Nonstandard Inner Products and Preconditioned Iterative Methods

Jennifer Pestana

Brasenose College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2011

Acknowledgements

I would like to sincerely thank my supervisor, Andy Wathen, for his constant enthusiasm, guidance and support. His curiosity and his ability to shed light on difficult topics has been invaluable, as has his perspective when I was particularly frustrated. His rapid and thorough proof-reading of this thesis deserves a special mention. I am also indebted to Nick Gould and Nick Trefethen for their advice, which has greatly improved this thesis.

It has been my good fortune to have had stimulating discussions with many visitors, including Zdenek Strakoš, Miroslav Rozložnìk and Eric de Sturler, for which I am very grateful. I wish to thank David Titley-Peloquin, whose input on nonsymmetric matrices has been very helpful, and Pedro Gonnet and Matthew Green who supplied code for the spectral discretization example, the workings of which Pedro patiently explained to me. My time at Oxford would not have been the same without the NA group and I wish to thank in particular the students and post docs, who have lent a friendly ear on many occasions. I could not have asked for nicer officemates than Almut and Edgar.

For their proof-reading and answers to questions, but primarily for their friendship, I would like to thank Dan Simpson and Cameron Hall. I would also like to thank the many people who have made Oxford special, especially those in the Brasenose HCR. Special thanks and love go to Caitlin, Lyd and Mel for many wonderful times and far too much cake.

It remains to thank those on whom I always rely and without whom this thesis would never have been written. My parents' love, support and belief have been unfaltering, and I cannot thank them enough. Finally, I thank Mark, who has patiently helped me through the difficult times and celebrated the good. I am both very lucky and very grateful.

Nonstandard Inner Products and Preconditioned Iterative Methods

Jennifer Pestana

Brasenose College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2011

By considering Krylov subspace methods in nonstandard inner products, we develop in this thesis new methods for solving large sparse linear systems and examine the effectiveness of existing preconditioners. We focus on saddle point systems and systems with a nonsymmetric, diagonalizable coefficient matrix.

For symmetric saddle point systems, we present a preconditioner that renders the preconditioned saddle point matrix nonsymmetric but self-adjoint with respect to an inner product and for which scaling is not required to apply a short-term recurrence method. The robustness and effectiveness of this preconditioner, when applied to a number of test problems, is demonstrated. We additionally utilize combination preconditioning (Stoll and Wathen. *SIAM J. Matrix Anal. Appl.* 2008; **30**:582–608) to develop three new combination preconditioners. One of these is formed from two preconditioners for which only a MINRES-type method can be applied, and yet a conjugate-gradient type method can be applied to the combination preconditioned system. Numerical experiments show that application of these preconditioners can result in faster convergence.

When the coefficient matrix is diagonalizable, but potentially nonsymmetric, we present conditions under which the pseudospectra of a preconditioner and coefficient matrix are identical and characterize the pseudospectra when this condition is not exactly fulfilled. We show that when the preconditioner and coefficient matrix are self-adjoint with respect to nearby symmetric bilinear forms the convergence of a particular minimum residual method is bounded by a term that depends on the spectrum of the preconditioned coefficient matrix and a constant that is small when the symmetric bilinear forms are close. An iteration-dependent bound for GMRES in the Euclidean inner product is presented that shows precisely why a standard bound can be pessimistic. We observe that for certain problems known, effective preconditioners are either self-adjoint with respect to the same symmetric bilinear form as the coefficient matrix or one that is nearby.

Contents

List of Figures

List of Algorithms

List of Tables

Notation

Fields and Vector Spaces

$\mathbb{C}$	The complex plane
$\mathbb{C}^n$	The space of n-dimensional complex vectors
$\mathbb{C}^{n\times n}$	The space of $n \times n$ complex matrices
$\mathbb{R}$	The real line
$\mathbb{R}^n$	The space of n-dimensional real vectors
$\mathbb{R}^{n\times n}$	The space of $n \times n$ real matrices

Krylov Subspaces

$\mathcal{K}_k(B, r_0)$	The Krylov subspace of dimension k of B and r_0
Π_k	The set of polynomials of degree k or lower
W-PCG	The preconditioned conjugate gradient method in $\langle\cdot,\cdot\rangle_W$
W-PGMRES	The preconditioned GMRES method in $\langle\cdot,\cdot\rangle_W$
W-PMINRES	The preconditioned MINRES method in $\langle\cdot,\cdot\rangle_W$

Linear Algebra

$\mathrm{diag}(a, b, c, \dots)$	A diagonal matrix with scalar entries $a, b, c, \dots$
$\mathrm{diag}(B, C, E, \dots)$	A block diagonal matrix with blocks $B, C, E, \dots$
$\dim(\mathcal{L})$	The dimension of the subspace $\mathcal{L}$
$\hat{Y}$	The reverse identity (or flip, or standard involutory) matrix
$\Lambda(B)$	The spectrum of the matrix B
$\Lambda_\epsilon(B)$	The ϵ-pseudospectrum of the matrix B
$\lambda_{max}(A)$	The largest eigenvalue of a symmetric matrix A
$\lambda_{min}(A)$	The smallest eigenvalue of a symmetric matrix A
$\mathcal{J}$	A real Jordan matrix in a real Jordan form
$\mathcal{J}(\mu \pm i\nu)$	A real Jordan block of the eigenvalues $\mu + i\nu$ and $\mu - i\nu$
$\mathrm{null}(B)$	The nullspace of the matrix B
$\mathrm{range}(B)$	The range of the matrix B
$\mathrm{rank}(B)$	The rank of the matrix B
$\sigma_{max}(B)$	The largest singular value of the matrix B
$\sigma_{min}(B)$	The smallest singular value of the matrix B
$\mathrm{span}\{v1, v2, \dots\}$	The span of the vectors $v_1, v_2, \dots$
$B \otimes C$	The Kronecker product of B and C
B^*	The Hermitian transpose of the matrix B

$B^\dagger$	The Moore-Penrose pseudoinverse of the matrix B
B^T	The transpose of the matrix B
J	A Jordan matrix in a Jordan canonical form
$J(\lambda)$	A Jordan block of the eigenvalue λ
$nnz(B)$	The number of nonzero entries in the matrix B

Saddle Point Preconditioners

BDI	The block diagonal preconditioner used with preconditioned MINRES in the Euclidean inner product
BDW	The block diagonal preconditioner used with $\mathcal{W}$-PMINRES
BP	The Bramble-Pasciak with Schur complement preconditioner
BP$^+$	The Bramble-Pasciak$^+$ preconditioner
SZ	The Schöberl-Zulehner preconditioner
SZ$^+$	The Schöberl-Zulehner$^+$ preconditioner

Symmetric Bilinear Forms

$\kappa_W(B)$	The condition number of the matrix B with respect to $\|\cdot\|_W$
$\sigma^W(B)$	A W-singular value of the matrix B
$\sigma^{M,W}(B)$	A $\{M, W\}$-singular value of the matrix B
$\langle\cdot,\cdot\rangle_W$	A nondegenerate symmetric bilinear form or inner product
$\|\cdot\|_W$	The norm induced by the inner product $\langle\cdot,\cdot\rangle_W$
$B^\star$	The W-adjoint of the matrix B
$y \perp_W \mathcal{L}$	The vector y is W-orthogonal to every vector in the space $\mathcal{L}$

The eigenvalues of a symmetric matrix A are labelled $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A)$, except where otherwise stated. Additionally, if $A \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{n\times n}$ are symmetric matrices, we use the Löwner partial ordering, so that $A > H$ if and only if $A - H$ is symmetric positive definite and $A \geq H$ if and only if $A - H$ is symmetric positive semidefinite.

Chapter 1

Introduction

Linear systems of the form

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

where $B \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, are ubiquitous in scientific computing arising, for example, in the solution of partial differential equations and in optimization. As computer hardware and software have improved, linear systems have become larger and applications more complex, with many problems requiring the solution of multiple systems. Consequently, the development of efficient algorithms for solving linear systems is an active area of research.

When the coefficient matrix is small, direct methods based on LU or Cholesky factorizations are usually favoured. These generally obtain the solution vector x to high accuracy and terminate in a finite number of operations. Furthermore, algorithms that exploit sparsity or other structure can be very fast. However, the $O(n^3)$ complexity for dense matrices and fill-in for sparse matrices make direct methods prohibitive for large matrices in general.

Larger linear systems are often sparse or structured in which case they are usually solved by iterative methods. In contrast to direct methods, these start from an initial approximation, x_0, to x and compute successive approximations that hopefully converge to the exact solution. Although some iterative methods have a finite termination property, the goal is to achieve a tolerably accurate solution in relatively few iterations. Simple iterations, such as the Jacobi and Gauss-Seidel methods, are easily implemented and analysed but often fail to converge. Conversely, more sophisticated methods are harder to analyse but converge for a much wider range of problems. Of these sophisticated methods, Krylov subspace methods are extremely

popular, although others such as multigrid can be very effective for particular classes of problems. We consider only Krylov subspace methods here.

Krylov subspace methods. Common to Krylov subspace methods is the approximation of the solution vector x from Krylov subspaces of B and the initial residual

$$r_0 = b - Bx_0, \tag{1.2}$$

where the Krylov subspace of dimension k of B and r_0 is given by

$$\mathcal{K}_k(B, r_0) = \text{span}\{r_0, Br_0, \ldots, B^{k-1}r_0\}. \tag{1.3}$$

Distinguishing different Krylov subspace methods is the approximation itself. Some methods enforce an optimality property, so that a quantity such as the error or residual decreases in some norm, or require only short-term recurrences, thereby minimizing the amount of work and storage at each iteration. If B is symmetric, there exist methods that have both an optimality condition and short-term recurrences. In contrast, methods for general nonsymmetric matrices can possess at most one of these properties. Krylov subspace methods are described in detail in, for example, [62, 118]. The behaviour of Krylov methods in finite precision arithmetic can differ from that in exact arithmetic. In particular, convergence of Krylov subspace methods could be delayed (see, for example, [101, 130]), although loss of orthogonality of the basis vectors with respect to an inner product may not cause significant difficulties for certain methods [35, 128]. A comprehensive overview of the effects of finite precision on Krylov subspace methods can be found in Zemke's thesis [155]. Throughout this thesis, we assume that exact arithmetic is used.

Preconditioning. When applied to (1.1), convergence of the chosen Krylov subspace method is often unacceptably slow. To remedy this, preconditioning is typically required, the aim of which is to replace the given linear system by one that is easier to solve. If the preconditioner $P \in \mathbb{R}^{n\times n}$ is applied to the left of (1.1), we solve the **left preconditioned** problem

$$P^{-1}Bx = P^{-1}b.$$

Two alternative approaches are to solve the **right preconditioned** problem

$$BP^{-1}y = b,$$

where $y = Px$, or the **split preconditioned** problem

$$P_1^{-1}BP_2^{-1}z = P_1^{-1}b,$$

where $z = P_2 x$. We focus here on left preconditioning in general, although we mention how certain results may be adapted for right preconditioning. When $P = P_1 P_2$ the three preconditioned coefficient matrices are similar. Insofar as convergence of the Krylov subspace method depends on the eigenvalues of the preconditioned coefficient matrix, the difference between left and right preconditioners is necessarily small. Although eigenvalues seem to determine convergence for certain problems (see, for example, [121]), in general convergence may depend on additional factors [15][118, Section 9.3].

Symmetric matrices. Both MINRES and SYMMLQ [107] are applicable when the coefficient matrix B is symmetric, although in this thesis we consider MINRES only. When B is additionally positive definite, the less expensive conjugate gradient method of Hestenes and Stiefel [71] is typically preferred. The popularity of the conjugate gradient method and MINRES can be attributed in part to their optimality properties and short-term recurrences. Another attraction is that the convergence of these methods is well understood; it typically depends on the distribution of the eigenvalues of B, although for certain right-hand side vectors convergence may be faster than is predicted by the eigenvalues.

As we have mentioned, it is often necessary to apply a preconditioner. If B is symmetric applying a conjugate gradient method or MINRES to the preconditioned linear system might be preferable to using a standard nonsymmetric method. The use of MINRES or the conjugate gradient method, however, restricts us to symmetric positive definite preconditioners even when effective preconditioners are symmetric indefinite or nonsymmetric.

Consider, for example, the symmetric saddle point problem

$$\mathcal{A}x = b, \tag{1.4}$$

where

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

$$x = \begin{bmatrix} u, & v \end{bmatrix}^T \quad \text{and} \quad b = \begin{bmatrix} f, & g \end{bmatrix}^T,$$

and $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite, $C \in \mathbb{R}^{m \times m}$ is symmetric positive semidefinite and $B \in \mathbb{R}^{m \times n}$ has full rank. The matrix $\mathcal{A}$ is symmetric indefinite and many effective preconditioners for this problem are block triangular (and, accordingly, nonsymmetric) [20, 34, 104, 127] or symmetric indefinite [9, 122, 157]. We would like to apply symmetric indefinite or nonsymmetric preconditioners to saddle

point problems (1.4) without losing the desirable properties of the conjugate gradient method and MINRES. For certain preconditioners this can be achieved by considering methods in nonstandard inner products, as we discuss below.

The Faber-Manteuffel Theorem and self-adjointness. In 1981 Gene Golub famously asked for which matrices it was possible to construct a conjugate gradient method, i.e., a method that has three-term recurrences and an optimality property. The more general characterization of the matrices for which it is possible to construct a method with an optimality property and recurrences of a fixed length was provided by Faber and Manteuffel in 1984 [49]. This result is connected to certain others pertaining to short-term recurrence methods [48, 90, 91]. A different proof of the Faber-Manteuffel theorem was recently provided [47].

It was shown by Faber and Manteuffel that three-term recurrence conjugate gradient methods exist for a matrix B if and only if its adjoint with respect to an inner product, $B^\star$, is a linear polynomial in B[1]. The simplest case is that $B^\star = B$, so that B is self-adjoint with respect to an inner product; much of this thesis concerns such self-adjoint matrices. We remark that a different three-term recurrence method is possible for matrices with at most four distinct eigenvalues, or whose eigenvalues are concyclic [11, 12, 88].

Returning to our preconditioned saddle point matrix (1.4), we see that if the preconditioned coefficient matrix is diagonalizable with real eigenvalues it is possible to apply a method that has an optimality property (in a nonstandard norm) and short-term recurrences. For a number of good preconditioners this inner product depends only on the blocks of $\mathcal{A}$ and the preconditioner, making it practical to use. Recently, it was shown that different preconditioners, for each of which the preconditioned saddle point matrix is self-adjoint, can be combined using combination preconditioning [135], which we describe in detail in Chapter 5. This increases further the number of preconditioned linear systems to which nonstandard Krylov subspace methods can be applied. In this thesis we aim to develop a greater understanding of when preconditioned saddle point matrices are self-adjoint with respect to an inner product and to propose new preconditioners such that the preconditioned coefficient matrix satisfies this property. We pay particular attention to combination preconditioning, which provides the tools to develop new, potentially better, preconditioners from existing ones.

[1]These matrices are precisely the diagonalizable matrices whose eigenvalues lie on a line in the complex plane.

Other factors for symmetric matrices. When the symmetric saddle point matrix involves discretizations of partial differential equations another motivation for nonstandard Krylov subspace methods is the notion of a "correct" norm with respect to which convergence should be measured [94, 153, 158]. This norm can be obtained by examining the partial differential equations themselves [94, 158] or by analysing the resulting discretizations [153]. Under certain circumstances, norms defined by certain block diagonal preconditioners, with blocks related to the blocks of $\mathcal{A}$, are appropriate. The idea of a correct norm is not explored further in this thesis, but a greater understanding of practical Krylov subspace methods in different inner products could prove useful in this context too.

Another reason for exploring Krylov subspace methods in nonstandard inner products is the understanding one can gain about the convergence of standard methods. For example, when the preconditioned biconjugate gradient algorithm, with a particular indefinite constraint preconditioner, is applied to the Stokes problem, insight into safeguarding strategies is gained by considering it as a conjugate gradient algorithm in a symmetric bilinear form (that is not an inner product) [115].

Nonsymmetric matrices. The difference between symmetric and nonsymmetric (or nonnormal[2]) matrices in the context of Krylov methods is striking. The first observation one can make regards the sheer number of existing methods. This makes it difficult to choose a method for a particular problem. Moreover, problems have been found for which each of CGN, GMRES [119], CGS [131], BiCG [51] and BiCGStab(ℓ) [129, 147] outperforms the others [83, 105]. Problems were also constructed for which IDR(s) [132] performed well or failed completely [83]. For none of the popular methods is convergence well understood for general nonnormal matrices. This makes choosing an effective preconditioner nontrivial.

Another difficulty is that methods for general nonsymmetric matrices do not have both short-term recurrences and optimality with respect to a well-understood inner product. Short-term recurrence methods like BiCG [51], IDR [132], and QMR [53]—and their variants—may minimize the error with respect to an inner product, but this inner product depends on the initial residual itself [10]. Unfortunately, whilst these methods may perform very well their convergence is not well understood in general, making it difficult to determine the properties a good preconditioner should possess.

[2]When referring to nonsymmetric matrices, we typically mean nonnormal nonsymmetric matrices, although the results usually apply in the normal case.

GMRES [119], on the other hand, minimizes the residual with respect to the Euclidean inner product. Convergence results exist for GMRES but they are not always as descriptive [44], or as easy to understand, as the convergence results for symmetric matrices[3]. For this reason, even in this case the determination of mathematical criteria for good preconditioners is difficult. Moreover, one must store and orthogonalize each new basis vector of the Krylov subspace against all previously computed Krylov subspace basis vectors, so that the work and storage required by GMRES grows at each step. An attempt to overcome this involves restarting the GMRES algorithm, taking the new initial guess to be the final iterate before the restart. However, if the restart occurs before enough information about the linear system has been gleaned by the Krylov subspaces, convergence may be very poor or complete stagnation may occur [62, page 77][78, 151]. We focus on full (unrestarted) GMRES in this thesis, since its convergence is better understood for general nonsymmetric matrices than that of other Krylov methods.

GMRES convergence. Let us discuss in greater detail the convergence of GMRES, since knowledge of the factors that affect its convergence would be helpful when choosing a good preconditioner. We recall that when B is symmetric (or, more generally, normal), convergence depends largely on its spectrum. Unfortunately, when B is nonnormal, this is not the case. Any nonincreasing convergence curve is possible for a coefficient matrix B, *regardless of its eigenvalues* [2, 64, 37]. Whether right-hand side vectors for practical problems result in poor convergence is not typically easy to ascertain, but bounds for GMRES must be able to account for this worst-case behaviour. Such bounds typically depend on the eigenvalues and eigenvectors [119], pseudospectra [143][145, Chapter 26] or the field of values [13, 38, 133][39, Corollary 6.2] of the coefficient matrix [44]. However, for certain matrices, each bound can be more descriptive than the others [44]. Other bounds certainly exist (see for example, [39, Theorem 6.1][148, 41, 76, 78, 87]), but these apply to a restricted set of matrices or are such that it can be hard to determine mathematical criteria for effective preconditioners from them.

Convection-diffusion equation. Convection-diffusion equations arise in many fields and their numerical solution has received much attention [103, 114, 152]. Although constant coefficient convection-diffusion operators on infinite domains are normal, the introduction of finite domains or variable coefficients introduce nonnormality [145, Chapter 12] that is inherited by the coefficient matrices of discretizations

[3]The exception is when B is normal for which, typically, convergence essentially depends on eigenvalues.

of them. Although we examine other test problems for nonsymmetric matrices, discretizations of the convection-diffusion equation

$$-\epsilon\nabla^2 u + w \cdot \nabla u = f \text{ in } \Omega$$

in one or two dimensions with suitable boundary conditions, wind w, viscosity ϵ and forcing term f appear throughout this thesis.

1.1 Aims and contributions

The primary goal of this thesis is to provide greater understanding of different ways in which nonstandard inner products are helpful in developing efficient preconditioned Krylov subspace solution methods. To this end we have:

- Introduced in Corollary 3.21 on page 61 an iteration dependent convergence bound for GMRES that can outperform the standard eigenvalue bound (3.27) on page 50. Figure 3.1 on page 65 shows that, significantly, it can capture superlinear behaviour of GMRES.

- In Section 4.3 derived the "Schöberl-Zulehner$^+$" (SZ$^+$) preconditioner for saddle point problems which was independently proposed by Krzyżanowski [85]. For this preconditioner it is possible to apply a MINRES method in a nonstandard inner product without the potentially costly exercise of scaling the preconditioner (see Theorem 4.15 on page 89). This is in contrast to established preconditioners of this sort, such as the Bramble-Pasciak (BP) [20] and Schöberl-Zulehner (SZ) [122, 157] preconditioners, for which scaling is typically required. This robustness sets the SZ$^+$ preconditioner apart from other similar preconditioners, as can be seen from Tables 4.4 and 4.6 and the discussion on scaling the SZ preconditioner on page 99. Table 4.4 additionally shows that the SZ$^+$ preconditioner can be more efficient than the Bramble-Pasciak$^+$ (BP$^+$) [135] preconditioner, particularly for difficult problems.

- In Section 5.1 identified a new combination preconditioner for symmetric saddle point problems, the Bramble-Pasciak$^+$-Block diagonal (BP$^+$-BDW) combination preconditioner, for which the BP$^+$-BDW-preconditioned MINRES method outperforms the better of the BP$^+$ and block diagonal (BDW) preconditioners by up to 34%, and by 19.8% on average, across a range of test problems (see Table 5.2 on page 112). Table 5.3 on page 114 also shows that the BP$^+$-BDW

combination conjugate gradient method also outperforms the BP$^+$-MINRES and BDW-MINRES methods by up to 31% and by 20.3% on average.

The BP$^+$ and BDW preconditioned saddle point matrices are each self-adjoint but indefinite with respect to a nonstandard inner product and the preconditioned system cannot, therefore, be reliably solved by a conjugate gradient method. However, our combination *can* be made positive definite with respect to a nonstandard inner product (see Theorem 5.5) so that a conjugate gradient method can be used. The lower computational cost of the conjugate gradient method makes this advantageous.

- In Section 5.2 derived the Bramble-Pasciak-Block diagonal (BP-BDW) combination preconditioner for symmetric saddle point problems, which outperforms the BP and block diagonal (BDW) preconditioners by up to 11% and by 4.3% on average, across a range of test problems (see Table 5.4 on page 118).

- In Section 5.3 proposed a new formulation of a Bramble-Pasciak-Schöberl-Zulehner (BP-SZ) combination preconditioner for symmetric saddle point problems. This outperforms the better of the BP and SZ preconditioners for a nonstandard preconditioned conjugate gradient method by up to 33%, and by 20.2% on average across a range of test problems, as shown in Table 5.5 on page 126. Since the combination parameters appear in our symmetric bilinear form (described by (5.32) on page 120), ensuring that this symmetric bilinear form is an inner product depends less on the scaling of the matrix A_0 (which may be costly) and more on choosing α and β than does the configuration in [135].

- In Theorem 6.1 on 132 presented a condition for pseudospectra of a preconditioner and coefficient matrix (or, more generally, two matrices) to be identical. In Theorem 6.6 on page 135 we describe the pseudospectra of matrices for which the condition is only approximately satisfied. The remarks proceeding Theorem 6.1 discusses some consequences for preconditioners.

- In Theorem 6.10 on page 152 quantified the dependence of the convergence of a Krylov subspace method in a nonstandard inner product on the eigenvalues of the preconditioned system.

1.2 Dissertation outline

For completeness, an introduction to nonstandard symmetric bilinear forms, inner products and norms is given in Chapter 2. Much of the chapter is devoted to the self-adjointness of a matrix with respect to a symmetric bilinear form or inner product.

Chapter 3 generalizes the conjugate gradient method, MINRES and GMRES for use with nonstandard inner products. We present algorithms for these methods, show how preconditioners can be incorporated and provide convergence results. Relationships between the convergence of minimum residual methods in different inner products are derived. These lead to an iteration-dependent convergence bound for standard GMRES that can be much sharper than the standard eigenvalue bound and can capture superlinear behaviour. This bound is new, as are the relationships between the convergence of different minimal residual methods. Lemma 3.18 is a generalization of a result in [45].

