin{bmatrix} 1 & 0 \end{bmatrix} Q_j = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and since $\pi_{h,j} = Q_j \pi_{o,j}$, we see that the first component of the biorthogonal and hierarchical projections are the same. Let these projections be denoted

$$\pi_{f,j} := \begin{bmatrix} 1 & 0 \end{bmatrix} \pi_{o,j} = \begin{bmatrix} 1 & 0 \end{bmatrix} \pi_{h,j} = \begin{bmatrix} 1 & -W_j \end{bmatrix} \pi_c. \tag{4.43}$$

For example, $\pi_{f,r}$ is the projection onto V_f along W_r. These projections will play a major role in the next section.

Define the biorthogonal representation of A as

$$[A]_o := \pi_{o,s}^{-*} A \pi_{o,r}^{-1} = Q_s^* [A]_h Q_r = \begin{bmatrix} S_f & \\ & \hat{A}_c \end{bmatrix} \tag{4.44}$$

where

$$S_f := A_f - \hat{A}_r \hat{A}_c^{-1} \hat{A}_s^* \tag{4.45}$$

is the Schur complement of the coarse operator $\hat{A}_c$. Inverting (4.44), we see that S_f obeys the relation

$$S_f^{-1} = \pi_{f,r} A^{-1} \pi_{f,s}^*. \tag{4.46}$$

A final useful form for S_f^{-1} is given by applying the Woodbury matrix identity to (4.45), producing

$$S_f^{-1} = A_f^{-1} + F_r S_c^{-1} F_s^*. \tag{4.47}$$

4.5 Projected Smoother

Here we shall show that the spectrum of the two-level iteration operator E, defined by (4.1), depends only on $\hat{B}_f$, the projection associated with the hierarchical decompositions of the smoother B to an "F-relaxation," i.e., a smoother which operates only on the F-variables. As the other components $\hat{B}_r$, $\hat{B}_s$, $\hat{B}_c$ of the smoother have no effect on the spectrum of E, they can be ignored in the subsequent analysis.

Define the representation of the iteration operator E with respect to the biorthogonal decomposition as

$$[E]_o := \pi_{o,r} E \pi_{o,r}^{-1} = (1 - [B]_o[A]_o)(1 - [P_r]_o \hat{A}_c^{-1} [P_s]_o^* [A]_o) \tag{4.48}$$

where

$$[B]_o := \pi_{o,r} B \pi_{o,s}^*, \qquad [P_j]_o := \pi_{o,j} P_j = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{4.49}$$

are the representations of the smoother and prolongation operators. Note that $[E]_o$ and E are related by a similarity transformation, and hence have identical spectra. Also note that the prolongation operators are trivial in this representation, by construction. We see that

$$\pi_{o,r}(1 - P_r \hat{A}_c^{-1} P_s^* A) \pi_{o,r}^{-1} = (1 - [P_r]_o \hat{A}_c^{-1} [P_s]_o^* [A]_o) = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix}. \tag{4.50}$$

That is, the coarse grid correction is simply the projection onto $\tilde{\mathsf{V}}_{f,r}$ along W_r. If we label the blocks of $[B]_o$ by

$$\begin{bmatrix} \tilde{B}_f & \tilde{B}_r \\ \tilde{B}_s^* & \tilde{B}_c \end{bmatrix} := [B]_o \tag{4.51}$$

then we easily compute

$$[E]_o = \begin{bmatrix} 1 - \tilde{B}_f S_f & 0 \\ -\tilde{B}_s^* S_f & 0 \end{bmatrix}, \tag{4.52}$$

and it follows immediately that

$$\sigma(E) = \sigma([E]_o) = \sigma(1 - \tilde{B}_f S_f) \cup \{0^{\dim \mathsf{V}_c}\}. \tag{4.53}$$

Here the superscript on the zero is intended as multiset notation, indicating the multiplicity of the zero eigenvalue. Recalling (4.43), we see that

$$\tilde{B}_f = \hat{B}_f = \pi_{f,r} B \pi_{f,s}^* = \begin{bmatrix} 1 & -W_r \end{bmatrix} [B]_c \begin{bmatrix} 1 & -W_s \end{bmatrix}^*. \tag{4.54}$$

That is, the V_f blocks of $[B]_o$ and $[B]_h$ are the same, and since $\pi_{f,r}$ and $\pi_{f,s}$ are projections onto V_f along W_r and W_s, we see that $\hat{B}_f$ is the projection associated with the hierarchical decompositions of the smoother to an "F-relaxation." We have just proved the following result.

Theorem 4.1. *The spectrum of E is equal to that of $1 - \hat{B}_f S_f$, but also includes the eigenvalue 0 associated with the geometric eigenspace $\mathsf{W}_r \subseteq \mathsf{N}(E)$, except when this space is trivial. That is,*

$$\sigma(E) = \sigma(1 - \hat{B}_f S_f) \cup \{0^{\dim \mathsf{W}_r}\}. \tag{4.55}$$

The spectrum of E controls the asymptotic behavior of the two-level iteration. Because the iteration is for a nonsymmetric problem, the phenomena of non-normality and transient growth are possible, and it seems worthwhile to also state an explicit norm bound for the error of the iteration. The preceding analysis allows us to state such a bound in terms of a norm of the projected iteration operator $1 - \hat{B}_f S_f$ or of its powers.

Corollary. *Suppose*

$$1 - BA = (1 - B_1 A)(1 - B_2 A). \tag{4.56}$$

Then the error e_n of the nth iteration, $n \geq 1$, of the two-level iteration with pre-smoother B_1 and post-smoother B_2, i.e.,

$$e_n = [(1 - B_2 A)(1 - P_r \hat{A}_c^{-1} P_s^* A)(1 - B_1 A)]^n e_0, \tag{4.57}$$

has norm bounded by

$$\begin{aligned} \|e_n\|_{\mathsf{X}} &\leq C \|(1 - \hat{B}_f S_f)^{n-1}\|_{\mathsf{X}_f} \|e_0\|_{\mathsf{X}} \\ &\leq C \|1 - \hat{B}_f S_f\|_{\mathsf{X}_f}^{n-1} \|e_0\|_{\mathsf{X}} \end{aligned} \tag{4.58}$$

where

$$C = \left\| (1 - B_2 A) \pi_{o,r}^{-1} \begin{bmatrix} 1_{\mathsf{V}_f} \\ 0 \end{bmatrix} \right\|_{\mathsf{X},\mathsf{X}_f} \|\pi_{f,r}(1 - B_1 A)\|_{\mathsf{X}_f,\mathsf{X}} \tag{4.59}$$

and where $\|\cdot\|_{\mathsf{X}}$ and $\|\cdot\|_{\mathsf{X}_f}$ denote arbitrary norms equivalent to $\|\cdot\|_{\mathsf{V}}$ and $\|\cdot\|_{\mathsf{V}_f}$, so that, e.g., the spaces X and V are topologically isomorphic.

Proof. Since $n \geq 1$, we can expand e_n as

$$e_n = (1 - B_2 A)\pi_{o,r}^{-1}\pi_{o,r}(1 - P_r \hat{A}_c P_s^* A)E^{n-1}\pi_{o,r}^{-1}\pi_{o,r}(1 - B_1 A)e_0. \tag{4.60}$$

We have that

$$\begin{aligned}\pi_{o,r}(1 - P_r \hat{A}_c^{-1} P_s^* A)E^{n-1}\pi_{o,r}^{-1} &= \begin{bmatrix} 1 & \\ & 0 \end{bmatrix} [E]_o^{n-1} \\ &= \begin{bmatrix} 1_{\mathsf{V}_f} \\ 0 \end{bmatrix} (1 - \hat{B}_f S_f)^{n-1} \begin{bmatrix} 1_{\mathsf{V}_f} & 0 \end{bmatrix}.\end{aligned} \tag{4.61}$$

The result now follows by substituting (4.61) into (4.60), taking norms, and recalling that $\pi_{f,r} := \begin{bmatrix} 1 & 0 \end{bmatrix} \pi_{o,r}$. □

Note that these results imply that a smoother other than the trivial choice A^{-1} may be "perfect" or "ideal." One such ideal smoother is the F-relaxation given by

$$[B]_c = \begin{bmatrix} S_f^{-1} & \\ & 0 \end{bmatrix} = \begin{bmatrix} A_f^{-1} + F_r S_c^{-1} F_s^* & \\ & 0 \end{bmatrix}, \tag{4.62}$$

since this choice gives $\hat{B}_f = S_f^{-1}$. Hence the resulting spectrum of E is just $\{0\}$, and indeed, as is evident from the block expansion of $[E]_o$ above, we have in this case $E^2 = 0$, i.e., a direct method. In implementation practice, neither S_f or S_c is available, much less their inverses. However, if the weight errors F_r and F_s are small, then the ideal projected smoother is close to A_f^{-1}. In the next section, this line of thought is pursued to develop convergence bounds that involve appropriate norms of F_r and F_s.

Before proceeding to develop these bounds, we shall examine for the remainder of this section the relationship of the above theorem, specifically in the matrix case, to results by Yvan Notay, particularly Theorem 2.1, in a paper [39] also analyzing the spectrum of the two-level iteration operator for nonsymmetric AMG. Notay's results, in turn, generalize earlier results for symmetric problems in the analysis by Falgout, Vassilevski, and Zikatanov [19], particularly Theorem 4.3.

The relationship to the main Theorem 2.1 of Notay's paper is as follows. Note that for every eigenvalue $\mu \neq 0$ of the generalized eigenproblem

$$S_f^{-1} g = \mu \hat{B}_f g, \tag{4.63}$$

the spectrum of the iteration operator E contains the eigenvalue $1 - \mu^{-1}$. Recalling (4.46), we may expand this generalized eigenproblem to

$$\pi_{f,r} A^{-1} \pi_{f,s}^* g = \mu \pi_{f,r} B \pi_{f,s}^* g. \tag{4.64}$$

This version of Theorem 4.1 is essentially part (5) of Notay's Theorem 2.1.

An equivalent eigenproblem is

$$A^{-1}(\pi_{f,s}^* \hat{B}_f^{-1} \pi_{f,r}) u = \mu u, \tag{4.65}$$

related by $g = \hat{B}_f^{-1} \pi_{f,r} u$, except for the additional nullspace $\mathsf{W}_r = \mathsf{N}(\pi_{f,r})$, a geometric eigenspace with eigenvalue 0 (see also Proposition 4.1). This last characterization of the spectrum is equivalent by Lemma 4.1 below to the main part (1) of Notay's theorem, which states that the μ are eigenvalues of, in our notation,

$$A^{-1}B^{-1}(1 - \pi_{B^{-1}}), \qquad \text{where} \quad \pi_X := P_r(P_s^* X P_r)^{-1} P_s^* X. \tag{4.66}$$

In the symmetric case, when additionally the smoother consists of a symmetrically paired pre- and post-smoother, this reduces to Theorem 4.3 of the analysis by Falgout et al. [19]. Note that in this last form, the result requires invertibility of B as a whole, and not just of its effective component $\hat{B}_f$.

Lemma 4.1. *If both B and $\hat{B}_f$ are invertible, then*

$$B^{-1}(1 - \pi_{B^{-1}}) = \pi_{f,s}^* \hat{B}_f^{-1} \pi_{f,r}. \tag{4.67}$$

Proof. It is convenient to work in the biorthogonal coordinates. We have

$$\pi_{o,s}^{-*} B^{-1}(1 - \pi_{B^{-1}}) \pi_{o,r}^{-1} = [B]_o^{-1} \{1 - [P_r]_o ([P_s]_o^* [B]_o^{-1} [P_r]_o)^{-1} [P_s]_o^* [B]_o^{-1}\}, \tag{4.68}$$

where, recall, $[P_r]_o = [P_s]_o = \begin{bmatrix} 0 & 1 \end{bmatrix}^*$. We may expand $[B]_o^{-1}$ using a Schur complement decomposition,

$$[B]_o^{-1} = \begin{bmatrix} 1 & -\tilde{B}_f^{-1} \tilde{B}_r \\ & 1 \end{bmatrix} \begin{bmatrix} \tilde{B}_f^{-1} & \\ & S_B^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\tilde{B}_f^{-*} \tilde{B}_s \\ & 1 \end{bmatrix}^*, \tag{4.69}$$

where $S_B := \tilde{B}_c - \tilde{B}_s^* \tilde{B}_f^{-1} \tilde{B}_r$. By straightforward computation,

$$\begin{aligned} [P_s]_o^* [B]_o^{-1} &= \begin{bmatrix} -S_B^{-1} \tilde{B}_s^* \tilde{B}_f^{-1} & S_B^{-1} \end{bmatrix}, \\ ([P_s]_o^* [B]_o^{-1} [P_r]_o)^{-1} &= S_B, \quad \text{and} \\ 1 - [P_r]_o ([P_s]^* [B]_o^{-1} [P_r]_o)^{-1} [P_s]_o^* [B]_o^{-1} &= \begin{bmatrix} 1 & 0 \\ \tilde{B}_s^* \tilde{B}_f^{-1} & 0 \end{bmatrix} \end{aligned} \tag{4.70}$$

so that

$$\pi_{o,s}^{-*} B^{-1}(1 - \pi_{B^{-1}}) \pi_{o,r}^{-1} = \begin{bmatrix} \tilde{B}_f^{-1} & \\ & 0 \end{bmatrix}. \tag{4.71}$$

The conclusion follows by recalling $\tilde{B}_f = \hat{B}_f$ and (4.43), that $\pi_{f,j} = \begin{bmatrix} 1 & 0 \end{bmatrix} \pi_{o,j}$. □

4.6 Convergence Bounds

While Theorem 4.1 and its corollary supply convergence bounds of a sort, these results are difficult to make use of directly in practice, as they involve $S_f = A_f - \hat{A}_r \hat{A}_c^{-1} \hat{A}_s^*$, which cannot readily be formed, due to the presence of the inverse of the coarse operator $\hat{A}_c$. However, if the weights are close to ideal, then S_f will be close to A_f, and we can use it as a surrogate. We develop this idea here, by finding bounds on $\|1 - \hat{B}_f S_f\|$ equal to $\|1 - \hat{B}_f A_f\|$ perturbed by a term proportional to the norms of F_r and F_s, the differences between the actual and ideal weights. Thus the convergence bound is separated into a smoothing property and two approximation properties. The presence of two separate approximation properties, one for the trial and one for the test coarse space, instead of a single approximation property combining both, is a novel feature of the theory.

We begin by using the Cauchy-Schwarz-type inequality (3.20), a property of the absolute value of the previous chapter, to bound the sizes of $A_f - S_f$ and $S_f^{-1} - A_f^{-1}$ by a product of norms of F_r and F_s (recall that $\hat{A}_r = -A_f F_r$ and $\hat{A}_s = -A_f^* F_s$).

Lemma 4.2. *For an arbitrary symmetric coercive form defined by $X \in \mathscr{B}(\mathsf{V}_f, \mathsf{V}_f^*)$,*

$$\|A_f - S_f\|_{X^{-1},X} \leq \|\hat{A}_r\|_{X^{-1},|\hat{A}_c|} \|\hat{A}_s\|_{X^{-1},|\hat{A}_c|}, \quad \textit{and} \tag{4.72}$$

$$\|S_f^{-1} - A_f^{-1}\|_{X,X^{-1}} \leq \|F_r\|_{X,|S_c|} \|F_s\|_{X,|S_c|}. \tag{4.73}$$

Proof. Note that $A_f - S_f = \hat{A}_r \hat{A}_c^{-1} \hat{A}_s^*$ by the definition of S_f, (4.45). We have, using the dual norm characterization (2.24),

$$\|\hat{A}_r \hat{A}_c^{-1} \hat{A}_s^*\|_{X^{-1},X} = \sup_{\substack{\|u\|_X=1, \\ \|v\|_X=1}} |(\hat{A}_r \hat{A}_c^{-1} \hat{A}_s^* u, v)|. \tag{4.74}$$

Using the Cauchy-Schwarz-type inequality (3.20), the fact that $|A^{-1}| = |A|^{-1}$ (Theorem 3.2), as well as the identity (2.26) for norms of adjoints, we see that

$$\begin{aligned} |(\hat{A}_r \hat{A}_c^{-1} \hat{A}_s^* u, v)| &= |(\hat{A}_c^{-1} \hat{A}_s^* u, \hat{A}_r^* v)| \\ &\leq \|\hat{A}_s^* u\|_{|\hat{A}_c|^{-1}} \|\hat{A}_r^* v\|_{|\hat{A}_c|^{-1}} \\ &\leq \|\hat{A}_s^*\|_{|\hat{A}_c|^{-1},X} \|\hat{A}_r^*\|_{|\hat{A}_c|^{-1},X} \|u\|_X \|v\|_X \\ &= \|\hat{A}_s\|_{X^{-1},|\hat{A}_c|} \|\hat{A}_r\|_{X^{-1},|\hat{A}_c|} \|u\|_X \|v\|_X. \end{aligned} \tag{4.75}$$

This establishes the first inequality. A similar argument establishes the second, starting from the identity $S_f^{-1} - A_f^{-1} = F_r S_c^{-1} F_s^*$, which follows from (4.47). □

The following theorem gives our first convergence bound using norms associated with A_f. Note that the statement of the theorem makes it evident that the bound is symmetric in the sense of §4.1: applying the theorem to the adjoint iteration gives exactly the same result.

Theorem 4.2. *The spectral radius of the iteration operator E is bounded by*

$$\rho(E) \leq \|1 - \hat{B}_f S_f\|_{|A_f|} \leq s + \gamma_r \gamma_s + s\gamma_r\gamma_s \tag{4.76}$$

where

$$s = \|1 - \hat{B}_f A_f\|_{|A_f|} = \|1 - \hat{B}_f^* A_f^*\|_{|A_f|} \tag{4.77}$$

$$\gamma_r = \|F_r\|_{|A_f|,|\hat{A}_c|}, \qquad \gamma_s = \|F_s\|_{|A_f|,|\hat{A}_c|}. \tag{4.78}$$

Proof. That $\|1 - \hat{B}_f A_f\|_{|A_f|} = \|1 - A_f \hat{B}_f\|_{|A_f|^{-1}} = \|1 - \hat{B}_f^* A_f^*\|_{|A_f|}$ follows from (3.17) and (2.26). Since, by (3.17), $\|1 - A_f^{-1} S_f\|_{|A_f|} = \|A_f - S_f\|_{|A_f|^{-1},|A_f|}$, it follows from the previous lemma that

$$\|1 - A_f^{-1} S_f\|_{|A_f|} \leq \|\hat{A}_r\|_{|A_f|^{-1},|\hat{A}_c|} \|\hat{A}_s\|_{|A_f|^{-1},|\hat{A}_c|} = \|F_r\|_{|A_f|,|\hat{A}_c|} \|F_s\|_{|A_f|,|\hat{A}_c|}, \tag{4.79}$$

where the final step follows from the identities (4.24), $\hat{A}_r = A_f F_r$, $\hat{A}_s = A_f^* F_s$, and from (3.17). From Theorem 4.1 we have $\rho(E) \leq \|1 - \hat{B}_f S_f\|_{|A_f|}$. The conclusion now follows from the identity

$$1 - \hat{B}_f S_f = (1 - \hat{B}_f A_f) + (1 - A_f^{-1} S_f) - (1 - \hat{B}_f A_f)(1 - A_f^{-1} S_f) \tag{4.80}$$

by taking norms and applying the triangle inequality, and using the submultiplicative property of the norm $\|\cdot\|_{|A_f|}$. □

Note that the theorem actually gives a bound on the norm of $1 - \hat{B}_f S_f$, which can be applied to the bound (4.58) on the norm of the error from the corollary to Theorem 4.1.

Corollary. *The norm of the error e_n of the nth iteration, $n \geq 1$, of the two-level iteration* (4.57) *is bounded by*

$$\|e_n\|_{\mathsf{X}} \leq C(s + \gamma_r\gamma_s + s\gamma_r\gamma_s)^{n-1} \|e_0\|_{\mathsf{X}} \tag{4.81}$$

where $\|\cdot\|_{\mathsf{X}}$ is any norm equivalent to $\|\cdot\|_{\mathsf{V}}$ and where C is as in (4.59) *with the norm $\|\cdot\|_{\mathsf{X}_f}$ taken to be $\|\cdot\|_{|A_f|}$.*

Theorem 4.2 proves convergence of the two-level iteration only when the bound in (4.76) happens to be less than 1, which is not guaranteed. However, particularly in the case of an F-relaxation, s can be made arbitrarily small by performing sufficiently many smoothing iterations, so that a convergence proof is possible whenever $\gamma_r\gamma_s < 1$. We state this formally in the following corollary.

Corollary. *Suppose that the smoother B corresponds to k steps of the F-relaxation defined by $B_f^{(1)}$ so that*

$$1 - BA = (1 - B^{(1)}A)^k \quad \textit{where} \quad [B^{(1)}]_c = \begin{bmatrix} B_f^{(1)} & \\ & 0 \end{bmatrix}. \tag{4.82}$$

Then, provided $\|1 - B_f^{(1)}A_f\|_{|A_f|} < 1$ and $\gamma_r\gamma_s = \|F_r\|_{|A_f|,|\hat{A}_c|}\|F_s\|_{|A_f|,|\hat{A}_c|} < 1$, there is a minimum number of smoothing steps k_{min} such that the two-level iteration converges whenever $k \geq k_{min}$.