Saddle point systems are the subject of Chapters 4 and 5. In the former, we characterize the symmetric bilinear forms with respect to which a fairly general preconditioned saddle point system is self-adjoint and use these conditions to obtain symmetric bilinear forms for certain preconditioners. The SZ^+ preconditioner is introduced and its properties investigated. Numerical results show that it is more robust than a number of other preconditioners, particularly for difficult problems. The new contributions are the description of the eigenvalues of a Krzyżanowski preconditioned saddle point matrix, conditions on symmetric bilinear forms and the description of the properties of the SZ^+ preconditioner. The SZ^+ preconditioner was derived independently by Krzyżanowski [85].

Combination preconditioners for symmetric saddle point matrices are addressed in Chapter 5. We present a new formulation of the BP-SZ combination preconditioner introduced in [135], and propose two new preconditioners, one of which combines two preconditioners for which a conjugate gradient method cannot be applied to create one for which it can. Requirements for the symmetric bilinear forms associated with each preconditioner to be an inner product, and for the preconditioned coefficient matrix to be self-adjoint with respect to this inner product, are outlined and we present numerical experiments that confirm their competitiveness. All the results in this chapter are novel, with the exception of the original formulation of the BP-SZ combination preconditioner that appeared in [135].

In Chapter 6 we explore preconditioners for nonsymmetric matrices in light of non-standard inner products. Relationships are found between the quality of the approximation of the eigenvectors of the preconditioner and the closeness of the symmetric bilinear forms with respect to which the preconditioner and coefficient matrix are self-adjoint. The closeness of the pseudospectra of the preconditioner and coefficient matrix is also discussed and a condition for these pseudospectra to be identical is derived. We show that certain effective preconditioners are self-adjoint with respect to the same symmetric bilinear forms as the coefficient matrix or one which is nearby. These results, to our knowledge, are novel.

Chapter 2

Symmetric bilinear forms, inner products and self-adjointness

In this chapter we examine symmetric bilinear forms and inner products, both of which feature prominently in subsequent chapters. While this thesis focuses on real matrices and vectors, for which symmetric bilinear forms are appropriate, the generalization of the results of this chapter to Hermitian sesquilinear forms and inner products on $\mathbb{C}^n$ typically only requires minor modifications; these are outlined where appropriate. We begin in Section 2.1 by defining a symmetric bilinear form and inner product and by introducing some of the concepts and tools required in later sections and chapters. Section 2.2 concentrates on self-adjoint matrices. In particular, the symmetric bilinear forms with respect to which a given matrix is self-adjoint are characterized, the computation of the matrices that define these forms is discussed and some of the properties of self-adjoint matrices, and of matrices that are positive definite with respect to an inner product, are presented. We find in Section 2.3 an inner product with respect to which a finite difference matrix for a convection-diffusion equation is self-adjoint and show that it is related to the inner product with respect to which the original differential operator is self-adjoint. This discretization, and variations of it, appear in later chapters. A summary of the key findings of this chapter is given in Section 2.4.

2.1 Properties of symmetric bilinear forms

Nonstandard symmetric bilinear forms, inner products and norms are generalizations of the familiar Euclidean inner product and norm. Many properties and concepts as-

sociated with the Euclidean inner product space transfer simply to these nonstandard Hilbert spaces.

Definition 2.1. A nondegenerate **symmetric bilinear form** $\langle \cdot, \cdot \rangle_W$ on $\mathbb{R}^n$ satisfies

$$\langle x, y \rangle_W = y^T W x \tag{2.1}$$

for all $x, y \in \mathbb{R}^n$ and an invertible symmetric matrix $W \in \mathbb{R}^{n \times n}$.

Definition 2.2. An **inner product** $\langle \cdot, \cdot \rangle_W$ on $\mathbb{R}^n$ is a symmetric bilinear form for which $W \in \mathbb{R}^{n \times n}$ is symmetric positive definite.

In the case that (2.1) is an inner product we can associate with W a vector norm and matrix norm.

Definition 2.3. If $\langle \cdot, \cdot \rangle_W$ is an inner product the associated ***W*-(vector) norm** for $x \in \mathbb{R}^n$ is

$$\|x\|_W = \sqrt{\langle x, x \rangle_W}.$$

Definition 2.4. If $\langle \cdot, \cdot \rangle_W$ is an inner product the associated ***W*-matrix norm** for $B \in \mathbb{R}^{n \times n}$ is

$$||B||_W = \max_{x \neq 0} \frac{\|Bx\|_W}{\|x\|_W}.$$

The most familiar inner product is the Euclidean inner product, here denoted the I-inner product for consistency, with its induced spectral norm $\|\cdot\|_I$. Any W-matrix-norm can be related to the I-norm since, for any factorization $W = S^T S$ with S invertible,

$$\|B\|_W = \|SBS^{-1}\|_I. \tag{2.2}$$

The W-matrix norm can be used to define the W-condition number of a matrix.

Definition 2.5. The ***W*-condition number** of an invertible matrix $B \in \mathbb{R}^{n \times n}$ is

$$\kappa_W(B) = \|B\|_W \|B^{-1}\|_W. \tag{2.3}$$

Given a symmetric bilinear form $\langle \cdot, \cdot \rangle_W$, the W-adjoint of B is defined as follows.

Definition 2.6. The ***W*-adjoint** of $B \in \mathbb{R}^{n \times n}$ is defined to be the unique matrix $B^\star \in \mathbb{R}^{n \times n}$ that satisfies

$$\langle Bx, y \rangle_W = \langle x, B^\star y \rangle_W \quad \forall x, y \in \mathbb{R}^n.$$

Definition 2.6 implies that $y^T W B x = (B^\star y)^T W x = y^T (B^\star)^T W x$ so that $WB = (B^\star)^T W$ or

$$B^\star = W^{-1} B^T W. \tag{2.4}$$

Since $B^\star$ and B^T are similar the eigenvalues of $B^\star$ and B coincide. Furthermore, for any two matrices B and C in $\mathbb{R}^{n\times n}$ and scalar α in $\mathbb{R}$ it is straightforward to show that

$$(B+C)^\star = B^\star + C^\star, \;\; (BC)^\star = C^\star B^\star, \;\; (B^{-1})^\star = (B^\star)^{-1} \quad \text{and} \quad (\alpha B)^\star = \alpha B. \tag{2.5}$$

Generalizations of symmetric, skew-symmetric, normal and positive definite matrices also exist.

Definition 2.7. A matrix $B \in \mathbb{R}^{n\times n}$ is ***W*-self-adjoint** if and only if

$$B^\star = W^{-1} B^T W = B.$$

Definition 2.8. A matrix $B \in \mathbb{R}^{n\times n}$ is ***W*-skew-adjoint** if and only if

$$B^\star = W^{-1} B^T W = -B.$$

Definition 2.9. A matrix $B \in \mathbb{R}^{n\times n}$ is ***W*-positive definite** if and only if

$$\langle Bx, x\rangle_W > 0 \;\; \text{for all} \;\; x \neq 0.$$

Definition 2.10. A matrix $B \in \mathbb{R}^{n\times n}$ is ***W*-normal** if and only if

$$BB^\star = B^\star B,$$

where $B^\star = W^{-1} B^T W$.

An important subset of W-normal matrices is the set of ***W*-polynomially-normal matrices**, for which $B^\star$ can be expressed as a polynomial in B and to which the W-self-adjoint and W-skew-adjoint matrices evidently belong. The symmetric bilinear forms with respect to which a matrix is polynomially-normal are characterized in [97].

The definition of orthogonal matrices can also be generalized. We utilize the following definition and note that this differs from the definition in, for example, Gohberg, Lancaster and Rodman [57].

Definition 2.11. A matrix $Z \in \mathbb{R}^{n\times n}$ is ***W*-orthogonal** if and only if

$$Z^T W Z = I. \tag{2.6}$$

It is well known that the eigenvector matrix U of an I-normal matrix can be chosen to be unitary, so that $U^*U = I$. More generally, if $B \in \mathbb{R}^{n\times n}$ is normal with respect to an inner product $\langle\cdot,\cdot\rangle_W$, the eigenvector matrix of B can be chosen so that $Z^*WZ = I$, i.e., so that Z is W-unitary. The proof is similar to that for polynomially-normal matrices in [91, Theorem 3.1].

Lemma 2.12. *Let $B \in \mathbb{R}^{n\times n}$ be normal with respect to the inner product $\langle\cdot,\cdot\rangle_W$. Then B is diagonalizable and its matrix of eigenvectors $Z \in \mathbb{C}^{n\times n}$ can be chosen so that $Z^*WZ = I$.*

Proof. If B is W-normal, then $BB^\star = B^\star B$, where $B^\star = W^{-1}B^TW$. Substituting for the W-adjoint gives that

$$BW^{-1}B^TW = W^{-1}B^TWB. \tag{2.7}$$

Now, since $W \in \mathbb{R}^{n\times n}$ defines an inner product it is symmetric positive definite and can be factored as $W = W^{\frac{1}{2}}W^{\frac{1}{2}}$. Premultiplying (2.7) by $W^{\frac{1}{2}}$ and postmultiplying by $W^{-\frac{1}{2}}$ gives that

$$W^{\frac{1}{2}}BW^{-1}B^TW^{\frac{1}{2}} = W^{-\frac{1}{2}}B^TWBW^{-\frac{1}{2}}$$

or, equivalently, that $CC^T = C^TC$, where $C = W^{\frac{1}{2}}BW^{-\frac{1}{2}}$. Thus, C is I-normal and has a unitary diagonalization $C = W^{\frac{1}{2}}BW^{-\frac{1}{2}} = U\Lambda U^*$, where $\Lambda, U \in \mathbb{C}^{n\times n}$, $U^*U = I$. It follows that $B = W^{-\frac{1}{2}}U\Lambda U^*W^{\frac{1}{2}}$, which proves that B is diagonalizable. If $Z = W^{-\frac{1}{2}}U$, then $Z^*WZ = U^*W^{-\frac{1}{2}}WW^{-\frac{1}{2}}U = U^*U = I$. □

With the concept of orthogonality it is possible to consider factorizations such as the polar decomposition or singular value decomposition in nonstandard symmetric bilinear forms. The ***W*-polar decomposition** is discussed in [74], where conditions on existence and uniqueness are given and an algorithm for its computation is presented. The singular value decomposition (SVD) can also be generalized, although we restrict our attention here, as in [149], to SVDs defined by inner products. This decomposition requires a notion of generalized singular values[1].

Definition 2.13. Let $W \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then, the ***W*-singular values** of a matrix $B \in \mathbb{R}^{n\times n}$ are given by

$$\sigma^W(B) = \left\{\sigma | \sigma \geq 0, \sigma \text{ a stationary value of } \frac{\|Bx\|_W}{\|x\|_W}\right\}. \tag{2.8}$$

[1]Van Loan [149] considered more general $\{M, W\}$-singular values. These are beyond the scope of this work but the results for these singular values are similarly obtained.

These generalized singular values, analogously to standard singular values, can be written in terms of eigenvalues.

Lemma 2.14. *Let $W \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then, the W-singular values of $B \in \mathbb{R}^{n\times n}$ are*

$$\sigma_i^W(B) = \sqrt{\lambda_i(W^{-1}B^TWB)} = \sqrt{\lambda_i(B^\star B)},\ i = 1,\dots,n. \tag{2.9}$$

Furthermore, if B is W-self-adjoint, its W-singular values are

$$\sigma_i^W(B) = |\lambda_i(B)|,\ i = 1,\dots,n,$$

where $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$.

Proof. From Definition 2.3 we deduce that any member of $\sigma^W(B)$ must be a stationary value of x^TB^TWBx, subject to $x^TWx = 1$, with corresponding stationary point $x \in \mathbb{R}^n$. By employing Lagrange multipliers, we find that x must also be a stationary point of

$$x^TB^TWBx - \lambda(x^TWx - 1), \tag{2.10}$$

the gradient of which is

$$\nabla_x(x^TB^TWBx - \lambda(x^TWx - 1)) = 2B^TWBx - 2\lambda Wx.$$

Clearly, this gradient is zero when

$$B^TWBx = \lambda Wx.$$

Consequently, a stationary point of (2.10) is an eigenvector of $W^{-1}B^TWB$ and the corresponding W-singular value is $\sqrt{\lambda(W^{-1}B^TWB)}$. Recall from Definition 2.6 that $B^\star = W^{-1}B^TW$. Thus,

$$\sigma^W(B) = \sqrt{\lambda(W^{-1}B^TWB)} = \sqrt{\lambda(B^\star B)}.$$

It follows that when B is W-self-adjoint, so that $B = B^\star$ (see Definition 2.7), the W-singular values of B are the absolute values of the eigenvalues of B. □

Given this definition of singular values, the ***W*-SVD** [149] is the following decomposition.

Definition 2.15. Let $W \in \mathbb{R}^{n\times n}$ be symmetric positive definite. The ***W*-SVD** of a matrix $B \in \mathbb{R}^{n\times n}$ is

$$B = S\Sigma T^{-1}, \tag{2.11}$$

where $S \in \mathbb{R}^{n\times n}$ and $T \in \mathbb{R}^{n\times n}$ are W-orthogonal and $\Sigma \in \mathbb{R}^{n\times n}$ is a diagonal matrix that has as its diagonal entries the W-singular values $\sigma_1^W \geq \sigma_2^W \geq \ldots \geq \sigma_n^W \geq 0$, with $\sigma_i^W \in \sigma^W(B)$, $i = 1, \ldots, n$.

As in the case of a standard polar decomposition, a W-polar decomposition can be constructed from a W-SVD.

The generalized singular values of a matrix B can be used to define its W-norm and W-condition number, similarly to the standard Euclidean case, as the following lemma shows.

Lemma 2.16. *Let $W \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Additionally, let B be nonsingular and have W-singular values $\sigma_1^W \geq \ldots \geq \sigma_n^W > 0$. Then*

$$\|B\|_W = \sigma_1^W \quad \textit{and} \quad \|B^{-1}\|_W = \frac{1}{\sigma_n^W(B)}.$$

The W-condition number of B is given by

$$\kappa_W(B) = \sqrt{\frac{\lambda_{max}(B^\star B)}{\lambda_{min}(B^\star B)}}$$

which, when B is W-self-adjoint, becomes

$$\kappa_W(B) = \frac{\max_{i=1,\ldots,n} |\lambda_i(B)|}{\min_{i=1,\ldots,n} |\lambda_i(B)|}. \tag{2.12}$$

Proof. The first part follows from Definition 2.4, since

$$\|B\|_W = \max_{x\neq 0} \frac{\|Bx\|_W}{\|x\|_W} = \max_{i=1,\ldots,n} \sigma_i^W(B) = \sigma_1^W(B).$$

Similarly,

$$\|B^{-1}\|_W = \max_{x\neq 0} \frac{\|B^{-1}x\|_W}{\|x\|_W} = \max_{y=B^{-1}x\neq 0} \frac{\|y\|_W}{\|By\|_W} = \frac{1}{\min_{y\neq 0} \frac{\|By\|_W}{\|y\|_W}} = \frac{1}{\sigma_n^W(B)}.$$

The W-condition number is, by definition, given by (2.3). It follows from (2.9) that

$$\kappa_W(B) = \|B\|_W\|B^{-1}\|_W = \frac{\sigma_1^W(B)}{\sigma_n^W(B)} = \sqrt{\frac{\lambda_{max}(B^\star B)}{\lambda_{min}(B^\star B)}}.$$

When B is W-self-adjoint, $B = B^\star$, which gives the simplification. □

Remark 2.17. More generally, (2.12) holds for any W-normal matrix B [5, Corollary 4.2].

In this section, attention was restricted to symmetric bilinear forms and inner products on $\mathbb{R}^n$. **Hermitian sesquilinear forms** on $\mathbb{C}^n$, in contrast, are related to an invertible Hermitian matrix $W \in \mathbb{C}^{n\times n}$ by

$$\langle x, y\rangle_W = y^* W x,$$

for all $x, y \in \mathbb{C}^n$, where y^* represents the conjugate transpose. When W is additionally positive definite, it induces an inner product on $\mathbb{C}^n$. Most of the properties of this section follow straightforwardly to the complex case. An exception is that in (2.5),

$$(\alpha B)^\star = \overline{\alpha} B^\star,$$

where $\overline{\alpha}$ represents the complex conjugate of α. The generalization of a W-orthogonal matrix is a ***W*-unitary matrix** $Z \in \mathbb{C}^{n\times n}$ for which

$$Z^* W Z = I.$$

A characterization of the Hermitian sesquilinear forms with respect to which matrices are polynomially normal is given in [98].

2.2 Self-adjointness

The condition for self-adjointness with respect to $\langle\cdot,\cdot\rangle_W$ in Definition 2.7 is that $B^\star = B$. By (2.4) this is equivalent to

$$WB = B^T W, \tag{2.13}$$

so that WB is symmetric. It is clear from (2.13) that a matrix is self-adjoint with respect to some symmetric bilinear form if and only if B and B^T are similar. Since every real matrix is similar to its transpose, every real matrix is self-adjoint with respect to some symmetric bilinear form.

Self-adjointness is closely related to the factorization of a matrix as the product of two symmetric matrices. In [138] Taussky states that every matrix $B \in \mathbb{R}^{n\times n}$ can be factored as

$$B = MT,$$

where $M, T \in \mathbb{R}^{n\times n}$ are symmetric. Provided B is nonsingular, M and T are nonsingular. Upon rearrangement of this equation we find that

$$M^{-1}B = T \tag{2.14}$$

which, since T is symmetric, shows that B is M^{-1}-self-adjoint. On the other hand, the symmetry of M gives that

$$BT^{-1} = M = T^{-1}B^T \tag{2.15}$$

which implies that B^T is T^{-1}-self-adjoint.

Equations (2.14) and (2.15) indicate that there is a connection between the symmetric bilinear forms with respect to which B and B^T are self-adjoint. This link is made precise in the following lemma.

Lemma 2.18. *$B \in \mathbb{R}^{n\times n}$ is self-adjoint with respect to the nondegenerate symmetric bilinear form $\langle\cdot,\cdot\rangle_W$ if and only if B^T is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{W^{-1}}$ and, provided B is invertible, if and only if B^{-1} is self-adjoint with respect to $\langle\cdot,\cdot\rangle_W$.*

Proof. Since $\langle\cdot,\cdot\rangle_W$ is a nondegenerate symmetric bilinear form, W is invertible (see Definition 2.1). To show that B^T is W^{-1}-self-adjoint, we multiply both sides of (2.13) by W^{-1}. This gives

$$W^{-1}B^T = BW^{-1}.$$

Comparison with (2.13) shows that B^T and W^{-1} satisfy the self-adjointness relationship and so B^T is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{W^{-1}}$. Premultiplying (2.13) by B^{-T} and postmultiplying by B^{-1} gives that

$$B^{-T}W = WB^{-1},$$

from which we conclude that B^{-1} is W-self-adjoint. In each case the reverse direction is similar. □

We remark that, given any matrix B and its W-adjoint $B^\star$, the matrix $B^\star B$ is always W-self-adjoint since $WB^\star B = B^T WB$, which is symmetric. The generalization of this result, and of Lemma 2.18, to Hermitian sesquilinear forms is straightforward. However, in contrast to real matrices, not every complex matrix is self-adjoint with respect to a Hermitian sesquilinear form as we shall see in the following section.

2.2.1 Characterizing the symmetric bilinear forms with respect to which a matrix is self-adjoint

We now focus on characterizing the symmetric bilinear forms with respect to which a given matrix is self-adjoint; this will be crucial in the following chapters. Throughout this section we will assume without loss of generality that the eigenvalues $\lambda_1, \ldots, \lambda_n$ of B are ordered so that $\lambda_1, \ldots, \lambda_j$ are real and $\lambda_{j+1} = \mu_j + i\nu_j, \overline{\lambda}_{j+1} = \mu_{j+1} - i\nu_{j+1}, \ldots, \lambda_k = \mu_k + i\nu_k, \overline{\lambda}_k = \mu_k - i\nu_k$ are nonreal, with $\lambda_{j+1}, \ldots, \lambda_k$ in the upper half-plane.

It is not difficult to show that the Jordan block of a real eigenvalue λ, $J(\lambda) \in \mathbb{R}^{n\times n}$,

$$J(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix},$$

is self-adjoint with respect to indefinite symmetric bilinear forms induced by Hankel matrices

$$W = \begin{bmatrix} & & & h_1 \\ & & \iddots & h_2 \\ & \iddots & \iddots & \vdots \\ h_1 & h_2 & \ldots & h_n \end{bmatrix}, \tag{2.16}$$

where $h_1, \ldots, h_n \in \mathbb{R}$, $h_1 \neq 0$. More generally, any persymmetric matrix[2] is self-adjoint with respect to the Hankel matrix

$$W = \alpha \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \tag{2.17}$$

where $\alpha \in \mathbb{R}$, $\alpha \neq 0$. Possibly the most common persymmetric matrix is the Toeplitz matrix

$$T = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \ldots & \ldots & a_{-n+1} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \ldots & \ldots & a_2 & a_1 & a_0 \end{bmatrix}. \tag{2.18}$$

[2] A persymmetric matrix is symmetric about its north-east to south-west diagonal.

When $\alpha = 1$, $W = \hat{Y}$ is known variously as the flip matrix, the reverse identity or the standard involutory permutation (sip) matrix.

Definition 2.19. The **reverse identity matrix** of dimension n is

$$\hat{Y}_n = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \tag{2.19}$$

where we drop the subscript n if the dimension is clear from the context.

The symmetric bilinear forms with respect to which Jordan blocks are self-adjoint can be used to construct symmetric bilinear forms with respect to which more general matrices are self-adjoint. We start with real diagonalizable matrices (possibly with complex conjugate eigenvalues).

Theorem 2.20. (adapted from [57, Theorem 6.1.5]) *Let $B \in \mathbb{R}^{n\times n}$ be diagonalizable with real Jordan form $B = R\mathcal{J}R^{-1}$, as defined in Appendix* A.1. *Then B is self-adjoint with respect to symmetric bilinear forms $\langle\cdot,\cdot\rangle_W$, where*

$$W = R^{-T}YR^{-1}, \tag{2.20}$$

$$Y = \operatorname{diag}\left(\epsilon_1, \ldots, \epsilon_j, \hat{Y}_2, \ldots, \hat{Y}_2\right) \tag{2.21}$$

with $\epsilon_i = \pm 1$, $i = 1, \ldots, j$ and where $\hat{Y}_2$ are 2×2 reverse identity matrices (2.19).

Conversely, if for some set of signs $\{\epsilon_1, \ldots, \epsilon_j\}$, the pairs (B, W) and (J, Y) are r-unitarily similar[3]*, then B is W-self-adjoint.*

The case that B has real eigenvalues only has been explored in more detail in, for example, [135, Lemma 3.8]. We begin with a well known result

Corollary 2.21. *A matrix $B \in \mathbb{R}^{n\times n}$ is self-adjoint with respect to some inner product $\langle\cdot,\cdot\rangle_W$ if and only if B is diagonalizable with real eigenvalues.*

Proof. See [139]. □

[3]Given the W_1-self-adjoint matrix B_1 and W_2-self-adjoint matrix B_2, where W_1, W_2, B_1 and B_2 are real $n \times n$ matrices, B_1 and B_2 are r-unitarily similar if $B_1 = S^{-1}B_2S$ and $W_1 = S^TW_2S$ for some real invertible matrix S [57, Section 6.1].