Proof. We have in this case

$$[B]_h = [B]_c = \begin{bmatrix} \hat{B}_f & \\ & 0 \end{bmatrix} \quad \text{where} \quad 1 - \hat{B}_f A_f = (1 - B_f^{(1)}A_f)^k. \tag{4.83}$$

That $[B]_h = [B]_c$ here is specific to the case of F-relaxations, and follows from (4.21). From (4.76) we see that two-level iteration converges provided

$$s < \frac{1 - \gamma_r\gamma_s}{1 + \gamma_r\gamma_s}, \tag{4.84}$$

and since

$$s = \|1 - \hat{B}_f A_f\|_{|A_f|} \leq \|1 - B_f^{(1)}A_f\|_{|A_f|}^k \tag{4.85}$$

it suffices to take $k_{\min}$ such that

$$\|1 - B_f^{(1)}A_f\|_{|A_f|}^{k_{\min}} < \frac{1 - \gamma_r\gamma_s}{1 + \gamma_r\gamma_s}. \tag{4.86}$$

□

Theorem 4.2 comes closest to generalizing known results for symmetric AMG, with an important discrepancy. In the result for the symmetric case, reproduced below as Theorem 4.4, the third term, $s\gamma_r\gamma_s$, appears with a negative sign. This change of sign appears not to be a deficiency in the theory, but rather fundamental to the nonsymmetry, capturing an essential difficulty the nonsymmetric case has over the symmetric case.

By basing the convergence on norms associated with the projected smoother $\hat{B}_f$ instead of A_f, we can avoid this last term. Again two forms of the bound are stated to emphasize that the bound is symmetric, giving an equal result when applied to the adjoint iteration.

Theorem 4.3. *If $\hat{B}_f$ has absolute value $|\hat{B}_f|$ then the spectral radius of the iteration operator E is bounded by*

$$\begin{aligned}\rho(E) \leq \|1 - \hat{B}_f S_f\|_{|\hat{B}_f|^{-1}} &\leq \|1 - A_f \hat{B}_f\|_{|\hat{B}_f|} + \|\hat{A}_r\|_{|\hat{B}_f|,|\hat{A}_c|}\|\hat{A}_s\|_{|\hat{B}_f|,|\hat{A}_c|} \\ &= \|1 - A_f^* \hat{B}_f^*\|_{|\hat{B}_f|} + \|\hat{A}_s\|_{|\hat{B}_f|,|\hat{A}_c|}\|\hat{A}_r\|_{|\hat{B}_f|,|\hat{A}_c|}.\end{aligned} \tag{4.87}$$

Proof. Using Theorem 4.1, (3.17), and the triangle inequality, we have

$$\begin{aligned}\rho(E) \leq \|1 - \hat{B}_f S_f\|_{|\hat{B}_f|^{-1}} &= \|\hat{B}_f^{-1} - S_f\|_{|\hat{B}_f|,|\hat{B}_f|^{-1}} \\ &\leq \|\hat{B}_f^{-1} - A_f\|_{|\hat{B}_f|,|\hat{B}_f|^{-1}} + \|A_f - S_f\|_{|\hat{B}_f|,|\hat{B}_f|^{-1}}.\end{aligned} \tag{4.88}$$

We simply apply Lemma 4.2 to the second term, while for the first term we have

$$\|\hat{B}_f^{-1} - A_f\|_{|\hat{B}_f|,|\hat{B}_f|^{-1}} = \|1 - A_f \hat{B}_f\|_{|\hat{B}_f|} = \|1 - \hat{B}_f A_f\|_{|\hat{B}_f|^{-1}} = \|1 - A_f^* \hat{B}_f^*\|_{|\hat{B}_f|} \tag{4.89}$$

as a consequence of (3.17) and (2.26). □

As was the case for the previous theorem, we have actually proved a bound on a norm of $1 - \hat{B}_f S_f$, and we can apply the corollary to Theorem 4.1 to get a bound on the norm of the error.

Corollary. *The norm of the error e_n of the nth iteration, $n \geq 1$, of the two-level iteration* (4.57) *is bounded by*

$$\|e_n\|_{\mathsf{X}} \leq C \left(\|1 - A_f \hat{B}_f\|_{|\hat{B}_f|} + \|\hat{A}_r\|_{|\hat{B}_f|,|\hat{A}_c|}\|\hat{A}_s\|_{|\hat{B}_f|,|\hat{A}_c|}\right)^{n-1} \|e_0\|_{\mathsf{X}} \tag{4.90}$$

where $\|\cdot\|_{\mathsf{X}}$ is any norm equivalent to $\|\cdot\|_{\mathsf{V}}$ and where C is as in (4.59) *with the norm $\|\cdot\|_{\mathsf{X}_f}$ taken to be $\|\cdot\|_{|\hat{B}_f|^{-1}}$.*

The specialized version of Theorem 4.2 that applies to symmetric problems (in which the third term in the bound appears with a negative sign) can be shown under mild assumptions on $\hat{B}_f$ to always give a bound that is less than one, so that it does in fact prove convergence for the two-level iteration. In contrast, the bounds of Theorems 4.2 and 4.3 are not guaranteed to be less than one. While we did manage to prove convergence in the corollary to Theorem 4.2, this was under the premise of the weights and the smoother satisfying certain conditions, conditions that are not necessary in the symmetric case. In the general case, in which we cannot show convergence, we can still at least prove a bound on a condition number of $\hat{B}_f S_f$.

Proposition 4.5. *Let $\kappa_{|A_f|}(L) := \|L\|_{|A_f|}\|L^{-1}\|_{|A_f|}$ denote the condition number of an operator $L \in \mathscr{B}(\mathsf{V}_f)$ with respect to the indicated absolute value norm. Then*

$$\begin{aligned}\kappa_{|A_f|}(\hat{B}_f S_f) \leq \kappa_{|A_f|}(\hat{B}_f A_f)(1 + \|F_r\|_{|A_f|,|\hat{A}_c|}\|F_s\|_{|A_f|,|\hat{A}_c|}) \\ (1 + \|F_r\|_{|A_f|,|S_c|}\|F_s\|_{|A_f|,|S_c|}).\end{aligned} \tag{4.91}$$

Proof. We have

$$\begin{aligned}\|\hat{B}_f S_f\|_{|A_f|} &\leq \|\hat{B}_f A_f\|_{|A_f|}\|A_f^{-1} S_f\|_{|A_f|} \\ &\leq \|\hat{B}_f A_f\|_{|A_f|}(1 + \|1 - A_f^{-1} S_f\|_{|A_f|}) \\ &\leq \|\hat{B}_f A_f\|_{|A_f|}(1 + \|F_r\|_{|A_f|,|\hat{A}_c|}\|F_s\|_{|A_f|,|\hat{A}_c|}),\end{aligned} \tag{4.92}$$

where the last step comes from the proof of Theorem 4.2. Similarly,

$$\begin{aligned}\|(\hat{B}_f S_f)^{-1}\|_{|A_f|} &\leq \|(\hat{B}_f A_f)^{-1}\|_{|A_f|}\|S_f^{-1} A_f\|_{|A_f|} \\ &\leq \|(\hat{B}_f A_f)^{-1}\|_{|A_f|}(1 + \|S_f^{-1} A_f - 1\|_{|A_f|}) \\ &= \|(\hat{B}_f A_f)^{-1}\|_{|A_f|}(1 + \|S_f^{-1} - A_f^{-1}\|_{|A_f|,|A_f|^{-1}}) \\ &\leq \|(\hat{B}_f A_f)^{-1}\|_{|A_f|}(1 + \|F_r\|_{|A_f|,|S_c|}\|F_s\|_{|A_f|,|S_c|}),\end{aligned} \tag{4.93}$$

where the last step is an application of Lemma 4.2. The result now follows from the product of these two inequalities. □

4.7 Symmetric Case

A major goal in developing the convergence theory of the previous section was to generalize the existing theory for symmetric problems. In this section we shall see how the nonsymmetric theory specializes in the symmetric case to reproduce some of the known convergence results from the literature. In this section we assume $A = A^*$, and that $W_r = W_s = W$, so that $P_r = P_s = P$, $F_r = F_s = F$. Note that the absolute value norms of the previous section reduce to the usual energy norms under these assumptions, as all of the operators (e.g., A_f and $\hat{A}_c$) define coercive symmetric forms.

The following general results will be very useful in this section. If both $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ and $B \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$ define coercive symmetric forms, then BA is a selfadjoint operator on both $\mathsf{H}_{B^{-1}}$ and H_A, as, e.g.,

$$(B^{-1}BAu, v) = (Au, v) = (B^{-1}u, BAv). \tag{4.94}$$

Therefore, the spectral radius is equal to the norm of the operator in either Hilbert space. Also, the convex hull of the spectrum $\sigma(BA)$ is the closure of the numerical range, the set of values of

$$\frac{(B^{-1}BAu, u)}{(B^{-1}u, u)} = \frac{\|u\|_A^2}{\|u\|_{B^{-1}}^2} \tag{4.95}$$

for $u \in \mathsf{V}$, which is a subset of the positive real axis. It follows that

$$\begin{aligned} &\sup_{\lambda\in\sigma(BA)} \lambda = \rho(BA) = \|BA\|_{B^{-1}} = \|BA\|_A = \|1\|^2_{A,B^{-1}} = \|1\|^2_{B,A^{-1}}, \quad \text{and} \\ &\inf_{\lambda\in\sigma(BA)} \lambda = \left[\sup_{u\in\mathsf{V}\backslash\{0\}} \frac{\|u\|^2_{B^{-1}}}{\|u\|^2_A}\right]^{-1} = \rho(A^{-1}B^{-1})^{-1}. \end{aligned} \tag{4.96}$$

The norms of the two identity operators are equal due to (2.26). We will make much use of these relationships below.

In the previous section, a few different norms of the weight errors F_r and F_s were used. In the symmetric case, there is a precise relationship between the energy-based ones (those not involving the smoother). The next proposition expresses several relevant norms in terms of a single angle $\phi \in (0, \pi/2]$, the abstract angle between the subspaces V_f and $\mathsf{W}_r = \mathsf{W}_s$.

Proposition 4.6. *In the symmetric case,*

$$\|1\|_{S_c,\hat{A}_c} \leq 1, \qquad \|1\|_{S_f,A_f} \leq 1, \tag{4.97}$$

and there exists an angle $\phi \in (0, \pi/2]$ such that

$$\csc\phi = \|1\|_{\hat{A}_c,S_c} = \|1\|_{A_f,S_f} \geq 1, \tag{4.98}$$

$$\cot\phi = \|F\|_{A_f,S_c}, \quad \textit{and} \tag{4.99}$$

$$\cos\phi = \|F\|_{A_f,\hat{A}_c} < 1. \tag{4.100}$$

Proof. In the symmetric case (4.26) reduces to $\hat{A}_c = S_c + F^* A_f F$, implying that $\|Fu\|^2_{A_f} = \|u\|^2_{\hat{A}_c} - \|u\|^2_{S_c}$ and $\|u\|^2_{\hat{A}_c} \geq \|u\|^2_{S_c}$, from which it follows that $\|1\|_{S_c,\hat{A}_c} \leq 1$. We also have that

$$\|F\|^2_{A_f,S_c} = \sup_{u\in\mathsf{V}_c\backslash\{0\}} \frac{\|u\|^2_{\hat{A}_c} - \|u\|^2_{S_c}}{\|u\|^2_{S_c}} = \|1\|^2_{\hat{A}_c,S_c} - 1, \quad \text{and} \tag{4.101}$$

$$\|F\|^2_{A_f,\hat{A}_c} = \sup_{u\in\mathsf{V}_c\backslash\{0\}} \frac{\|u\|^2_{\hat{A}_c} - \|u\|^2_{S_c}}{\|u\|^2_{\hat{A}_c}} = 1 - \|1\|^{-2}_{\hat{A}_c,S_c}. \tag{4.102}$$

Similarly, (4.47) reduces to $S_f^{-1} = A_f^{-1} + F S_c^{-1} F^*$. Note that S_f^{-1}, and therefore S_f, is positive symmetric as $S_f^{-1} \geq A_f^{-1}$. It also follows that $\|1\|_{A_f^{-1},S_f^{-1}} = \|1\|_{S_f,A_f} \leq 1$. Finally, $\|F^* f\|^2_{S_c^{-1}} = \|f\|^2_{S_f^{-1}} - \|f\|^2_{A_f^{-1}}$ and

$$\begin{aligned} \|F\|^2_{A_f,S_c} = \|F^*\|^2_{S_c^{-1},A_f^{-1}} &= \sup_{f\in\mathsf{V}_f^*\backslash\{0\}} \frac{\|f\|^2_{S_f^{-1}} - \|f\|^2_{A_f^{-1}}}{\|f\|^2_{A_f^{-1}}} \\ &= \|1\|^2_{S_f^{-1},A_f^{-1}} - 1 = \|1\|^2_{A_f,S_f} - 1. \end{aligned} \tag{4.103}$$

The proposition now follows by taking $\phi = \operatorname{arccot} \|F\|_{A_f, S_c}$ and applying basic trigonometric identities. $\square$

Let us remark that the constant $\gamma = \cos\phi = \|F\|_{A_f, \hat{A}_c} = \|\hat{A}_r\|_{A_f^{-1}, \hat{A}_c} < 1$ has long played an important role in the analysis of multilevel methods, due to its presence in the inequality

$$|(\hat{A}_r u, v)| \leq \|\hat{A}_r u\|_{A_f^{-1}} \|v\|_{A_f} \leq \gamma \|u\|_{\hat{A}_c} \|v\|_{A_f} \qquad \text{for all } u \in \mathsf{V}_c, v \in \mathsf{V}_f, \tag{4.104}$$

which can equivalently be written

$$|(Au, v)| \leq \gamma \|u\|_A \|v\|_A \qquad \text{for all } u \in \mathsf{W}_r = \mathsf{W}_s, v \in \mathsf{V}_f. \tag{4.105}$$

This is known as the "strengthened" Cauchy-Schwarz inequality, and is the context in which γ is seen as the cosine of an abstract angle between subspaces. See, for example, the paper by Eijkhout and Vassilevski [15] and the references therein. Let us also note that the pair of results $\csc\phi = \|1\|_{\hat{A}_c, S_c}$ and $\|1\|_{S_c, \hat{A}_c} \leq 1$ from the above proposition are essentially Lemma 2.1 of the two-level analysis by Falgout, Vassilevski, and Zikatanov [19]. To our knowledge, the use of S_f, the Schur complement of the coarse operator, and the results involving it, have not appeared before in the literature.

Proposition 4.6 provides a satisfying geometric picture in the symmetric case. All of the listed norms of interest are expressible in terms of a single angle ϕ, the abstract angle between the subspaces V_f and $\mathsf{W}_r = \mathsf{W}_s$. Note that the ideal weights case $F = 0$ corresponds to full orthogonality ($\phi = \pi/2$). There is no analog to this proposition relating the various norms in the general nonsymmetric case. Neither is there an analog to the bound $\|F\|_{A_f, \hat{A}_c} < 1$.

It is also worth emphasizing the result $\|1\|_{S_c, \hat{A}_c} \leq 1$, which means simply that $\|u\|_{\hat{A}_c}^2 \geq \|u\|_{S_c}^2$ for arbitrary u. This is due to the fact that the difference $A_c - S_c$ is $F_s^* A_f F_r$, which defines a coercive symmetric form in the symmetric case. In the general nonsymmetric case, not only is A_f not symmetric, but more importantly there is no reason to assume that F_r and F_s are correlated in any way, not even if we choose $W_s = W_r$. Summarizing, the result $\|1\|_{S_c, \hat{A}_c} \leq 1$, as well as the similar result regarding A_f and S_f, is quite specific to the symmetric case.

All of the groundwork is now in place to prove the following convergence bound, which is similar to, but improves, Theorem 4.2. The improvements can be seen to be due to the special features of the symmetric case just highlighted.

Theorem 4.4. *Let $\hat{B}_f$ be symmetric and let*

$$s = \|1 - \hat{B}_f A_f\|_{A_f}, \quad \textit{and} \quad \gamma = \|F\|_{A_f, \hat{A}_c} = \cos\phi. \tag{4.106}$$

If $s < 1$, then the spectral radius of the iteration operator E is bounded by

$$\begin{aligned} \rho(E) &\leq s + \gamma^2 - s\gamma^2 \\ &= 1 - (1-s)\sin^2\phi < 1. \end{aligned} \tag{4.107}$$

Proof. First, note that the assumption $s < 1$ implies that $\hat{B}_f$ is positive, since for any $f \in \mathsf{V}_f^*$, letting $u = A_f^{-1} f$, we have

$$\begin{aligned} (f, \hat{B}_f f) &= (A_f u, \hat{B}_f A_f u) \\ &= (A_f u, u) - (A_f u, [1 - \hat{B}_f A_f] u) \\ &\geq \|u\|_{A_f}^2 - \|1 - \hat{B}_f A_f\|_{A_f} \|u\|_{A_f}^2 \\ &= (1-s)\|f\|_{A_f^{-1}}^2. \end{aligned} \tag{4.108}$$

This inequality also implies $\|(\hat{B}_f A_f)^{-1}\|_{A_f} = \|1\|_{A_f^{-1}, \hat{B}_f}^2 \leq (1-s)^{-1}$. The inequality (4.93) becomes

$$\begin{aligned} \|(\hat{B}_f S_f)^{-1}\|_{A_f} &\leq \|(\hat{B}_f A_f)^{-1}\|_{A_f} (1 + \|F\|_{A_f, S_c}^2) \\ &\leq (1-s)^{-1}(1 + \cot^2\phi) \\ &= [(1-s)\sin^2\phi]^{-1}, \end{aligned} \tag{4.109}$$

while (4.92) improves to

$$\|\hat{B}_f S_f\|_{A_f} \leq \|\hat{B}_f A_f\|_{A_f} \|A_f^{-1} S_f\|_{A_f} \leq 1 + s \tag{4.110}$$

where we have used that $\|A_f^{-1} S_f\|_{A_f} = \|1\|_{S_f, A_f}^2 \leq 1$, by the previous proposition. Hence, for $\mu \in \sigma(B_f S_f)$,

$$(1-s)\sin^2\phi \leq \rho(S_f^{-1}\hat{B}_f^{-1})^{-1} \leq \mu \leq \rho(\hat{B}_f S_f) \leq 1 + s. \tag{4.111}$$

Since by Theorem 4.1, $\sigma(E) \cup \{0\} = \sigma(1 - \hat{B}_f S_f) \cup \{0\}$, we have that

$$-s \leq \lambda \leq 1 - (1-s)\sin^2\phi \tag{4.112}$$

for $\lambda \in \sigma(E)$, and the result follows. □

This theorem reproduces precisely the same convergence bound as Theorem 4.2 of the Falgout et al. analysis [19] mentioned earlier, though that theorem is restricted to symmetrically paired pre- and post-smoothers consisting of F-relaxations. On the other

hand, Theorem 4.4 applies not just for F-relaxations but rather for general (symmetric) smoothers, and also, for example, in the case of a symmetric post-smoother and no pre-smoother, or vice versa. We used Theorem 4.1 to reduce the general case to the analysis of F-relaxations, whereas Falgout, Vassilevski, and Zikatanov pursue a different line of analysis to handle the general case, although a significant amount of the treatment was unified. Actually, Theorem 4.3 of their analysis, which applies to general smoothers (but still a symmetrically paired pre- and post-smoother), is equivalent to Theorem 4.1 above. The equivalence of Theorem 4.1 to a result in Notay's nonsymmetric analysis, specifically part (1) of Theorem 2.1, was discussed at the end of §4.5. Notay's result, in turn, directly generalizes Theorem 4.3 of Falgout et al.

Chapter 5

Application to a new AMG method

In this chapter we will give an example of how the abstract convergence theory of the previous chapter can be applied to develop AMG methods for nonsymmetric problems. First we will lay out an overall framework inspired by the theory, before discussing the design of each AMG component in turn. The approach to designing AMG methods presented in this chapter is merely one among many possible approaches employing the convergence theory of the previous chapter. It is intended as an example and as a proof-of-concept that the abstract theory *can* be used to inform the design of concrete methods. We will close the chapter with numerical tests of the newly developed methods, including some numerical demonstrations of how well the convergence theory describes actual concrete methods.

The particular approach taken here will be to aim for robustness, in particular, prioritizing robustness over the speed of set-up. While set-up run times that scale polynomially with problem size are out of the question, we will not mind a large constant in a linear or log-linear scaling. Another prime concern will be for both the set-up and execution of the method to be parallelizable, at least in principle. The biggest constraint this adds is to rule out W-cycles, which are well-known to scale poorly, as they imply a latency cost that is exponential in the number of levels, as opposed to linear for V-cycles. This makes the problem significantly harder, as W-cycles are one convenient way to make up for "poor" quality interpolation and restriction operators, which can otherwise have many significant benefits, for example, in limiting stencil growth.

5.1 Independent Quality Measures

Based on the analysis of the previous chapter, we propose quantifying the "quality" of each component of the two-grid iteration as follows.