Remark 2.22. When B is diagonalizable with real eigenvalues, the eigenvectors of B can be taken to be real and the representation of W in (2.20) is related to that given in [135, Lemma 3.8] in which $W = Z^{-T}\Theta Z^{-1}$ with Θ a real diagonal matrix. To see this, first note that $\Theta = |\Theta|^{\frac{1}{2}}Y|\Theta|^{\frac{1}{2}}$, where $Y = \text{sign}(\Theta)$ and $|\Theta|$ has as its entries the absolute values of the corresponding entries in Θ. Then, the representation $W = Z^{-T}\Theta Z^{-1}$ is equivalent to $W = (Z|\Theta|^{\frac{1}{2}})^{-T}Y(Z|\Theta|^{\frac{1}{2}})^{-1}$ and Θ serves only to scale the columns of Z. Moreover, whenever B has diagonalization $B = Z\Lambda Z^{-1}$, since diagonal matrices commute $B = Z|\Theta|^{\frac{1}{2}}\Lambda(Z|\Theta|^{\frac{1}{2}})^{-1}$ is a second diagonalization. Without loss of generality we take $\Theta = I$ in the following.

Remark 2.23. If B is diagonalizable with real eigenvalues, the inner product $\langle\cdot,\cdot\rangle_W$, where $W = Z^{-*}Z^{-1}$, is invariant under permutation of the columns of Z. To see this, let P be a matrix that permutes these columns. Then $W = Z^{-*}Z^{-1} = Z^{-*}PP^TZ^{-1} = (ZP)^{-*}(ZP)^{-1}$.

Remark 2.24. Whenever B has real eigenvalues there is some freedom to choose the inertia of W by choosing the signs of ϵ_i, $i = 1, \ldots j$. However, the inertia of each $Y_i \in \mathbb{R}^{p\times p}$ associated with a complex conjugate pair of eigenvalues is fixed: it has one eigenvalue at 1 and the other at -1. In the context of the preconditioned Krylov subspace methods discussed in subsequent chapters it might be useful to exploit this freedom. For simplicity, we tend to choose $\epsilon_i = 1$, $i = 1\ldots, j$ so that

$$Y = \text{diag}(1, \ldots, 1, \hat{Y}_2, \ldots, \hat{Y}_2), \tag{2.22}$$

which renders W as near to positive definite as possible.

Theorem 2.20 characterizes all the symmetric bilinear forms with respect to which a given real diagonalizable matrix is self-adjoint. The symmetric bilinear forms with respect to which a nondiagonalizable matrix is self-adjoint are similarly defined.

Theorem 2.25. (adapted from [57, 6.1.5]) *Let $B, W \in \mathbb{R}^{n\times n}$ with $W = W^T$, W invertible and let B be W-self-adjoint. Then there is a real invertible $R \in \mathbb{R}^{n\times n}$ such that $B = R\mathcal{J}R^{-1}$ and $W = R^{-T}Y_{\epsilon,J}R^{-1}$, where*

$$\mathcal{J} = \text{diag}(J(\lambda_1), \ldots, J(\lambda_j), \mathcal{J}(\mu_{j+1} \pm i\nu_{j+1}), \ldots, \mathcal{J}(\mu_k \pm i\nu_k))$$

is a real Jordan normal form of B (defined in Appendix A.1*) and*

$$Y_{\epsilon,J} = \text{diag}(\epsilon_1\hat{Y}^{(1)}, \ldots, \epsilon_j\hat{Y}^{(j)}, \hat{Y}^{(j+1)}, \ldots, \hat{Y}^{(k)}),$$

where $\hat{Y}^{(i)}$, $i = 1, \ldots, k$ *are reverse identity matrices* (2.19) *with the size of each equal to that of the corresponding Jordan block in* $\mathcal{J}$ *and* $\epsilon = \{\epsilon_1, \ldots, \epsilon_j\}$ *is an ordered set of signs* ± 1*. The set* ϵ *is uniquely determined by* (B, W) *up to permutation of signs corresponding to equal Jordan blocks.*

Conversely, if for some set of signs ϵ*, the pairs* (B, W) *and* $(J, Y_{\epsilon,J})$ *are r-unitarily similar, then* B *is* W*-self-adjoint.*

This is a straightforward generalization of Theorem 2.20. In the extreme case that B is a Jordan block, W has as many positive eigenvalues as negative eigenvalues (or one more positive eigenvalue when the dimension is odd) and certainly does not define an inner product. However, the matrices B with which we will primarily be concerned will be diagonalizable and will, therefore, have symmetric bilinear forms $\langle \cdot, \cdot \rangle_W$ defined by (2.16) whose inertia depends on the number of complex conjugate pairs of eigenvalues of B; the greater this number, the further from definite the symmetric bilinear form.

Finally, we discuss the most general case, that $B \in \mathbb{C}^{n \times n}$ is self-adjoint with respect to a Hermitian sesquilinear form. A detailed discussion of this topic can be found in [57] but we summarize the key results here. The first result details properties of the spectrum of a W-self-adjoint matrix B.

Proposition 2.26. [57, Proposition 4.2.3] *The spectrum* $\Lambda(B)$ *of a* W*-self-adjoint matrix* B *is symmetric relative to the real axis, i.e.,* $\lambda_0 \in \Lambda(B)$ *implies* $\overline{\lambda}_0 \in \Lambda(B)$*. Moreover, in the Jordan normal form of* B*, the sizes of the Jordan blocks with eigenvalue* λ_0 *are equal to the sizes of Jordan blocks with eigenvalue* $\overline{\lambda}_0$*.*

Proposition 2.26 implies that only those matrices that are similar to their conjugate transposes are self-adjoint with respect to a Hermitian sesquilinear form.

As a consequence of Proposition 2.26, without loss of generality we can assume that

$$J = \operatorname{diag}(J(\lambda_1), \ldots J(\lambda_j), J(\lambda_{j+1}), \cdots, J(\lambda_k)),$$

where $\lambda_1, \ldots, \lambda_j$ are the real eigenvalues of B, $\lambda_{j+1}, \ldots, \lambda_k$ are the nonreal eigenvalues in the upper half-plane and $J(\lambda_i)$, $i = j+1, \ldots, k$ is a direct sum of the Jordan blocks corresponding to λ_i and $\overline{\lambda}_i$.

The following theorem characterizes the Hermitian sesquilinear forms with respect to which B is self-adjoint.

Theorem 2.27. (adapted from [57, Theorem 5.1.1]) *Let B, $W \in \mathbb{C}^{n\times n}$ and let B be W-self-adjoint. Then there is an invertible $Z \in \mathbb{C}^{n\times n}$ such that $B = ZJZ^{-1}$ and*

$$W = Z^{-*}Y_{\epsilon,J}Z^{-1}, \tag{2.23}$$

where

$$J = \operatorname{diag}(J(\lambda_1), \dots J(\lambda_j), J(\lambda_{j+1}), \cdots, J(\lambda_k))$$

is a Jordan normal form for B, and

$$Y_{\epsilon,J} = \operatorname{diag}(\epsilon_1\hat{Y}_1, \dots, \epsilon_j\hat{Y}_j, \hat{Y}_{j+1}, \dots, \hat{Y}_k),$$

where $\hat{Y}_i$, $i = 1, \dots, k$ are reverse identity matrices (2.19) *with the sizes of $J(\lambda_i)$, $i = 1, \dots, k$ respectively, and $\epsilon = \{\epsilon_1, \dots, \epsilon_j\}$ is an ordered set of signs ± 1. The set ϵ is uniquely determined by (B, W) up to permutation of signs corresponding to repeated Jordan blocks.*

Conversely, if for some set of signs ϵ, the pairs (B, W) and $(J, Y_{\epsilon,J})$ are unitarily similar[4]*, then B is W-self-adjoint.*

This general case is interesting in its own right, although it is beyond this thesis to examine complex, nondiagonalizable matrices. Significantly, however, Theorem 2.27 can be used to show that the symmetric bilinear forms with respect to which a real diagonalizable matrix B is self-adjoint can be written in terms of the (possibly complex) diagonalization, as long as we are willing to use complex arithmetic during an intermediate step.

Corollary 2.28. *Let $B \in \mathbb{R}^{n\times n}$ have diagonalization $B = Z\Lambda Z^{-1}$, $\Lambda, Z \in \mathbb{C}^{n\times n}$. Then B is self-adjoint with respect to $\langle\cdot,\cdot\rangle_W$, where $W = Z^{-*}YZ^{-1}$ is real and Y is of the form* (2.22)*. Additionally, $M = Z^{-*}Z^{-1}$ is real.*

Proof. Let B have j real eigenvalues $\lambda_1, \dots, \lambda_j$ and $2(k-j)$ nonreal eigenvalues $\lambda_{j+1}, \overline{\lambda}_{j+1}, \dots, \lambda_k, \overline{\lambda}_k$. We assume without loss of generality that

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_j, \Lambda_{j+1}, \dots, \Lambda_k),$$

where

$$\Lambda_i = \begin{bmatrix} \lambda_i & 0 \\ 0 & \overline{\lambda}_i \end{bmatrix}, \ i = j+1, \dots, k,$$

[4]If B_1 is W_1-self-adjoint and B_2 is W_2-self-adjoint then we say that B_1 and B_2 are **unitarily similar** if $B_1 = T^{-1}B_2T$ with $W_1 = T^*W_2T$ [57, Section 4.5].

contains a pair of complex conjugate eigenvalues. Then, the corresponding matrix of eigenvectors can be written as

$$Z = \left[z_1, \ldots, z_j, z_{j+1}, \overline{z}_{j+1}, \ldots, z_k, \overline{z}_k\right].$$

Now, from Theorem 2.27, $W = Z^* Y Z^{-1}$ defines a Hermitian sesquilinear form with respect to which B is self-adjoint. We now show that, in fact, W is real and symmetric and, therefore, defines a symmetric bilinear form. To begin, note that

$$ZY = \left[z_1, \ldots, z_j, \overline{z}_{j+1}, z_{j+1}, \ldots, \overline{z}_k, z_k\right],$$

so that the effect of Y is to swap the order of the two eigenvectors corresponding to a complex conjugate pair of eigenvalues and to leave unchanged the order of eigenvectors corresponding to real eigenvalues.

By (2.23), $W = Z^{-*} Y Z^{-1}$. Since Y is involutory,

$$\begin{aligned} W^{-1} = ZY^{-1}Z^* = ZYZ^* &= \left[z_1, \ldots, z_j, \overline{z}_{j+1}, z_{j+1}, \ldots, \overline{z}_k, z_k\right] \begin{bmatrix} z_1^T \\ \vdots \\ z_j^T \\ \overline{z}_{j+1}^T \\ z_{j+1}^T \\ \vdots \\ \overline{z}_k^T \\ z_k^T \end{bmatrix} \\ &= \sum_{i=1}^{j} z_i z_i^T + \sum_{i=j+1}^{k} \left(\overline{z}_i \overline{z}_i^T + z_i z_i^T\right) \\ &= \sum_{i=1}^{j} z_i z_i^T + \sum_{i=j+1}^{k} \left(\overline{z_i z_i^T} + z_i z_i^T\right) \\ &= \sum_{i=1}^{j} z_i z_i^T + 2 \sum_{i=j+1}^{k} \Re\left(z_i z_i^T\right) \end{aligned}$$

which is real. Thus, W is also real.

Similarly, since $M = Z^{-*}Z^{-1}$,

$$M^{-1} = ZZ^* = \begin{bmatrix} z_1, \ldots, z_j, z_{j+1}, \overline{z}_{j+1}, \ldots, z_k, \overline{z}_k \end{bmatrix} \begin{bmatrix} z_1^T \\ \vdots \\ z_j^T \\ \overline{z}_{j+1}^T \\ z_{j+1}^T \\ \vdots \\ \overline{z}_k^T \\ z_k^T \end{bmatrix}$$

$$= \sum_{i=1}^{j} z_i z_i^T + \sum_{i=j+1}^{k} \left(z_i \overline{z}_i^T + \overline{z}_i z_i^T \right)$$

$$= \sum_{i=1}^{j} z_i z_i^T + \sum_{i=j+1}^{k} \left(z_i \overline{z}_i^T + \overline{z_i \overline{z}_i^T} \right)$$

$$= \sum_{i=1}^{j} z_i z_i^T + 2 \sum_{i=j+1}^{k} \Re \left(z_i \overline{z}_i^T \right).$$

This shows that M^{-1} is real. Thus, M is real. □

Corollary 2.28 shows that it is not necessary to use the real Jordan form to describe the symmetric bilinear forms with respect to which a real diagonalizable matrix is self-adjoint. We tend to employ the complex diagonalization in the following chapters, except where using real arithmetic simplifies the exposition.

2.2.2 Computing symmetric bilinear forms

The above characterizations of symmetric bilinear forms all depend on the eigenvectors or generalized eigenvectors R of B. These may be extremely ill-conditioned for a nonsymmetric matrix. A more computationally stable method of computing the symmetric bilinear forms with respect to which a matrix is self-adjoint becomes apparent when one observes that (2.13) is a Sylvester equation [59, page 367] in disguise, since $WB - B^T W = 0$. Its solution cannot be unique since B and B^T have common eigenvalues. This is to be expected since Theorem 2.27 implies that simply scaling or reordering the eigenvectors of B leads to a different matrix that still satisfies (2.13). Consequently, solving this Lyapunov equation is nontrivial. If B has small dimension, one approach to finding W is to transform the Lyapunov equation into the matrix equation

$$(I \otimes B^T - B^T \otimes I) \operatorname{vec}(W) = 0$$

via Kronecker products and vectorization. Both operators are described in Appendix A.2, together with pertinent properties of Kronecker products. The appropriate solutions $\text{vec}(W)$ lie in the nullspace of $(I \otimes B^T - B^T \otimes I)$ and can be reshaped to recover W.

Computations of W by this nullspace method can be simplified by employing the real Schur form. Specifically, if B has Schur form $B = UTU^T$, where U is orthogonal and T is block upper triangular, then

$$\begin{aligned}\left((UT^TU^T) \otimes I - (UT^TU^T) \otimes I\right) \text{vec}(W) &= 0 \\ \left((UT^TU^T) \otimes (UU^T) - (UTU^T) \otimes (UU^T)\right) \text{vec}(W) &= 0 \\ (U \otimes U)(T^T \otimes I - I \otimes T^T)(U^T \otimes U^T) \text{vec}(W) &= 0,\end{aligned}$$

where the last line follows from the mixed-product property. Now, $U \otimes U$ is nonsingular since $\text{rank}(U \otimes U) = \text{rank}(U)^2$. Thus,

$$(T^T \otimes I - I \otimes T^T)(U^T \otimes U^T) \text{vec}(W) = 0.$$

The inverse of $U^T \otimes U^T$ is $(U \otimes U)$ and so if y is any vector in the nullspace of $T^T \otimes I - I \otimes T^T$, then $\text{vec}(W) = (U \otimes U)y$, from which it is possible to recover W. Clearly this Kronecker product approach is infeasible for larger problems unless more sophisticated techniques are employed. We note that this method is similar to that used by Carrette [24] to compute commuting matrices.

2.2.3 Properties of self-adjoint matrices

We conclude this section by examining properties of self-adjoint matrices. We stated previously that every real matrix was self-adjoint with respect to a symmetric bilinear form. If this matrix is diagonalizable, it is also normal with respect to an inner product.

Lemma 2.29. *Let $B \in \mathbb{R}^{n\times n}$ be diagonalizable with diagonalization $B = Z\Lambda Z^{-1}$, $\Lambda, Z \in \mathbb{C}^{n\times n}$. Then B is M-normal, where $M = Z^{-*}Z^{-1}$.*

Proof. If B is M-normal then $BB^\star = B^\star B$ by Definition 2.10, where the adjoint is taken with respect to the M inner product. Substituting for M and B gives that

$$\begin{aligned}BB^\star - B^\star B &= (Z\Lambda Z^{-1})(ZZ^*Z^{-*}\Lambda^* Z^* Z^{-*} Z^{-1}) - (ZZ^*Z^{-*}\Lambda^* Z^* Z^{-*} Z^{-1})(Z\Lambda Z^{-1}) \\ &= Z\Lambda^*\Lambda Z^{-1} - Z\Lambda\Lambda^* Z^{-1} \\ &= 0\end{aligned}$$

since diagonal matrices commute. It follows that B is M-normal. □

However, B may be normal with respect to an inner product $\langle \cdot, \cdot \rangle_M$, where M is not of the form $M = Z^{-*}Z^{-1}$.

Remark 2.30. From Corollary 2.28, it is clear that M is real. Thus, any real diagonalizable matrix is self-adjoint with respect to a symmetric bilinear form on $\mathbb{R}^n$ (also by Corollary 2.28) and is normal with respect to an inner product on $\mathbb{R}^n$.

Lemma 2.29 will be helpful in Chapter 3, where it will be shown that the convergence of Krylov subspace methods in $\langle \cdot, \cdot \rangle_M$ are bounded by a term involving only the eigenvalues of B.

When $\langle \cdot, \cdot \rangle_W$ is an inner product it is often important to know whether a W-self-adjoint matrix is W-positive definite, i.e., whether $\langle Bx, x \rangle_W > 0$ for all nonzero vectors x. The following lemma is useful for this purpose.

Lemma 2.31. *Let $B \in \mathbb{R}^{n \times n}$ be self-adjoint with respect to an inner product $\langle \cdot, \cdot \rangle_W$. Then, B is positive definite with respect to $\langle \cdot, \cdot \rangle_W$ if and only if WB is positive definite with respect to the Euclidean inner product.*

Proof. Since B is W-self-adjoint, $WB = T$ for some symmetric matrix $T = T^T$. Let $x \neq 0$. Then if $\langle Bx, x \rangle_W > 0$ it follows that

$$\begin{aligned} & x^T B^T W x > 0 \\ \Rightarrow & x^T W B x > 0 \\ \Rightarrow & x^T T x > 0. \end{aligned}$$

On the other hand, if $x^T T x > 0$,

$$\begin{aligned} & x^T W B x > 0 \\ \Rightarrow & (Bx)^T W x > 0 \\ \Rightarrow & \langle Bx, x \rangle_W > 0. \end{aligned}$$

□

Additionally, when B is self-adjoint and positive definite with respect to an inner product, all its eigenvalues are real and positive as the next lemma shows.

Lemma 2.32. *Let $B \in \mathbb{R}^{n\times n}$ be self-adjoint with respect to an inner product $\langle\cdot,\cdot\rangle_W$. Then, B is positive definite with respect to $\langle\cdot,\cdot\rangle_W$ if and only if every eigenvalue of B is real and positive.*

Proof. If B is self-adjoint with respect to the inner product $\langle\cdot,\cdot\rangle_W$, then B is diagonalizable with real eigenvalues (see Corollary 2.21). Consequently, by Theorem 2.27, B has diagonalization $B = Z\Lambda Z^{-1}$, where $\Lambda \in \mathbb{R}^{n\times n}$, $Z \in \mathbb{C}^{n\times n}$ and $W = Z^{-*}Z^{-1}$. Note that by Sylvester's law of inertia (Theorem A.14), $Y = I$ in Theorem 2.27.

From Lemma 2.31 we know that B is W-positive definite if and only if WB is positive definite in the Euclidean inner product, i.e., if for all $x \in \mathbb{R}^{n\times n}$, $x \neq 0$, $x^TWBx > 0$. However, $WB = Z^{-*}\Lambda Z^{-1}$ is congruent to Λ, so Sylvester's law of inertia implies that WB is positive definite if and only if Λ is positive definite. □

We additionally find that matrices that are self-adjoint with respect to an inner product $\langle\cdot,\cdot\rangle_W$ possess many of the properties of symmetric matrices. For example, it is possible to show that the stationary points of the W-Rayleigh quotient

$$\rho_W(x) = \frac{\langle Bx, x\rangle_W}{\langle x, x\rangle_W}$$

are the eigenvectors of B and that the value of ρ_W at an eigenvector is the corresponding eigenvalue. Additionally, a generalization of the Courant-Fischer theorem (in Appendix A.4) holds, a consequence of which is that

$$\lambda_{min}(B) \leq \frac{\langle Bx, x\rangle_W}{\langle x, x\rangle_W} \leq \lambda_{max}(B). \tag{2.24}$$

We note that all of the properties described in this section apply to Hermitian sesquilinear forms. The Kronecker product approach can also be used to compute these sesquilinear forms.

2.3 A convection-diffusion example

Many of the applications in the following chapters are discretizations of partial differential equations. Accordingly, we are interested in the symmetric bilinear forms with respect to which such discretized matrices are self-adjoint since knowledge of these symmetric bilinear forms may aid the design of efficient Krylov subspace methods for the discretizations. Intuitively, we might expect that these symmetric bilinear

forms are related in some way to those with respect to which the original differential operator is self-adjoint. If this is true, one approach to preconditioning could be to precondition the coefficient matrix resulting from one discretization with the matrix formed from a lower-order discretization. An example of this form of preconditioning in practice is the use of low-order finite difference discretizations to precondition higher-order discretizations from, say, spectral methods [144, Chapter 40] or a coarse grid approximation of the same operator. Both of these are explored in Chapter 6. Here we obtain an inner product with respect to which an upwind finite difference discretization of a one-dimensional convection-diffusion operator is self-adjoint. We also show that this inner product is related to that with respect to which the original convection-diffusion operator is self-adjoint.

Specifically, the convection-diffusion equation we consider is

$$-\epsilon u'' + u' = f(x), \quad u(a) = u(b) = 0, \quad \epsilon > 0.$$

Letting $h = (b-a)/(n+1)$; $x_j = a + jh$; u_j be the approximation of u at the point x_j and $f_j = f(x_j)$, $j = 0, \ldots, n+1$; the upwind finite difference equations for this differential equation are of the form

$$-\left(\frac{\epsilon}{h^2} + \frac{1}{h}\right) u_{j-1} + \left(\frac{2\epsilon}{h^2} + \frac{1}{h}\right) u_j - \frac{\epsilon}{h^2} u_{j+1} = f_j, \quad j = 1, \ldots, n$$

with $u_0 = u(a) = 0$ and $u_{n+1} = u(b) = 0$.

The equivalent matrix formulation is $Bx = b$, where

$$B = \begin{bmatrix} \frac{2\epsilon}{h^2} + \frac{1}{h} & -\frac{\epsilon}{h^2} & & & & & \\ -\left(\frac{\epsilon}{h^2} + \frac{1}{h}\right) & \frac{2\epsilon}{h^2} + \frac{1}{h} & -\frac{\epsilon}{h^2} & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -\left(\frac{\epsilon}{h^2} + \frac{1}{h}\right) & \frac{2\epsilon}{h^2} + \frac{1}{h} & -\frac{\epsilon}{h^2} & \\ & & & & -\left(\frac{\epsilon}{h^2} + \frac{1}{h}\right) & \frac{2\epsilon}{h^2} + \frac{1}{h} \end{bmatrix}, \tag{2.25}$$

$$u = \begin{bmatrix} u_1, & u_2, & \ldots, & u_n \end{bmatrix}^T$$

and

$$f = \begin{bmatrix} f_1, & f_2, & \ldots, & f_n \end{bmatrix}^T.$$

As shown at the start of Section 2.2.1, since B is a Toeplitz matrix it is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\hat{Y}}$, where $\hat{Y}$ is the reverse identity matrix (2.19). However, since B is diagonalizable with real eigenvalues it is also self-adjoint with respect to an inner product.

Theorem 2.33. *Let $B \in \mathbb{R}^{n\times n}$ be defined by* (2.25)*. Then B is self-adjoint with respect to $\langle\cdot,\cdot\rangle_W$, where*

$$W = \frac{2}{n+1}\begin{bmatrix} \frac{\epsilon}{\epsilon+h} & & & \\ & (\frac{\epsilon}{\epsilon+h})^2 & & \\ & & \ddots & \\ & & & (\frac{\epsilon}{\epsilon+h})^n \end{bmatrix}. \tag{2.26}$$

Proof. Since B is a tridiagonal Toeplitz matrix, its eigenvalues and eigenvectors are known [18, Theorem 2.4]. The eigenvalues are the real numbers

$$\lambda_j = \frac{2\epsilon+h}{h^2} + \frac{2}{h^2}\sqrt{\epsilon(\epsilon+h)}\cos\frac{j\pi}{n+1},\ j = 1,\dots,n,$$

while the corresponding eigenvectors are

$$v_j = \begin{bmatrix} \left(\frac{\epsilon+h}{\epsilon}\right)^{\frac{1}{2}}\sin\frac{j\pi}{n+1} & \left(\frac{\epsilon+h}{\epsilon}\right)\sin\frac{2j\pi}{n+1} & \dots & \left(\frac{\epsilon+h}{\epsilon}\right)^{\frac{n}{2}}\sin\frac{nj\pi}{n+1} \end{bmatrix}^T.$$

Thus, B has diagonalization $B = Z\Lambda Z^{-1}$, where $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$ and

$$Z = \Theta^{-\frac{1}{2}}S, \tag{2.27}$$

with

$$\Theta = \operatorname{diag}\left(\left(\frac{\epsilon}{\epsilon+h}\right), \left(\frac{\epsilon}{\epsilon+h}\right)^2, \dots, \left(\frac{\epsilon}{\epsilon+h}\right)^n\right)$$

and

$$S = \begin{bmatrix} \sin\frac{\pi}{n+1} & \sin\frac{2\pi}{n+1} & \cdots & \sin\frac{n\pi}{n+1} \\ \sin\frac{2\pi}{n+1} & \sin\frac{4\pi}{n+1} & \cdots & \sin\frac{2n\pi}{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sin\frac{n\pi}{n+1} & \sin\frac{2n\pi}{n+1} & \cdots & \sin\frac{n^2\pi}{n+1} \end{bmatrix}.$$

The matrix S is the discrete sine transform (DST-I) matrix (see [150, Section 4.4]) and its inverse is simply

$$S^{-1} = \frac{2}{n+1}S.$$

It follows that

$$Z^{-T} = \frac{2}{n+1}\Theta^{\frac{1}{2}}S$$

and so

$$W = Z^{-T}Z^{-1} = \frac{2}{n+1}\Theta^{\frac{1}{2}}SS^{-1}\Theta^{\frac{1}{2}} = \frac{2}{n+1}\Theta.$$

□

Remark 2.34. The condition number of W is simply the ratio of the largest diagonal entry to the smallest. Thus,

$$\kappa_I(W) = \left(1 + \frac{h}{\epsilon}\right)^{n-1} \tag{2.28}$$

and the conditioning of the inner product (and the eigenvectors of B) worsens with decreasing ϵ.

Remark 2.35. The symmetric bilinear form defined by the reverse identity matrix $\hat{Y}$ can also be obtained from the eigenvectors Z in (2.27). Specifically, $\hat{Y} = Z^{-T} D Z^{-1}$, where

$$D = \frac{n+1}{2}\left(\frac{\epsilon}{\epsilon+h}\right)^{\frac{n}{2}} \operatorname{diag}\left(1, -1, 1, -1, \ldots, (-1)^{n-2}, (-1)^{n-1}\right).$$

Let us consider the inner product in Theorem 2.33 and its relationship to the inner product with respect to which the original convection-diffusion operator is self-adjoint. If $u = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}^T$ and $w = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T$, where $u_i = u(x_i)$ and similarly for w_i, then

$$\langle u, w \rangle_W = \frac{2}{n+1} \sum_{i=1}^{n} \left(\frac{\epsilon}{\epsilon+h}\right)^i u_i w_i. \tag{2.29}$$

Consider, on the other hand, the inner product

$$\langle u, w \rangle = \int_a^b e^{-\frac{x}{\epsilon}} u w dx, \tag{2.30}$$

for any $u, w \in C[a, b]$. If the convection-diffusion operator is denoted by $\mathcal{L} = -\epsilon u'' + u'$ then, with $u = w = 0$ at $x = a$ and $x = b$,

$$\begin{aligned}
\langle \mathcal{L}u, w \rangle = \int_a^b e^{-\frac{x}{\epsilon}} \left(-\epsilon u'' + u'\right) w dx &= -\epsilon \int_a^b \left(e^{-\frac{x}{\epsilon}} u'' - \frac{1}{\epsilon} e^{-\frac{x}{\epsilon}} u'\right) w dx \\
&= -\epsilon \int_a^b \frac{d}{dx}\left(u' e^{-\frac{x}{\epsilon}}\right) w dx \\
&= -\epsilon e^{-\frac{x}{\epsilon}} u' w \Big|_{x=a}^{b} + \epsilon \int_a^b e^{-\frac{x}{\epsilon}} u' w' dx \\
&= \epsilon \int_a^b e^{-\frac{x}{\epsilon}} u' w' dx.
\end{aligned}$$

Meanwhile,

$$\begin{aligned}\langle u, \mathcal{L}w\rangle = \int_a^b e^{-\frac{x}{\epsilon}} u\left(-\epsilon w'' + w'\right) dx &= -\epsilon \int_a^b u\left(e^{-\frac{x}{\epsilon}} w'' - \frac{1}{\epsilon} e^{-\frac{x}{\epsilon}} w'\right) dx \\ &= -\epsilon \int_a^b u \frac{d}{dx}\left(w' e^{-\frac{x}{\epsilon}}\right) dx \\ &= -\epsilon e^{-\frac{x}{\epsilon}} u w'\Big|_{x=a}^b + \epsilon \int_a^b e^{-\frac{x}{\epsilon}} u' w' dx \\ &= \epsilon \int_a^b e^{-\frac{x}{\epsilon}} u' w' dx \\ &= \langle \mathcal{L}u, w\rangle\end{aligned}$$

and so $\mathcal{L}$ is a self-adjoint operator with respect to the inner product (2.30).

We now wish to show that in the limit as $h \to 0$, or $n \to \infty$, the inner product with respect to which the finite difference matrix B is self-adjoint becomes the inner product with respect to which the differential operator is self-adjoint. For simplicity we let $a = 0$ and $b = 1$, so that $u(0) = w(0) = u(1) = w(1) = 0$. Recall from (2.29) that the inner product with respect to which the finite difference discretization is self-adjoint is given by

$$\begin{aligned}\frac{1}{2}\langle u, w\rangle &= \frac{1}{n+1} \sum_{i=1}^n \left(\frac{\epsilon}{\epsilon + h}\right)^i u\left(\frac{i}{n+1}\right) w\left(\frac{i}{n+1}\right) \\ &= \frac{1}{n+1} \sum_{i=0}^n \left(1 + \frac{1}{\epsilon(n+1)}\right)^{-i} u\left(\frac{i}{n+1}\right) w\left(\frac{i}{n+1}\right) \\ &= \frac{1}{n+1} \sum_{i=0}^n \left[\left(1 + \frac{1}{\epsilon(n+1)}\right)^{\epsilon(n+1)}\right]^{-\frac{i}{\epsilon(n+1)}} u\left(\frac{i}{n+1}\right) w\left(\frac{i}{n+1}\right)\end{aligned}$$

since $u(0) = w(0) = 0$. Thus,

$$\lim_{n\to\infty} \frac{1}{2}\langle u, w\rangle = \lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^n \left[\left(1 + \frac{1}{\epsilon(n+1)}\right)^{\epsilon(n+1)}\right]^{-\frac{i}{\epsilon(n+1)}} u\left(\frac{i}{n+1}\right) w\left(\frac{i}{n+1}\right).$$

Now, if $x = \frac{1}{\epsilon(n+1)}$ then

$$\left[\left(1 + \frac{1}{\epsilon(n+1)}\right)^{\epsilon(n+1)}\right]^{-\frac{i}{\epsilon(n+1)}} = \left[(1+x)^{\frac{1}{x}}\right]^{-ix}.$$

Also, as $x \to 0$ (so that $n \to \infty$) we find that

$$\log\left((1+x)^{\frac{1}{x}}\right) = \frac{1}{x}(x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + O(x^4)) = 1 - \frac{1}{2}x + \frac{1}{3}x^2 + O(x^3)$$

and that

$$(1+x)^{\frac{1}{x}} = e^{1-\frac{1}{2}x+\frac{1}{3}x^2+O(x^3)}.$$

Therefore,

$$\begin{aligned}\left[(1+x)^{\frac{1}{x}}\right]^{-ix} = e^{-ix+\frac{1}{2}ix^2-\frac{1}{3}ix^3+O\left(x^4\right)} &= \sum_{j=0}^{\infty}\frac{1}{j!}\left(-ix+\frac{1}{2}ix^2-\frac{1}{3}ix^3+O\left(x^4\right)\right)^j \\ &= e^{-ix}+O\left(x^2\right).\end{aligned}$$

Thus,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{2}\langle u,w\rangle &= \lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n}\left[e^{-\frac{i}{\epsilon(n+1)}}+O\left(\frac{1}{(n+1)^2}\right)\right]u\left(\frac{i}{n+1}\right)w\left(\frac{i}{n+1}\right) \\ &= \lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n}e^{-\frac{i}{\epsilon(n+1)}}\ u\left(\frac{i}{n+1}\right)w\left(\frac{i}{n+1}\right) \\ &\quad+\lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n}O\left(\frac{1}{(n+1)^2}\right) \\ &= \lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n}e^{-\frac{i}{\epsilon(n+1)}}u\left(\frac{i}{n+1}\right)w\left(\frac{i}{n+1}\right).\end{aligned}$$

The Euler-Maclaurin summation formula can be written as

$$\begin{aligned}\sum_{j=0}^{n}A\left(jn^{-1}\right) = n\int_0^1 A(y)dy + \frac{A(0)+A(1)}{2} \\ +\sum_{k=2}^{K}\frac{B_k}{n^{(k-1)}k!}(A^{(k-1)}(1)-A^{(k-1)}(0))+O\left(n^{-K}\right),\end{aligned}$$

where $A \in C^K[0,1]$ and B_k is the kth Bernoulli number. Assume that $u, w \in C^2[0,1]$ and let

$$f(x) = e^{-\frac{x}{\epsilon}}u(x)w(x).$$

Then, applying the Euler-Maclaurin formula to our problem gives that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{2}\langle u,w\rangle &= \lim_{n\to\infty}\left[\int_0^1 f(y)dy + \frac{f(0)+f(1)}{2}\right.\\ &\quad\left. + \frac{B_2}{2n}(f'(1)-f'(0)) + O\left(\frac{1}{(n+1)^2}\right)\right]\\ &= \lim_{n\to\infty}\left[\int_0^1 e^{-\frac{y}{\epsilon}}u(y)w(y)dy + O\left(\frac{1}{(n+1)^2}\right)\right]\\ &= \int_0^1 e^{-\frac{y}{\epsilon}}u(y)w(y)dy,\end{aligned}$$

which is precisely the form of the inner product of the differential operator (2.30).

Thus the finite difference approximation B is self-adjoint with respect to an inner product that becomes the inner product with respect to which the original differential operator is self-adjoint in the limit as $n \to \infty$, i.e., as the approximation becomes more accurate. At least for this example, there is a link between the differential operator and its discretization. One might conjecture that a different discretization of this convection-diffusion operator might be self-adjoint with respect to a different inner product that also approaches that with respect to which the operator is self-adjoint. If so, this idea could potentially be exploited for preconditioning. Moreover, this link between the inner products with respect to which the operator and its discretization are self-adjoint might be more generally true, making this preconditioning strategy more widely applicable.

2.4 Conclusions

In this chapter we have given an overview of the properties of symmetric bilinear forms and inner products that are required in later chapters. Symmetric bilinear forms allow us to generalize the idea of a symmetric matrix, since every real matrix is self-adjoint with respect to some symmetric bilinear form. The symmetric bilinear forms with respect to which a given matrix is self-adjoint have been characterized and some of the properties possessed by self-adjoint matrices have been discussed. We will draw on these characterizations and properties throughout this thesis. Finally, we have shown that a finite difference matrix for a PDE is self-adjoint with respect to a symmetric bilinear form that is related to that with respect to which the differential operator is self-adjoint. We will return to this example in Chapter 4 and to a related two-dimensional example in Chapter 6.

Chapter 3

Iterative methods

In the introduction we suggested several reasons why one might consider Krylov subspace methods in nonstandard inner products. For a symmetric saddle point matrix, for example, an effective preconditioner may cause the preconditioned coefficient matrix to be nonsymmetric but self-adjoint with respect to a nonstandard inner product. If it is practical to compute with this inner product, solving the linear system by a short-term recurrence method in a nonstandard inner product may be preferable to solving it by a standard method for nonsymmetric matrices. Preconditioners of this type are presented in Chapters 4 and 5. Another reason, discussed more in Chapter 6, is that examining the effect of preconditioners for nonsymmetric matrices on the convergence of Krylov subspace methods in nonstandard inner products may provide some insight into good preconditioners for standard Krylov subspace methods.

In this chapter, we introduce the three Krylov subspace methods that will be used in the remainder of the thesis: the W-conjugate gradient method, W-MINRES and W-GMRES. These are standard in the literature but are typically presented with either the Euclidean inner product $\langle \cdot, \cdot \rangle_I$ (for MINRES and GMRES) or the energy inner product $\langle \cdot, \cdot \rangle_B$ (for the conjugate gradient method). We require these methods in nonstandard inner products and so they are presented here in greater generality. Section 3.1 covers the class of orthogonal error methods, in which lies the conjugate gradient method [71] described in Section 3.1.1. Minimum residual methods, which are the subject of Section 3.2, include the MINRES algorithm [107] discussed in Section 3.2.1 and its generalization to nonsymmetric matrices, GMRES [119], presented in Section 3.2.2. For each of the three methods we adapt the original algorithm for use with a nonstandard inner product and provide convergence results. In Chapter 6 it will be necessary to relate convergence bounds for minimal residual methods in

nonstandard inner products, derived in Section 3.2.3, to bounds for GMRES and MINRES convergence in the Euclidean inner product and in Section 3.2.4 we derive relationships for this purpose. Conclusions are drawn in Section 3.3. All of the results in this chapter hold for complex matrices and vectors but, for ease of exposition, we discuss only the real case here.

We remark that other well known Krylov subspace methods such as BiCG [51] and QMR [53] can also be generalized in a fairly straightforward manner. Since we do not require these methods in later chapters, no details are given here. We additionally note that QMR in $\langle \cdot, \cdot \rangle_I$ simplifies when the coefficient matrix is self-adjoint with respect to a symmetric bilinear form. Specifically, if B is self-adjoint with respect to $\langle \cdot, \cdot \rangle_W$, where $\langle \cdot, \cdot \rangle_W$ is a symmetric bilinear form, SQMR [52, 54] replaces multiplications with B^T, required in the standard formulation of QMR, by multiplications with W.

3.1 Orthogonal error methods

Orthogonal error (OE) methods select iterates x_k from shifted Krylov subspaces, where $\mathcal{K}_k(B, r_0)$ is the Krylov subspace

$$\mathcal{K}_k(B, r_0) = \text{span}\{r_0, Br_0, \ldots, B^{k-1}r_0\}$$

of dimension k of B and the initial residual $r_0 = b - Bx_0$. The error $e_k = x - x_k$ is orthogonal, with respect to an inner product, to a subspace. This inner product may be independent of the initial guess x_0, in which case it is a **fixed metric**, following Barth and Manteuffel [10]; it is otherwise a **variable metric**. Specifically, orthogonal error methods are defined as follows.

Definition 3.1. Iterates of an **orthogonal error (OE) method** satisfy

$$x_k \in x_0 + \mathcal{K}_k(B, r_0), \tag{3.1}$$

$$e_k \perp_W \mathcal{K}_k(B, r_0), \tag{3.2}$$

where $y \perp_W \mathcal{K}_k(B, r_0)$ means that $\langle y, z \rangle_W = 0 \; \forall z \in \mathcal{K}_k(B, r_0)$.

Equivalent to the orthogonality condition (3.2) is the minimization property [66]

$$\|e_k\|_W = \min_{y \in x_0 + \mathcal{K}_k(B, r_0)} \|x - y\|_W. \tag{3.3}$$

The Faber-Manteuffel result [49] discussed in the introduction describes the matrices to which a conjugate gradient algorithm that satisfies (3.1) and (3.2) with a fixed

inner product can be applied. These comprise a small set [47, 49], although it is one to which certain preconditioned saddle point matrices belong. Other short-term recurrence methods, such as those introduced by Barth and Manteuffel [11, 12], are beyond the scope of this chapter but these apply to a similarly limited set of matrices. Below we examine this conjugate gradient algorithm, which will be used in Chapters 4 and 5 to solve preconditioned saddle point systems.

3.1.1 The W-conjugate gradient method

The I-conjugate gradient (I-CG) method of Hestenes and Stiefel [71] requires only three-term recurrences and—provided B is symmetric positive definite, so that $\langle\cdot,\cdot\rangle_B$ defines an inner product—minimizes the B-norm of the error over the Krylov subspace, i.e.,

$$x_k = \operatorname{argmin}_{y\in x_0+\mathcal{K}_k(B,r_0)} \|x-y\|_B.$$

Equivalently, I-CG satisfies the orthogonality condition

$$e_k \perp_B \mathcal{K}_k(B,r_0).$$

This I-conjugate gradient method can be generalized to one in which a different inner product $\langle\cdot,\cdot\rangle_W$ is used [5, 6, 22, 66, 67, 89, 135]. Indeed, certain standard Krylov methods such as the conjugate residual method [71] can be cast as conjugate gradient methods in nonstandard inner products [5, 6, 22, 66]. The W-conjugate gradient (W-CG) method selects iterates $x_k \in x_0 + \mathcal{K}_k(B,r_0)$ such that

$$e_k \perp_{WB} \mathcal{K}_k(B,r_0). \tag{3.4}$$

The equivalent minimization condition (3.3), when B is self-adjoint and positive definite with respect to the inner product $\langle\cdot,\cdot\rangle_W$ or, equivalently, when W and WB are symmetric positive definite, is

$$\|e_k\|_{WB} = \min_{y\in x_0+\mathcal{K}_k(B,r_0)} \|x-y\|_{WB}.$$

Similarly to I-CG, the W-CG method can be derived from the W-Lanczos procedure with an orthogonality (or equivalent minimization) condition, where the **W-Lanczos method** for a W-self-adjoint matrix B is given in Algorithm 3.1. The W-Lanczos vectors $v_1,\ldots,v_k$ generate a W-orthogonal basis of the kth Krylov subspace of B and r, $\mathcal{K}_k(B,r)$ and if the W-Lanczos vectors form the columns of the matrix

$$V_k = \begin{bmatrix} v_1 & v_2 & \ldots & v_k \end{bmatrix} \tag{3.5}$$

then the Lanczos recurrences can be represented by

$$BV_k = V_{k+1}\widehat{T}_k, \tag{3.6}$$

where

$$\widehat{T}_k = \begin{bmatrix} T_k \\ \beta_k e_k^T \end{bmatrix} \text{ with } T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix} \tag{3.7}$$

and e_k is the kth unit vector. Moreover, the W-Lanczos vectors are W-orthonormal and so

$$V_k^T W V_k = I_k \text{ and } V_k^T W B V_k = T_k. \tag{3.8}$$

Algorithm 3.1 W-Lanczos algorithm

Choose r. Set $v_1 = r/\|r\|_W$. Set $\beta_0 = 0$ and $v_0 = 0$.
for $k = 1, 2, \ldots$ **do**
  $\alpha_k = \langle Bv_k, v_k\rangle_W$
  $\tilde{v}_{k+1} = Bv_k - \alpha_k v_k - \beta_{k-1} v_{k-1}$
  $\beta_k = \|v_{k+1}\|_W$
  $v_{k+1} = v_{k+1}/\beta_k$
end for

We now examine how these W-Lanczos relationships can be used to construct iterates satisfying (3.1) and (3.4). If the W-Lanczos method in Algorithm 3.1 is applied to $r_0 = b - Bx_0$, then the W-Lanczos vectors form a basis of $\mathcal{K}_k(B, r_0)$. Any W-CG iterate $x_k \in x_0 + \mathcal{K}_k(B, r_0)$ is of the form

$$x_k = x_0 + V_k y_k, \text{ where } y_k \in \mathbb{R}^k,$$

and the corresponding residual is $r_k = Be_k = r_0 - BV_k y_k$. This means that the orthogonality condition for W-CG can be enforced by choosing x_k so that

$$0 = V_k^T W B e_k = V_k^T W r_k = V_k^T W (r_0 - BV_k y_k). \tag{3.9}$$

By utilizing (3.8) and by observing that the initial residual r_0 is equal to $\beta V_{k+1} e_1$, where $\beta = \|r_0\|_W$, we find that (3.9) simplifies to

$$T_k y_k = \beta e_1,$$

from which it is possible to find y_k and hence the iterate x_k.

By using a procedure similar to that for I-CG in, for example, [62, Section 2.5] and [118, Section 6.7] it is possible to compute x_k without storing V_k and T_k. The result is the W-CG algorithm in Algorithm 3.2 that requires only three-term (or coupled two-term) recurrences. The W-CG algorithm differs from the standard I-CG method only in the computation of α_k and β_k, where nonstandard inner products are used. We note that the W-conjugate gradient method can also be expressed as a preconditioned conjugate gradient method in the I-norm [34] but we prefer to separate the roles of the preconditioner and the inner product. One reason for this is that this separation allows us to develop preconditioners for saddle point matrices that may be used either with W-CG or with W-MINRES (described in the next section).

Algorithm 3.2 W-conjugate gradient method

Choose x_0, compute $r_0 = b - Bx_0$, set $p_0 = r_0$
for $k = 1, 2, \ldots$ **do**
$\alpha_k = \dfrac{\langle r_k, r_k \rangle_W}{\langle Bp_k, p_k \rangle_W}$
$x_{k+1} = x_k + \alpha_k p_k$
$r_{k+1} = r_k - \alpha_k B p_k$
<Test for convergence>
$\beta_k = \dfrac{\langle r_{k+1}, r_{k+1} \rangle_W}{\langle r_k, r_k \rangle_W}$
$p_{k+1} = r_{k+1} + \beta_k p_k$
end for

The W-CG residuals and direction vectors satisfy the standard conjugacy properties [6, 89] with respect to the W-inner product, i.e.,

$$\begin{aligned} &\langle r_k, p_j \rangle_W = \langle r_k, r_j \rangle_W = 0, \ j < k, \\ &\langle Bp_k, p_j \rangle_W = 0, \ j < k, \\ &\text{span}\{r_0, r_1, \ldots, r_{k+1}\} = \text{span}\{p_0, p_1, \ldots, p_{k+1}\} = \mathcal{K}_k(B, r_0). \end{aligned} \tag{3.10}$$

Significantly, convergence results for W-CG can be obtained similarly to the standard conjugate gradient method. These are helpful when determining whether convergence of W-CG will be rapid and when choosing preconditioners. To bound the W-CG error as in [89] we utilize the connection between Krylov subspaces and matrix polynomials in the following theorem.

Theorem 3.2. *Let $B \in \mathbb{R}^{n \times n}$ be self-adjoint and positive definite with respect to the inner product $\langle \cdot, \cdot \rangle_W$. Then the error, $e_k = x - x_k$, of the kth W-CG iterate is bounded by*

$$\frac{\|e_k\|_{WB}}{\|e_0\|_{WB}} \leq \min_{p \in \Pi_k, p(0)=1} \max_{\lambda \in \Lambda(B)} |p(\lambda)|. \tag{3.11}$$

Proof. Since $x_k \in x_0 + \mathcal{K}_k(B, r_0)$, we have that $x_k = x_0 + q_{k-1}(B)r_0$, where $q_{k-1}(z)$ is a polynomial of degree at most $k-1$. It follows that

$$e_k = x - x_k = x - (x_0 + q_{k-1}(B)r_0) = (I - Bq_{k-1}(B))\, e_0 = p_k(B)e_0, \tag{3.12}$$

where $p_k(z) = 1 - zq_{k-1}(z)$.