Table 5.1: AMG Component Quality Measures

Component	Quality	Cost
Coarsening	$\|1 - D_f^{-1} A_f\|_{\|D_f\|}$	n_c/n
Trial weights	$\|\hat{A}_r\|_{X_f^{-1},\|\hat{A}_c\|}$	$\mathrm{nnz}(W_r)$
Test weights	$\|\hat{A}_s\|_{X_f^{-1},\|\hat{A}_c\|}$	$\mathrm{nnz}(W_s)$
Smoother	$\|1 - \hat{B}_f A_f\|_{X_f}$	cost of applying B

$$X_f = |A_f| \text{ or } |\hat{B}_f|^{-1}$$
$$\hat{A}_r := A_f W_r + A_r, \qquad \hat{A}_s := A_f^* W_s + A_s$$
$$\hat{B}_f := \begin{bmatrix} I & -W_r \end{bmatrix} [B]_c \begin{bmatrix} I & -W_s \end{bmatrix}^*$$

For easy reference, we have repeated definitions (4.22) and (4.54) from the previous chapter of the coupling operators $\hat{A}_r$ and $\hat{A}_s$ and the projection of the smoother to an F-relaxation $\hat{B}_f$. The matrix D_f here, and in the remainder of the chapter, denotes the diagonal part of A_f. In all cases, smaller numbers in the table are better. Before explaining the theoretical justification of the quality measures, let us highlight a few properties. First, when we take $X_f = |A_f|$, the measures have no forward dependencies. That is, the norm indicating coarsening quality is independent of the weights and the smoother, and the norms for the weights in this case do not depend on the smoother. This feature will allow us to develop heuristics for these components targeting each quality measure, without having to worry about the components yet to be constructed. Second, the quality indicator in each row is in opposition to the cost indicator. Improving the quality by lowering the norm will generally increase the cost, and vice versa. For example, choosing the ideal weights $-A_f^{-1} A_r$ makes the norm of $\hat{A}_r$ zero, but also maximizes the number of nonzeros $\mathrm{nnz}(W_r)$. The challenge of constructing good components is to balance the quality against the cost.

The choice of the norms in the last three rows is explained by their presence in the convergence bounds of Theorems 4.2 and 4.3 (depending on the choice of X_f); lowering any of them improves the corresponding bound.

The norm indicating the quality of the coarsening is a bound on the convergence rate of the Jacobi iteration applied to F-variables. The idea is that, in order for it to be possible to construct good quality weights and smoothers that also have low associated cost, it needs to be "easy" to solve systems governed by A_f (and A_f^*). In the case of the smoother, this is direct. In particular, if $\|1 - D_f^{-1} A_f\|_{|D_f|}$ is small, then the F-relaxation consisting of a few steps of Jacobi suffices to make $1 - \hat{B}_f A_f = (1 - D_f^{-1} A_f)^k$

small. In the case of the weights W_r, the situation for W_s being similar, we see that they need to be a sparse approximate solution to $\hat{A}_r = A_f W_r + A_r = 0$. Finding a good sparse approximation in general is only possible if systems governed by A_f are easy to solve. To be precise, the following proposition shows that the entries of the matrix A_f^{-1}, and therefore the ideal weights $-A_f^{-1}A_r$, decay exponentially at a rate controlled by $\|1 - D_f^{-1}A_f\|_{|D_f|}$.

Proposition 5.1. *Let $\mathbf{e}_i$ denote the ith coordinate vector, and let $[A]_{ij} := (A\mathbf{e}_j, \mathbf{e}_i)$ denote the entry of the matrix A in row i and column j. Define the graph distance*

$$d_A(i,j) := \min\{k \mid k \geq 0, [X^k]_{ij} \neq 0\} \quad \textit{where} \quad [X]_{ij} = |[A]_{ij}|. \tag{5.1}$$

If A and D are nonsingular matrices of the same size with D diagonal, then

$$|[A^{-1}]_{ij}| \leq \frac{Cq^{d_A(i,j)}}{\sqrt{|[D]_{ii}[D]_{jj}|}} \tag{5.2}$$

where

$$\begin{aligned} q &= \|1 - D^{-1}A\|_{|D|} \qquad \textit{and} \\ C &= \|A^{-1}\|_{|D|,|D|^{-1}} \leq (1-q)^{-1}, \end{aligned} \tag{5.3}$$

with the final inequality holding when $q < 1$.

Proof. We have that

$$\begin{aligned} (1 - D^{-1}A)^n A^{-1} &= A^{-1} - \sum_{k=1}^{n} \binom{n}{k} (-D^{-1}A)^{k-1} D^{-1} \\ &= A^{-1} + p(D^{-1}A)D^{-1} \end{aligned} \tag{5.4}$$

where p is a polynomial of degree $n-1$. Since $[(D^{-1}A)^k]_{ij}$ is necessarily zero when $[X^k]_{ij}$ is, it follows that

$$[A^{-1}]_{ij} = [(1 - D^{-1}A)^n A^{-1}]_{ij} \qquad \text{for } n \geq d_A(i,j). \tag{5.5}$$

The main result now follows by applying the generalized Cauchy-Schwarz inequality,

$$\begin{aligned} |[(1 - D^{-1}A)^n A^{-1}]_{ij}| &= ([1 - D^{-1}A]^n A^{-1}\mathbf{e}_j, \mathbf{e}_i) \\ &\leq \|(1 - D^{-1}A)^n A^{-1}\mathbf{e}_j\|_{|D|} \|\mathbf{e}_i\|_{|D|^{-1}} \\ &\leq \|1 - D^{-1}A\|_{|D|}^n \|A^{-1}\|_{|D|,|D|^{-1}} \|\mathbf{e}_j\|_{|D|^{-1}} \|\mathbf{e}_i\|_{|D|^{-1}} \\ &= |[D]_{jj}|^{-1/2} |[D]_{ii}|^{-1/2} \|A^{-1}\|_{|D|,|D|^{-1}} q^n. \end{aligned} \tag{5.6}$$

Note that, in the case of a diagonal matrix, the form absolute value of Chapter 3 coincides with the entry-wise absolute value (as a consequence of coinciding in the scalar case). It remains to show that $\|A^{-1}\|_{|D|,|D|^{-1}} \leq (1-q)^{-1}$ when $q < 1$, but

$$\begin{aligned}\|A^{-1}\|_{|D|,|D|^{-1}} &= \sup_{f\neq 0} \frac{\|A^{-1}f\|_{|D|}}{\|f\|_{|D|^{-1}}} = \sup_{g\neq 0} \frac{\|A^{-1}AD^{-1}g\|_{|D|}}{\|AD^{-1}g\|_{|D|^{-1}}} \\ &= \sup_{g\neq 0} \frac{\|g\|_{|D|^{-1}}}{\|[1-(1-AD^{-1})]g\|_{|D|^{-1}}} \\ &\leq \sup_{g\neq 0} \frac{\|g\|_{|D|^{-1}}}{|1-\|1-AD^{-1}\|_{|D|^{-1}}|\|g\|_{|D|^{-1}}} = \frac{1}{1-q}\end{aligned} \tag{5.7}$$

since $\|1-AD^{-1}\|_{|D|^{-1}} = \|1-D^{-1}A\|_{|D|}$. □

The inspiration for this proposition comes from a similar result by Demko, Moss, and Smith [13] about exponential decay of the entries of the inverse of a sparse symmetric positive definite matrix. That result is stronger, owing to the fact that significantly better approximating polynomials for A^{-1} are available in the symmetric case (e.g., Chebyshev polynomials). The significance of these results in the context of algebraic multigrid was pointed out by Brannick and Zikatanov [9], who attribute the contribution to a private communication from Vassilevski.

Our notion of coarsening quality bears a strong resemblance to Brandt's concept of coarsening by "compatible relaxation" [5], as it is indeed a bound on the convergence rate of the Jacobi iteration on F-variables. However, we do not explicitly assume a smoother of this type; rather, our notion of coarsening quality is independent of the smoother. It was in the context of compatible relaxation that Brannick and Zikatanov [9] brought up the result of Demko, Moss, and Smith [13], and indeed they argue that the conditioning of A_f and the convergence rate of compatible relaxation are correlated.

5.2 Coarsening

The coarsening procedure is responsible for partitioning the unknowns into the C- and F-variables. According to Table 5.1, a good procedure should bound $\|1-D_f^{-1}A_f\|_{|D_f|}$ without selecting more C-variables than necessary.

As a first step, let us construct a procedure that bounds the spectral radius

$$\rho(1-D_f^{-1}A_f) = \rho(1-D_f^{-1/2}A_fD_f^{-1/2}). \tag{5.8}$$

Call the second matrix X, and let S be its entry-wise absolute value.

$$X = 1 - D_f^{-1/2} A_f D_f^{-1/2}, \qquad [S]_{ij} = |[X]_{ij}| \tag{5.9}$$

The absolute row and column sums of X are given by

$$r_i = \mathbf{e}_i^* S \mathbf{1}, \qquad c_j = \mathbf{1}^* S \mathbf{e}_j, \tag{5.10}$$

where $\mathbf{1}$ is the vector of all ones and $\mathbf{e}_i$ is the ith coordinate vector. Because the diagonal elements of X are all 0, these row and column sums are also the radii of the Gershgorin discs associated with X and X^*, which are all centered at 0. It follows that $\rho(X)$ is bounded by both the largest row sum and largest column sum. In fact, Ostrowski's generalization to the Gershgorin circle theorem [41] implies that

$$\rho(X) \leq \max_i r_i^\alpha c_i^{1-\alpha} \quad \text{for} \quad 0 \leq \alpha \leq 1. \tag{5.11}$$

Let g_i be the geometric mean of the ith row and column sum,

$$g_i := r_i^{1/2} c_i^{1/2}, \tag{5.12}$$

so that Ostrowski's result, applied with $\alpha = 1/2$, implies

$$\rho(1 - D_f^{-1} A_f) = \rho(X) \leq \max_i g_i. \tag{5.13}$$

A simple greedy strategy to select a set of C-variables such that $\rho(1 - D_f^{-1} A_f)$ is no greater than a given value ρ is to start with no C-variables, then repeatedly eliminate F-variables with the largest Ostrowski radii by changing them to C-variables, until all remaining radii are at most ρ. Obviously, removing the largest radius will lower the maximum, but there is a secondary effect. The row sum for a given variable is also the sum of that variable's contributions to the column sums of its neighbors, a similar statement applying to its column sum. Hence, eliminating variables with the largest Ostrowski radii can be expected to be very effective in reducing the radii of the remaining variables. This gives us the following algorithm, which bounds $\rho(1 - D_f^{-1} A_f)$ by the input parameter ρ.

Note that the matrix S is defined on the F-variables so that it shrinks on each loop iteration. By "locally maximal" in line 7, we mean $g_i \geq g_j$ for all $j \in I$ that are neighbors in the adjacency graph, i.e., such that $S_{ij} \neq 0$ or $S_{ji} \neq 0$. Ties can be resolved by referring to the arbitrary ordering of unknowns. A problem that arises with the denser matrices that show up at the coarser levels, which contain many very small entries, is that the neighbor sets become quite large, needlessly slowing the

Algorithm 1 Ostrowski coarsening with spectral bound

function $C \leftarrow$ COARSEN1(A, ρ)
  $C \leftarrow \{\}, \quad F \leftarrow \{1, \ldots, n\}$
  loop
    $g_i \leftarrow (\mathbf{e}_i^* S \mathbf{1})^{1/2} (\mathbf{1}^* S \mathbf{e}_i)^{1/2}$ for each $i \in F$
    $I \leftarrow \{i \in F \mid g_i \geq \rho\}$
    if $I = \{\}$ **then stop**
    $C \leftarrow C \cup \{i \in I \mid g_i$ locally maximal among $I\}$
    $F \leftarrow \{1, \ldots, n\} \setminus C$
  end loop
end function

procedure down. A remedy for this is to take the neighbor set to be, instead, those j such that $S_{ij} \geq \theta \max_j \max(S_{ij}, S_{ji})$, where θ is a parameter we typically take to be 0.1.

Let us highlight some features of this simple algorithm. The use of the Ostrowski bound ensures that the procedure behaves identically when applied to a matrix A or its adjoint A^*. Positive off-diagonal entries pose no difficulty and require no special treatment. The matrix S resembles the "strength of connection" heuristic, and g_i resembles the "number of strong connections" heuristic. However g_i is not integral, and does not depend on classifying connections as being either "strong" or "weak." The procedure is inherently parallelizable and resembles the CLJP coarsening procedure of Cleary et al. [12]; but because g_i is not integral, it is not necessary to add a random number to it to create local maxima (although that's a possible alternative strategy for resolving ties).

As it stands, our procedure bounds

$$\rho(1 - D_f^{-1} A_f) = \lim_{n \to \infty} \|(1 - D_f^{-1} A_f)^n\|^{1/n}, \tag{5.14}$$

which holds for any induced norm. Because we are dealing with nonsymmetric matrices, there is a real possibility of transient growth: for small n, $\|(1 - D_f^{-1} A_f)^n\|$ may increase with n, perhaps quite dramatically, even when the spectral radius is less than 1. Hence, it would be preferable to have a procedure that bounds $\|1 - D_f^{-1} A_f\|_{|D_f|}$. Let us investigate how the above procedure can be modified to achieve this.

First, let us note that

$$\|1 - D_f^{-1} A_f\|_{|D_f|} = \|1 - D_f^{-1/2} A_f D_f^{-1/2}\|_2 = \|X\|_2, \tag{5.15}$$

where $\|\cdot\|_2$ is the matrix 2-norm. This is more or less immediate in the case of real A_f, but holds even when A_f has complex entries. To see this, let us write $D_f = e^{i\Theta} |D_f|$

where Θ is diagonal with real entries, and such that $D_f^{1/2} = e^{i\Theta/2}|D_f|^{1/2}$. Because A is assumed to have positive real part, the diagonal entries of D_f are necessarily in the right half of the complex plane, so that we can pick the entries of Θ to lie in the interval $(-\pi/2, \pi/2)$, which makes the above square root principal. Then we can write

$$\begin{aligned} \|1 - D_f^{-1}A_f\|_{|D_f|} &= \||D_f|^{1/2}(1 - D_f^{-1}A_f)|D_f|^{-1/2}\|_2 \\ &= \|e^{-i\Theta/2}(1 - D_f^{-1/2}A_f D_f^{-1/2})e^{i\Theta/2}\|_2 \\ &= \|1 - D_f^{-1/2}A_f D_f^{-1/2}\|_2 = \|X\|_2. \end{aligned} \tag{5.16}$$

We can use the following result by Nikiforov [38] to bound $\|X\|_2$.

Theorem 5.1 (Nikiforov [38])**.** *For all $m \times n$ matrices S with nonnegative real entries, and integers $r \geq 0$, $p \geq 1$,*

$$\|S\|_2^{2p} \leq \max_{i, w_i^{(r)} \neq 0} \frac{w_i^{(r+p)}}{w_i^{(r)}}, \quad \textit{where} \quad \mathbf{w}^{(r)} := (SS^*)^r \mathbf{1}. \tag{5.17}$$

The bound typically improves as r increases. In fact, Nikiforov proves that, for most matrices, the bound approaches $\|S\|_2^{2p}$ in the limit $r \to \infty$ (the exception being matrices for which $\mathbf{1}$ is orthogonal to the eigenspace of SS^* associated to the largest singular value of S). Also, as Nikiforov notes, the bound may be applied to any complex matrix A by taking S to be its entry-wise absolute value, as $\|A\|_2 \leq \|S\|_2$ always holds.

Algorithm 2 Ostrowski coarsening with norm bound

function $C \leftarrow$ COARSEN2(A, ρ)
  $C \leftarrow \{\}, \quad F \leftarrow \{1, \ldots, n\}$
  loop
    $\mathbf{v}^{(1)} \leftarrow S^*S\mathbf{1}, \quad \mathbf{v}^{(2)} \leftarrow S^*S\mathbf{v}^{(1)}$
    $\mathbf{w}^{(1)} \leftarrow SS^*\mathbf{1}, \quad \mathbf{w}^{(2)} \leftarrow SS^*\mathbf{w}^{(1)}$
    $I \leftarrow \{i \in F \mid v_i^{(1)} \neq 0, v_i^{(2)}/v_i^{(1)} \geq \rho^2\} \cup \{i \in F \mid w_i^{(1)} \neq 0, w_i^{(2)}/w_i^{(1)} \geq \rho^2\}$
    if $I = \{\}$ **then stop**
    $g_i \leftarrow (\mathbf{e}_i^* S\mathbf{1})^{1/2}(\mathbf{1}^* S\mathbf{e}_i)^{1/2}$ for each $i \in F$
    $C \leftarrow C \cup \{i \in I \mid g_i$ locally maximal among $I\}$
    $F \leftarrow \{1, \ldots, n\} \setminus C$
  end loop
end function

Returning to our application, we have $\|X\|_2 \leq \|S\|_2$, with Nikiforov's bound applying to S. Algorithm 2 applies the bound, to both S and S^*, with $r = 1$ and $p = 1$, thereby choosing a set of C-variables such that $\|I - D_f^{-1}A_f\|_{|D_f|}$ is bounded

by the input parameter ρ. While it is possible to make use of the bounds given by higher values of r, doing so increases the cost of the procedure for diminishing returns. The $r = 0$, $p = 1$ bound is already significantly stronger than the Ostrowski radius for a symmetric matrix (in which case $\rho(S) = \|S\|_2$, making the bounds comparable), and the $r = 1$ bound is stronger still. This procedure differs from the previous one only in the selection of the set of candidate indices I from which C-variables are chosen. In particular, it is still those variables with the largest Ostrowski radii that are selected to be C-variables. This was found to work better, in the sense of resulting in fewer C-variables, than using indicators based on the Nikiforov bound. An explicit (unoptimized) Matlab implementation is shown in Figure 5.1.

Shown in Figure 5.3 is an example hierarchy produced by Algorithm 2 with the parameter $\rho = 0.8$ for an advection-diffusion problem, the double glazing problem featured previously with advection velocity as shown in Figure 5.2, discretized on a 32 by 32 Chebyshev grid using the streamline diffusion scheme introduced in §1.2.3. Figure 5.3 illustrates several features of the coarsening algorithm. First, as we might expect, the procedure automatically semi-coarsens in the direction of strong coupling, which is along streamlines in this case. Notably, the final three C-variables on the bottom level are adjacent to each other. That the hierarchy bottoms out at three degrees of freedom is not arbitrary but rather determined by the coarsening algorithm itself: Algorithm 2 produces zero C-variables when run on the final 3×3 matrix, indicating that a simple Jacobi iteration is sufficient to solve this problem, and that there is no need to recurse further.

5.3 Interpolation

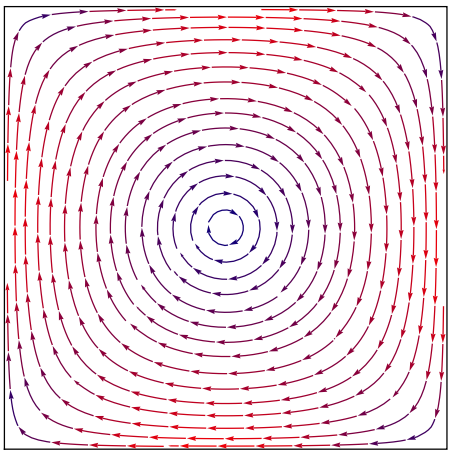
As indicated in Table 5.1, we would like the AMG interpolation procedure to construct the coarse trial space weights W_r with as few nonzeros as possible such that

$$\|\hat{A}_r\|_{X_f^{-1},|\hat{A}_c|} = \sup_{u \in \mathsf{V}_c \setminus \{0\}} \frac{\|(A_f W_r + A_r)u\|_{X_f^{-1}}}{\|u\|_{|\hat{A}_c|}} \leq \gamma, \tag{5.18}$$

where γ is a given parameter, and X_f is either $|A_f|$ or $|\hat{B}_f|^{-1}$. The problem for the coarse test space weights W_s reads similarly, but with A_f^* in place of A_f and A_s in place of A_r. The above norm is difficult even to estimate, much less optimize, due to the presence of $|\hat{A}_c|$. This remains true even in the symmetric case, without the absolute value.

```
function C = coarsen2( A, rho, theta )
%COARSEN2 Partitions unknowns into C and F variables
%   Returns the set of C variables as a logical vector
%   Ensures ||1 - Df \ Af||_|Df| <= rho

if nargin<2; rho = 0.7; end
if nargin<3; theta = 0.1; end

[n,~] = size(A);
C = false(n,1); F = true(n,1); id = (1:n)';
D = diag(sparse(1./sqrt(diag(A))));
S = abs(D*A*D); S = S - diag(diag(S));
H = max(S,S');
while true
    r = F.*(S *F) ; c = F.*(F'*S)'; g = sqrt(r.*c) ;

    v1 = F.*(r'*S)';  v2 = F.*((F.*(S*v1))'*S)';
    v = v2./v1; v(v1==0) = 0;

    w1 = F.*(S*c);  w2 = F.*(S*(F.*(w1'*S)'));
    w = w2./w1; w(w1==0) = 0;

    if min(max([v1 v w1 w])) < rho^2; break; end

    % construct neighbor sets as columns of nbr
    m = 1 ./ max(H * diag(sparse(F)), [], 2); m(C) = 0;
    nbr = diag(sparse(m)) * H > theta; nbr(:,C)=0;

    mask = max(v,w) > rho^2; % among i where
                             %   v_i or w_i > rho^2
    ng = max(diag(sparse(mask.*g)) * nbr)';
    mask = mask & (g >= ng); % select local maxima
    nid = max(diag(sparse(mask.*id)) * nbr)';
    mask = mask & (id > nid);% with highest id
    C = C | mask; F = xor(F, mask);
end

end
```

Figure 5.1: Matlab implementation of Algorithm 2



$$c_x = \quad 2y(1-x^2)$$
$$c_y = -2x(1-y^2)$$

Figure 5.2: Advection velocity for 2D example

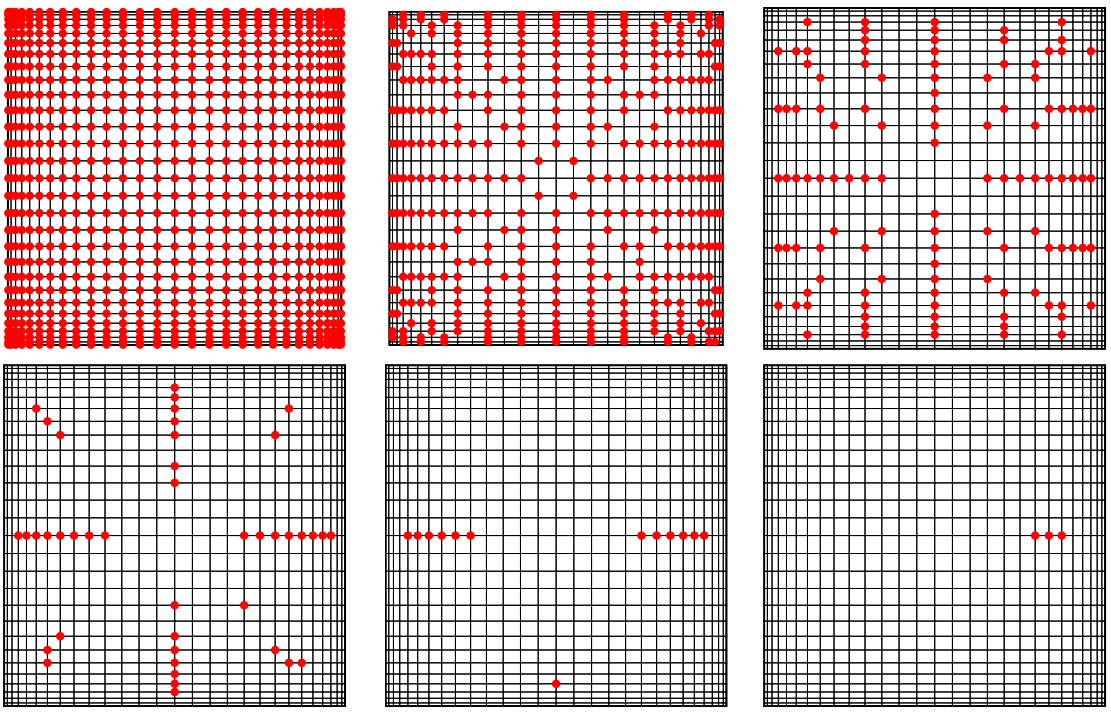


Figure 5.3: Example coarsening

We will follow an approach for approximately solving the above problem inspired by the energy-minimizing weights originally suggested by Wan, Chan, and Smith [52]; see also the article by Xu and Zikatanov [54]. In some sense, what follows is an attempt to generalize energy-minimizing interpolation to the nonsymmetric case.