The matrix B is self-adjoint with respect to an inner product and is, therefore, diagonalizable (c.f. Corollary 2.21 in Section 2.2). Accordingly, the eigenvectors of B, which we denote u_k, $k = 1, \ldots, n$, form a basis of $\mathbb{R}^n$. The initial error e_0 can be expressed in this basis as

$$e_0 = \sum_{j=1}^{n} \gamma_j u_j,$$

for some scalars γ_j, $j = 1, \ldots, n$. From the minimization property and the polynomial representation of the error (3.12) we have that

$$\begin{aligned}
\|e_k\|_{WB} = \min_{p\in\Pi_k, p(0)=1} \|p(B)e_0\|_{WB} &= \min_{p\in\Pi_k, p(0)=1} \left\| \sum_{j=1}^{n} \gamma_j p(B) u_j \right\|_{WB} \\
&= \min_{p\in\Pi_k, p(0)=1} \left\| \sum_{j=1}^{n} \gamma_j p(\lambda_j) u_j \right\|_{WB} \\
&\leq \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)| \left\| \sum_{j=1}^{n} \gamma_j u_j \right\|_{WB} \\
&= \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)| \|e_0\|_{WB},
\end{aligned}$$

where Π_k is the set of polynomials of degree at most k. □

Furthermore, since $r_k = b - Bx_k = B(x - x_k) = Be_k$, we have that

$$\|e_k\|_{WB}^2 = e_k^T W B e_k = r_k^T B^{-T} W B B^{-1} r_k = r_k^T B^{-T} W r_k.$$

By Lemma 2.18, B^{-1} is also W-self-adjoint and so $B^{-T}W = WB^{-1}$. Thus,

$$\|e_k\|_{WB}^2 = r_k^T W B^{-1} r_k = \|r_k\|_{WB^{-1}}^2 \quad \text{and} \quad \|e_0\|_{WB} = \|r_0\|_{WB^{-1}}$$

which, when combined with (3.11) in Theorem 3.2 shows that the kth residual satisfies

$$\frac{\|r_k\|_{WB^{-1}}}{\|r_0\|_{WB^{-1}}} \leq \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|.$$

Theorem 3.2 indicates that, as in the case of I-CG, eigenvalues influence the rate of convergence of this method. It is our experience that eigenvalues generally dictate

convergence of the W-conjugate gradient method. However, in obtaining the upper bound (3.11) we have separated the right-hand side from the matrix polynomial. The consequence is that the bound (3.11) may predict slower convergence than is observed in practice. Nevertheless, since a *sufficient* condition for fast convergence is that the eigenvalues of B are nicely distributed, a criterion for choosing preconditioners for use with a conjugate gradient method in a nonstandard inner product is that the preconditioned coefficient matrix has nicely distributed eigenvalues.

Since W-self-adjoint, W-positive definite matrices have real, positive eigenvalues (see Lemma 2.32), the polynomial term in (3.11) may be bounded by the W-condition number of B by using Chebyshev polynomials.

Corollary 3.3. *Let $B \in \mathbb{R}^{n\times n}$ be self-adjoint and positive definite with respect to the inner product $\langle\cdot,\cdot\rangle_W$. Then the kth error of the W-conjugate gradient method, e_k, is bounded by*

$$\frac{\|e_k\|_{WB}}{\|e_0\|_{WB}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \tag{3.13}$$

where

$$\kappa = \frac{\lambda_{max}(B)}{\lambda_{min}(B)}$$

is the W-norm condition number of B in Definition 2.5.

Proof. The proof is the same as for I-CG in, for example, [62, Theorem 3.1.1], where we have used that the W-norm condition number of a W-self-adjoint matrix is related to its eigenvalues (see Lemma 2.16) and that the eigenvalues of a W-self-adjoint, W-positive definite matrix are real and positive (see Lemma 2.32). □

Remark 3.4. From Corollary 3.3, we might expect that when the eigenvalues of B are clustered, the W-conjugate gradient algorithm will perform well in exact arithmetic.

As previously discussed, preconditioning is often required for acceptably rapid convergence of Krylov subspace methods. A preconditioner may be straightforwardly introduced into the conjugate gradient algorithm in Algorithm 3.2 if the preconditioned coefficient matrix $P^{-1}B$ is self-adjoint and positive definite with respect to some inner product $\langle\cdot,\cdot\rangle_W$, so that $WP^{-1}B$ is symmetric positive definite. When this condition is satisfied, the ***W*-PCG method** given in Algorithm 3.3 can be applied. In Chapters 4 and 5 we examine specific combinations of W, P and B for which W-PCG is applicable.

Algorithm 3.3 W-PCG method

Choose x_0, compute $\widehat{r}_0 = P^{-1}(b - Bx_0)$, set $p_0 = \widehat{r}_0$
for $k = 1, 2, \ldots$ **do**
$\alpha_k = \dfrac{\langle \widehat{r}_k, \widehat{r}_k \rangle_W}{\langle P^{-1}Bp_k, p_k \rangle_W}$
$x_{k+1} = x_k + \alpha_k p_k$
$\widehat{r}_{k+1} = \widehat{r}_k - \alpha_k P^{-1}Bp_k$
<Test for convergence>
$\beta_k = \dfrac{\langle \widehat{r}_{k+1}, \widehat{r}_{k+1} \rangle_W}{\langle \widehat{r}_k, \widehat{r}_k \rangle_W}$
$p_{k+1} = \widehat{r}_{k+1} + \beta_k p_k$
end for

Algorithm 3.3 differs from the standard I-PCG method in, for example, [62, page 121]. In I-PCG *both* the preconditioner and the coefficient matrix are symmetric positive definite (with respect to the I-inner product). The preconditioner is factorized[1] as $P = LL^T$ and applied as a split preconditioner, so that the preconditioned coefficient matrix $L^{-1}BL^{-T}$ remains symmetric positive definite. Modifications to the algorithm can then be made so that, in practice, only left or right preconditioning is required. Our motivation in Chapters 4 and 5 is different since the original saddle point matrix has positive and negative eigenvalues and so is never positive definite with respect to an inner product (see Lemma 2.32). We require a method for which only the preconditioned coefficient matrix need be self-adjoint and positive definite with respect to an inner product.

The quantity that is minimized by the W-PCG method is $\|e_k\|_{WP^{-1}B}$. The corresponding error bound, which is based on Theorem 3.2, is given in the following corollary.

Corollary 3.5. *Let $P^{-1}B \in \mathbb{R}^{n\times n}$ be self-adjoint and positive definite with respect to the inner product $\langle \cdot, \cdot \rangle_W$. Then the error, e_k, of the kth W-PCG iterate is given by*

$$\frac{\|e_k\|_{WP^{-1}B}}{\|e_0\|_{WP^{-1}B}} \leq \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)|. \tag{3.14}$$

Remark 3.6. The standard convergence bound for the I-PCG method mentioned above is [43, page 79]

$$\frac{\|e_k\|_B}{\|e_0\|_B} \leq \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)|.$$

[1]The positive definiteness of P guarantees that a factorization of this sort exists.

The preconditioner does not appear in the norm on the left-hand side of this convergence bound unlike in (3.14) in Corollary 3.5. This disparity is caused by the difference between the way our algorithm is formulated and the way the standard I-PCG method is derived, as discussed above. We note, however, that given a W-self-adjoint and positive definite preconditioner for a W-self-adjoint and positive definite coefficient matrix it is possible to devise a conjugate gradient method for the preconditioned system for which the error is minimized with respect to $\|\cdot\|_{WB}$. We do not elucidate on this here.

From Corollary 3.5 we discern that the eigenvalues of the preconditioned matrix play an important part in determining the convergence of this Krylov subspace method. In particular, convergence will be rapid if a preconditioned coefficient matrix has eigenvalues for which polynomials, satisfying $p(0) = 1$, can be found that are small at each eigenvalue. This is the motivation behind choosing preconditioners that achieve a good eigenvalue distribution, similarly to I-PCG.

Let us recap the properties of the W-PCG method. It is applicable to nonsymmetric preconditioned matrices that are self-adjoint and positive definite with respect to an inner product. Such preconditioned matrices are diagonalizable with real, positive eigenvalues (see Corollary 2.21 and Lemma 2.32) but may be nonsymmetric. The W-PCG method minimizes the error with respect to some norm, with convergence typically determined by the eigenvalues of the preconditioned coefficient matrix. It additionally requires only three-term recurrences. In contrast, standard methods for nonsymmetric matrices either require more work as iterations progress, like GMRES, or have poorly understood convergence in general, such as BiCG [51] or QMR [53].

The properties of W-PCG provide an impetus for seeking an inner product with respect to which a preconditioned matrix is self-adjoint and positive definite since, provided the inner product is not too costly to apply, the overall time required to solve the preconditioned system by W-PCG may be less than that required by standard methods. Finding such preconditioners and inner products is the subject of Chapters 4 and 5.

3.2 Minimal residual methods

In contrast to orthogonal error methods, iterates of minimal residual (MR) methods are chosen so that the corresponding residual vector

$$r_k = b - Bx_k \in r_0 + B\mathcal{K}_k(B, r_0)$$

is minimized with respect to some norm. This leads to the following definition.

Definition 3.7. Iterates of a **minimal residual (MR) method** satisfy

$$x_k \in x_0 + \mathcal{K}_k(B, r_0), \tag{3.15}$$

$$\|r_k\|_W = \min_{y \in x_0 + \mathcal{K}_k(B, r_0)} \|b - By\|_W. \tag{3.16}$$

The MINRES method of Paige and Saunders [107] is one of the best known minimal residual methods. W-MINRES can be applied when B is self-adjoint (although not necessarily positive definite) with respect to an inner product, and will be used in Chapters 4 and 5 to solve preconditioned saddle point problems. An extension of MINRES for non-self-adjoint matrices is GMRES, which was proposed by Saad and Schultz [119]. Below we outline W-GMRES, which will be employed in Chapter 6. Collectively, these are W-MR methods. We additionally describe below the convergence of minimum residual methods and formulate expressions that relate the convergence of W-MR to that of I-MR methods. These will also be required in Chapter 6.

3.2.1 W-MINRES

Analogously to the original I-MINRES algorithm of Paige and Saunders [107] that minimizes the I-norm of the residual, the W-MINRES method in [135] minimizes

$$\|r_k\|_W = \min_{y \in x_0 + \mathcal{K}_k(B, r_0)} \|b - By\|_W. \tag{3.17}$$

Its reliable application requires that B is self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_W$.

W-MINRES can be implemented with W-orthogonal vectors generated by the W-Lanczos method in Algorithm 3.1 applied to B and r_0, the result of which is that only three-term (or coupled two-term) recurrences are required. If these basis vectors

form columns of the matrix V_k as in (3.5) then, since W-MINRES iterates lie in the affine space defined by (3.15),

$$x_k^W = x_0 + V_k y_k,$$

for some $y_k \in \mathbb{R}^k$. The corresponding residual vectors are given by

$$r_k^W = r_0 - BV_k y_k \in r_0 + B\mathcal{K}_k(B, r_0)$$

and it follows from the Lanczos relationship (3.6) that the W-norm of r_k^W satisfies

$$\begin{aligned} \left\|r_k^W\right\|_W &= \|r_0 - BV_k y_k\|_W \\ &= \left\| \|r_0\|_W v_1 - V_{k+1}\widehat{T}_k y_k \right\|_W \\ &= \left\| V_{k+1}(\|r_0\|_W e_1 - \widehat{T}_k y_k) \right\|_W, \end{aligned}$$

where $e_1 \in \mathbb{R}^{k+1}$ is the first canonical basis vector . Squaring the right-hand side gives

$$\left\| V_{k+1}\left(\|r_0\|_W e_1 - \widehat{T}_k y_k\right) \right\|_W^2 = (\|r_0\|_W e_1 - \widehat{T}_k y_k)^T V_{k+1}^T W V_{k+1}(\|r_0\|_W e_1 - \widehat{T}_k y_k).$$

Since V_{k+1} is W-orthogonal, $V_{k+1}^T W V_{k+1} = I$ and so

$$\left\|r_k^W\right\|_W^2 = (\|r_0\|_W e_1 - \widehat{T}_k y_k)^T(\|r_0\|_W e_1 - \widehat{T}_k y_k) = \left\| \|r_0\|_W e_1 - \widehat{T}_k y_k \right\|_I^2$$

or

$$\left\|r_k^W\right\|_W = \left\| \|r_0\|_W e_1 - \widehat{T}_k y_k \right\|_I. \tag{3.18}$$

Consequently, minimizing the residual vector in $\|\cdot\|_W$ is equivalent to solving an I-norm least squares problem regardless of the original inner product $\langle \cdot, \cdot \rangle_W$. The implementation of a minimum residual method in a nonstandard inner product is straightforward, since the machinery of I-norm least squares problems can be used; weighted least-squares problems are never solved. We stress, however, that the actual entries of $\widehat{T}_k$ vary with the inner product.

As for W-CG, it would be advantageous to compute r_k^W without forming the matrices V_k and $\widehat{T}_k$, defined by (3.5) and (3.7). This is achieved at the expense of storing a small number of auxiliary scalars and vectors, as we now describe.

The tridiagonal matrix $\widehat{T}_k$ can be factored into the I-orthogonal $(k+1) \times (k+1)$ matrix Q_{k+1}^T and the $(k+1) \times k$ upper triangular matrix R_{k+1} by k Givens rotations. As a result, y_k can be expressed as

$$\begin{aligned} y_k = \text{argmin}_{y \in \mathbb{R}^k} \left\| \widehat{T}_k y - \beta e_1 \right\|_I &= \text{argmin}_{y \in \mathbb{R}^k} \left\| Q_{k+1}^T R_{k+1} y - \beta e_1 \right\|_I \\ &= \text{argmin}_{y \in \mathbb{R}^k} \left\| Q_{k+1}^T (R_{k+1} y - \beta Q_{k+1} e_1) \right\|_I \end{aligned}$$

so that, since Q_{k+1} is orthogonal,

$$y_k = \text{argmin}_{y \in \mathbb{R}^k} \left\| R_{k+1} y - \beta Q_{k+1} e_1 \right\|_I. \tag{3.19}$$

Let

$$R_{k+1} = \begin{bmatrix} R_{k \times k} \\ 0^T \end{bmatrix}, \quad \text{where} \quad R_{k \times k} = \begin{bmatrix} \delta_1 & \epsilon_1 & \zeta_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \zeta_{k-2} \\ & & & \ddots & \epsilon_{k-1} \\ & & & & \delta_k \end{bmatrix},$$

and let

$$c_{k+1} = \beta Q_{k+1} e_1 = \begin{bmatrix} c_{k \times 1} \\ a_k \end{bmatrix}, \tag{3.20}$$

where a_k is the $(k+1)$th entry of c_{k+1}. Then the solution of the least-squares problem (3.19) is

$$y_k = R_{k \times k}^{-1} c_{k \times 1}. \tag{3.21}$$

Furthermore, the QR factorization can be cheaply updated at each step. If the QR decomposition of $\widehat{T}_{k-1}$ is $\widehat{T}_{k-1} = Q_k R_k$, then

$$\begin{bmatrix} Q_k^T & 0 \\ 0^T & 1 \end{bmatrix} \widehat{T}_k = \begin{bmatrix} Q_k^T & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \widehat{T}_k & d_k \\ 0^T & \beta_k \end{bmatrix} = \begin{bmatrix} R_k & Q_k^T d_k \\ 0^T & \beta_k \end{bmatrix},$$

where $d_k = \alpha_k e_k + \beta_{k-1} e_{k-1}$, and only one Givens rotation, to zero out β_k, is needed to form the QR decomposition of $\widehat{T}_k$. The right-hand side vector c_{k+1} is obtained from c_k by altering only the last two elements.

Iterates x_k can be computed by short-term recurrences by introducing an additional set of vectors $p_0, \ldots, p_{k-1}$, where $P_k = \begin{bmatrix} p_0, \ldots, p_{k-1} \end{bmatrix} = V_k R_{k \times k}^{-1}$. To see this, note that

$$x_k = x_0 + V_k y_k = x_0 + P_k c_{k \times 1},$$

where we have used (3.21). It is clear from (3.20) that this can be split as

$$x_k = x_0 + P_{k-1} c_{k-1 \times 1} + a_k p_{k-1}.$$

Since $P_k = V_k R_{k\times k}^{-1}$ and $R_{k\times k}$ has only three nonzero diagonals, the auxiliary vectors satisfy the recurrence relationship

$$p_{k-1} = \frac{1}{\delta_k}(v_k - \epsilon_{k-1}p_{k-2} - \zeta_{k-1}p_{k-3}),$$

which shows that it is not necessary to store $\widehat{T}_k$ or V_k. Algorithm 3.4 implements W-MINRES with Givens rotations.

Algorithm 3.4 W-MINRES

$v_0 = 0, w_0 = 0, w_1 = 0$
Choose x_0, compute $v_1 = b - Bx_0$, set $\gamma_1 = \|v_1\|_W$
Set $\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 0$
for $k = 1, 2, \ldots$ **do**
  $v_k = v_k/\gamma_k$
  $\delta_k = \langle Bv_k, v_k\rangle_W$
  $v_{k+1} = Bv_k - \delta_k v_k - \gamma_k v_{k-1}$
  $\gamma_{k+1} = \|v_{k+1}\|_W$
  $\alpha_0 = c_k\delta_k - c_{k-1}s_k\gamma_k$
  $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{k+1}^2}$
  $\alpha_2 = s_k\delta_k + c_{k-1}c_k\gamma_k$
  $\alpha_3 = s_{k-1}\gamma_k$
  $c_{k+1} = \alpha_0/\alpha_1$, $s_{k+1} = \gamma_{k+1}/\alpha_1$
  $w_{k+1} = (v_k - \alpha_3 w_{k-1} - \alpha_2 w_k)/\alpha_1$
  $x_k = x_{k-1} + c_{k+1}\eta w_{k+1}$
  $\eta = -s_{k+1}\eta$
  <Test for convergence>
end for

If preconditioning is employed and $P^{-1}B$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_W$, then the ***W*-PMINRES** algorithm may be used. This method differs only slightly from Algorithm 3.4: v_1 is replaced by $v_1 = P^{-1}(b - Bx_0)$ and, at each iteration, $\delta_k = \langle P^{-1}Bv_k, v_k\rangle_W$ and

$$v_{k+1} = P^{-1}Bv_k - \delta_k v_k - \gamma_k v_{k-1}.$$

In Section 3.2.3 below we will discuss the convergence of W-PMINRES and will show that nicely distributed eigenvalues are sufficient for fast convergence.

Let us summarize the key properties of W-PMINRES, a three-term recurrence method that requires little change from the I-MINRES algorithm. Although W-PMINRES demands slightly more computation at each iteration than W-PCG it can be applied when $P^{-1}B$ is not W-positive definite, merely W-self-adjoint. Recall that these are precisely the matrices that are diagonalizable with real eigenvalues (see Corollary

2.21). Section 3.2.3 shows that its convergence depends largely on the eigenvalues of the preconditioned coefficient matrix. As discussed for W-PCG at the end of the previous section, these properties mean that a preconditioner that allows W-PMINRES to be applied with relatively cheap inner products can lead to efficient methods. Preconditioners of this sort for saddle point matrices are discussed in Chapters 4 and 5.

3.2.2 W-GMRES

The previous sections of this chapter show that when B is self-adjoint with respect to an inner product it is possible to devise an algorithm that satisfies an optimality condition and performs a fixed amount of work per step. Otherwise, one must generally choose between a method with an optimality property in a fixed inner product for which the work required grows at each step and a method for which the work at each iteration is constant but for which the optimality property is defined with respect to a variable metric. W-GMRES [45, 121], which is based on the I-GMRES algorithm of Saad and Schultz [119], falls into the first category. Like W-MINRES, it minimizes the W-norm of the residual vector so that

$$x_k \in x_0 + \mathcal{K}_k(B, r_0), \tag{3.22}$$

where

$$\|r_k\|_W = \min_{x_k \in x_0 + \mathcal{K}_k(B,r_0)} \|b - Bx_k\|_W . \tag{3.23}$$

The algorithm utilizes the W-orthogonal basis $\{v_i\}_{i=1,\ldots k}$ of $\mathcal{K}_k(B, r_0)$ generated by the ***W*-Arnoldi method** in Algorithm 3.5 for which we have the relationships

$$BV_k = V_{k+1}\widehat{H}_k \quad \text{and} \quad V_k^T W V_k = I, \tag{3.24}$$

where

$$\widehat{H}_k = \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1k} \\ h_{21} & h_{22} & \ldots & h_{2k} \\ & \ddots & \ddots & \vdots \\ \vdots & & h_{k,k-1} & h_{kk} \\ 0 & \ldots & 0 & h_{k+1,k} \end{bmatrix} \quad \text{and} \quad V_k = \begin{bmatrix} v_1, & v_2, & \ldots, & v_k \end{bmatrix}. \tag{3.25}$$

Algorithm 3.5 W-Arnoldi method

Choose $v_1 \in \mathbb{R}^n$ with $\|v_1\|_W = 1$
for $k = 1, 2, \ldots$ **do**
  $w = Bv_k$
  for $j = 1, 2, \ldots, k$ **do**
    $h_{jk} = \langle v_j, w \rangle_W$
    $w = w - h_{jk} v_j$
  end for
  $h_{k+1,k} = \|w\|_W$
  $v_{k+1} = w / h_{k+1,k}$
end for

Iterates satisfying (3.15), similarly to those of W-MINRES, are of the form $x_k = x_0 + V_k y_k$. As for W-MINRES, the kth W-GMRES residual satisfies

$$\left\|r_k^W\right\|_W = \min_{y \in \mathbb{R}^k} \|r_0 - BV_k y\|_W = \left\| \|r_0\|_W v_1 - V_{k+1} \widehat{H}_k y \right\|_W$$
$$= \left\| V_{k+1} (\|r_0\|_W e_1 - \widehat{H}_k y) \right\|_W .$$

Since V_{k+1} is W-orthonormal we again find that

$$\|r_k^W\|_W = \min_{y \in \mathbb{R}^k} \left\| \|r_0\|_W e_1 - \widehat{H}_k y \right\|_I , \tag{3.26}$$

from which $x_k = x_0 + V_k y_k$ can be found. As discussed for W-MINRES, it is advantageous that only I-norm least squares problems require solution because efficient methods for these are known. For example, (3.26) can be updated efficiently at each step by employing Givens rotations, similarly to W-MINRES. The W-GMRES algorithm is outlined in Algorithm 3.6.

Algorithm 3.6 W-GMRES

Choose x_0
Compute $r_0 = b - Bx_0$ and set $v_1 = r_0 / \|r_0\|_W$
for $k = 1, 2, \ldots$ **do**
  Do step k of the W-Arnoldi method
  Update the QR factorization of $\widehat{H}_k^W$ in (3.25)
  Solve $y = \arg\min \| \|r_0\|_W e_1 - \widehat{H}_k^W y \|_I$
  Set $x_k = x_0 + V_k y$
  <Test for convergence>
end for

The application of a left preconditioner requires only minor modifications to W-GMRES. In the ***W*-PGMRES method** the initial residual is replaced by the preconditioned residual $r_0 = P^{-1}(b - Bx_0)$ and at each step of the W-Arnoldi method (Algorithm 3.5) the vector w becomes $w = P^{-1}Bv_k$.

In summary, both W-GMRES and I-GMRES can be applied to any matrix. Moreover, the algorithms are similar, with the only additional work required by the former being the W-orthogonalization of the basis vectors. However, changing the inner product affects the convergence behaviour and convergence bounds. Some bounds may be more descriptive than others, a feature we exploit in Chapter 6 for W-PGMRES. Unfortunately, the inner products of interest will typically be too costly to find or compute with in practice. However, an understanding of the convergence of the nonstandard minimum residual method and relationships between the convergence of the nonstandard method and I-GMRES, will enable us to say something about convergence and preconditioners for I-GMRES.

3.2.3 Convergence of minimum residual methods

We now investigate the convergence of W-minimal residual (W-MR) Krylov subspace methods, which will prove useful when evaluating preconditioners. Only diagonalizable matrices are considered, since only these will be considered in subsequent chapters. The results are generalizations of convergence bounds for I-MR methods, which we first state.