As a first step towards finding a tractable substitute problem to solve, let us replace $|\hat{A}_c|$ with $\hat{D}_c$, the diagonal matrix with entries on the diagonal equal to the absolute value of those of $\hat{A}_c$. This is a rather drastic step, particularly for those modes u such that $\|u\|_{|\hat{A}_c|}/\|u\|_{\hat{D}_c}$ is small. Taking our cue from the energy-minimizing weights for symmetric AMG, we can counter this by adding the constraint that a given representative such vector be interpolated perfectly. Our new problem then reads: find W_r with the minimum number of nonzeros such that

$$\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c} \leq \gamma \qquad \text{and} \qquad W_r u = -A_f^{-1} A_r u =: v \tag{5.19}$$

where u is the given "near null space" vector. For our advection-diffusion problems, the appropriate choice for u is the discretized constant function, $u = \mathbf{1}$. Note that the coarsening procedure described in the last section ensures that $v := -A_f^{-1} A_r u$ can be computed quickly with a (potentially Krylov-accelerated) Jacobi iteration. In practice, we use a GMRES iteration. The soundness of the above step, that of replacing $|\hat{A}_c|$ with $\hat{D}_c$ and adding the constraint, amounts to an assumption about the target application. While appropriate for symmetric problems like Poisson's equation, the context where the idea originated, it is less clear that it is appropriate for advection-diffusion, as we shall see in our numerical tests.

We can divide the above problem into two subproblems: (1) given a fixed support (i.e., sparsity pattern or skeleton) of W_r, finding the numerical entries that minimize $\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c}$ subject to the constraint $W_r u = v$, and (2) determining the minimal support such that the norm is smaller than γ. The approximate solution procedure we will describe expands the support iteratively, (approximately) solving subproblem (1) at each step until (an approximation to) the norm is small enough. As such, let us look at subproblem (1) first.

5.3.1 Interpolation Weights (Diagonal)

Even though we have replaced $|\hat{A}_c|$ with a diagonal matrix, minimizing the norm $\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c}$ remains quite difficult, in part because the dependence of the norm on W_r is not smooth. Since

$$\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c} = \|X_f^{-1/2} \hat{A}_r \hat{D}_c^{-1/2}\|_2, \tag{5.20}$$

the latter norm being the matrix 2-norm, we propose to minimize the (square of the) corresponding Frobenius norm,

$$\|X_f^{-1/2}\hat{A}_r\hat{D}_c^{-1/2}\|_F^2 = \sum_{j=1}^{n} \alpha_j^{-1}\|\hat{A}_r\mathbf{e}_j\|_{X_f^{-1}}^2 \quad \text{where } \alpha_j = [\hat{D}_c]_{jj}. \tag{5.21}$$

Here $\mathbf{e}_j$ is the jth standard basis vector and n is the number of columns of W_r, i.e., the number of C-variables. Hence, our surrogate problem reads

$$\text{minimize} \quad \sum_{j=1}^{n} \alpha_j^{-1}\|\hat{A}_r\mathbf{e}_j\|_{X_f^{-1}}^2 \quad \text{subject to} \quad W_r u = v. \tag{5.22}$$

In the absence of the constraint $W_r u = v$, the minimizer is independent of $\hat{D}_c$ (the scalars α_j), as it independently minimizes the norm of each column. The constraint couples the column norm minimization problems, and the values α_j determine the relative importance of minimizing each column norm. In the symmetric case, if we take $X_f = |A_f| = A_f$ and if we take the scalars α_j to all be 1 instead of $[\hat{D}_c]_{jj}$, then the above problem would be equivalent to minimizing the trace of $P_r^* A P_r$ subject to the constraint, which precisely describes the energy-minimizing weights suggested by Wan, Chan, and Smith [52]; see also the article by Xu and Zikatanov [54]. Indeed, the energy-minimizing weights of Wan et al. inspired the development of the weights presented here.

Let us suppose for now that X_f^{-1} is available exactly. This is quite reasonable, for example, if $X_f^{-1} = |\hat{B}_f|$ and $\hat{B}_f$ corresponds to a single step of Jacobi, for in this case X_f^{-1} is a known diagonal matrix. In the next section, we will consider approximations we can make when $X_f = |A_f|$.

To express the fixed sparsity constraint, let R_j be the matrix that "restricts" to the support of the jth column of W_r. That is, R_j consists of those rows of the $m \times m$ identity matrix corresponding to the nonzero entries of $W_r\mathbf{e}_j$, where m is the number of rows of W_r, i.e., the number of F-variables. Define

$$W_r\mathbf{e}_j =: R_j^* w_j, \quad X_j := R_j A_f^* X_f^{-1} A_f R_j^*, \quad b_j := -R_j A_f^* X_f^{-1} A_r \mathbf{e}_j. \tag{5.23}$$

Here w_j is the to-be-determined vector of nonzeros in column j of W_r. Then,

$$\begin{aligned} \|\hat{A}_r\mathbf{e}_j\|_{X_f^{-1}}^2 &= \|A_f W_r\mathbf{e}_j + A_r\mathbf{e}_j\|_{X_f^{-1}}^2 \\ &= \|w_j\|_{X_j}^2 - 2\,\mathrm{Re}\,(b_j, w_j) + \|A_r\mathbf{e}_j\|_{X_f^{-1}}^2 \\ &= \|X_j w_j - b_j\|_{X_j^{-1}}^2 - \|b_j\|_{X_j^{-1}}^2 + \|A_r\mathbf{e}_j\|_{X_f^{-1}}^2, \end{aligned} \tag{5.24}$$

and the constraint $W_r u = v$ becomes $\sum_{j=1}^n R_j^* w_j u_j = v$, where u_j is the jth component of the near nullspace vector u.

By introducing the Lagrange multiplier λ, a (dual) vector of length m, the minimizer we seek can be expressed as a stationary point of the functional

$$\phi(W_r, \lambda) = \sum_{j=1}^{n} \alpha_j^{-1} \{ \tfrac{1}{2} \|w_j\|_{X_j}^2 - \operatorname{Re}(b_j, w_j) \} + \operatorname{Re} \left(\lambda, v - \sum_{j=1}^{n} R_j^* w_j u_j \right), \tag{5.25}$$

leading to the symmetric saddle point system

$$\begin{bmatrix} \alpha_1^{-1} X_1 & & & -\overline{u}_1 R_1 \\ & \ddots & & \vdots \\ & & \alpha_n^{-1} X_n & -\overline{u}_n R_n \\ -u_1 R_1^* & \dots & -u_n R_n^* & \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} \alpha_1^{-1} b_1 \\ \vdots \\ \alpha_n^{-1} b_n \\ -v \end{bmatrix}. \tag{5.26}$$

Once λ is known, the solutions for w_j are given by

$$X_j w_j^0 = b_j, \qquad w_j = w_j^0 + \alpha_j \overline{u}_j X_j^{-1} R_j \lambda. \tag{5.27}$$

Here w_j^0 is the solution for w_j in the problem without the constraint. Let W_0 be the corresponding interpolation matrix (so that $W_0 \mathbf{e}_j = R_j^* w_j^0$). Then we can expand the constraint as

$$v = \sum_{j=1}^{n} R_j^* w_j u_j = W_0 u + \sum_{j=1}^{n} \alpha_j |u_j|^2 R_j^* X_j^{-1} R_j \lambda,$$

and we see that λ must solve

$$S\lambda = v - W_0 u \quad \text{where} \quad S := \sum_{j=1}^{n} \alpha_j |u_j|^2 R_j^* X_j^{-1} R_j. \tag{5.28}$$

Notice that the solution is structured as the individual column minimizers W_0 plus a term depending on λ that enforces the constraint. If we substitute W_0 for W_r in (5.24), we see that the minimum possible norm of column j without the constraint is

$$\|(A_f W_0 + A_r)\mathbf{e}_j\|_{X_f^{-1}}^2 = \|A_r \mathbf{e}_j\|_{X_f^{-1}}^2 - \|b_j\|_{X_j^{-1}}^2. \tag{5.29}$$

As the support set for column j is made bigger, this norm must decrease (or stay the same), eventually becoming 0 once the support includes all F-variables. We can rewrite (5.24) in terms of this minimum column norm.

$$\begin{aligned} \|\hat{A}_r \mathbf{e}_j\|_{X_f^{-1}}^2 &= \|(A_f W_0 + A_r)\mathbf{e}_j\|_{X_f^{-1}}^2 + \|X_j w_j - b_j\|_{X_j^{-1}}^2 \\ &= \|(A_f W_0 + A_r)\mathbf{e}_j\|_{X_f^{-1}}^2 + \alpha_j^2 |u_j|^2 \|R_j \lambda\|_{X_j^{-1}}^2. \end{aligned} \tag{5.30}$$

We see that the norm squared of a column of $\hat{A}_r$ is a sum of two components: the minimum norm squared for the column without the constraint, and a contribution from the term added to W_0 to satisfy the constraint. Moreover, this added term is determined in such a way as to minimize the sum of its contributions to the column norms, scaled by α_j^{-1}, i.e.,

$$\sum_{j=1}^{n} \alpha_j^{-1} \|X_j w_j - b_j\|_{X_j^{-1}}^2 = \|\lambda\|_S^2. \tag{5.31}$$

But minimizing this sum, a consequence of switching from the 2-norm to the Frobenius norm, is not actually what we want—we actually want to minimize the 2-norm. We can therefore treat the values of α_j as parameters which we can tune to better minimize the 2-norm. For example, if we have reason to think that decreasing the norm of column j will help decrease the 2-norm, we can decrease α_j. Conversely, if we think that the norm of column j can be increased without affecting the 2-norm, we can increase α_j. The procedure in §5.3.3 will do just this.

Let us remark on the implementation aspects of solving the above minimization problem. Note that the matrices X_j are small, their size being the number of specified nonzeros in the corresponding column of W_r. So, it is practical to compute their Cholesky factorizations, which can be used both to solve the required subproblems for w_j^0 and w_j, once λ has been determined, and to form S. To determine λ, we propose using the preconditioned conjugate gradients (PCG) method, with Jacobi preconditioner. The reason we can expect S to be reasonably well approximated by its diagonal is that A_f is well approximated by its diagonal, which the coarsening procedure guarantees. To see this, note that without the scalar factors $\alpha_j|u_j|^2$, S would be a standard additive overlapping Schwarz preconditioner for the matrix $A_f^* X_f^{-1} A_f$, which in turn should be well approximated by a diagonal matrix (since A_f is). We can moreover expect a diagonal scaling to account for the scalar factors we just ignored. In practice, we typically see 10–20 iterations of PCG to compute λ to full machine precision.

The observation that the Schur complement governing the Lagrange multiplier is related to Schwarz preconditioners was made, in the context of the energy-minimizing weights for symmetric problems, by Brannick and Zikatanov [9]. However, they used a larger Lagrange multiplier that included all variables and not just the F-variables, resulting in the governing matrix being an overlapping Schwarz preconditioner for A instead of A_f. The combination of restricting the Lagrange multiplier to the F-variables and using a coarsening procedure guaranteeing A_f is well approximated

by its diagonal is very advantageous, allowing us to be assured that problem (5.28) for the Lagrange multiplier is readily solved by diagonally preconditioned conjugate gradients.

5.3.2 Interpolation Weights (Absolute Value)

The method for determining the numerical values of the weights in the last section is probably only practical for diagonal X_f, and certainly not for $X_f = |A_f|$. In this section we will examine approximations we can apply in the latter case. The heuristic argument to follow will be fairly technical and can be skimmed, as a more direct and simple heuristic justification for the approximation we come up with will be given at the end of the section.

Let $\tilde{R}_j$ be a restriction matrix like R_j, but including additional rows of the identity, in particular those rows for which $A_f R_j^*$ and $A_r \mathbf{e}_j$ contain nonzeros. The matrix $\tilde{R}_j$ restricts to a "local neighborhood" of the support of $W_r \mathbf{e}_j$. Then

$$W_r \mathbf{e}_j = R_j^* w_j, \qquad \tilde{R}_j^* \tilde{R}_j A_f R_j^* = A_f R_j^*, \qquad \tilde{R}_j^* \tilde{R}_j A_r \mathbf{e}_j = A_r \mathbf{e}_j \tag{5.32}$$

and

$$\begin{aligned} \hat{A}_r \mathbf{e}_j = A_f W_r \mathbf{e}_j + A_r \mathbf{e}_j &= \tilde{R}_j^* (\tilde{R}_j A_f R_j^* w_j + \tilde{R}_j A_r \mathbf{e}_j) \\ &= \tilde{R}_j^* (\tilde{R}_j A_f \tilde{R}_j^* \tilde{R}_j R_j^* w_j + \tilde{R}_j A_r \mathbf{e}_j). \end{aligned} \tag{5.33}$$

Let

$$\tilde{A}_j := \tilde{R}_j A_f \tilde{R}_j^*, \qquad \tilde{f}_j := -\tilde{R}_j A_r \mathbf{e}_j, \qquad \tilde{e}_j := \tilde{R}_j R_j^* w_j - \tilde{A}_j^{-1} \tilde{f}_j \tag{5.34}$$

so that

$$\begin{aligned} \hat{A}_r \mathbf{e}_j &= \tilde{R}_j^* \tilde{A}_j \tilde{e}_j \\ &= A_f \tilde{R}_j^* \tilde{e}_j - (1 - \tilde{R}_j^* \tilde{R}_j) A_f \tilde{R}_j^* \tilde{e}_j \\ &= A_f \tilde{R}_j^* \tilde{e}_j - (1 - \tilde{R}_j^* \tilde{R}_j) A_f \tilde{R}_j^* (-\tilde{A}_j^{-1} \tilde{f}_j) \\ &= A_f \tilde{R}_j^* \tilde{e}_j - (\tilde{R}_j^* - A_f \tilde{R}_j^* \tilde{A}_j^{-1}) \tilde{f}_j \end{aligned} \tag{5.35}$$

where we used that

$$(1 - \tilde{R}_j^* \tilde{R}_j) A_f \tilde{R}_j^* (\tilde{R}_j R_j^* w_j) = (1 - \tilde{R}_j^* \tilde{R}_j) A_f R_j^* w_j = 0 \tag{5.36}$$

in order to replace $\tilde{e}_j$ with $-\tilde{A}_j^{-1} \tilde{f}_j$. We can thus use the triangle inequality to bound

$$\begin{aligned} \|\hat{A}_r \mathbf{e}_j\|_{|A_f|^{-1}} &= \|\tilde{A}_j \tilde{e}_j\|_{\tilde{R}_j |A_f|^{-1} \tilde{R}_j^*} \\ &\leq \|A_f \tilde{R}_j^* \tilde{e}_j\|_{|A_f|^{-1}} + \|(\tilde{R}_j^* - A_f \tilde{R}_j^* \tilde{A}_j^{-1}) \tilde{f}_j\|_{|A_f|^{-1}} \\ &= \|\tilde{e}_j\|_{\tilde{R}_j |A_f| \tilde{R}_j^*} + \|(\tilde{R}_j^* - A_f \tilde{R}_j^* \tilde{A}_j^{-1}) \tilde{f}_j\|_{|A_f|^{-1}}. \end{aligned} \tag{5.37}$$

Since $\tilde{f}_j$ does not depend on the weights, we minimize this upper bound of $\|\hat{A}_r \mathbf{e}_j\|_{|A_f|^{-1}}$ by minimizing $\|\tilde{e}_j\|_{\tilde{R}_j|A_f|\tilde{R}_j^*}$. Doing so still requires $|A_f|$ to be computed. However, by Proposition 3.1,

$$\|\tilde{R}_j^* \tilde{e}_j\|_{H_f} \leq \frac{\|\tilde{e}_j\|_{\tilde{R}_j|A_f|\tilde{R}_j^*}}{\|\tilde{e}_j\|_{|\tilde{R}_j A_f \tilde{R}_j^*|}} \leq \|\tilde{R}_j^* \tilde{e}_j\|_{H_f}^{1/2} \|A_f \tilde{R}_j^* \tilde{e}_j\|_{H_f^{-1}}^{1/2} \tag{5.38}$$

where $H_f = \frac{1}{2}(A_f + A_f^*)$. This suggests that $\|\tilde{e}_j\|_{|\tilde{R}_j A_f \tilde{R}_j^*|} = \|\tilde{e}_j\|_{|\tilde{A}_j|}$ should be a good surrogate for $\|\tilde{e}_j\|_{\tilde{R}_j|A_f|\tilde{R}_j^*}$. Now, $\|\tilde{e}_j\|_{|\tilde{A}_j|}^2$ expands to

$$\|\tilde{e}_j\|_{|\tilde{A}_j|}^2 = \|\tilde{R}_j R_j^* w_j\|_{|\tilde{A}_j|}^2 - 2\,\mathrm{Re}(\tilde{f}_j, |\tilde{A}_j|^{-1} \tilde{A}_j \tilde{R}_j R_j^* w_j) + \|\tilde{f}_j\|_{|\tilde{A}_j|^{-1}}^2, \tag{5.39}$$

so, if we take

$$\begin{aligned} X_j &:= R_j \tilde{R}_j^* |\tilde{A}_j| \tilde{R}_j R_j^* \qquad \text{and} \\ b_j &:= R_j \tilde{R}_j^* \tilde{A}_j^* |\tilde{A}_j|^{-1} \tilde{f}_j = -R_j \tilde{R}_j^* \tilde{A}_j^* |\tilde{A}_j|^{-1} \tilde{R}_j A_r \mathbf{e}_j, \end{aligned} \tag{5.40}$$

then

$$\|\tilde{e}_j\|_{|\tilde{A}_j|}^2 = \|X_j w_j - b_j\|_{X_j^{-1}}^2 - \|b_j\|^2 + \|\tilde{f}_j\|_{|\tilde{A}_j|^{-1}}^2. \tag{5.41}$$

Therefore, minimizing $\sum_{j=1}^n \alpha_j^{-1} \|\tilde{e}_j\|_{|\tilde{A}_j|}^2$ is equivalent to minimizing

$$\sum_{j=1}^n \alpha_j^{-1} \|X_j w_j - b_j\|_{X_j^{-1}}^2. \tag{5.42}$$

We can solve this minimization problem, including the constraint $W_r u = v$, exactly as in the previous section.

Because

$$\begin{aligned} X_j w_j - b_j &= (R_j \tilde{R}_j^* \tilde{A}_j^* |\tilde{A}_j|^{-1} \tilde{R}_j)(A_f R_j^* w_j + A_r \mathbf{e}_j) \\ &= (R_j \tilde{R}_j^* \tilde{U}_j^* \tilde{R}_j) \hat{A}_r \mathbf{e}_j \end{aligned} \tag{5.43}$$

where $\tilde{U}_j$ is the unitary factor of the polar decomposition $\tilde{A}_j = |\tilde{A}_j| \tilde{U}_j$, we can alternatively directly justify minimizing $\|X_j w_j - b_j\|_{X_j^{-1}}$ as a surrogate for minimizing $\|\hat{A}_r \mathbf{e}_j\|_{|A_f|^{-1}}$. Whereas our earlier reasoning relied on a particular choice of $\tilde{R}_j$, this equation for $X_j w_j - b_j$ holds for any choice of restriction matrix $\tilde{R}_j$ provided only that it at least includes the rows of R_j. This suggests that this method may be reasonable for other choices of $\tilde{R}_j$. Taking $\tilde{R}_j = R_j$, for example, would significantly reduce the computational cost.

Figure 5.4 summarizes the procedure for determining the numerical values of the interpolation weights, showing the key equations that define the weights, as well as

Method	X_j	b_j
Diagonal	$R_j A_f^* D_f^{-1} A_f R_j^*$	$-R_j A_f^* D_f^{-1} A_r \mathbf{e}_j$
Abs, exact	$R_j \lvert A_f \rvert R_j^*$	$-R_j A_f^* \lvert A_f \rvert^{-1} A_r \mathbf{e}_j$
Abs, extended	$R_j \tilde{R}_j^* \lvert \tilde{A}_j \rvert \tilde{R}_j R_j^*$	$-R_j \tilde{R}_j^* \tilde{A}_j^* \lvert \tilde{A}_j \rvert^{-1} \tilde{R}_j A_r \mathbf{e}_j$
Abs, restricted	$\lvert A_j \rvert$	$-A_j^* \lvert A_j \rvert^{-1} R_j A_r \mathbf{e}_j$

$$
\begin{aligned}
&W_r \mathbf{e}_j = R_j^* w_j, && W_0 \mathbf{e}_j = R_j^* w_j^0, \\
&\quad w_j = w_j^0 + \alpha_j \overline{u}_j X_j^{-1} R_j \lambda, && X_j w_j^0 = b_j, \\
&S\lambda = v - W_0 u, && S := \sum_{j=1}^{n} \alpha_j |u_j|^2 R_j^* X_j^{-1} R_j
\end{aligned}
$$

Figure 5.4: Interpolation weights summary

showing four different choices for X_j and b_j from this and the previous section. The first two, "diagonal" and "abs, exact," are from the previous section with $X_f = D_f$ and $X_f = |A_f|$. "Abs, extended" is the main choice of this section, while "abs, restricted" corresponds to taking $\tilde{R}_j = R_j$ as suggested in the previous paragraph. A nice property of the three absolute value methods is that, if we set the scalars α_j to 1, then in the symmetric case each of them reduces to the energy-minimizing weights of Wan et al. Note that "abs, exact" is impractical but is included for reference purposes; for small problems we will be able to compare it to the heuristics based on it from this section. On the other hand, the other two absolute value based methods are practical, as they only require computing absolute values of small matrices: of the size of the support for the column in the restricted case, and of the size of the "local neighborhood" of the support in the extended case. These can be computed efficiently using the scaled iteration of §3.6.3. There is a danger that the "local neighborhoods" can become too large on coarser levels where the matrices are denser and contain many small nonzeros. To deal with this, when determining $\tilde{R}_j$, i.e., the local neighborhood, in practice we ignore the "small" entries of A_f and A_r, where "small" means the corresponding entry of $D_f^{-1} A_f$ or $D_f^{-1} A_r$ is less than some factor (we use 0.1) of the largest in the column, excluding the (unit) diagonal of $D_f^{-1} A_f$.

5.3.3 Interpolation Support

In order to adaptively determine the minimal support required to make $\|\hat{A}_r\|_{X_f^{-1}, \hat{D}_c}$ smaller than the given parameter γ, we can make use of Nikiforov's bound on the

matrix 2-norm, Theorem 5.1, as

$$\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c} = \|X_f^{-1/2}\hat{A}_r\hat{D}_c^{-1/2}\|_2. \tag{5.44}$$

Therefore, let us define R_r as the entry-wise absolute value of the matrix inside the matrix 2-norm, and R_s similarly,

$$[R_r]_{ij} := |[X_f^{-1/2}\hat{A}_r\hat{D}_c^{-1/2}]_{ij}| = |[X_f^{-1/2}(A_fW_r + A_r)\hat{D}_c^{-1/2}]_{ij}|, \tag{5.45}$$

$$[R_s]_{ij} := |[X_f^{-1/2}\hat{A}_s\hat{D}_c^{-1/2}]_{ij}| = |[X_f^{-1/2}(A_f^*W_s + A_s)\hat{D}_c^{-1/2}]_{ij}|, \tag{5.46}$$

so that, e.g., $\|\hat{A}_r\|_{X_f^{-1},\hat{D}_c} \le \|R_r\|_2$. These are scaled "residuals" for the weights W_r and W_s. In the case $X_f = D_f$, as in the "diagonal" method in Figure 5.4, R_r and R_s are readily computed. For the "abs, extended" and "abs, restricted" cases, for which $X_f = |A_f|$, we propose replacing $X_f^{-1/2}$ in the definitions of R_r and R_s with the symmetric approximation

$$\begin{aligned} T &= \tfrac{1}{2}(C^{-1}Y + YC^{-1}) \approx |A_f|^{-1/2}, \\ Y &= \sum_{j=1}^{n} \tilde{R}_j^*|\tilde{A}_j|^{-1/2}\tilde{R}_j, \quad [C]_{ii} = \sum_{j=1}^{n}\|\tilde{R}_j\mathbf{e}_i\|_2^2, \end{aligned} \tag{5.47}$$

where $\tilde{R}_j = R_j$ in the "abs, restricted" case. Here C is a diagonal matrix whose ith diagonal entry is simply the number of "local neighborhoods" that include the ith F-variable, i.e., the number of the $\tilde{R}_j$ matrices that include the ith row of the identity matrix. This approximation is quite similar to an additive overlapping Schwarz preconditioner, with C serving to average results (instead of simply summing them) in regions of overlap. The cost of forming the above approximation is comparable to the cost of forming the matrix S of Figure 5.4 governing the Lagrange multiplier to enforce the constraint. An alternative option is just to approximate $|A_f|^{-1/2}$ by $D_f^{-1/2}$.

Algorithm 3 below starts with skeletons S_r and S_s for the weights W_r and W_s with just enough entries such that the constraint is enforcable, and adaptively expands the skeletons until the Nikiforov bounds (with $p = 1$ and $r = 0, 1$) on the norms $\|R_r\|_2$ and $\|R_s\|_2$ are at most γ. The support is expanded in the calls to the function ExpandSupport, which is described in the next paragraph. The routine MinSkel constructs a skeleton with an entry in each row at the position of the maximum entry in the corresponding row of its input matrix. This is enough to ensure that the constraints $W_ru = v_r$ and $W_su = v_s$ can be enforced. The routine SolveWeights implements the procedure for determining the numerical values of the weights given

the skeletons, as described in the previous sections and summarized in Figure 5.4. In addition to the weights, it also returns the unconstrained best weights W_0 as well as the approximation T to $|A_f|^{-1/2}$ described above (or $D_f^{-1/2}$ for the "diagonal" method). The final parameter is intended to be a relative tolerance controlling the number of iterations of PCG done in solving for the Lagrange multiplier λ. Algorithm 3 waits until the bound is satisfied and the skeletons finalized before solving the weights to full precision. A final comment to make is that on line 20 the scalars $\alpha_{r,j}$ and $\alpha_{s,j}$ are set to something other than $[\hat{D}_c]_{jj}$, following the suggestion in §5.3.1. The particular choice made was arrived at by trial and error; the basic idea is that SolveWeights should try harder to keep those column norms small where $w_{r,j}^{(2)}$ and $w_{s,j}^{(2)}$ are large, as these are the columns causing the Nikiforov bound to be large.

Algorithm 3 Coarse trial and test space weights construction

1: **function** $(W_r, W_s) \leftarrow$ Weights$(A,\, C,\, \gamma,\, u)$
2: $\quad v_r \leftarrow -A_f^{-1}A_r u, \quad v_s \leftarrow -A_f^{-*}A_s u$
3: $\quad S_r \leftarrow \textsc{MinSkel}(D_f^{-1/2}A_r D_c^{-1/2}), \qquad S_s \leftarrow \textsc{MinSkel}(D_f^{-1/2}A_s D_c^{-1/2})$
4: $\quad \alpha_{r,j} \leftarrow [D_c]_{jj}, \quad \alpha_{s,j} \leftarrow [D_c]_{jj}, \quad 1 \le j \le n$
5: $\quad$ **loop**
6: $\quad\quad (W_r, W_{r,0}, T_r) \leftarrow \textsc{SolveWeights}(A_f, A_r, S_r, \alpha_r, u, v_r, 10^{-4})$
7: $\quad\quad (W_s, W_{s,0}, T_s) \leftarrow \textsc{SolveWeights}(A_f^*, A_s, S_s, \alpha_s, u, v_s, 10^{-4})$
8: $\quad\quad \hat{A}_r \leftarrow A_f W_r + A_r, \qquad \hat{A}_{r,0} \leftarrow A_f W_{r,0} + A_r$
$\quad\quad\quad \hat{A}_s \leftarrow A_f^* W_s + A_s, \qquad \hat{A}_{s,0} \leftarrow A_f^* W_{s,0} + A_s$
9: $\quad\quad \hat{A}_c \leftarrow W_s^* \hat{A}_r + A_s^* W_r + A_c$
10: $\quad\quad [R_r]_{ij} \leftarrow |[T_r \hat{A}_r \hat{D}_c^{-1/2}]_{ij}|, \qquad [R_{r,0}]_{ij} \leftarrow |[T_r \hat{A}_{r,0} \hat{D}_c^{-1/2}]_{ij}|$
$\quad\quad\quad [R_s]_{ij} \leftarrow |[T_s \hat{A}_r \hat{D}_c^{-1/2}]_{ij}|, \qquad [R_{s,0}]_{ij} \leftarrow |[T_s \hat{A}_{s,0} \hat{D}_c^{-1/2}]_{ij}|$
11: $\quad\quad \mathbf{w}_r^{(1)} \leftarrow R_r^* R_r \mathbf{1}, \quad \mathbf{w}_r^{(2)} \leftarrow R_r^* R_r \mathbf{w}_r^{(1)}$
12: $\quad\quad \mathbf{w}_s^{(1)} \leftarrow R_s^* R_s \mathbf{1}, \quad \mathbf{w}_s^{(2)} \leftarrow R_s^* R_s \mathbf{w}_s^{(1)}$
13: $\quad\quad \gamma_{r,0}^2 \leftarrow \max_j w_{r,j}^{(1)}, \quad \gamma_{r,1}^2 \leftarrow \max_{j, w_{r,j}^{(1)} \neq 0} w_{r,j}^{(2)} / w_{r,j}^{(1)}$
14: $\quad\quad \gamma_{s,0}^2 \leftarrow \max_j w_{s,j}^{(1)}, \quad \gamma_{s,1}^2 \leftarrow \max_{j, w_{s,j}^{(1)} \neq 0} w_{s,j}^{(2)} / w_{s,j}^{(1)}$
15: $\quad\quad$ **if** $\min(\gamma_{r,0}^2, \gamma_{r,1}^2) \le \gamma^2$ and $\min(\gamma_{s,0}^2, \gamma_{s,1}^2) \le \gamma^2$ **then**
16: $\quad\quad\quad (W_r, \_, \_) \leftarrow \textsc{SolveWeights}(A_f, A_r, S_r, \alpha_r, u, v_r, 10^{-16})$
17: $\quad\quad\quad (W_s, \_, \_) \leftarrow \textsc{SolveWeights}(A_f^*, A_s, S_s, \alpha_s, u, v_s, 10^{-16})$
18: $\quad\quad\quad$ **return**
19: $\quad\quad$ **end if**
20: $\quad\quad \alpha_{r,j} \leftarrow [\hat{D}_c]_{jj} / \max(w_{r,j}^{(2)}, 10^{-6}), \qquad \alpha_{s,j} \leftarrow [\hat{D}_c]_{jj} / \max(w_{s,j}^{(2)}, 10^{-6})$
21: $\quad\quad S_r \leftarrow \textsc{ExpandSupport}(S_r, R_r, R_{r,0}, \gamma)$
22: $\quad\quad S_s \leftarrow \textsc{ExpandSupport}(S_s, R_s, R_{s,0}, \gamma)$
23: $\quad$ **end loop**
24: **end function**

Algorithm 4 Adaptive support expansion

1: **function** $S' \leftarrow$ ExpandSupport(S, R, R_0, γ)
2: $S_0 \leftarrow$ FindSupport(R, γ)
3: $S' \leftarrow S \cup S_0$
4: set R_0 to 0 on S
5: **for** i such that row i of $S \cap S_0$ is nonempty **do**
6: set $j_1, j_2, \ldots$ so that $[R_0]_{i,j_1} \geq [R_0]_{i,j_2} \geq \cdots > 0$
7: $p \leftarrow \min\{p \mid \sum_{k=1}^{p} [R_0]_{i,j_k} \geq \frac{1}{2} \sum_{j=1}^{n} [R_0]_{ij}\}$
8: $S' \leftarrow S' \cup (i, j_1) \cup \cdots \cup (i, j_p)$
9: **end for**
10: **end function**

The heuristic we propose for ExpandSupport is Algorithm 4. The routine takes as input the current skeleton, the weights "residual" R_r or R_s, as well as the corresponding unconstrained weights "residual" $R_{r,0}$ or $R_{s,0}$, based on, e.g., $W_{r,0}$ instead of W_r, and finally the parameter γ. The routine begins by invoking the auxiliary routine FindSupport, Algorithm 5. FindSupport returns an approximately minimal skeleton S_0 such that setting the corresponding entries of R to 0 would give $\|R\|_2 \leq \gamma$. The remainder of the routine heuristically adds entries to the skeleton of the weights in an attempt to make these entries of R small after the next call to SolveWeights.

Recall that, e.g., R_r is the entry-wise absolute value of $T_r \hat{A}_r \hat{D}_c^{-1/2}$, which is approximately (or exactly) $D_f^{-1} \hat{A}_r \hat{D}_c^{-1/2}$. The same holds for $R_{r,0}$ with $\hat{A}_{r,0} = A_f W_{r,0} + A_r$ in place of $\hat{A}_r$, i.e., with the unconstrained weights in place of the constrained weights. Recall also that $\hat{A}_r$ is equal to $\hat{A}_{r,0}$ plus a term linear in λ. Hence we can view as R_r as decomposing into $R_{r,0}$ and a second component due to the constraint. Now, we expect $\hat{A}_{r,0}$ and hence $R_{r,0}$ to be small at the entries corresponding to the skeleton of W_r. Indeed, for the "abs, restricted" method, these entries of $\hat{A}_{r,0}$ are exactly 0. Hence, in order to reduce $R_{r,0}$ and thereby R_r at positions not already in the skeleton of the weights W_r, it makes sense to simply add these positions to the skeleton. This is accomplished by line 3 of Algorithm 4.

However, there may remain entries of S_0 corresponding to entries of R we would like to reduce which are at positions already present in the skeleton of the weights. We can conclude that these entries of R must be large as a result of enforcing the constraint. Note that the constraint decomposes into separate constraints on each row. For example, the constraint on row i of W_r is

$$\sum_{j=1}^{n} [W_r]_{ij} u_j = [v_r]_i. \tag{5.48}$$

The constraint on a given row becomes easier to satisfy as we add more entries to the skeleton in that row. In particular, if we add all the entries to row i, then the exact values $[-A_f^{-1}A_r]_{ij}$ (which do not contribute to the column norms) could be used to satisfy the constraint, as we took $v_r = -A_f^{-1}A_r u$. The strategy implemented in Algorithm 4 is to add, in any row in which we wish to reduce R at a position already in the skeleton, the smallest set of new entries to the row such that, if R_0 were to become zero at these positions, the row sum would reduce by at least half. This heuristic was arrived at by some trial and error. Experience indicates that it allows Algorithm 3 to finish in a small number (3–6) of loop iterations, which is important as each call to SolveWeights is rather expensive. In contrast, the alternative strategy of only adding one new entry to the row (at the position of the largest value of R_0), while somewhat reducing the final number of nonzeros in the weights, can cause Algorithm 3 to take a very large number of iterations, especially at coarser levels with denser matrices.

Algorithm 5 Find set S of nonzeros of R whose elimination gives $\|R\|_2 \leq \gamma$

```
1: function S ← FindSupport(R, γ)
2:     S ← {},      θ ← 1/2
3:     loop
4:         r ← R1,   w^(1) ← R*r,   w^(2) ← R*Rw^(1)
5:         if max_{j,w_j^(1)≠0} w_j^(2)/w_j^(1) ≤ γ^2 or max_j w_j^(1) ≤ γ^2 then return
6:         while max_j w_j^(1) ≤ (1+θ)γ^2 do
7:             θ ← θ/2
8:         end while
9:         for j ∈ {j | w_j^(1) > (1+θ)γ^2} do
10:            i ← arg max_i r_i R_ij
11:            S ← S ∪ (i, j),   R_ij ← 0
12:        end for
13:    end loop
14: end function
```

Finally, we turn to the implementation of FindSupport, Algorithm 5, which is responsible for finding the small set of positions of nonzeros of R such that setting R to 0 at these positions would give $\|R\|_2 \leq \gamma$. First, we note that the stopping criteria in line 5 are based on the same Nikiforov bounds used by Algorithm 3. This is important, as the two routines need to agree on when more entries are needed in the skeleton. In particular, if FindSupport used a different bound that happened to already be satisfied, it would not suggest any new entries to add, and as a result

Algorithm 3 would loop forever. Algorithm 5 works by reducing the entries of $\mathbf{w}^{(1)}$, the largest of which gives the square of the $r = 0, p = 1$ Nikiforov bound, until they are all below $(1 + \theta)\gamma^2$. The idea is that this may well be sufficient to make the $r = 1, p = 1$ Nikiforov bound, which is usually smaller, less than or equal to γ. If not, the algorithm reduces θ and tries again.

In each iteration of its main loop, the algorithm reduces each $w_j^{(1)} > (1 + \theta)\gamma^2$ by eliminating the single entry of R effecting the greatest reduction of $w_j^{(1)}$. This can be determined explicitly as follows. Define R_1 as R with the entry at position (i, k) set to zero,

$$R_1 := R - \mathbf{e}_i R_{ik} \mathbf{e}_k^*. \tag{5.49}$$

Then

$$R_1 \mathbf{1} = R\mathbf{1} - \mathbf{e}_i R_{ik}, \tag{5.50}$$

$$R_1^* R_1 \mathbf{1} = R^* R\mathbf{1} - (\mathbf{e}_i^* R)^* R_{ik} - \mathbf{e}_k R_{ik} \mathbf{e}_i^* R\mathbf{1} + \mathbf{e}_k R_{ik}^2, \tag{5.51}$$

and, letting $r_i = \mathbf{e}_i^* R\mathbf{1}$ be the ith row sum of R, the reduction of $w_j^{(1)}$ is

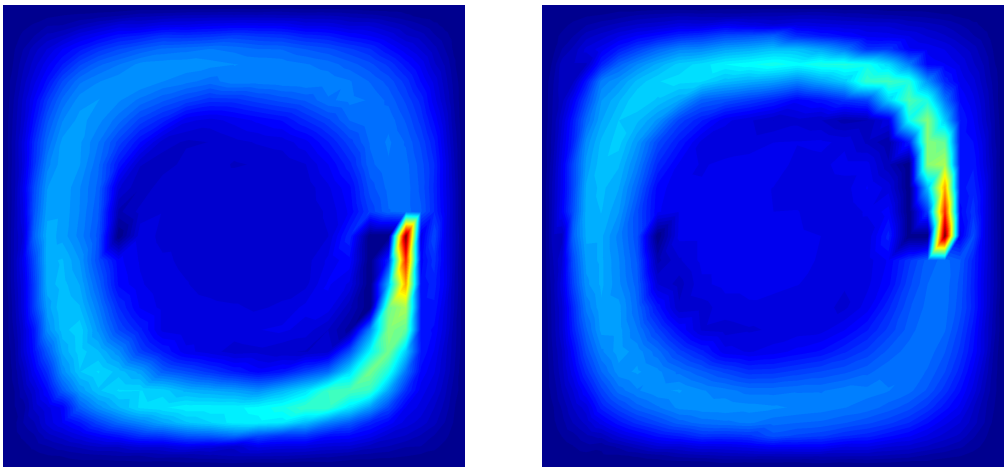
$$\begin{aligned} \mathbf{e}_j^*(R^*R - R_1^*R_1)\mathbf{1} &= R_{ij}R_{ik} + \delta_{jk}(r_i R_{ik} - R_{ik}^2) \\ &= \begin{cases} r_i R_{ik} & j = k \\ R_{ij} R_{ik} & j \neq k \end{cases} \end{aligned} \tag{5.52}$$

As r_i is a sum of nonnegative values that include R_{ij}, the above expression is always maximized when $j = k$. That is, the entry of R whose elimination reduces $w_j^{(1)}$ the most is always in column j. Further, it is in the row i such that $r_i R_{ij}$ is maximized. This is precisely how Algorithm 5 chooses which entry of R to eliminate, in line 10.

Figure 5.5 shows the coarse trial and test space vector for one of the final three degrees of freedom for the hierarchy illustrated in Figure 5.3. The plot on the left, for example, is $P_{r,1} \cdots P_{r,5} \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$. The interpolation and restriction operators were produced by Algorithm 3 using the (rather tight) tolerance $\gamma = 0.3$. We can see that interpolation (P_r) propagates information downstream while, symmetrically, restriction (P_s^*) gathers information from upstream. Also, note that for the adjoint problem, in which the direction of advection is reversed, P_r and P_s simply swap.