If B is diagonalizable with diagonalization $B = Z\Lambda Z^{-1}$ then the kth residual of an I-MR method applied to B satisfies the **eigenvalue bound** [119]

$$\frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \kappa_I(Z) \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|. \tag{3.27}$$

Since the eigenvectors of a normal matrix can be chosen to be orthogonal, (3.27) tells us that convergence for normal matrices depends essentially on eigenvalues. However, if B is highly nonnormal, so that $\kappa_I(Z)$ is large, (3.27) may be pessimistic [62, page 56]. Furthermore, this bound is not applicable if B is not diagonalizable.

Pseudospectra, which are described in Appendix A.3, can be used to construct alternative bounds for minimal residual methods that address the drawbacks of (3.27). For any matrix B, we have the **pseudospectral bound** [143][145, Chapter 26]

$$\frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p(z)|, \tag{3.28}$$

where Γ_ϵ is a contour or union of contours enclosing the ϵ-pseudospectrum of B and L_ϵ is the arc length of Γ_ϵ. The advantage of (3.28) is that different bounds are obtained with different values of ϵ. However, it can be difficult to know a priori which values of ϵ give the most informative bounds.

The eigenvalue bound (3.27) and pseudospectral bound (3.28) are related to the **ideal GMRES polynomial** [65],

$$p_*(B) = \arg \min_{p\in\Pi_k, p(0)=1} \|p(B)\|_I. \tag{3.29}$$

The spectral norm of $p_*(B)$ is an upper bound on the **worst-case GMRES polynomial** which solves

$$\max_{\|r_0\|_I=1} \min_{p\in\Pi_k, p(0)=1} \|p(B)r_0\|_I.$$

Although there are matrices for which this inequality is strict [46, 141], for many matrices equality holds. These include normal matrices [63, 79], upper triangular Toeplitz matrices with $p_*(B) = 1$ [46] and matrices for which the ideal polynomial p_* has a simple maximal singular value [65]. Additionally, for arbitrary matrices equality holds at the first step ($k = 1$) [63, 79].

The worst-case GMRES polynomial and, therefore, the ideal GMRES polynomial, provide upper bounds on I-GMRES convergence since, for every r_0 with $\|r_0\|_I = 1$, the kth I-GMRES residual, r_k^I, satisfies

$$\|r_k^I\|_I = \min_{p\in\Pi_k, p(0)=1} \|p(B)r_0\|_I \le \max_{\|r_0\|_I=1} \min_{p\in\Pi_k, p(0)=1} \|p(B)r_0\|_I \le \min_{p\in\Pi_k, p(0)=1} \|p(B)\|_I.$$

The following generalizations of the pseudospectral and eigenvalue bounds for I-MR are similarly linked to ideal GMRES polynomials in a nonstandard inner product $\langle\cdot,\cdot\rangle_W$, i.e.,

$$p_*^W(B) = \arg \min_{p\in\Pi_k, p(0)=1} \|p(B)\|_W.$$

We now generalize the eigenvalue and pseudospectral I-MR bounds (3.27) and (3.28), starting with the former.

Theorem 3.8. *Let $B \in \mathbb{R}^{n\times n}$ have diagonalization $B = Z\Lambda Z^{-1}$, where $\Lambda, Z \in \mathbb{C}^{n\times n}$ and let $W = C^*C$ for some nonsingular matrix $C \in \mathbb{C}^{n\times n}$ for which W is real. Then r_k^W, the k^{th} residual vector of W-MR applied to $Bx = b$, $b \in \mathbb{R}^n$, satisfies*

$$\frac{\|r_k^W\|_W}{\|r_0\|_W} \le \kappa_I(CZ) \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|. \tag{3.30}$$

Proof. By definition, the W-MR residual vector is given by $r_k^W = p_k(B)r_0$, where $p_k(z)$ is the unique polynomial of degree less than or equal to k with $p(0) = 1$ that minimizes $\|r_k^W\|_W$. Using (2.2), it follows that

$$\begin{aligned}\|r_k^W\|_W = \|p_k(B)r_0\|_W &\le \|CZp_k(\Lambda)Z^{-1}C^{-1}\|_I\|r_0\|_W \\ &\le \|CZ\|_I\|(CZ)^{-1}\|_I \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|\|r_0\|_W.\end{aligned}$$

□

Remark 3.9. Theorem 3.8 is a generalization of the standard I-MR bound (3.27), since with $W = C = I$, (3.30) and (3.27) are identical.

Clearly $\kappa_I(CZ)$ is smaller for some matrices C than for others. Consequently, the bound (3.30) may be more descriptive for certain inner products $\langle \cdot, \cdot \rangle_W$. As stated in the following corollary, the smallest bound is obtained when $C = Z^{-1}$, so that $W = Z^{-*}Z^{-1}$. This matrix W defines an inner product with respect to which B is normal (see Lemma 2.29), so that the result is a generalization of a well known result for I-orthogonal matrices and I-MR methods; convergence for a normal matrix depends largely on eigenvalues.

Corollary 3.10. *Let $B \in \mathbb{R}^{n\times n}$ have diagonalization $B = Z\Lambda Z^{-1}$, $\Lambda, Z \in \mathbb{C}^{n\times n}$, so that B is normal with respect to $\langle \cdot, \cdot \rangle_W$, where $W = Z^{-*}Z^{-1}$. Then r_k^W, the k^{th} residual vector of W-MR applied to $Bx = b$, $b \in \mathbb{R}^n$, is bounded by*

$$\frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \min_{p\in\Pi_k,\, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|. \tag{3.31}$$

Remark 3.11. When B has real eigenvalues, B is not only W-normal but is also W-self-adjoint (see Corollary 2.28). In this case, we can apply W-MINRES to the linear system $Bx = b$, which requires only three-term recurrences.

We will use the idea behind Corollary 3.10 numerous times in subsequent chapters and so we recapitulate the key result. When $B = Z\Lambda Z^{-1}$, W-MR convergence, when measured in the W norm, depends primarily on the eigenvalues of B. We note that the eigenvectors still influence convergence, as these define the W-norm.

If the coefficient matrix B is real, it may be desirable to avoid complex arithmetic in Theorem 3.8 and Corollary 3.10. Both results can be alternatively derived using real arithmetic by employing the real Jordan form outlined in Definition A.1. To give some flavour of how this is done we provide an alternative result to Corollary 3.10, which will be used in Section 6.3.1.

Corollary 3.12. *Let $\lambda_1, \ldots, \lambda_j$ be the real eigenvalues of $B \in \mathbb{R}^{n\times n}$ and let $\lambda_{j+1} = \mu_j + i\nu_j, \overline{\lambda}_{j+1} = \mu_{j+1} - i\nu_{j+1}, \ldots, \lambda_k = \mu_k + i\nu_k, \overline{\lambda}_k = \mu_k - i\nu_k$ be the $2(k-j)$ nonreal eigenvalues. Additionally, let B have real Jordan form $B = R\mathcal{J}R^{-1}$, $R \in \mathbb{R}^{n\times n}$, where $\mathcal{J} \in \mathbb{R}^{n\times n}$ is of the form*

$$\mathcal{J} = \operatorname{diag}(\mathcal{J}_1, \ldots, \mathcal{J}_j, \mathcal{J}_{j+1}, \ldots, \mathcal{J}_k)$$

with

$$\mathcal{J}_\ell = \begin{cases} \lambda_\ell & \ell = 1, \ldots, j \\ \begin{bmatrix} \mu_\ell & \nu_\ell \\ -\nu_\ell & \mu_\ell \end{bmatrix} & \ell = j+1, \ldots, k. \end{cases}$$

If $W = R^{-T}R^{-1}$, then the kth residual r_k^W of W-MR applied to $Bx = b$, $b \in \mathbb{R}^n$, satisfies

$$\begin{aligned} \|r_k^W\|_W = \|p_k(B)r_0\|_W &\le \|R^{-1}Rp_k(\mathcal{J})R^{-1}R\|_I \|r_0\|_W \\ &= \min_{p\in\Pi_k, p(0)=1} \max_{\ell=1,\ldots,k} \|p(\mathcal{J}_\ell)\| \|r_0\|_W. \end{aligned} \tag{3.32}$$

We choose to use the (possibly complex) canonical Jordan form over the real Jordan form except where it is easier to avoid complex arithmetic.

One advantage of (3.28) over (3.27) is that different ϵ lead to different bounds. The same is true of pseudospectral bounds for W-MR methods.

Theorem 3.13. *Let $W = C^*C$ for some nonsingular matrix C for which W is real. Then r_k^W, the k^{th} residual vector of W-MR applied to $Bx = b$, $b \in \mathbb{R}^n$, satisfies*

$$\frac{\|r_k^W\|_W}{\|r_0\|_W} \le \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|, \tag{3.33}$$

where Γ_ϵ is a contour or union of contours enclosing the ϵ-pseudospectra of CBC^{-1} and L_ϵ is the arc length of Γ_ϵ.

Proof. We know that

$$\|r_k^W\|_W = \|p_k(B)r_0\|_W \le \|p_k(B)\|_W \|r_0\|_W,$$

where p_k is the unique polynomial of degree k with $p(0) = 1$ for which $\|p_k(B)r_0\|_W$ is minimized. The polynomial term $\|p_k(B)\|_W\|r_0\|_W$ can be represented by a Cauchy integral to give that, for any contour Γ enclosing the spectrum of B,

$$\begin{aligned} \|p_k(B)\|_W &= \left\| \frac{1}{2\pi i} \int_\Gamma p_k(z)(zI - B)^{-1} dz \right\|_W \\ &\le \frac{1}{2\pi} \int_\Gamma |p_k(z)| \left\|(zI - B)^{-1}\right\|_W |dz| \\ &= \frac{1}{2\pi} \int_\Gamma |p_k(z)| \left\|(zI - CBC^{-1})^{-1}\right\|_I |dz|. \end{aligned}$$

Comparing with (A.1) in Appendix A.3 we see the by choosing $\Gamma = \Gamma_\epsilon$ we obtain

$$\|p_k(B)\|_W \le \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|,$$

which is the result. □

As for the eigenvalue bound in Theorem 3.8, the pseudospectral bound (3.33) depends on C. The choice $C = Z^{-1}$, where Z is an eigenvector matrix of B gave the bound in Corollary 3.10. The same choice of B here means that CBC^{-1} is diagonal. This choice of C corresponds to $W = Z^{-*}Z^{-1}$, which again defines an inner product with respect to which B is normal. Accordingly, the ϵ-pseudospectrum of B with respect to $\|\cdot\|_W$ consists of open ϵ-balls around the eigenvalues of B (see Lemma A.9).

Corollary 3.14. *Let $B \in \mathbb{R}^{n\times n}$ be diagonalizable with matrix of eigenvectors $Z \in \mathbb{C}^{n\times n}$ and let $W = Z^{-*}Z^{-1}$. Then, r_k^W, the k^{th} residual vector of W-MR applied to $Bx = b$, $b \in \mathbb{R}^n$, is bounded by*

$$\frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|, \tag{3.34}$$

where Γ_ϵ is any curve or union of curves that encloses $\Lambda(B)+\triangle_\epsilon$, where $\triangle_\epsilon$ is defined by (A.4) *in Appendix* A.3*, and L_ϵ is the arc length of Γ_ϵ.*

Proof. As in the proof of Theorem 3.13, with $C = Z^{-1}$, where Z diagonalizes B, we have that

$$\|p_k(B)\|_W \leq \frac{1}{2\pi}\int_\Gamma |p_k(z)|\|(zI-B)^{-1}\|_W|dz|.$$

By choosing $\Gamma = \Gamma_\epsilon$ and using Lemma A.9 we obtain the result. □

Corollary 3.14 again shows that, since B is W-normal, it is the eigenvalues of B that largely determine the convergence of this W-MR method, when measured with respect to the W-norm. This idea is used throughout the remainder of this thesis. We stress that when B is diagonalizable with real eigenvalues, B is W-self-adjoint (and, trivially W-normal) and in this case a three-term recurrence method—W-MINRES—can be used to solve the linear system.

The focus of subsequent chapters is on choosing good preconditioners for use with Krylov subspace methods. To do so, it is helpful to understand the factors that drive convergence of W-PMINRES and W-PGMRES. Let us, therefore, consider how Corollaries 3.10 and 3.14 apply to left preconditioned linear systems.

Corollary 3.15. *Let P and B be real $n \times n$ matrices and let $P^{-1}B = Z\Lambda Z^{-1}$ be diagonalizable, where $\Lambda, Z \in \mathbb{C}^{n\times n}$. If $W = Z^{-*}Z^{-1}$, then the preconditioned W-MR residuals $P^{-1}r_k^W = P^{-1}(b-Bx_k^W)$ for the system $P^{-1}Bx = P^{-1}b$, $b \in \mathbb{R}^n$, are bounded by*

$$\frac{\|P^{-1}r_k^W\|_W}{\|P^{-1}r_0\|_W} \leq \min_{p\in\Pi_k,\, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)| \tag{3.35}$$

and by

$$\frac{\|P^{-1}r_k\|_W}{\|P^{-1}r_0\|_W} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|, \tag{3.36}$$

where Γ_ϵ *is any curve or union of curves that encloses* $\Lambda(P^{-1}B) + \triangle_\epsilon$, $\triangle_\epsilon$ *is defined by* (A.4) *in Appendix* A.3, *and* L_ϵ *is the arc length of* Γ_ϵ.

When $P^{-1}B = Z\Lambda Z^{-1}$ is diagonalizable it is M-self-adjoint and W-normal. Thus, convergence of W-PMINRES or W-PGMRES depends only on the eigenvalues of $P^{-1}B$. A sufficient condition for fast convergence is that polynomials can be found that are small at all of the eigenvalues, which will occur when the eigenvalues are nicely distributed. Corollary 3.15 provides some guidance for choosing preconditioners for saddle point problems in Chapters 4 and 5 and aids in understanding preconditioners for nonsymmetric matrices in Chapter 6. When $P^{-1}B$ is self-adjoint with respect to an inner product rather than a symmetric bilinear form, i.e., when it is diagonalizable with real eigenvalues, the preconditioned linear system can be solved by a three-term recurrence MINRES method.

One can develop similar bounds for right preconditioned systems.

Corollary 3.16. *Let* $P \in \mathbb{R}^{n\times n}$ *and* $B \in \mathbb{R}^{n\times n}$ *and let* $BP^{-1} = Z\Lambda Z^{-1}$ *be diagonalizable, where* $\Lambda, Z \in \mathbb{C}^{n\times n}$. *If* $W = Z^{-*}Z^{-1}$, *then the* W*-MR residuals* $r_k^W = b - Bx_k^W$ *for the system* $BP^{-1}y = b$, $y = Px$, $b \in \mathbb{R}^n$, *are bounded by*

$$\frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \min_{p\in\Pi_k,\, p(0)=1} \max_{\lambda\in\Lambda(BP^{-1})} |p(\lambda)| \tag{3.37}$$

and by

$$\frac{\|r_k\|_W}{\|r_0\|_W} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|, \tag{3.38}$$

where Γ_ϵ *is any curve or union of curves that encloses* $\Lambda(BP^{-1}) + \triangle_\epsilon$, $\triangle_\epsilon$ *is defined by* (A.4) *in Appendix* A.3, *and* L_ϵ *is the arc length of* Γ_ϵ.

Whether it is useful to apply a preconditioned method in a nonstandard inner product is determined by the availability of an appropriate inner product with which it is cheap to compute. Even when the inner product W is not explicitly known, or when computations with $\langle\cdot,\cdot\rangle_W$ are impractical, the viewpoint provided by Krylov methods in nonstandard inner products can give some insight into the convergence of, and good preconditioners for, standard Krylov methods. This is discussed in Chapter 6.

3.2.4 Relationships between the convergence of minimum residual methods in different inner products

Recall that if $B = Z\Lambda Z^{-1}$ is diagonalizable then it is M-self-adjoint, where $M = Z^{-*}YZ^{-1}$ and W-normal, where $W = Z^{-*}Z^{-1}$. By Corollaries 3.10 and 3.14, the convergence of a W-MR method applied to a linear system with B as its coefficient matrix is bounded by a term that only involves the eigenvalues of B. These results are generally descriptive in our experience and are useful when choosing preconditioners. However, it can be infeasible to obtain and use the inner product $\langle \cdot, \cdot \rangle_W$. Fortunately, the convergence of minimum residual methods in different inner products can be related. One such link was proposed by Essai [45], who showed that the convergence of GMRES and FOM in a nonstandard inner product—defined by a diagonal matrix prescribed by the initial residual—could be related to the convergence of GMRES and FOM in the Euclidean inner product. The key to Essai's result is a QR decomposition in a nonstandard inner product, which we employ below and which has also been exploited by, for example, Sadok and Szyld [120] in their comparison of CMRH and GMRES, by Philippe and Reichel to implement the Arnoldi process with Chebyshev and Newton polynomial bases [110] and by Simoncini and Szyld [128, Theorem 3.6] to relate W-MR convergence, in any inner product $\langle \cdot, \cdot \rangle_W$, to I-MR convergence. Approaches that do not explicitly exploit a QR relationship can also be used to connect W-MR and I-MR convergence as shown in [39, Section 4.3]. Similar results for conjugate gradient methods are given in [5].

To describe how the convergence of a minimum residual method in a nonstandard inner product $\langle \cdot, \cdot \rangle_W$ and that of a minimum residual method in the Euclidean inner product compare we extend the results of Essai. All of these results can be generalized to compare two methods in nonstandard inner products, say $\langle \cdot, \cdot \rangle_W$ and $\langle \cdot, \cdot \rangle_X$, but for clarity we do not do this here. The analysis is restricted to MR methods for which the basis vectors are orthogonal with respect to the associated inner product in the sense of (2.6). However, the most popular implementations of minimum residual methods—MINRES and GMRES—satisfy this requirement. Without loss of generality, we assume in the following that the initial iterate is the zero vector[2].

A W-MR method selects iterates $x_k \in \mathcal{K}_k(B, r_0)$ for which

$$x_k^W = \arg \min_{x \in x_0 + \mathcal{K}_k(B, r_0)} \|b - Bx\|_W.$$

[2]If $x_0 \neq 0$ we can always solve the shifted linear system $B(x - x_0) = b - Bx_0 = r_0$, from which x can be recovered.

Similarly, iterates $x_k \in \mathcal{K}_k(B, r_0)$ of an I-MR method are chosen so that

$$x_k^I = \arg \min_{x \in x_0 + \mathcal{K}_k(B, r_0)} \|b - Bx\|_I.$$

The Arnoldi relations (3.24) for W-MR and I-MR are

$$BV_k^W = V_{k+1}^W \widehat{H}_k^W \tag{3.39}$$

and

$$BV_k^I = V_{k+1}^I \widehat{H}_k^I, \tag{3.40}$$

where the columns of V_k^W and V_k^I are the W-orthogonal basis vectors for W-MR and the I-orthogonal basis vectors for I-MR, respectively, and $\widehat{H}_k^W$ and $\widehat{H}_k^I$ are $(k+1) \times k$ upper Hessenberg matrices.

The norms of the W-MR and I-MR residuals can be expressed as

$$\|r_k^W\|_W = \min_{z \in \mathbb{R}^k} \|\|r_0\|_W e_1 - \widehat{H}_k^W z\|_I, \tag{3.41}$$

and

$$\|r_k^I\|_I = \min_{y \in \mathbb{R}^k} \|\|r_0\|_I e_1 - \widehat{H}_k^I y\|_I \tag{3.42}$$

by using (3.18) for MINRES or (3.26) for GMRES.

Although the basis vectors of W-MR and I-MR differ, W-MR and I-MR form bases of the same nested Krylov subspaces. Thus, assuming that these methods do not terminate at or before the k^{th} step,

$$V_k^I = V_k^W R_k, \tag{3.43}$$

where $R_k \in \mathbb{R}^{k \times k}$ is upper triangular and nonsingular; this is a QR decomposition of V_k^I in the $\langle \cdot, \cdot \rangle_W$ inner product. From this fundamental equation we can obtain useful relationships between V_k^W and V_k^I, and between $\widehat{H}_k^W$ and $\widehat{H}_k^I$.

Lemma 3.17. (Proposition 1 in [45]) *Let $\langle \cdot, \cdot \rangle_W$ be an inner product. Additionally, let $V_k^W \in \mathbb{R}^{n \times k}$ and $\widehat{H}_k^W \in \mathbb{R}^{(k+1) \times k}$ satisfy* (3.39) *and $V_k^I \in \mathbb{R}^{n \times k}$ and $\widehat{H}_k^I \in \mathbb{R}^{(k+1) \times k}$ satisfy* (3.40)*, with V_k^W having W-orthogonal columns and V_k^I having I-orthogonal columns. Then, if*

$$V_k^I = V_k^W R_k$$

for some nonsingular matrix $R_k \in \mathbb{R}^{k \times k}$, it follows that

$$R_k = (V_k^W)^T W V_k^I, \quad R_k^{-1} = (V_k^I)^T V_k^W \quad \textit{and} \quad \widehat{H}_k^I = R_{k+1}^{-1} \widehat{H}_k^W R_k. \tag{3.44}$$

Proof. Since V_k^W is W-orthogonal, premultiplying both sides of (3.43) by $(V_k^W)^T W$ gives that

$$R_k = (V_k^W)^T W V_k^I.$$

An equation for R_k^{-1} is derived by instead premultiplying (3.43) by $(V_k^I)^T$ to obtain that $I = (V_k^I)^T V_k^W R_k$ or that

$$R_k^{-1} = (V_k^I)^T V_k^W. \tag{3.45}$$

To relate $\widehat{H}_k^I$ and $\widehat{H}_k^W$ note that, by (3.43), $BV_k^W = BV_k^I R_k^{-1}$ so that (3.39) can be expressed as $BV_k^I = V_{k+1}^W \widehat{H}_k^W R_k$. Substituting this into (3.40) shows that $V_{k+1}^I \widehat{H}_k^I = V_{k+1}^W \widehat{H}_k^W R_k$ and after premultiplying both sides by $(V_{k+1}^I)^T$ we obtain that

$$\widehat{H}_k^I = (V_{k+1}^I)^T V_{k+1}^W \widehat{H}_k^W R_k = R_{k+1}^{-1} \widehat{H}_k^W R_k,$$

where we have used (3.45) to simplify the right-hand side. □

By utilizing these relationships, it is possible to express I-MR residuals in terms of quantities associated with W-MR as the following lemma, which generalizes a result in [45, Section 4], demonstrates.

Lemma 3.18. *Let r_k^I be the k^{th} residual of I-MR. Then*

$$\|r_k^I\|_I = \min_{\bar{y} \in \mathbb{R}^k} \| \|r_0\|_W e_1 - \widehat{H}_k^W \bar{y} \|_{R_{k+1}^{-T} R_{k+1}^{-1}}, \tag{3.46}$$

where $\widehat{H}_k^W$ is the upper Hessenberg matrix in (3.39) *and R_{k+1} is the upper triangular matrix in* (3.43)*.*

Proof. The I-MR residual satisfies

$$\|r_k^I\|_I = \min_{y \in \mathbb{R}^k} \left\| \|r_0\|_I e_1 - \widehat{H}_k^I y \right\|_I = \min_{\bar{y} \in \mathbb{R}^k} \left\| \|r_0\|_I e_1 - \widehat{H}_k^I R_k^{-1} \bar{y} \right\|_I.$$

Now,

$$r_0 = \|r_0\|_I V_{k+1}^I e_1 = \|r_0\|_W V_{k+1}^W e_1$$

and so, since V_{k+1}^I is I-orthogonal,

$$\|r_0\|_I e_1 = (V_{k+1}^I)^T \|r_0\|_W V_{k+1}^W e_1 = \|r_0\|_W R_{k+1}^{-1} e_1 \tag{3.47}$$

by Lemma 3.17. We also have from Lemma 3.17 that

$$\widehat{H}_k^I = R_{k+1}^{-1} \widehat{H}_k^W R_k. \tag{3.48}$$

Substituting (3.47) and (3.48) into (3.47) gives that

$$\|r_k^I\|_I = \min_{\bar{y}\in\mathbb{R}^k} \left\| R_{k+1}^{-1}(\|r_0\|_W e_1 - \widehat{H}_k^W \bar{y}) \right\|_I = \min_{\bar{y}\in\mathbb{R}^k} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W \bar{y} \right\|_{R_{k+1}^{-T}R_{k+1}^{-1}}$$

which is the stated result. □

Comparison with the W-MR residual (3.41) shows that the quantities minimized by I-MR and W-MR are the same but the norms with respect to which they are minimized differ. This is the key relationship that allows us to link I-MR and W-MR convergence in the following theorem.