5.4 Smoother

According to Table 5.1, a good smoother is one that is inexpensive and for which $\|1 - \hat{B}_f A_f\|_{X_f}$ is small (where, again, $X_f = |A_f|$ or $|\hat{B}_f|^{-1}$). If the smoother is an



P_r P_s

Figure 5.5: Example coarse trial and test space basis vector

F-relaxation, that is, if

$$[B]_c = \begin{bmatrix} B_f & \\ & 0 \end{bmatrix}, \tag{5.53}$$

then in this case $[B]_h = [B]_c$ so that $\hat{B}_f = B_f$. Thus, an obvious approach is to use an F-relaxation with B_f chosen as a suitable preconditioner for A_f. In particular, the coarsening procedure of §5.2 ensures that plain (undamped) Jacobi is a suitable choice of F-relaxation, since the procedure guarantees that $\|1 - D_f^{-1} A_f\|_{|D_f|} \leq \rho$ for input parameter ρ (our default choice being 0.7). For k steps of Jacobi iteration on the F-variables, we expect $\|1 - \hat{B}_f A_f\|$ to behave in the limit of increasing k like ρ^k or better, for any norm.

One might potentially improve on plain Jacobi by introducing a damping parameter. Instead, we have elected to go with the parameter-free diagonal sparse approximate inverse (SPAI-0) of Bröker and Grote [10], which is optimal in a certain Frobenius norm, and is easily computed, the ith diagonal entry being given by

$$[B_f^{(1)}]_{ii} = [A_f^*]_{ii} / \|A_f \mathbf{e}_i\|_2^2. \tag{5.54}$$

This alternative to damped Jacobi is especially attractive in the original setting as a full smoother, in which case the damping is essential, unlike in the F-relaxation setting here. The use of sparse approximate inverse preconditioners as multigrid smoothers was suggested originally by Tang and Wan [49], motivated largely by their inherent parallelism, especially as compared to popular alternatives such as Gauss-Seidel.

Finally, when doing multiple smoothing steps, we can consider using Krylov-subspace methods to accelerate this basic linear iteration. Options include using

GMRES as well as the Chebyshev iteration, whose application to nonsymmetric problems was investigated in depth by Manteuffel [36]. Two features make the Chebyshev iteration attractive when compared with GMRES: GMRES requires global inner products, whereas Chebyshev does not, and the use of GMRES renders the overall multigrid cycle nonlinear, whereas with Chebyshev the cycle remains linear. The improvement over simple linear iteration by Chebyshev acceleration is most dramatic for symmetric problems, and, unfortunately, degrades rapidly in the presence of even modest nonsymmetry. The added complexity for marginal gains therefore made the Chebyshev iteration not seem worthwhile to implement (whereas it would clearly be worthwhile in a symmetric setting). Instead, we implemented both plain linear and GMRES-accelerated iterations for smoothers. The fact that we saw, as reported in the next section, only marginal gains with the GMRES-accelerated F-relaxations confirms that the gains with Chebyshev would also have been small.

Traditionally, algebraic multigrid methods have favored the use of full smoothers over F-relaxations. For example, the general two-level analysis of Falgout et al. [19] distinguishes between two-grid methods and two-level hierarchical basis methods, the latter of which can be seen to be an instance of the former in which the smoothers are F-relaxations. Granting this distinction, multigrid methods are clearly "better" in that they are more general, allowing full smoothers and not just F-relaxations.

The choice then to use F-relaxations anyway, given the freedom to use anything, deserves some comment. First, note that Theorem 4.1 and its corollary imply that an arbitrary smoother can be replaced by its projection to an F-relaxation while preserving the two-level iteration behavior (e.g., the spectrum of the iteration operator). Effecting this replacement could even be implemented reasonably efficiently, though for no obvious gains. Hence, in principle, the only possible advantages of a general smoother over an F-relaxation must stem from the difference between the idealized two-level iteration employing an exact coarse solve and a multilevel iteration. For example, a particularly heavy-duty smoother could potentially pick up the slack if the coarse solve is not as effective as it should be on some of the coarse error modes. However, this is not an example of a smoother being better at *smoothing*.

Furthermore, while Theorem 4.1 only indicates that the F-relaxation chosen for a smoother (or more generally the projected smoother) needs to be an approximate inverse of the Schur complement S_f of the coarse operator $\hat{A}_c$, in our particular framework—as laid out in Table 5.1 and based on the convergence bounds of Theorems 4.2 and 4.3—we have chosen to identify interpolation and restriction as being "good" when A_f is a good approximation of S_f, as precisely captured by Lemma 4.2.

This is an enormous advantage, as we have reduced the problem of finding a good smoother to that of finding a good solver for A_f, to which we can apply any of a wide variety of general techniques, e.g., Krylov subspace methods, as discussed above. In contrast, it is far from obvious how to use Krylov techniques to accelerate general smoothers, as they are not designed to ensure optimal *smoothing*. Adams et al. [1] investigated the use of Chebyshev polynomials for (symmetric) multigrid smoothers, and used an ad hoc smallest eigenvalue parameter to ensure the polynomial does not target the slowly decaying smooth modes. By contrast, as discussed above, we can use standard Krylov methods to make $\|1 - \hat{B}_f A_f\|$ small.

5.5 Numerical Results

To test the AMG method(s) comprised of the heuristics described in the previous sections, we chose as a test problem the advection-diffusion equation

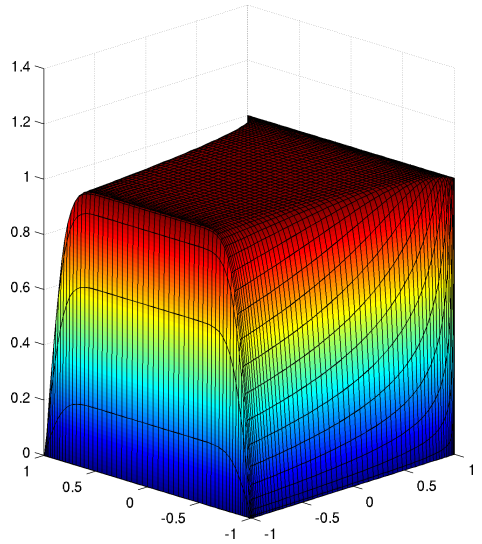
$$-\nabla \cdot (p\nabla u) + \mathbf{c} \cdot \nabla u = f \tag{5.55}$$

on the domains $[-1, 1]^2$ and $[-1, 1]^3$, discretized with bi- and tri-linear finite elements using the streamline diffusion method, as described in §1.2.3. Four different advection fields $\mathbf{c}$ were considered, listed in Table 5.2. The boundary conditions for all of the 2D problems are Dirichlet. The first problem, "Poisson," is the case of no advection, reducing to Poisson's equation, which we included to test whether the nonsymmetric AMG methods would remain efficient in the degenerate symmetric case. The second problem, "2D-const," is the case of constant advection in the horizontal direction. Figure 5.6 illustrates the solution for the case $p = 0.005$, with inflow boundary condition $u = 1$ and remaining boundaries set to $u = 0$; it features both characteristic boundary layers, along the top and bottom, as well as an exponential boundary layer at the outflow boundary. Note that, while this particular boundary data is not regular enough for the continuous solution to be in $H^1(\Omega)$, we will only be concerned with error in the approximation to the discrete solution, and not with the error of the discrete solution itself, making the issue moot. At any rate, the boundary data is part of the right hand side f and does not affect the matrix A or the constructed AMG hierarchy. The third problem, illustrated in Figure 5.7 and simply called "2D," is the double glazing problem that was first introduced in §1.2.3, and which has been used in this and previous chapters for illustration purposes. The final problem, "3D" was created simply by extending the double glazing problem to 3D, using periodic boundary conditions in the new z-direction, and adding a small z-component to the

advection, which has the effect of changing the streamlines from the closed loops of Figure 5.7 to slowly winding helices.

Table 5.2: Test problems

Problem	$\mathbf{c}$
Poisson	$(0, 0)$
2D-const	$(-1, 0)$
2D	$(2y(1-x^2), -2x(1-y^2))$
3D	$(2y(1-x^2), -2x(1-y^2), 0.05)$

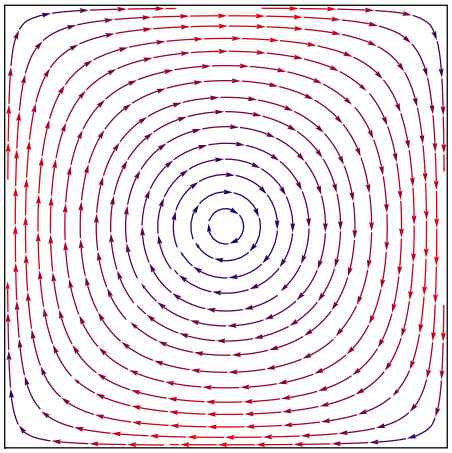


$p = 0.005$

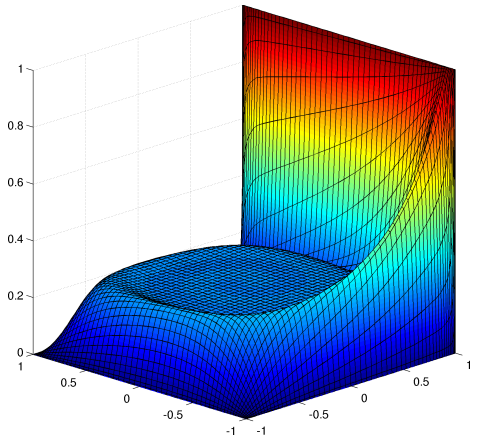
Figure 5.6: Test problem "2D-const"

All of the runs reported in this section used Algorithm 2 for coarsening with parameter $\rho = 0.7$ (thus guaranteeing $\|1 - D_f^{-1} A_f\|_{|D_f|} \leq 0.7$). Except where explicitly mentioned (specifically for the runs reported in Table 5.8), the "Abs, restricted" variant of the interpolation scheme of §5.3 was used, with varying parameter γ. Finally, the following very simple strategy was employed to combat the problem of stencil growth: any entry A_{ij} for which $|A_{ij}| \leq 10^{-4} |A_{ii}|^{1/2} |A_{jj}|^{1/2}$ was set to zero, and its original value added to the diagonal entry A_{ii} in order to preserve the row sums, or equivalently, the action of A on the constant vector $\mathbf{1}$.

We will first examine a few specific cases in some detail. Tables 5.3 and 5.4 present some information for the AMG hierarchies constructed with interpolation tolerances $\gamma = 0.5$ and 0.9 for the "2D" test problem on a 63×63 Chebyshev grid with diffusion parameter $p = 1/200$. The column n_c/n is the coarsening ratio, θ is the sectorial



$$c_x = \quad 2y(1-x^2)$$
$$c_y = -2x(1-y^2)$$



$p = 0.005$

Figure 5.7: Test problem "2D" (the double glazing problem)

half-angle of A, θ_f of A_f, and nnz/n is the average number of nonzeros per row. Note that the nonsymmetry of A_f, as measured by the sectorial half-angle, is modest relative to that of the whole matrix. This is expected as coarsening ensures that A_f is well approximated by its diagonal, which is symmetric as A is real. The column $\|E_{tg}^{(\infty)}\|$ is $\|1 - A_f^{-1}S_f\|_{|A_f|}$, which is a norm of the two-grid iteration operator when $\hat{B}_f = A_f^{-1}$, which holds in the limit as the number of F-relaxation steps approaches ∞ (as the column label is meant to convey). The remaining columns are the norm $\|E_{mg}^{(m,m)}\|_{|A|}$ and spectral radius of the full multigrid V-cycle iteration matrix, with m pre- and post-smoothing steps. Recall that a single smoothing step here is a diagonal F-relaxation, as described in the previous section.

Comparing the two tables reveals the effect of the interpolation parameter γ. Tightening the tolerance from $\gamma = 0.9$ to $\gamma = 0.5$ improves the quality of the interpolation, which is reflected in a lower norm $\|1 - A_f^{-1}S_f\|_{|A_f|}$ (i.e., $\|E_{tg}^{(\infty)}\|$), but at the expense of denser matrices.

Figures 5.8 to 5.11 show residual norm histories for multigrid V-cycle iterations corresponding to these two AMG hierarchies, for the cases of 1, 2, and 4 pre- and post-smoothing steps at every level. Figures 5.8 and 5.9 plainly show that accelerating the plain multigrid V-cycle with the GMRES method (i.e., using the V-cycle as a preconditioner) results in substantial improvement. We also see that the $\gamma = 0.9$ case sees comparatively little gain in increasing the number of pre- and post-smoothing steps from 2 to 4, compared to the $\gamma = 0.5$. This reflects the structure of the convergence bounds of Theorems 4.2 and 4.3, which are sums of separate terms for the interpolation

Table 5.3: AMG Hierarchy for 2D, $p = 1/200$, 63×63 Chebyshev, $\gamma = 0.5$

Level	n	n_c/n	θ (°)	nnz/n	θ_f (°)	$\|E_{tg}^{(\infty)}\|$	$\|E_{mg}^{(4,4)}\|$	$\|E_{mg}^{(1,1)}\|$	$\rho(E_{mg}^{(1,1)})$
1	3721	0.47	85.8	8.8	27.1	0.16	0.19	0.84	0.56
2	1742	0.37	83.4	14.0	30.3	0.11	0.17	0.72	0.56
3	639	0.33	78.1	22.1	29.7	0.090	0.15	0.62	0.55
4	213	0.34	66.8	25.7	26.7	0.079	0.14	0.58	0.55
5	73	0.30	43.4	28.8	28.8	0.072	0.13	0.56	0.55
6	22	0.23	4.57	21.5	4.55	0.12	0.12	0.54	0.54
7	5	0	0.304	5	0.30	0	0.083	0.54	0.54

Table 5.4: AMG Hierarchy for 2D, $p = 1/200$, 63×63 Chebyshev, $\gamma = 0.9$

Level	n	n_c/n	θ (°)	nnz/n	θ_f (°)	$\|E_{tg}^{(\infty)}\|$	$\|E_{mg}^{(4,4)}\|$	$\|E_{mg}^{(1,1)}\|$	$\rho(E_{mg}^{(1,1)})$
1	3721	0.47	85.8	8.8	27.1	0.45	0.65	1.49	0.54
2	1742	0.35	83.5	10.2	31.3	0.63	0.57	1.33	0.53
3	617	0.31	78.3	14.8	32.3	0.32	0.43	1.02	0.51
4	193	0.33	65	20.3	28.4	0.27	0.34	0.79	0.50
5	64	0.27	37.8	26.2	26.8	0.31	0.29	0.59	0.48
6	17	0.18	7.5	16.9	5.47	0.27	0.27	0.49	0.47
7	3	0	2.54	3	2.54	0	0.04	0.45	0.45

and the smoother. Taking more smoother steps lowers the term for the smoother, but this does little good if the interpolation term is already dominant. Figures 5.10 and 5.11 show the effect of accelerating the smoother iteration itself using GMRES, in addition to putting the whole V-cycle inside a GMRES iteration; this is what the GMRES^2 label is intended to indicate. Our implementation of GMRES is actually FGMRES, which is required here as the use of GMRES for the smoother renders the overall V-cycle nonlinear. While there does seem to be some improvement with the GMRES-accelerated F-relaxation, it appears marginal, especially in the $\gamma = 0.9$ case.

Tables 5.5 and 5.6 look in detail at the interpolation on each level, again for the two hierarchies presented in Tables 5.3 and 5.4. The unsubscripted norms $\|F_r\|$ and $\|F_s\|$ in these tables are $\|F_r\|_{|A_f|,|\hat{A}_c|}$ and $\|F_s\|_{|A_f|,|\hat{A}_c|}$, while $\|F_r\|_D$ is $\|F_r\|_{|A_f|,|\hat{D}_c|}$. Recall that, in the development of the interpolation heuristics of §5.3, which would ideally find, e.g., weights W_r so that $\|F_r\| = \|F_r\|_{|A_f|,|\hat{A}_c|}$ is no more than γ, we took the drastic step of replacing this norm with $\|F_r\|_D = \|F_r\|_{|A_f|,|\hat{D}_c|}$ and assuming this would suffice provided we ensure constants are handled correctly. Comparing the values of the two norms in these tables, it would appear that this assumption is borne out, at least in these two cases. However, this is not always so. Table 5.7 is the same as Table 5.6 but for the case of a uniformly instead of Chebyshev spaced grid. There is a

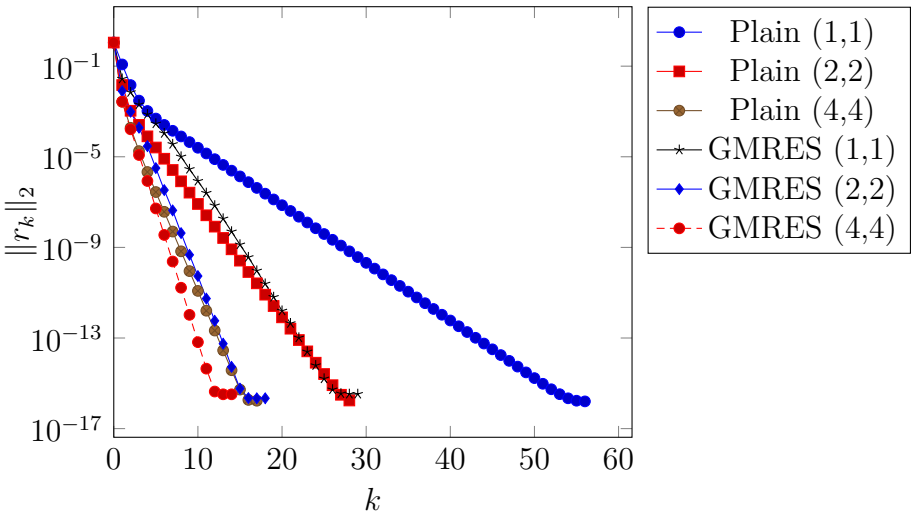


Figure 5.8: Top level residual histories for Table 5.3 ($\gamma = 0.5$)

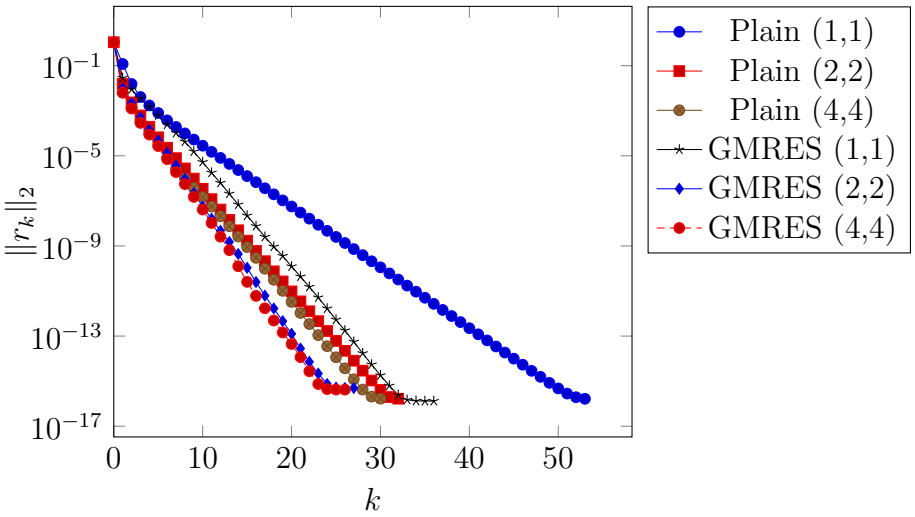


Figure 5.9: Top level residual histories for Table 5.4 ($\gamma = 0.9$)

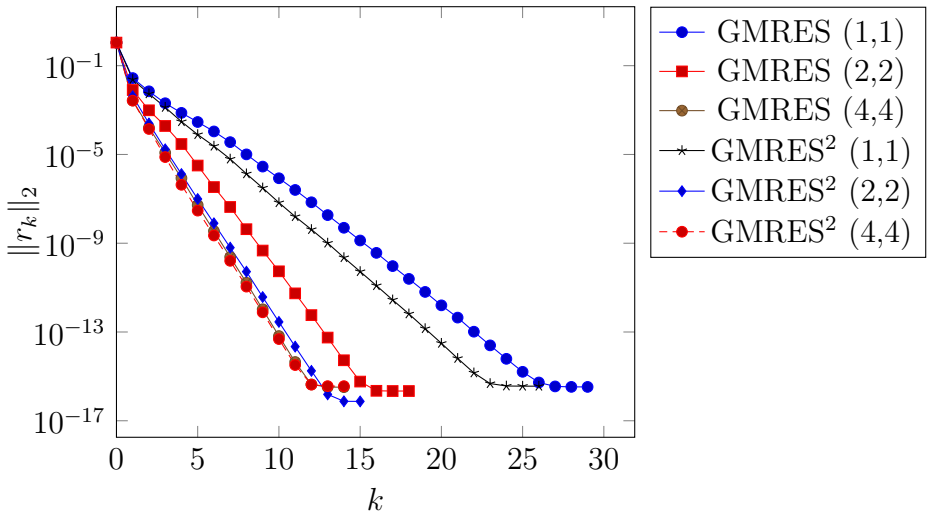


Figure 5.10: Top level residual histories for Table 5.3 ($\gamma = 0.5$)

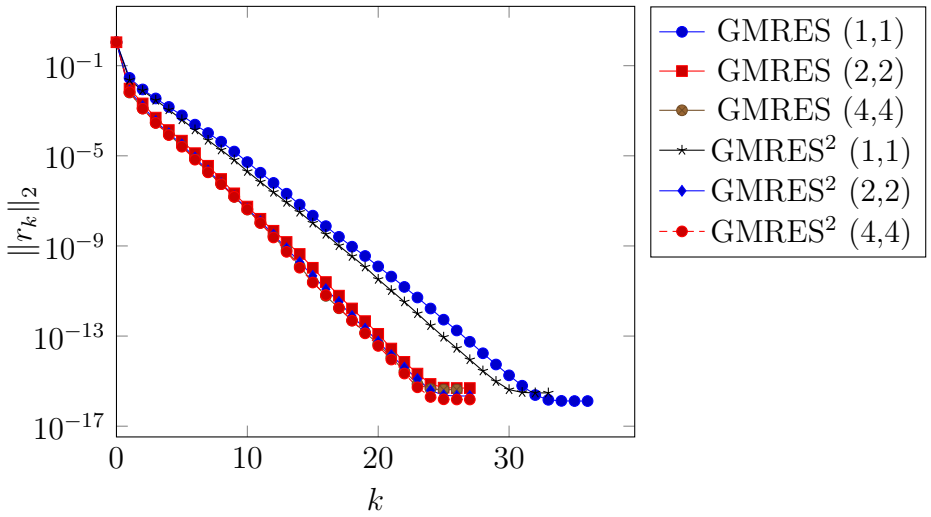


Figure 5.11: Top level residual histories for Table 5.4 ($\gamma = 0.9$)

notable discrepancy between the two norms for the second level. One hypothesis that may explain this is that, for strongly advection-dominated problems, the near-null space is not limited just to global constants, but includes also any function constant along streamlines. Addressing this shortcoming is left for future work. At any rate, the heuristics appear to have done quite well in making the norm they actually targeted, $\|F_r\|_{|A_f|,|\hat{D}_c|}$, be lower than or at least not much more than γ. As a final observation on these tables, Lemma 4.2 states exactly that the values in the third column, $\|E_{tg}^{(\infty)}\|$, i.e., $\|1 - A_f^{-1} S_f\|_{|A_f|}$, must be less than or equal to those in the fourth, the product of the norms of the trial and test weight errors, and indeed we can see that to be the case, including for the problematic second level of Table 5.7.