Theorem 3.19. *Let W be symmetric positive definite. Then, the k^{th} I-MR residual r_k^I and the k^{th} W-MR residual r_k^W satisfy*

$$\sigma_{min}(R_{k+1}^{-1})\|r_k^W\|_W \leq \|r_k^I\|_I \leq \sigma_{max}(R_{k+1}^{-1})\|r_k^W\|_W, \tag{3.49}$$

where R_{k+1} is defined in (3.43).

Proof. Let

$$z_k = \arg\min_{z\in\mathbb{R}^k} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W z \right\|_I.$$

Then, from (3.46),

$$\|r_k^I\|_I = \min_{\bar{y}\in\mathbb{R}^k} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W \bar{y} \right\|_{R_{k+1}^{-T}R_{k+1}^{-1}} \leq \left\| \|r_0\|_W e_1 - \widehat{H}_k^W z_k \right\|_{R_{k+1}^{-T}R_{k+1}^{-1}}$$

since $\bar{y}$ achieves the minimum value. Accordingly,

$$\|r_k^I\|_I \leq \left\| \|r_0\|_W e_1 - \widehat{H}_k^W z_k \right\|_{R_{k+1}^{-T}R_{k+1}^{-1}} \leq \sqrt{\lambda_{max}(R_{k+1}^{-T}R_{k+1}^{-1})} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W z_k \right\|_I,$$

where the last step follows from application of the Courant-Fischer theorem (see Appendix A.4) to $R_{k+1}^{-T}R_{k+1}^{-1}$. Comparison with (3.41) shows that this is equivalent to

$$\|r_k^I\|_I \leq \sqrt{\lambda_{max}(R_{k+1}^{-T}R_{k+1}^{-1})}\|r_k^W\|_W = \sigma_{max}(R_{k+1}^{-1})\|r_k^W\|_W.$$

To obtain the lower bound, we first let

$$y_k = \arg\min_{\bar{y}\in\mathbb{R}^k} \left\| \|r_0\|_I e_1 - \widehat{H}_k^I \bar{y} \right\|_{R_{k+1}^{-T}R_{k+1}^{-1}}$$

Then, from (3.41),

$$\|r_k^W\|_W = \min_{z\in\mathbb{R}^k} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W z \right\|_I \leq \left\| \|r_0\|_W e_1 - \widehat{H}_k^W y_k \right\|_I$$

and by again employing the Courant-Fischer theorem we find that

$$\|r_k^W\|_W \leq \left\| \|r_0\|_W e_1 - \widehat{H}_k^W y_k \right\|_I \leq \frac{1}{\sigma_{min}(R_{k+1}^{-1})} \left\| \|r_0\|_W e_1 - \widehat{H}_k^W y_k \right\|_{R_{k+1}^{-T} R_{k+1}^{-1}}.$$

Lemma 3.18 states that

$$\|r_k^I\|_I = \left\| \|r_0\|_W e_1 - \widehat{H}_k^W y_k \right\|_{R_{k+1}^{-T} R_{k+1}^{-1}}$$

and so

$$\sigma_{min}(R_{k+1}^{-1}) \|r_k^W\|_W \leq \|r_k^I\|_I,$$

which is the lower bound. □

However, Theorem 3.19 depends on the singular values of R_{k+1}, the computation of which requires both W-MR and I-MR basis vectors. This motivates us to consider alternative representations of (3.49). The first of these is given in the following corollary.

Corollary 3.20. *If $W = C^*C$, where $C \in \mathbb{C}^{n\times n}$ and $W \in \mathbb{R}^{n\times n}$, the k^{th} I-MR residual r_k^I and the k^{th} W-MR residual r_k^W satisfy*

$$\frac{1}{\sigma_{max}(CV_{k+1}^I)} \|r_k^W\|_W \leq \|r_k^I\|_I \leq \frac{1}{\sigma_{min}(CV_{k+1}^I)} \|r_k^W\|_W, \tag{3.50}$$

while the corresponding relative residuals satisfy

$$\frac{\sigma_{min}(C)}{\sigma_{max}(CV_{k+1}^I)} \frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \frac{\sigma_{max}(C)}{\sigma_{min}(CV_{k+1}^I)} \frac{\|r_k^W\|_W}{\|r_0\|_W}. \tag{3.51}$$

Proof. From (3.43) and the W-orthogonality of V_{k+1}^W,

$$(V_{k+1}^I)^T W V_{k+1}^I = R_{k+1}^T (V_{k+1}^W)^T W V_{k+1}^W R_{k+1} = R_{k+1}^T R_{k+1}.$$

Taking inverses establishes that

$$\left[(V_{k+1}^I)^T W V_{k+1}^I\right]^{-1} = R_{k+1}^{-1} R_{k+1}^{-T}$$

from which it is clear that the eigenvalues of $\left[(V_{k+1}^I)^T W V_{k+1}^I\right]^{-1}$ and the singular values of R_{k+1}^{-1} are related by

$$\lambda\left(\left[(V_{k+1}^I)^T W V_{k+1}^I\right]^{-1}\right) = \lambda(R_{k+1}^{-1} R_{k+1}^{-T}) = \sigma(R_{k+1}^{-1})^2.$$

We therefore look for an expression for the eigenvalues of $\left[(V_{k+1}^I)^T W V_{k+1}^I\right]^{-1}$ in order to re-write (3.49). Since $W = C^*C$ we have that

$$R_{k+1}^T R_{k+1} = (V_{k+1}^I)^T W V_{k+1}^I = (V_{k+1}^I)^T C^* C V_{k+1}^I = (CV_{k+1}^I)^*(CV_{k+1}^I).$$

Thus,

$$\sigma_k(R_{k+1}^{-1})^2 = \lambda_k\left([(V_{k+1}^I)^T W V_{k+1}^I]^{-1}\right) = \frac{1}{\lambda_{n-k}((V_{k+1}^I)^T W V_{k+1}^I)} = \frac{1}{\sigma_{n-k}^2(CV_{k+1}^I)}.$$

In particular,

$$\sigma_{min}(R_{k+1}^{-1}) = \frac{1}{\sigma_{max}(CV_{k+1}^I)} \quad \text{and} \quad \sigma_{max}(R_{k+1}^{-1}) = \frac{1}{\sigma_{min}(CV_{k+1}^I)}$$

so that (3.49) is equivalent to

$$\frac{1}{\sigma_{max}(CV_{k+1}^I)}\|r_k^W\|_W \leq \|r_k^I\|_I \leq \frac{1}{\sigma_{min}(CV_{k+1}^I)}\|r_k^W\|_W.$$

The relative residual bound is obtained from (3.50) by dividing by $\|r_0\|_I$, which gives that

$$\frac{1}{\sigma_{max}(CV_{k+1}^I)}\frac{\|r_k^W\|_W}{\|r_0\|_I} \leq \frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \frac{1}{\sigma_{min}(CV_{k+1}^I)}\frac{\|r_k^W\|_W}{\|r_0\|_I}.$$

By the Courant-Fischer theorem (see Appendix A.4),

$$\lambda_{min}(W)\|r_0\|_I^2 \leq \|r_0\|_W^2 \leq \lambda_{max}(W)\|r_0\|_I^2$$

or

$$\sigma_{min}(C)\frac{1}{\|r_0\|_W} \leq \frac{1}{\|r_0\|_I} \leq \sigma_{max}(C)\frac{1}{\|r_0\|_W}$$

and so

$$\frac{\sigma_{min}(C)}{\sigma_{max}(CV_{k+1}^I)}\frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \frac{\sigma_{max}(C)}{\sigma_{min}(CV_{k+1}^I)}\frac{\|r_k^W\|_W}{\|r_0\|_W}.$$

□

Of particular interest in this thesis is the inner product $W = Z^{-*}Z^{-1}$, where B has diagonalization $B = Z\Lambda Z^{-1}$. With this choice, Corollary 3.20 leads to an iteration-dependent bound for the relative residuals of I-MR that can be much more descriptive than the standard eigenvalue bound (3.27).

Corollary 3.21. *If $B \in \mathbb{R}^{n\times n}$ has diagonalization $B = Z\Lambda Z^{-1}$, $\Lambda, Z \in \mathbb{C}^{n\times n}$, then I-MR residuals r_k^I for the system $Bx = b$, $b \in \mathbb{R}^n$, satisfy*

$$\frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \frac{\sigma_{max}(Z^{-1})}{\sigma_{min}(Z^{-1}V_{k+1}^I)} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|. \tag{3.52}$$

Proof. Applying (3.50) with $W = Z^{-*}Z^{-1}$ gives that

$$\|r_k^I\|_I \leq \frac{1}{\sigma_{min}(Z^{-1}V_{k+1}^I)}\|r_k^W\|_W. \tag{3.53}$$

This choice of W also allows us to bound the W-MR residuals using Corollary 3.10 and we find that

$$\begin{aligned}\|r_k^I\|_I &\leq \frac{1}{\sigma_{min}(Z^{-1}V_{k+1}^I)} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|\|r_0\|_W \\ &\leq \frac{\sigma_{max}(Z^{-1})}{\sigma_{min}(Z^{-1}V_{k+1}^I)} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|\|r_0\|_I\end{aligned} \tag{3.54}$$

since $\|r_0\|_W^2 = r_0^T Z^{-*}Z^{-1}r_0 \leq \sigma_{max}^2(Z^{-1})^2\|r_0\|_I^2$. □

It is possible to obtain a tighter bound than (3.52) in Corollary 3.21[3] by starting from (3.54) and applying the lower bound in (3.50) to $\|r_0\|_W$, to obtain that $\|r_0\|_W \leq \sigma_{max}(Z^{-1}V_1^I)\|r_0\|_I$. Doing so gives that

$$\|r_k^I\|_I \leq \frac{\sigma_{max}(Z^{-1}V_1)}{\sigma_{min}(Z^{-1}V_{k+1})} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|\|r_0\|_I. \tag{3.55}$$

By Corollary A.12 in Appendix A a slightly weaker, but more elegant bound than (3.55), is[3]

$$\begin{aligned}\frac{\|r_k^I\|_I}{\|r_0\|_I} &\leq \frac{\sigma_{max}(Z^{-1}V_{k+1})}{\sigma_{min}(Z^{-1}V_{k+1})} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)| \\ &= \kappa_I(Z^{-1}V_{k+1}) \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|.\end{aligned} \tag{3.56}$$

From Corollary A.12 in Appendix A it is clear that $\sigma_{min}(Z^{-1}) \leq \sigma_{min}(Z^{-1}V_{k+1}^I)$ and $\sigma_{max}(Z^{-1}V_{k+1}^I) \leq \sigma_{min}(Z^{-1})$. Thus, comparing (3.52) in Corollary 3.21 and (3.56) to (3.27), the standard eigenvalue bound for I-MR, we see that the former two are tighter when $\sigma_{min}(Z^{-1})$ is smaller than $\sigma_{min}(Z^{-1}V_{k+1}^I)$ and can never be larger. If the difference is pronounced, (3.52) may be a significantly better predictor of convergence.

We now present the second of our alternatives to Theorem 3.19, which will be useful when the singular values of V_{k+1}^W can be obtained or estimated.

Corollary 3.22. *Let W be symmetric positive definite. Then, the k^{th} I-MR residual r_k^I and the k^{th} W-MR residual r_k^W satisfy*

$$\sigma_{min}(V_{k+1}^W)\|r_k^W\|_W \leq \|r_k^I\|_I \leq \sigma_{max}(V_{k+1}^W)\|r_k^W\|_W. \tag{3.57}$$

[3]D. Titley-Peloquin, personal communication, June 2011.

Proof. If R_{k+1}^{-1} has as its singular value decomposition $R_{k+1}^{-1} = U\Sigma V^T$ then

$$V_{k+1}^W = V_{k+1}^I R_{k+1}^{-1} = (V_{k+1}^I U)\Sigma V^T.$$

Since V_{k+1}^I and U are each I-orthogonal, so is their product. Thus, $(V_{k+1}^I U)\Sigma V^T$ is a reduced SVD of V_{k+1}^W and the singular values of R_{k+1}^{-1} are equal to the singular values of V_{k+1}^W. It follows that (3.49) can be replaced by

$$\sigma_{min}(V_{k+1}^W)\|r_k^W\|_W \leq \|r_k^I\|_I \leq \sigma_{max}(V_{k+1}^W)\|r_k^W\|_W.$$

□

The bounds (3.49), (3.50) and (3.57) depend on the iteration number, k. The following is independent of k but—although there could be certain right-hand sides for which it is tight—it may be pessimistic in general particularly when k is small. This topic is discussed in [128].

Theorem 3.23. *Let W be symmetric positive definite. Then, the residuals of I-MR and W-MR satisfy*

$$\frac{1}{\sqrt{\lambda_{max}(W)}}\|r_k^W\|_W \leq \|r_k^I\|_I \leq \frac{1}{\sqrt{\lambda_{min}(W)}}\|r_k^W\|_W, \tag{3.58}$$

while the corresponding relative residuals satisfy

$$\frac{1}{\sqrt{\kappa_I(W)}}\frac{\|r_k^W\|_W}{\|r_0\|_W} \leq \frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \sqrt{\kappa_I(W)}\frac{\|r_k^W\|_W}{\|r_0\|_W}, \tag{3.59}$$

where $\lambda_{min}(W)$ and $\lambda_{max}(W)$ are the minimum and maximum eigenvalues of W and $\kappa_I(W)$ is its I-norm condition number.

Proof. Let the positive square root of $W = S^2$ be S. Then for any vector $y \in \mathbb{R}^n$,

$$\begin{aligned}\|y\|_I^2 = y^T y = y^T S S^{-2} S y &\leq \lambda_{max}(S^{-2})\|Sy\|_2^2 \\ &= \lambda_{max}(W^{-1})\|y\|_W^2 \\ &= \frac{1}{\lambda_{min}(W)}\|y\|_W^2.\end{aligned}$$

The lower bound is obtained analogously.

The relative residual bounds are obtained similarly to those in Corollary 3.20. □

Remark 3.24. When $B = Z\Lambda Z^{-1}$ and $W = Z^{-*}Z^{-1}$, combining (3.59) with the bound in Corollary 3.10 gives that

$$\frac{\|r_k^I\|_I}{\|r_0\|_I} \leq \sqrt{\kappa_I(W)} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)| = \kappa_I(Z) \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|,$$

the standard eigenvalue bound (3.27) for GMRES. This shows that (3.59) in Theorem 3.23 is a generalization of (3.27).

We illustrate by an example the difference between the standard bound (3.27) (that can be obtained from (3.59) as discussed in the remark above) and the iteration-dependent bound (3.52). We consider a finite difference discretization of a two-dimensional convection-diffusion operator (the unpreconditioned linear system in Example 6.2 in Section 6.2.2). Discretizations of this operator have been examined extensively in the literature, with a Fourier analysis given in [42] and studies of GMRES convergence of an SUPG finite element discretization the subject of [36, 42, 50, 92]. We choose $n_x = 7$, $n_y = 28$ and $\epsilon = 0.0325$. Applying GMRES to the linear system gives the solid convergence curve in Figure 3.1.

The size of the matrix makes solving

$$\min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)|$$

difficult but the eigenvalues are all real and lie in $(17.8, 217.5)$. We therefore take the approximation

$$\min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(B)} |p(\lambda)| \leq \min_{p\in\Pi_k, p(0)=1} \max_{z\in[\lambda_{min},\lambda_{max}]} |p(z)| \leq 2\left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^k, \qquad (3.60)$$

where λ_{min} and λ_{max} are the extreme eigenvalues of B and $\eta = \frac{\lambda_{max}}{\lambda_{min}}$ (see, for example, [118, Section 6.11.3] for details).

Using (3.60) in the standard eigenvalue bound (3.27) for I-GMRES convergence gives the green dashed line in Figure 3.1, which clearly overestimates the relative residuals. The bound (3.52) is denoted by the red dot-dashed line and (3.55) is given by the cyan line, where (3.60) has again been used to approximate the polynomial term. Both are much tighter than the standard eigenvalue bound (3.27) and, significantly, both capture the end of the stagnation phase. We note that (3.52) and (3.55) are close, although this will not be the case if $\sigma_{max}(Z^{-1})$ and $\sigma_{max}(Z^{-1}V_1)$ differ greatly. This example, and the examples we present in Chapter 6 highlight that (3.27) can be misleading.

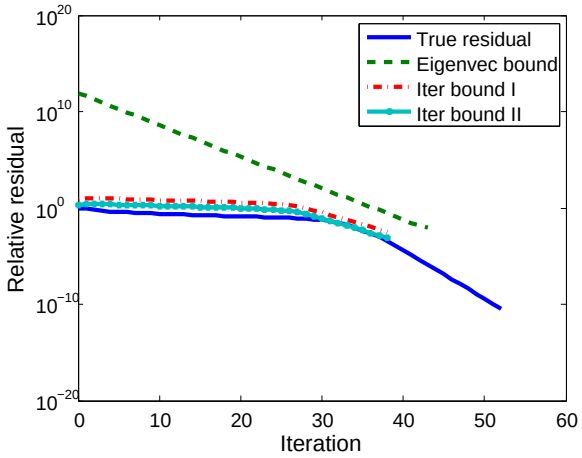


Figure 3.1: Convergence bounds for a finite difference discretization of a convection-diffusion problem. The dotted line is (3.27), the dashed line is our iteration-dependent bound (3.52).

3.3 Conclusions

In this chapter we have shown that the conjugate gradient method, MINRES and GMRES can be generalized for use with nonstandard inner products and that the resulting algorithms require little modification of the standard methods. Bounds on the relative error for W-PCG and the relative preconditioned residual for W-PMINRES were obtained that depend only on the eigenvalues of the coefficient matrix. Similar eigenvalue-dependent bounds exist for relative preconditioned W-PGMRES residuals for a particular inner product with respect to which the coefficient matrix is normal. Otherwise, the bounds we derived based on pseudospectra or eigenvalues can be applied. We related the convergence of a minimum residual method in a nonstandard inner product to that of standard minimum residual methods in $\langle \cdot, \cdot \rangle_I$ and derived a new iteration-dependent bound for I-GMRES that can be more descriptive than the standard eigenvalue bound (3.27). The convergence results presented here provide a starting point for examining effective preconditioned Krylov subspace methods, a topic to which subsequent chapters are devoted.

Chapter 4

Preconditioners for saddle point systems

Generalized saddle point problems of the form

$$\mathcal{A}x = b \quad \text{or} \quad \begin{bmatrix} A & B^T \\ D & -C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{4.1}$$

with $A \in \mathbb{R}^{n\times n}$, $B, D \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{m\times m}$, $n \geq m$ arise frequently in scientific computing, with two major applications being mixed finite element methods and interior point algorithms. (A comprehensive survey of the fields in which saddle point problems arise can be found in [16].) When partitioned, any matrix satisfies (4.1). However, we are interested generalized saddle point matrices with specific properties. As in [16] we explicitly exclude the cases where A, B or D is zero and we assume that the saddle point matrix satisfies at least one of the following properties.

Property 4.1. A matrix

$$\begin{bmatrix} A & B^T \\ D & -C \end{bmatrix},$$

$A \in \mathbb{R}^{n\times n}$, $B, D \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{m\times m}$, $n \geq m$ will be called a saddle point matrix if at least one of the following properties holds.

1. $A = A^T$;
2. the symmetric part of A, $H \equiv \frac{1}{2}(A + A^T)$, is positive semidefinite;
3. $B = D$;
4. C is symmetric positive semidefinite;
5. $C = 0$.

Note that (5) is a special case of (4).

When $\mathcal{A}$ is large and sparse, (4.1) is commonly solved by preconditioned Krylov subspace methods. Consequently, we are interested in choosing $\mathcal{P}$ so that

$$\widehat{\mathcal{A}} = \mathcal{P}^{-1}\mathcal{A}x = \mathcal{P}^{-1}b$$

can be rapidly solved by a Krylov method. If $\mathcal{A}$ is nonsymmetric then so is the preconditioned coefficient matrix in general and either a nonsymmetric Krylov subspace method or a method in a nonstandard inner product can be applied. For a large class of saddle point systems, however, $\mathcal{A}$ is symmetric indefinite, with A symmetric positive definite, C symmetric positive semidefinite, and $D = B$, i.e.,

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}. \tag{4.2}$$

Preconditioned MINRES in the Euclidean inner product can be used when $\mathcal{P}$ is symmetric positive definite but a number of effective preconditioners are indefinite or nonsymmetric. For these, nonstandard inner products may prove useful since, whenever $\mathcal{P}$ is real, the preconditioned coefficient matrix is self-adjoint with respect to a nonstandard symmetric bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. For certain choices of P, $\mathcal{W}$ depends only on A, B and approximations to A and the **(negative) Schur complement**

$$S = DA^{-1}B^T + C; \tag{4.3}$$

this may make computations with $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ inexpensive. Moreover, under certain additional conditions $\mathcal{W}$ defines an inner product and $\mathcal{W}$-PMINRES (see Section 3.2.1) can be applied. The less computationally expensive $\mathcal{W}$-PCG method in Algorithm 3.3 is usually preferred when the preconditioned coefficient matrix is symmetric and positive definite with respect to this inner product.

In this chapter we examine the role of nonstandard inner products in the development of effective preconditioned Krylov subspace methods for saddle point problems. Much of this chapter is devoted to symmetric problems, which constitute a large subset of saddle point systems. However, some of our results apply to nonsymmetric matrices. Section 4.1 is a brief (and certainly not exhaustive) overview of existing preconditioners for saddle point problems. Many of these are examples of the parameter-dependent Krzyżanowski preconditioner [85], that we also describe. To select the optimal combination of preconditioner and symmetric bilinear form (or inner product) it is helpful to have a thorough understanding of the symmetric bilinear forms with respect to which a Krzyżanowski preconditioned saddle point matrix is

self-adjoint. Moreover, one might wish to apply a Krzyżanowski-like preconditioner to a nonsymmetric problem. For both these reasons we present in Section 4.2 a characterization of the symmetric bilinear forms with respect to which a generalized Krzyżanowski preconditioned coefficient matrix is self-adjoint. This characterization is applied to two Murphy-Golub-Wathen [104] preconditioned saddle point matrices, and another preconditioned nonsymmetric saddle point matrix. In Section 4.3 a modification of a well known preconditioner is proposed and its properties derived while in Section 4.4 we present numerical experiments that appraise its effectiveness and robustness. The main findings of this chapter are summarized in Section 4.5.

Throughout, our assumptions are that A is nonsingular and that B and D have full rank. When A and C are symmetric we assume that they are positive definite and positive semidefinite, respectively. By $\mathcal{A}$ we denote the symmetric saddle point matrix in (4.1) or (4.2), by $\mathcal{P}$ a left preconditioner and by $\mathcal{W}$ the matrix that defines the symmetric bilinear form with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint. An approximation of A is represented by A_0 and is symmetric positive definite if A is symmetric positive definite. Similarly, S_0 is an approximation to the Schur complement S in (4.3) that is symmetric positive definite when S is symmetric positive definite. If the saddle point matrix is nonsymmetric we assume that A and $DA^{-1}B^T$ are nonsingular. Sufficient conditions for nonsingularity of the Schur complement can be found in [16].

4.1 Examples

The first part of this section is concerned with the description of preconditioners for symmetric matrices, each of which is an example of the Krzyżanowski preconditioner [85]. In the remainder, preconditioners for nonsymmetric saddle point systems are discussed.