Table 5.5: Interpolation for 2D, $p = 1/200$, 63×63 Chebyshev, $\gamma = 0.5$

Level	n	$\|F_r\|_D$	$\|F_r\|$	$\|F_s\|$	$\|E_{tg}^{(\infty)}\|$	$\|F_r\|\|F_s\|$
1	3721	0.44	0.46	0.46	0.16	0.21
2	1742	0.36	0.36	0.36	0.11	0.13
3	639	0.32	0.30	0.33	0.09	0.098
4	213	0.32	0.31	0.31	0.079	0.098
5	73	0.30	0.27	0.28	0.072	0.074
6	22	0.38	0.34	0.35	0.12	0.12

Table 5.6: Interpolation for 2D, $p = 1/200$, 63×63 Chebyshev, $\gamma = 0.9$

Level	n	$\|F_r\|_D$	$\|F_r\|$	$\|F_s\|$	$\|E_{tg}^{(\infty)}\|$	$\|F_r\|\|F_s\|$
1	3721	0.81	0.74	0.74	0.45	0.55
2	1742	0.78	0.89	0.84	0.63	0.75
3	617	0.61	0.59	0.61	0.32	0.36
4	193	0.53	0.53	0.52	0.27	0.27
5	64	0.62	0.61	0.57	0.31	0.34
6	17	0.66	0.56	0.50	0.27	0.28

Table 5.8 reports figures summarizing the AMG hierarchies constructed for the main case we have been considering using the four different interpolation methods from §5.3 and listed in Figure 5.4, including the impractical "Abs, exact" method, which is only possible here as the problem size is small. "Levels" in the table is the total number of levels in the hierarchy. "Growth" is the sum of the count of nonzeros of the A matrices from all levels divided by the number of nonzeros of the original matrix. The last three columns are then the spectral radius of the V-cycle with 1 pre- and post-smoothing step, and the average convergence rate of the GMRES accelerated

Table 5.7: Interpolation for 2D, $p = 1/200$, 63×63 uniform, $\gamma = 0.9$

Level	n	$\|F_r\|_D$	$\|F_r\|$	$\|F_s\|$	$\|E_{tg}^{(\infty)}\|$	$\|F_r\|\|F_s\|$
1	3721	0.75	0.74	0.74	0.43	0.55
2	1851	0.92	1.96	1.95	3.02	3.82
3	842	0.74	0.87	0.88	0.61	0.77
4	349	0.61	0.62	0.61	0.31	0.38
5	129	0.58	0.56	0.52	0.24	0.29
6	43	0.54	0.53	0.54	0.21	0.29
7	14	0.76	0.64	0.62	0.40	0.40

V-cycles with 1 and 4 pre- and post-smoothing steps. The main conclusion from the table is that the particular choice of interpolation method does not appear to make much difference.

Table 5.8: 2D on 63 by 63 Chebyshev grid, $p = 1/200$

				$\rho(E)$	GMRES	
Interpolation	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
Diagonal	0.5	7	2.4	0.55	0.29	0.072
Abs, exact	0.5	7	2.47	0.48	0.28	0.067
Abs, extended	0.5	7	2.45	0.55	0.29	0.065
Abs, restricted	0.5	7	2.42	0.56	0.26	0.065
Diagonal	0.9	8	2	0.55	0.35	0.25
Abs, exact	0.9	7	2.04	0.55	0.34	0.22
Abs, extended	0.9	7	2	0.6	0.34	0.24
Abs, restricted	0.9	7	2	0.54	0.34	0.26

For our last detailed look at this particular test problem, we turn to examining the effectiveness of the convergence bounds proved in the last chapter. Computed values of two norms of the two-level projected iteration matrix $1 - \hat{B}_f S_f$ along with the bounds of these norms from Theorems 4.2 and 4.3 are reported in Table 5.9, for the third level of the $\gamma = 0.9$ case we have been considering. These theorems used Lemma 4.2 to split one of the terms of the bound into a product of two norms, one of the trial weight errors and one of the test weight errors. The intermediate bounds in the table were obtained by skipping this step. The original motivation for developing the bound of Theorem 4.3 was to avoid the final term of the bound of Theorem 4.2, and indeed we see that Theorem 4.3 tends to be substantially better, especially at the smaller number of smoothing steps. The two bounds converge as the number of smoothing steps increases; this is simply because $\hat{B}_f^{-1} \to A_f$ in this limit.

Table 5.9: Convergence bounds for level 3 of Table 5.4

Smoother steps	1	2	3	4	8	16
$\|1 - \hat{B}_f S_f\|_{\|A_f\|}$	0.752	0.567	0.425	0.342	0.318	0.317
Bound (Int)	1.24	0.926	0.694	0.536	0.344	0.318
Bound (Thm 4.2)	1.31	0.986	0.747	0.583	0.385	0.359
$\|1 - \hat{B}_f S_f\|_{\|\hat{B}_f\|^{-1}}$	0.733	0.549	0.415	0.342	0.318	0.317
Bound (Int)	0.956	0.772	0.609	0.487	0.337	0.318
Bound (Thm 4.3)	1.03	0.806	0.644	0.525	0.378	0.359

Next we look at a few larger test problems. Table 5.10 shows the hierarchy for the "2D" problem on a 511×511 Chebyshev grid with diffusion coefficient $p = 1/3200$, constructed with interpolation tolerance $\gamma = 0.8$. There is some missing data in the table, as there is no easy way to compute the form absolute value norms of the iteration matrices of such sizes. One number that stands out is the large value of $\|E_{tg}^{(\infty)}\| = \|1 - A_f^{-1} S_f\|_{|A_f|}$ in level two, indicating that the interpolation heuristic did not work in this case, just as we saw in Table 5.7. In this case, the bad interpolation seriously degrades the performance of the overall V-cycle, in some cases even diverging, as Figure 5.12 shows. We also see, however, that the use of GMRES restores the performance of the method.

Table 5.10: AMG Hierarchy for 2D, $p = 1/3200$, 511×511 Chebyshev, $\gamma = 0.8$

Level	n	n_c/n	θ (°)	nnz/n	θ_f (°)	$\|E_{tg}^{(\infty)}\|$	$\|E_{mg}^{(4,4)}\|$	$\|E_{mg}^{(1,1)}\|$	$\rho(E_{mg}^{(1,1)})$
1	259081	0.51	89.6	8.98	29.9	0.46	—	—	0.92
2	131023	0.45	89.3	10.3	35.3	4.75	—	—	0.85
3	59563	0.42	88.9	15.5	33.7	0.46	—	—	0.61
4	25027	0.39	87.9	23.4	33.9	0.22	—	—	0.61
5	9783	0.39	85.7	33.2	32.7	0.19	—	—	0.61
6	3858	0.39	80.7	44.0	30.6	0.15	0.41	0.71	0.60
7	1518	0.37	68.9	54.6	28.8	0.15	0.39	0.63	0.60
8	568	0.4	46.9	56.9	21	0.17	0.37	0.61	0.60
9	230	0.37	8.19	57.1	7.24	0.33	0.35	0.60	0.59
10	86	0.35	4.01	38.4	3.7	0.22	0.28	0.58	0.57
11	30	0.27	3.9	23.3	3.56	0.24	0.24	0.53	0.53
12	8	0.12	3	8	2.58	0	0.068	0.51	0.51
13	1	0	0	1	0	0	0	0	0

Tables 5.11 and 5.12 summarize the hierarchies constructed for the 3D test problem with $p = 1/200$ and $p = 1/2000$. One obvious feature illustrated by these tables is that the problem of stencil growth is substantially worse in 3D than in 2D. It is also

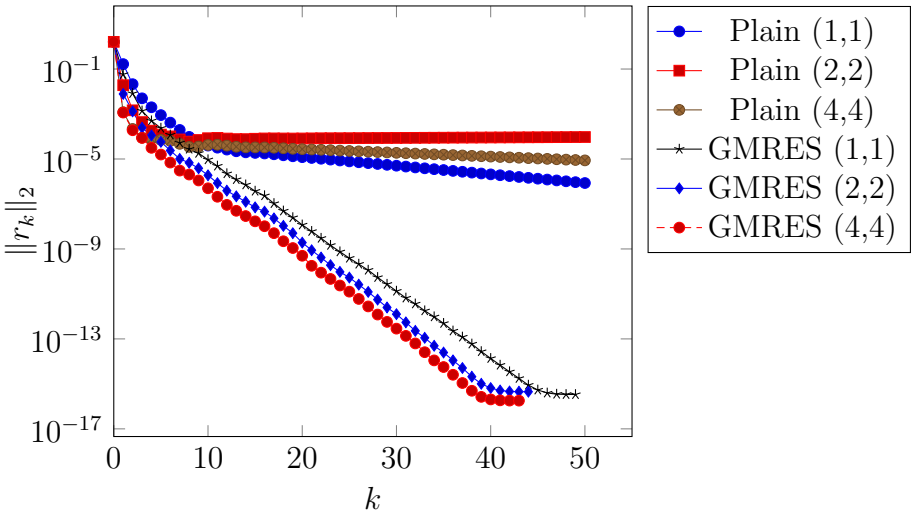


Figure 5.12: Top level residual histories for Table 5.10

interesting to note how reducing the diffusion coefficient by a factor of 10 dramatically increased the number of multigrid levels. The streamlines are particularly long for this example (warped slowly winding helices), so that the strategy of semi-coarsening along streamlines can continue for a large number of steps. When the diffusion coefficient is larger, the coarsening heuristic is sooner able to switch to a more aggressive strategy, whereas with a smaller coefficient, the degrees of freedom along different streamlines remain very weakly coupled even far down in the hierarchy.

Table 5.11: AMG Hierarchy for 3D, $p = 1/200$, $63 \times 63 \times 63$, $\gamma = 1.2$

Level	n	n_c/n	θ (°)	nnz/n	θ_f (°)	$\|E_{tg}^{(\infty)}\|$	$\|E_{mg}^{(4,4)}\|$	$\|E_{mg}^{(1,1)}\|$	$\rho(E_{mg}^{(1,1)})$
1	230702	0.55	85.8	26.4	22.6	0.83	—	—	0.62
2	127342	0.46	83.5	54.6	25.8	0.45	—	—	0.48
3	58077	0.38	78.8	111.8	27.7	0.55	—	—	0.44
4	21845	0.33	69.4	186.9	28.0	0.53	—	—	0.43
5	7228	0.32	56.5	340.1	29.4	0.36	—	—	0.43
6	2330	0.28	45.7	554.9	19.2	0.35	0.30	0.43	0.40
7	647	0.21	28.0	399.2	19.8	0.47	0.31	0.39	0.37
8	138	0.087	8.66	135.3	8.51	0.098	0.047	0.21	0.21
9	12	0	1.42	12	1.42	0	$< 10^{-8}$	0.0082	0.0081

Tables 5.13 to 5.18 summarize the results for all four test problems, varying problem size, the diffusion coefficient, and the interpolation tolerance, and including the cases we have already seen. Table 5.13 in particular shows that the AMG method continues to function well in the degenerate symmetric case. We can see that there is a consistent trade-off between the interpolation tolerance and the sparsity of the

Table 5.12: AMG Hierarchy for 3D, $p = 1/2000$, $63 \times 63 \times 63$, $\gamma = 1.2$

Level	n	n_c/n	θ (°)	nnz/n	θ_f (°)	$\|E_{tg}^{(\infty)}\|$	$\|E_{mg}^{(4,4)}\|$	$\|E_{mg}^{(1,1)}\|$	$\rho(E_{mg}^{(1,1)})$
1	230702	0.64	87.7	26.4	25.7	0.48	—	—	0.59
2	148079	0.58	86.8	39.8	27.9	0.62	—	—	0.50
3	85352	0.59	86.5	67.1	25.6	0.73	—	—	0.50
4	50424	0.6	86	99.0	23.8	0.50	—	—	0.50
5	30321	0.6	85.1	132.3	20.8	0.47	—	—	0.51
6	18075	0.58	83.9	164.4	19.9	0.44	—	—	0.50
7	10468	0.57	82.2	198.5	21.2	0.40	—	—	0.50
8	5920	0.55	79.4	240.9	21.4	0.46	—	—	0.50
9	3255	0.51	74.7	280.3	21.1	0.37	0.37	0.54	0.50
10	1671	0.46	65.9	292.2	23.7	0.26	0.37	0.53	0.50
11	777	0.4	47.5	283.4	26.2	0.31	0.37	0.52	0.50
12	314	0.34	16.3	246.0	12.3	0.42	0.37	0.50	0.48
13	107	0.28	6.79	104.9	6.42	0.47	0.28	0.48	0.46
14	30	0.13	7.88	30	7.71	0.22	0.13	0.31	0.30
15	4	0	1.16	4	1.16	0	10^{-6}	0.033	0.033

matrices, and also that it is not worth expending too much effort on the smoother unless the interpolation quality is also high.

Table 5.13: Poisson on Chebyshev grid

					$\rho(E)$	GMRES	
N	n	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	841	0.8	4	1.59	0.54	0.18	0.11
63	3721	0.8	5	2	0.6	0.21	0.14
127	15625	0.8	6	2.2	0.61	0.22	0.15
255	64009	0.8	7	2.27	0.64	0.24	0.16
511	259081	0.8	8	2.33	0.66	0.26	0.17

Table 5.14: 2D-const on Chebyshev grid

							$\rho(E)$	GMRES	
N	n	p	θ (°)	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	841	1/200	83.3	0.3	5	1.83	0.27	0.15	0.088
31	841	1/200	83.3	0.5	5	1.67	0.28	0.16	0.095
31	841	1/200	83.3	0.7	5	1.6	0.28	0.15	0.12
31	841	1/200	83.3	0.8	5	1.58	0.3	0.14	0.15
31	841	1/200	83.3	0.9	5	1.57	0.33	0.15	0.15
63	3721	1/400	86.7	0.3	7	2.15	0.3	0.21	0.13
63	3721	1/400	86.7	0.5	6	1.92	0.3	0.24	0.19
63	3721	1/400	86.7	0.7	7	1.8	0.3	0.24	0.21
63	3721	1/400	86.7	0.8	6	1.74	0.32	0.25	0.24
63	3721	1/400	86.7	0.9	7	1.71	0.44	0.25	0.27
63	3721	1/200	86.4	0.3	7	2.41	0.29	0.17	0.043
63	3721	1/200	86.4	0.5	6	2.12	0.3	0.24	0.23
63	3721	1/200	86.4	0.7	6	1.95	0.3	0.26	0.28
63	3721	1/200	86.4	0.8	8	1.87	0.36	0.26	0.28
63	3721	1/200	86.4	0.9	7	1.83	0.36	0.25	0.24
127	15625	1/800	88.4	0.3	10	2.44	0.3	0.22	0.15
127	15625	1/800	88.4	0.5	8	2.14	0.34	0.27	0.27
127	15625	1/800	88.4	0.7	10	2.01	0.36	0.37	0.37
127	15625	1/800	88.4	0.8	11	1.95	0.37	0.36	0.35
127	15625	1/800	88.4	0.9	8	1.89	0.47	0.38	0.38
127	15625	1/400	88.2	0.3	7	2.79	0.33	0.27	0.066
127	15625	1/400	88.2	0.5	9	2.42	0.34	0.29	0.25
127	15625	1/400	88.2	0.7	8	2.17	0.36	0.3	0.3
127	15625	1/400	88.2	0.8	7	2.07	0.4	0.34	0.33
127	15625	1/400	88.2	0.9	8	1.99	0.47	0.35	0.38
127	15625	1/200	87.9	0.3	7	2.97	0.37	0.24	0.036
127	15625	1/200	87.9	0.5	7	2.66	0.39	0.28	0.1
127	15625	1/200	87.9	0.7	7	2.4	0.39	0.3	0.25
127	15625	1/200	87.9	0.8	7	2.25	0.41	0.35	0.33
127	15625	1/200	87.9	0.9	7	2.16	0.51	0.4	0.4
255	64009	1/1600	89.2	0.8	10	2.02	0.41	0.5	0.5
255	64009	1/200	88.6	0.8	7	2.46	0.5	0.37	0.21
511	259081	1/3200	89.6	0.6	10	2.24	0.42	0.52	0.53
511	259081	1/3200	89.6	0.7	11	2.18	0.48	0.56	0.57
511	259081	1/3200	89.6	0.8	12	2.12	1.38	0.62	0.62
511	259081	1/200	88.7	0.6	7	2.7	0.5	0.38	0.19
511	259081	1/200	88.7	0.7	7	2.57	0.54	0.4	0.25
511	259081	1/200	88.7	0.8	7	2.48	0.57	0.44	0.3

Table 5.15: 2D-const on uniform grid

							$\rho(E)$	GMRES	
N	n	p	θ (°)	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	841	1/200	83.6	0.3	5	2.51	0.18	0.078	0.03
31	841	1/200	83.6	0.5	6	1.97	0.26	0.14	0.1
31	841	1/200	83.6	0.7	6	1.76	0.26	0.12	0.064
31	841	1/200	83.6	0.8	6	1.71	0.26	0.12	0.075
31	841	1/200	83.6	0.9	6	1.71	0.26	0.11	0.08
63	3721	1/400	86.9	0.3	7	2.86	0.17	0.09	0.037
63	3721	1/400	86.9	0.5	7	2.14	0.27	0.2	0.18
63	3721	1/400	86.9	0.7	6	1.9	0.27	0.19	0.21
63	3721	1/400	86.9	0.8	7	1.82	0.29	0.21	0.25
63	3721	1/400	86.9	0.9	7	1.77	0.34	0.22	0.24
63	3721	1/200	86.7	0.3	6	2.47	0.16	0.11	0.054
63	3721	1/200	86.7	0.5	6	2.42	0.15	0.083	0.053
63	3721	1/200	86.7	0.7	6	1.88	0.34	0.29	0.3
63	3721	1/200	86.7	0.8	6	1.81	0.31	0.29	0.31
63	3721	1/200	86.7	0.9	6	1.8	0.32	0.29	0.3
127	15625	1/800	88.5	0.3	8	2.72	0.18	0.11	0.038
127	15625	1/800	88.5	0.5	8	2.04	0.32	0.26	0.24
127	15625	1/800	88.5	0.7	7	1.96	0.3	0.35	0.31
127	15625	1/800	88.5	0.8	8	1.88	0.73	0.38	0.37
127	15625	1/800	88.5	0.9	8	1.85	0.74	0.4	0.38
127	15625	1/400	88.4	0.3	7	2.63	0.2	0.2	0.054
127	15625	1/400	88.4	0.5	7	2.53	0.23	0.29	0.09
127	15625	1/400	88.4	0.7	7	1.99	0.34	0.36	0.34
127	15625	1/400	88.4	0.8	7	1.94	0.36	0.36	0.34
127	15625	1/400	88.4	0.9	7	1.92	0.36	0.37	0.35
127	15625	1/200	88.2	0.3	6	2.88	0.21	0.31	0.029
127	15625	1/200	88.2	0.5	6	2.58	0.23	0.32	0.08
127	15625	1/200	88.2	0.7	6	2.49	0.29	0.4	0.16
127	15625	1/200	88.2	0.8	6	2.45	0.26	0.39	0.16
127	15625	1/200	88.2	0.9	6	2.4	0.31	0.39	0.28

Table 5.16: 2D on Chebyshev grid

							$\rho(E)$	GMRES	
N	n	p	θ (°)	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	841	1/200	83.1	0.3	7	2.31	0.52	0.19	0.025
31	841	1/200	83.1	0.5	7	1.98	0.44	0.25	0.082
31	841	1/200	83.1	0.7	7	1.85	0.43	0.26	0.14
31	841	1/200	83.1	0.8	7	1.83	0.51	0.29	0.19
31	841	1/200	83.1	0.9	7	1.76	0.54	0.32	0.26
63	3721	1/400	86.6	0.3	9	2.78	0.43	0.23	0.034
63	3721	1/400	86.6	0.5	9	2.36	0.44	0.27	0.084
63	3721	1/400	86.6	0.7	9	2.17	0.47	0.28	0.15
63	3721	1/400	86.6	0.8	9	2.07	0.54	0.33	0.23
63	3721	1/400	86.6	0.9	9	2.02	0.63	0.37	0.28
63	3721	1/200	85.8	0.3	8	2.81	0.46	0.25	0.029
63	3721	1/200	85.8	0.5	7	2.42	0.56	0.26	0.065
63	3721	1/200	85.8	0.7	7	2.18	0.52	0.31	0.15
63	3721	1/200	85.8	0.8	7	2.1	0.56	0.32	0.18
63	3721	1/200	85.8	0.9	7	2	0.54	0.34	0.26
127	15625	1/800	88.3	0.3	10	3.05	0.5	0.28	0.051
127	15625	1/800	88.3	0.5	10	2.64	0.51	0.31	0.12
127	15625	1/800	88.3	0.7	10	2.36	0.53	0.35	0.21
127	15625	1/800	88.3	0.8	10	2.27	0.6	0.38	0.31
127	15625	1/800	88.3	0.9	10	2.2	0.69	0.45	0.39
127	15625	1/400	87.9	0.3	9	3.14	0.46	0.28	0.036
127	15625	1/400	87.9	0.5	9	2.7	0.47	0.3	0.1
127	15625	1/400	87.9	0.7	9	2.41	0.55	0.34	0.18
127	15625	1/400	87.9	0.8	9	2.31	0.57	0.35	0.24
127	15625	1/400	87.9	0.9	9	2.22	0.55	0.38	0.33
127	15625	1/200	87.2	0.3	7	3.38	0.53	0.29	0.034
127	15625	1/200	87.2	0.5	8	2.85	0.48	0.32	0.096
127	15625	1/200	87.2	0.7	7	2.53	0.56	0.35	0.15
127	15625	1/200	87.2	0.8	7	2.42	0.58	0.37	0.17
127	15625	1/200	87.2	0.9	8	2.35	0.58	0.4	0.23
255	64009	1/1600	89.2	0.8	12	2.42	0.61	0.44	0.39
255	64009	1/200	87.9	0.8	8	2.48	0.55	0.41	0.21
511	259081	1/3200	89.6	0.6	13	2.74	0.56	0.36	0.29
511	259081	1/3200	89.6	0.7	13	2.61	0.59	0.4	0.33
511	259081	1/3200	89.6	0.8	13	2.5	0.92	0.5	0.47
511	259081	1/200	88.2	0.6	9	2.75	0.54	0.36	0.15
511	259081	1/200	88.2	0.7	9	2.6	0.55	0.38	0.23
511	259081	1/200	88.2	0.8	9	2.48	0.57	0.41	0.24

Table 5.17: 2D on uniform grid

							$\rho(E)$	GMRES	
N	n	p	θ (°)	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	841	1/200	84.8	0.3	7	2.87	0.53	0.18	0.038
31	841	1/200	84.8	0.5	8	2.43	0.41	0.2	0.12
31	841	1/200	84.8	0.7	8	2.23	0.48	0.25	0.18
31	841	1/200	84.8	0.8	8	2.16	0.44	0.3	0.22
31	841	1/200	84.8	0.9	8	2.06	0.57	0.35	0.28
63	3721	1/400	87.5	0.3	9	2.74	0.44	0.19	0.093
63	3721	1/400	87.5	0.5	9	2.36	0.5	0.23	0.23
63	3721	1/400	87.5	0.7	9	2.19	0.52	0.32	0.32
63	3721	1/400	87.5	0.8	9	2.17	0.62	0.38	0.37
63	3721	1/400	87.5	0.9	9	2.12	0.67	0.43	0.41
63	3721	1/200	86.7	0.3	8	2.66	0.44	0.2	0.043
63	3721	1/200	86.7	0.5	8	2.48	0.48	0.25	0.16
63	3721	1/200	86.7	0.7	8	2.19	0.49	0.3	0.23
63	3721	1/200	86.7	0.8	8	2.11	0.55	0.37	0.31
63	3721	1/200	86.7	0.9	8	2.04	1.2	0.51	0.46
127	15625	1/800	88.8	0.3	11	3.02	0.49	0.19	0.13
127	15625	1/800	88.8	0.5	10	2.47	0.51	0.31	0.3
127	15625	1/800	88.8	0.7	11	2.28	0.57	0.41	0.42
127	15625	1/800	88.8	0.8	11	2.24	0.59	0.48	0.46
127	15625	1/800	88.8	0.9	11	2.16	0.89	0.56	0.54
127	15625	1/400	88.4	0.3	10	2.88	0.44	0.23	0.055
127	15625	1/400	88.4	0.5	10	2.55	0.5	0.26	0.23
127	15625	1/400	88.4	0.7	10	2.29	0.59	0.34	0.26
127	15625	1/400	88.4	0.8	9	2.21	0.8	0.48	0.43
127	15625	1/400	88.4	0.9	10	2.14	1.9	0.6	0.56
127	15625	1/200	87.7	0.3	8	3.34	0.43	0.3	0.028
127	15625	1/200	87.7	0.5	8	2.89	0.48	0.32	0.09
127	15625	1/200	87.7	0.7	8	2.57	0.52	0.33	0.23
127	15625	1/200	87.7	0.8	8	2.39	0.51	0.37	0.34
127	15625	1/200	87.7	0.9	8	2.25	0.55	0.41	0.39

Table 5.18: 3D on Chebyshev grid

							$\rho(E)$	GMRES	
N	n	p	θ (°)	γ	Levels	Growth	(1,1)	(1,1)	(4,4)
31	25230	1/200	83.1	0.8	9	4.37	0.48	0.24	0.11
31	25230	1/200	83.1	1.2	9	3.31	0.57	0.35	0.26
63	230702	1/2000	87.7	1.2	15	5.73	0.59	0.4	0.34
63	230702	1/200	85.8	1.2	9	4.53	0.62	0.38	0.32

Chapter 6

Conclusions

This work focused on analysis of methods for solving "positive" nonsymmetric problems, with an emphasis on application to algebraic multigrid (AMG) in particular. Specifically, the focus was on problems of the form

$$Au = f \tag{6.1}$$

with $A \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ defining a bounded sesquilinear form on the complex reflexive Banach space V and having positive real part $H = \frac{1}{2}(A + A^*)$. An archetypal example of such a problem is the weak form of the steady advection-diffusion equation $-p\nabla^2 u + \mathbf{c} \cdot \nabla u = f_S$ under suitable assumptions on the domain and coefficients, as presented in §1.2. In this case V is a suitable infinite-dimensional function space. Another example is given by any suitable discretization of this problem, like the one described in §1.2.3. In this case $\mathsf{V} = \mathbb{C}^n$ and A is a large sparse matrix. This latter example is a suitable target for AMG.

A key tool in the analysis of methods for symmetric problems, i.e., when $A = A^*$, is the energy norm $\|u\|_A^2 = (Au, u)$, which is characterized by, among other properties, having optimal boundedness and coercivity constants (both equal to 1):

$$\begin{aligned} |(Au, v)| &\leq \|u\|_A \|v\|_A \quad \text{for all } u, v \in \mathsf{V}, \quad \text{and} \\ (Au, u) &\geq \|u\|_A^2 \qquad\quad \text{for all } u \in \mathsf{V}. \end{aligned} \tag{6.2}$$

For nonsymmetric A, when the real part H of A is positive, it provides a norm. However, the corresponding boundedness and coercivity inequalities are

$$\begin{aligned} |(Au, v)| &\leq \sec\theta \, \|u\|_H \|v\|_H \quad \text{for all } u, v \in \mathsf{V}, \quad \text{and} \\ \sup_{v \in \mathsf{V} \setminus \{0\}} \frac{|(Au, v)|}{\|v\|_H} &\geq \|u\|_H \qquad\qquad\quad \text{for all } u \in \mathsf{V}, \end{aligned} \tag{6.3}$$

where θ is the sectorial half-angle of A, a measure of the nonsymmetry of A. In particular, for highly nonsymmetric problems, such as the advection-dominated cases

of advection-diffusion, θ is close to $\pi/2$ and $\sec\theta$ is very large. The implication is that the norm $\|\cdot\|_H$ is not suitable by itself for analysis of highly nonsymmetric problems.

In Chapter 3 we showed that, under the above assumptions on A, there is a unique $|A| \in \mathscr{B}(\mathsf{V}, \mathsf{V}^*)$ defining a coercive symmetric form on V satisfying boundedness and coercivity inequalities

$$\begin{aligned} |(Au, v)| &\le \|u\|_{|A|}\|v\|_{|A|} \quad \text{for all } u, v \in \mathsf{V}, \quad \text{and} \\ \sup_{v\in\mathsf{V}\setminus\{0\}} \frac{|(Au, v)|}{\|v\|_{|A|}} &\ge \|u\|_{|A|} \qquad\qquad \text{for all } u \in \mathsf{V}, \end{aligned} \tag{6.4}$$

that is, with the optimal constants, both equal to 1. Indeed this can be taken as a definition of what we have called the "form" absolute value. Alternately, $|A|$ can be defined as the unique positive symmetric solution to $A^*|A|^{-1}A = |A|$, or as the positive factor of a special polar decomposition $A = |A|U$ with $U \in \mathscr{B}(\mathsf{V})$ satisfying $U^*|A|U = |A|$. That is, $A = |A|U$ is the polar decomposition of A in the Hilbert space associated with the inner product defined by $|A|$ itself. The norm $\|\cdot\|_{|A|}$ provides a replacement in the nonsymmetric case of the energy norm in the symmetric case, a key tool in the analysis of solution methods. In particular, $|A|$ reduces to A in the symmetric case, and $\|\cdot\|_{|A|}$ reduces to the usual energy norm.

Chapter 4 presented a convergence theory for general two-level iterations for solving (6.1) corresponding to the iteration operator

$$(1 - B_2A)(1 - P_r\hat{A}_c^{-1}P_s^*A)(1 - B_1A), \tag{6.5}$$

with pre- and post-smoothers $B_1, B_2 \in \mathscr{B}(\mathsf{V}^*, \mathsf{V})$, prolongation operators $P_r, P_s \in \mathscr{B}(\mathsf{V}_c, \mathsf{V})$ defined on a closed subspace V_c of V, and the Petrov-Galerkin coarse operator $\hat{A}_c := P_s^*AP_r$. Our first main result, Theorem 4.1, was that the convergence of the above iteration is controlled by the effective part of the combined smoother B, defined by $1 - BA = (1 - B_1A)(1 - B_2A)$, given by the projection of B to an "F-relaxation." This result is equivalent, in a non-trivial way, to prior known results, in particular to a result by Notay [39] generalizing an earlier result for symmetric multigrid by Falgout et al. [19]. We then gave convergence bounds, Theorems 4.2 and 4.3, in terms of a smoothing property and separate approximation properties for the coarse trial and test spaces. These made key use of the new form absolute value and associated norm.

The heuristics presented in Chapter 5 demonstrated that the abstract convergence theory of Chapter 4 could be put directly to use in designing practical AMG methods. Indeed, the methods presented in that chapter worked very well on a wide variety of problems. Furthermore, where they had shortcomings, particularly in the assumptions

made in the interpolation heuristics, the convergence theory indicates exactly where the method goes wrong, and points the way to ideas for improvement.

There are a few ways to further develop the AMG heuristics of Chapter 5. Most pressing, the assumption that preserving a single vector, the discretized constant function, is sufficient to ensure good interpolation quality, needs to be revisited. Another issue worth investigating is that, while the Nikiforov bounds used in several places are quite good when applied to non-negative matrices (usually converging to the true norm as r increases), there is quite a penalty in bounding the two-norm of a matrix by that of its entry-wise absolute value (which we do in order to apply the bound). As a result, the Nikiforov bounds can sometimes be fairly pessimistic, and more so, it appeared, on coarser levels. This results in more expensive interpolation than necessary.

In the future, it would be interesting to investigate whether the form absolute value norm could be put to use in a multilevel (not just two-level) analysis. More generally, the form absolute value might perhaps find other applications besides in the analysis of iterative solution methods. For example, a natural question is whether the associated norm could be useful in the theory of errors for discretizations of nonsymmetric problems.

Bibliography

[1] Mark Adams, Marian Brezina, Jonathan J. Hu, and Raymond S. Tuminaro. Parallel multigrid smoothing: polynomial versus Gauss-Seidel. *J. Comp. Phys.*, 188(2):593–610, 2003.

[2] Yuri Bolshakov, Cornelis V. M. van der Mee, André M. Ran, Boris Reichstein, and Leiba Rodman. Polar decompositions in finite dimensional indefinite scalar product spaces: General theory. *Linear Algebra and its Applications*, 261:91–141, 1997.

[3] James H. Bramble, Joseph E. Pasciak, Junping Wang, and Jinchao Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Mathematics of Computation*, 57(195):23–45, 1991.

[4] Achi Brandt. Algebraic multigrid theory: The symmetric case. *Applied Mathematics and Computation*, 19(1–4):23–56, 1986.

[5] Achi Brandt. General highly accurate algebraic coarsening. *Electron. Trans. Numer. Anal.*, 10:1–20, 2000.

[6] Achi Brandt, Steve F. McCormick, and J. W. Ruge. Algebraic multigrid (AMG) for automatic multigrid ssolution with application to geodetic computations. Technical report, Institute for Computational Studies, Coloroda State University, 1982.

[7] Achi Brandt, Steve F. McCormick, and J. W. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In David J. Evans, editor, *Sparsity and its Applications*, pages 257–284. Cambridge University Press, 1985.

[8] James Brannick, Marian Brezina, S. MacLachlan, Thomas A. Manteuffel, S. F. McCormick, and J. W. Ruge. An energy-based AMG coarsening strategy. *Numer. Linear Algebra Appl.*, 13:133–148, 2006.

[9] James Brannick and Ludmil Zikatanov. Algebraic multigrid methods based on compatible relaxation and energy minimization. In *Proc. of the 16th Int. Conf. on Domain Decomposition Methods*, 2005.

[10] Oliver Bröker and Marcus J. Grote. Sparse approximate inverse smoothers for geometric and algebraic multigrid. *Appl. Numer. Math.*, 41(1):61–80, 2002.

[11] Alexander N. Brooks and Thomas J.R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32(1–3):199–259, 1982.

[12] Andrew J. Cleary, Robert D. Falgout, Van Emden Henson, and Jim E. Jones. *Solving Irregularly Structured Problems in Parallel*, volume 1457 of *Lecture Notes in Computer Science*, chapter Coarse-grid selection for parallel algebraic multigrid, pages 104–115. Springer, 1998.

[13] Stephen Demko, William F. Moss, and Philip W. Smith. Decay rates for inverses of band matrices. *Math. Comp.*, 43(168):491–499, October 1984.

[14] William F. Donoghue. The interpolation of quadratic norms. *Acta Mathematica*, 118(1):251–270, 1967.

[15] Victor Eijkhout and Panayot Vassilevski. The role of the strengthened Cauchy-Buniakowskii-Schwarz inequality in multilevel methods. *SIAM Rev.*, 33(3):405–419, August 1991.

[16] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2005.

[17] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 1998.

[18] Robert D. Falgout and Panayot S. Vassilevski. On generalizing the algebraic multigrid framework. *SIAM J. Numer. Anal.*, 42(4):1669–1693, 2004.

[19] Robert D. Falgout, Panayot S. Vassilevski, and Ludmil T. Zikatanov. On two-grid convergence estimates. *Numer. Linear Algebra Appl.*, 12(5–6):471–494, 2005.

[20] Radii Petrovich Fedorenko. A relaxation method for solving elliptic difference equations. *USSR Comput. Math. Math. Phys.*, 1:1092–1096, 1962.

[21] Radii Petrovich Fedorenko. The rate of convergence of an iterative process. *USSR Comput. Math. Math. Phys.*, 4:227–235, 1964.

[22] Radii Petrovich Fedorenko. Iterative methods for elliptic difference equations. *Russian Mathematical Surveys*, 28:129–195, 1973.

[23] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, third edition, 1996.

[24] Michael Griebel. Multilevel algorithms considered as iterative methods on semidefinite systems. *SIAM J. Sci. Comput.*, 15(3):547–565, 1994.

[25] Wolfgang Hackbusch. *Multi-Grid Methods and Applications*. Springer Series in Computational Mathematics. Springer, 1985.

[26] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[27] Nick Higham, Christian Mehl, and Françoise Tisseur. The canonical generalized polar decomposition. *Siam J. Matrix Anal. Appl.*, 31(4):2163–2180, 2010.

[28] Roger A. Horn and Vladimir V. Sergeichuk. Canonical forms for complex matrix congruence and *congruence. *Linear Algebra and its Applications*, 416(2–3):1010–1032, 2006.

[29] Charles R. Johnson and Susana Furtado. A generalization of Sylvester's law of inertia. *Linear Algebra and its Applications*, 338:287–290, 2001.

[30] Jim E. Jones and Panayot S. Vassilevski. AMGe based on element agglomeration. *SIAM J. Sci. Comp.*, 23(1):109–133, 2001.

[31] Tosio Kato. *Perturbation theory for linear operators*. Springer, 1966.

[32] Hwan Ho Kim, , Jinchao Xu, and Ludmil Zikatanov. A multigrid method based on graph matching for convectiondiffusion equations. *Numerical Linear Algebra with Applications*, 10(1-2):181–195, 2003.

[33] Fumio Kubo and Tsuyoshi Ando. Means of positive linear operators. *Math. Ann.*, 246(3):205–224, 1980.

[34] Jimmie D. Lawson and Yongdo Lim. The geometric mean, matrices, metrics, and more. *The American Mathematical Monthly*, 108(9):797–812, 2001.

[35] Jan Mandel, Steve F. McCormick, and John W. Ruge. An algebraic theory for multigrid methods for variational problems. *SIAM Journal on Numerical Analysis*, 25(1):91–110, 1988.

[36] Thomas A. Manteuffel. The Tchebychev iteration for nonsymmetric linear systems. *Numer. Math.*, 28:307–327, 1977.

[37] John E. McCarthy. Geometric interpolation between Hilbert spaces. *Arkiv för Matematik*, 30:321–330, 1992.

[38] Vladimir Nikiforov. Revisiting Schur's bound on the largest singular value, 2007. Available at http://arxiv.org/abs/math/0702722.

[39] Yvan Notay. Algebraic analysis of two-grid methods: The nonsymmetric case. *Numer. Linear Algebra Appl.*, 17:73–96, 2010.

[40] Yvan Notay. Aggregation-based algebraic multigrid for convection-diffusion equations. *SIAM Journal on Scientific Computing*, 34(4):A2288–A2316, 2012.

[41] Alexander Ostrowski. Über das Nichtverschwinden einer Klasse von Determinanten und die Lokalisierung der charakteristischen Wurzeln von Matrizen. *Compositio Math.*, 9:209–226, 1951.

[42] W. Pusz and S. L. Woronowicz. Functional calculus for sesquilinear forms and the purification map. *Reports on Mathematical Physics*, 8(2):159–170, 1975.

[43] Alfio Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations*. Springer Series in Computational Mathematics. Springer, 1994.

[44] Ioan Roşca. On the Babuška Lax Milgram theorem. *An. Univ. Bucureşti*, 38(3):61–65, 1989.

[45] Ioan Roşca. Bilinear coercive and weakly coercive operators. *An. Univ. Bucureşti, Mat*, (2):183–188, 2002.

[46] J. W. Ruge and Klaus Stüben. Algebraic multigrid. In S. F. McCormick, editor, *Multigrid Methods*, pages 73–130. SIAM, 1987.

[47] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2 edition, April 2003.

[48] Klaus Stüben. Algebraic multigrid (AMG): An introduction with applications. In Ulrich Trottenberg, Cornelius W. Oosterlee, and Anton Schüller, editors, *Multigrid*. Academic Press, 2001.

[49] Wei-Pai Tang and Wing Lok Wan. Sparse approximate inverse smoother for multigrid. *SIAM J. Matrix Anal. Appl.*, 21(4):1236–1252, 2000.

[50] Petr Vaněk. Acceleration of convergence of a two-level algorithm by smoothing transfer operators. *Appl. Math.*, 37(4):265–274, 1992.

[51] Panayot S. Vassilevski. *Multilevel Block Factorization Preconditioners: Matrix-based Analysis and Algorithms for Solving Finite Element Equations*. Springer, 2008.

[52] W. L. Wan, Tony F. Chan, and Barry Smith. An energy-minimizing interpolation for robust multigrid methods. *SIAM J. Sci. Comp.*, 21(4):1632–1649, 1999.

[53] Jinchao Xu and Ludmil Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *Journal of the American Mathematical Society*, 15(3):573–597, 2002.

[54] Jinchao Xu and Ludmil Zikatanov. On an energy minimizing basis for algebraic multigrid methods. *Comput. Vis. Sci.*, 7(3–4):121–127, 2004.