One of the best known Krylov subspace methods in a nonstandard inner product is the **Bramble-Pasciak preconditioned conjugate gradient method** [20] for symmetric saddle point problems (4.2) with symmetric positive definite A and symmetric positive semidefinite C [1, 7, 8, 25, 40, 56, 80, 95, 109, 112, 113, 157]. Although $\mathcal{A}$ is indefinite, the application of a block triangular preconditioner renders the (nonsymmetric) preconditioned system self-adjoint and positive definite with respect to a nonstandard symmetric bilinear form. Under additional conditions this symmetric bilinear form becomes an inner product $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ and a $\mathcal{W}$-PCG method can be used reliably.

Specifically,

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}$$

is left-preconditioned by

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ B & -I \end{bmatrix},$$

where A_0 is a symmetric approximation to A. The preconditioned saddle point matrix is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \begin{bmatrix} A - A_0 & 0 \\ 0 & I \end{bmatrix}$$

defines an inner product when $A > A_0$, i.e., when $A - A_0$ is symmetric positive definite. We note that any symmetric positive definite matrix A_0 can be scaled by a positive scalar α so that $A - \alpha A_0 > 0$ or, equivalently, $\lambda_{min}(\frac{1}{\alpha} A_0^{-1} A) > 0$, although determining an appropriate scaling factor can be expensive. However, when $\mathcal{W}$ is positive definite, $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ [20] and a $\mathcal{W}$-PCG method, which can be efficiently implemented [112, 134], is robust. This coincidence of desirable properties makes the Bramble-Pasciak approach popular.

The notion of combining a preconditioner with a nonstandard inner product has been developed in several ways (some of which can be found in [33, 34, 89] and the following examples). The **Bramble-Pasciak with Schur complement (BP) preconditioner**

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ B & -S_0 \end{bmatrix} \tag{4.4}$$

for symmetric saddle point systems in [80, 102, 127] incorporates a symmetric positive definite approximation S_0 to the Schur complement (4.3) into the BP preconditioner that may give a better approximation to $\mathcal{A}$ when S is far from the identity matrix[1].

The preconditioned coefficient matrix $\widehat{\mathcal{A}} = \mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to the symmetric bilinear form defined by

$$\mathcal{W} = \begin{bmatrix} A - A_0 & 0 \\ 0 & S_0 \end{bmatrix}, \tag{4.5}$$

which defines an inner product when $A - A_0 > 0$ and S_0 is positive definite. When $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ is an inner product it is one with respect to which $\widehat{\mathcal{A}}$ is always positive definite [80], similarly to the original Bramble-Pasciak preconditioner. Thus, a preconditioned

[1] Zulehner [157] considered a similar idea in the context of inexact Uzawa methods, which are connected to conjugate gradient methods [17, 100].

conjugate gradient method in $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ can be reliably applied to this preconditioned system when $\mathcal{W} > 0$.

The major drawback of the above methods is that often A_0 must be scaled for $\mathcal{W}$ to define an inner product. This typically requires an approximation to the smallest eigenvalue of $A_0^{-1}A$, the computation of which can be costly. To circumvent this difficulty, Stoll and Wathen [135] proposed the **Bramble-Pasciak$^+$ (BP$^+$) preconditioner** for symmetric saddle point systems

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ -B & S_0 \end{bmatrix} \tag{4.6}$$

that negates the second block-row of the BP preconditioner (4.4). The matrix $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \begin{bmatrix} A + A_0 & 0 \\ 0 & S_0 \end{bmatrix}. \tag{4.7}$$

When A, A_0 and S_0 are symmetric positive definite, $\mathcal{W}$ is symmetric positive definite and defines an inner product.

However, positive definiteness of $\mathcal{W}$ is achieved at the expense of positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. In contrast to the Bramble-Pasciak preconditioners above, $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ [135] and a $\mathcal{W}$-PCG method cannot be reliably applied. However, it is always possible to use $\mathcal{W}$-PMINRES. Bounds on the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ which do not depend on the norms of the eigenvectors of $\mathcal{P}^{-1}\mathcal{A}$ are given in Appendix B. Since these bounds do not contain norms of the eigenvectors of $\mathcal{P}^{-1}\mathcal{A}$, which are typically costly to compute, they may be easier to apply than the eigenvalue bounds in [134, Section 3.4].

Each of these Bramble-Pasciak-style preconditioners is block triangular. An alternative is a constraint preconditioner, such as the **Schöberl-Zulehner (SZ) preconditioner** [122, 157]

$$\mathcal{P} = \begin{bmatrix} A_0 & B^T \\ B & BA_0^{-1}B^T - S_0 \end{bmatrix}. \tag{4.8}$$

It is suitable for symmetric saddle point matrices with zero $(2,2)$ block, particularly those arising from optimal control problems [70, 108, 122]. The preconditioned saddle point matrix is self-adjoint with respect to the symmetric bilinear form defined by

$$\mathcal{W} = \begin{bmatrix} A_0 - A & 0 \\ 0 & BA_0^{-1}B^T - S_0 \end{bmatrix}, \tag{4.9}$$

which is positive definite when $A_0 > A$ and $BA_0^{-1}B^T > S_0$. Again, when $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ is an inner product, it is one with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite [122].

It is imperative that the SZ preconditioner can be applied cheaply. Schöberl and Zulehner [122] outline its efficient application to a vector and we reproduce their results here for completeness. Let $y = \begin{bmatrix} w^T, & q^T \end{bmatrix}^T$. Then if $y = \mathcal{P}^{-1}x$ we have that $\mathcal{P}y = x$ or, in block form,

$$\begin{bmatrix} A_0 & B^T \\ B & BA_0^{-1}B^T - S_0 \end{bmatrix} \begin{bmatrix} w \\ q \end{bmatrix} = \begin{bmatrix} s \\ v \end{bmatrix}.$$

To determine y, one must solve

$$A_0 w + B^T q = s \quad \text{and} \quad Bw + BA_0^{-1}B^T q - S_0 q = v$$

for w and q. This can be accomplished in three stages.

I. With $\widehat{w} = w + A_0^{-1}B^T q$ the first equation becomes

$$A_0 \widehat{w} = s,$$

from which $\widehat{w}$ can be obtained.

II. The second equation can be rearranged to give that

$$S_0 q = B(w + A_0^{-1}B^T q) - v = B\widehat{w} - v,$$

which can be solved for q.

III. Then, w can be recovered from $\widehat{w}$ since

$$A_0 w = A_0 \widehat{w} - B^T q = s - B^T q.$$

This shows that applying $\mathcal{P}$ to a vector requires two linear solves with A_0 and one with S_0. Inner products with $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ require matrix-vector products and one linear solve with A_0, although there is potential for computational savings, as described in [85].

Perhaps the simplest preconditioner for a symmetric saddle point matrix, however, is the **block diagonal preconditioner (BDW)** [58, 81, 126]

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \tag{4.10}$$

for which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to the symmetric bilinear form defined by

$$\mathcal{W} = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix}. \tag{4.11}$$

Indeed, the self-adjointness requirement (2.13) that $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ is symmetric is trivially satisfied since

$$\mathcal{W}\mathcal{P}^{-1}\mathcal{A} = \mathcal{A},$$

the original symmetric saddle point matrix. The indefiniteness of $\mathcal{A}$, however, means that $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$. When the approximation of A and the Schur complement in (4.10) is exact, i.e., when $A_0 = A$ and $S_0 = S$, the resulting linear system can be solved by a Krylov subspace method in three steps [104] (since A is nonsingular) as we discuss in greater detail for nonsymmetric saddle point problems later in this section. In practice a cheaper approximation of A and S is used and rapid convergence is often maintained.

Each of the preconditioners described so far in this section is a **Krzyżanowski preconditioner** [85]

$$\mathcal{P} = \begin{bmatrix} I & 0 \\ cBA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & dA_0^{-1}B^T \\ 0 & I \end{bmatrix} \tag{4.12}$$

$$= \begin{bmatrix} A_0 & 0 \\ cB & S_0 \end{bmatrix} \begin{bmatrix} I & dA_0^{-1}B^T \\ 0 & I \end{bmatrix}, \tag{4.13}$$

with $|c|, |d| \leq 1$. The preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \delta \begin{bmatrix} A_0 - cA & 0 \\ 0 & S_0 + cdBA_0^{-1}B^T + dC \end{bmatrix}, \tag{4.14}$$

and $\delta = \pm 1$. Table 4.1 contains the parameter choices that lead to the formerly mentioned preconditioners. The most general Krzyżanowski preconditioner, i.e., with $c, d \neq 0$, can be applied with two solves with A_0 and one with S_0. This is most clearly seen from (4.13) and the procedure is similar to that for the SZ preconditioner, described earlier in this section. Specifically, if we wish to solve $y = \mathcal{P}^{-1}x$, where $y = \begin{bmatrix} w^T, & q^T \end{bmatrix}^T$ then

$$\begin{bmatrix} A_0 & B^T \\ B & BA_0^{-1}B^T - S_0 \end{bmatrix} \begin{bmatrix} w \\ q \end{bmatrix} = \begin{bmatrix} s \\ v \end{bmatrix}.$$

To determine y, it is again necessary to solve

$$A_0 w + dB^T q = s \quad \text{and} \quad cBw + cdBA_0^{-1}B^T q + S_0 q = v$$

for w and q which can be performed in three steps

I. If $\widehat{w} = w + dA_0^{-1}B^T q$ the first equation becomes

$$A_0\widehat{w} = s,$$

from which $\widehat{w}$ can be obtained.

Preconditioner	A_0 approximation	S_0 approximation	c	d	δ
Block diagonal	A_0	S_0	0	0	1
Bramble-Pasciak	A_0	$-I$	1	0	-1
BP	A_0	$-S_0$	1	0	-1
BP$^+$	A_0	S_0	-1	0	1
SZ	A_0	$-S_0$	1	1	1

Table 4.1: Krzyżanowski parameters for saddle point preconditioners.

II. Then q can be found from the second equation

$$S_0 q = v - B\widehat{w},$$

III. Finally, w can be recovered from $\widehat{w}$ since

$$A_0 w = A_0\widehat{w} - dB^T q = s - dB^T q.$$

Eigenvalues typically describe the convergence of a $\mathcal{W}$-PCG or $\mathcal{W}$-PMINRES method applied to the Krzyżanowski preconditioned system (see Corollaries 3.5 and 3.15). We prove below that the eigenvalues of the Krzyżanowski preconditioned saddle point matrix depend on certain generalized Rayleigh quotients when $C = 0$.

Theorem 4.2. *Let the symmetric saddle point matrix*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $B \in \mathbb{R}^{m\times n}$ has full rank be left preconditioned by the Krzyżanowski preconditioner (4.12)*, with $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ symmetric positive definite. Additionally, let $\mathcal{W}$ in* (4.14) *be positive definite and let λ be an eigenvalue of $\mathcal{P}^{-1}\mathcal{A}$ with associated eigenvector*

$$\begin{bmatrix} u^T, & v^T \end{bmatrix}^T.$$

Then λ is real, $\lambda \neq 0$ and, when c is nonzero, $\lambda \neq 1/c$. If $Bu = 0$ then

$$\lambda = \frac{u^T A u}{u^T A_0 u}. \tag{4.15}$$

Otherwise,

$$\lambda = \frac{1 - (c+d)\omega \pm \sqrt{(1-(c+d)\omega)^2 + 4(\psi - cd\omega)\omega}}{2(\psi - cd\omega)}, \tag{4.16}$$

where

$$\psi = \frac{u^T A_0 u}{u^T A u} \tag{4.17}$$

and

$$\omega = \frac{u^T B^T (cdBA_0^{-1}B^T + S_0)^{-1} Bu}{u^T Au}. \tag{4.18}$$

Proof. The eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are those of the generalized eigenvalue problem

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} A_0 & dB^T \\ cB & cdBA_0^{-1}B^T + S_0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

which, in component form is

$$Au + B^T v = \lambda A_0 u + \lambda dB^T v \tag{4.19}$$

$$Bu = \lambda cBu + \lambda(cdBA_0^{-1}B^T + S_0)v. \tag{4.20}$$

Since $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to an inner product its eigenvalues are real (see Corollary 2.21 in Section 2.2). If $\lambda = 0$, (4.20) implies that $Bu = 0$. Premultiplying (4.19) by u^T gives that $u^T Au = 0$ which, since A is positive definite, can only be true if $u = 0$. However, then (4.19) reduces to $B^T v = 0$. The matrix B has full rank and it follows that $v = 0$. Thus, zero is not an eigenvalue of $\mathcal{P}^{-1}\mathcal{A}$. More generally, $u \neq 0$ since the definiteness of $S_0 + cdBA_0^{-1}B^T$—guaranteed by the positive definiteness of $\mathcal{W}$—means that when $u = 0$ (4.20) can only be satisfied when $v = 0$.

If $c \neq 0$ and $\lambda = \frac{1}{c}$ we again find from (4.20) that $v = 0$. It follows that (4.19) can be simplified and rearranged to give that $(A_0 - cA)u = 0$. However, positive definiteness of $\mathcal{W}$ ensures that $A_0 - cA$ is definite, which shows that $u = 0$. Thus, $\lambda \neq \frac{1}{c}$.

Let us now determine the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$. When $Bu = 0$, (4.20) becomes

$$\lambda(cdBA_0^{-1}B^T + S_0)v = 0$$

which implies, since $cdBA_0^{-1}B^T + S_0$ is definite and $\lambda \neq 0$, that $v = 0$. Substituting $v = 0$ into (4.19) yields that

$$Au = \lambda A_0 u$$

which, because $u \neq 0$, shows that

$$\lambda = \frac{u^T Au}{u^T A_0 u}.$$

When $Bu \neq 0$, rearranging (4.20) gives that

$$v = \left(\frac{1}{\lambda} - c\right) \tilde{S}^{-1} Bu,$$

where $\tilde{S} = cdBA_0^{-1}B^T + S_0$. By substituting for v in (4.19) we obtain the equation

$$Au + (1 - \lambda d)\left(\frac{1}{\lambda} - c\right) B^T \tilde{S}^{-1} Bu = \lambda A_0 u$$

for λ. Premultiplying by λu^T and dividing by $u^T Au$ leads to the quadratic equation

$$(\psi - cd\omega)\lambda^2 + ((c+d)\omega - 1)\,\lambda - \omega = 0,$$

the roots of which are

$$\lambda = \frac{1 - (c+d)\omega \pm \sqrt{(1-(c+d)\omega)^2 + 4(\psi - cd\omega)\omega}}{2(\psi - cd\omega)} \tag{4.21}$$

which completes the proof. □

Remark 4.3. The equation (4.16) is complicated but simplifies in each of the cases in Table 4.1 and so provides useful information in these specific instances, particularly if $c = 0$ or $d = 0$. Knowledge of the extreme values of (4.17) and (4.18) allow the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ to be bounded and a priori information about the convergence of a $\mathcal{W}$-Krylov subspace method to be obtained. Appendix B provides such bounds for the BP$^+$ preconditioner.

Block diagonal, block triangular and constraint preconditioners have also been proposed for nonsymmetric matrices. The resulting linear systems are typically solved by a nonsymmetric Krylov subspace method in the Euclidean inner product [23, 32, 82, 93, 124, 156] which reflects the difficulty in obtaining a suitable nonstandard inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint. An exception is Krzyżanowski's approach [84], which is to convert the problem into one that is self-adjoint with respect to an inner product, similarly to the normal equations. To this new system $\mathcal{W}$-PCG or $\mathcal{W}$-PMINRES can be applied. We assume in this chapter that A, $DA^{-1}B^T$ and $C + DA^{-1}B^T$ are nonsingular.

The **Murphy-Golub-Wathen block diagonal preconditioner** [104] for nonsymmetric saddle point problems with $C = 0$ takes the form

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & DA^{-1}B^T \end{bmatrix}. \tag{4.22}$$

The preconditioned matrix $\widehat{\mathcal{A}} = \mathcal{P}^{-1}\mathcal{A}$ satisfies

$$\widehat{\mathcal{A}}(\widehat{\mathcal{A}} - I)(\widehat{\mathcal{A}}^2 - \widehat{\mathcal{A}} - I) = 0$$

from which we discern that $\widehat{\mathcal{A}}$ is diagonalizable and has at most four distinct eigenvalues, $0, 1, \frac{1}{2} \pm \frac{\sqrt{5}}{2}$ (or three when A is invertible). Application of any Krylov subspace method that minimizes the error or residual to the preconditioned system will, therefore, converge in three or four steps. Since $\mathcal{P}^{-1}\mathcal{A}$ is diagonalizable with real eigenvalues it is self-adjoint with respect to an inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, which we describe in Section 4.2.

Application of the **Murphy-Golub-Wathen upper triangular preconditioner**[2]

$$\mathcal{P} = \begin{bmatrix} A & B^T \\ 0 & DA^{-1}B^T \end{bmatrix} \tag{4.23}$$

in [104] to a saddle point problem with zero (2,2) block gives a preconditioned coefficient matrix that has eigenvalues at ± 1. The solution to the preconditioned system is, therefore, obtained after two steps of a Krylov subspace method such as GMRES. The symmetric bilinear forms with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint are also described in Section 4.2.

While the Murphy-Golub-Wathen preconditioners are optimal, they are often prohibitively expensive to apply. However, if good approximations of A and the Schur complement can be found, rapid convergence might still be expected. If $A = E - F$ is a matrix splitting with E and $DE^{-1}B^T$ nonsingular one approximation to the block diagonal preconditioner (4.22) is [32]

$$\mathcal{P} = \begin{bmatrix} E & 0 \\ 0 & DE^{-1}B^T \end{bmatrix}. \tag{4.24}$$

In the case $D = B$ we identify symmetric bilinear forms with respect to which the resulting preconditioned matrix is self-adjoint in Section 4.2. The matrix-splitting idea is extended for matrices with a nonzero (2,2) block in [124], where the Schur complement approximation $C + DE^{-1}B^T$ replaces the (2,2) block of (4.24), provided $C + DE^{-1}B^T$ is nonsingular.

The number of competing preconditioners for saddle point systems indicates that there is at present no definitive way to precondition a saddle point problem. We feel that it would be useful to better understand how preconditioners and inner products can be chosen to achieve efficient Krylov subspace methods. To this end, we characterize in the next section the symmetric bilinear forms with respect to which a preconditioned linear system is self-adjoint, where the preconditioner is a generalization of the Krzyżanowski preconditioner.

[2] An alternative block triangular preconditioner, that we do not consider here, can be obtained from (4.23) by negating the (2,2) block. The resulting preconditioned coefficient matrix has a sole eigenvalue at one but is not diagonalizable.

4.2 Characterization of symmetric bilinear forms

All of the preconditioners for symmetric saddle point problems in the last section are examples of the Krzyżanowski [85] preconditioner (4.12), i.e., $\mathcal{P}$ is of the form

$$\mathcal{P} = \begin{bmatrix} I & 0 \\ cBA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & dA_0^{-1}B^T \\ 0 & I \end{bmatrix}.$$

Although a symmetric bilinear form with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self adjoint is given in [85], this is not the only choice. Additionally, we may want to consider nonstandard inner products and preconditioners for nonsymmetric saddle point systems. Accordingly, we provide a complete description of the symmetric bilinear forms with respect to which a nonsymmetric matrix of the form (4.1), preconditioned by a nonsymmetric version of $\mathcal{P}$, is self adjoint. The preconditioner we consider is

$$\mathcal{P} = \begin{bmatrix} I & 0 \\ cDA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & dA_0^{-1}B^T \\ 0 & I \end{bmatrix}$$

which we assume to be invertible. The application of $\mathcal{P}$ to the nonsymmetric saddle point matrix

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ D & -C \end{bmatrix},$$

gives

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{bmatrix}, \tag{4.25}$$

where

$$\begin{aligned} \widehat{A}_{11} &= A_0^{-1}A + cdA_0^{-1}B^T S_0^{-1}DA_0^{-1}A - dA_0^{-1}B^T S_0^{-1}D \\ \widehat{A}_{12} &= A_0^{-1}B^T + dA_0^{-1}B^T S_0^{-1}(cDA_0^{-1}B^T + C) \\ \widehat{A}_{21} &= -cS_0^{-1}DA_0^{-1}A + S_0^{-1}D \\ \widehat{A}_{22} &= -S_0^{-1}(cDA_0^{-1}B^T + C). \end{aligned} \tag{4.26}$$

The symmetric bilinear forms with respect to which the preconditioned coefficient matrix is self-adjoint are given in the following lemma.

Lemma 4.4. *Let the Krzyánowski-type left preconditioner*

$$\mathcal{P} = \begin{bmatrix} I & 0 \\ cDA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & dA_0^{-1}B^T \\ 0 & I \end{bmatrix}, \tag{4.27}$$

be invertible and let

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ D & -C \end{bmatrix}$$

where $A \in \mathbb{R}^{n\times n}$ is invertible, $B, D \in \mathbb{R}^{m\times n}$ have full rank, $C \in \mathbb{R}^{m\times m}$ and Property 4.1 holds. Then $\mathcal{P}^{-1}\mathcal{A}$, given by (4.25) *and* (4.26) *above, is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where*

$$\mathcal{W} = \begin{bmatrix} W_{11} & W_{12}^T \\ W_{12} & W_{22} \end{bmatrix}$$

and W_{11}, W_{12} and W_{22} satisfy the equations

$$X_{11} = X_{11}^T, \quad X_{12} = X_{21}^T, \quad \text{and} \quad X_{22} = X_{22}^T, \tag{4.28}$$

where

$$\begin{aligned} X_{11} = W_{11}\widehat{A}_{11} + W_{12}\widehat{A}_{21}, \quad X_{12} = W_{11}\widehat{A}_{12} + W_{12}\widehat{A}_{22}, \\ X_{21} = W_{21}\widehat{A}_{11} + W_{22}\widehat{A}_{21} \quad \text{and} \quad X_{22} = W_{21}\widehat{A}_{12} + W_{22}\widehat{A}_{22} \end{aligned} \tag{4.29}$$

and $\hat{A}_{11}$, $\hat{A}_{12}$, $\hat{A}_{21}$ and $\hat{A}_{22}$ are defined by (4.26)*.*

Proof. The preconditioned coefficient matrix is

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{bmatrix}.$$

If $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, then $\mathcal{W}$, $\mathcal{P}^{-1}$ and $\mathcal{A}$ satisfy (2.13), i.e., $\mathcal{W}(\mathcal{P}^{-1}\mathcal{A}) = (\mathcal{P}^{-1}\mathcal{A})^T\mathcal{W}$, so that $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ is symmetric. Letting

$$\mathcal{X} = \mathcal{W}\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} X_{11} & X_{12}^T \\ X_{12} & X_{22} \end{bmatrix}$$

we find that

$$\begin{aligned} X_{11} = W_{11}\widehat{A}_{11} + W_{12}\widehat{A}_{21}, \qquad & X_{12} = W_{11}\widehat{A}_{12} + W_{12}\widehat{A}_{22}, \\ X_{21} = W_{21}\widehat{A}_{11} + W_{22}\widehat{A}_{21}, \qquad & X_{22} = W_{21}\widehat{A}_{12} + W_{22}\widehat{A}_{22} \end{aligned}$$

and the symmetry condition $\mathcal{X} = \mathcal{X}^T$ gives the desired result. □

When A, A_0 and S_0 are symmetric positive definite and C is positive semidefinite the conditions in Lemma 4.4 are certainly satisfied by (4.14). However, Lemma 4.4 provides a framework for examining the symmetric bilinear forms with respect to which other preconditioned saddle point matrices are self-adjoint.

We now use Lemma 4.4 to identify symmetric bilinear forms with respect to which the Murphy-Golub-Wathen preconditioners (4.22) and (4.23) described in Section 4.1 are self-adjoint. Although these preconditioners require A and S, the approach used here is instructive and can be applied to more practical preconditioners, as we show at the end of this section. First, we present some technical results that will be needed to describe these symmetric bilinear forms.

Lemma 4.5. *Assume $n \geq m$ and let $B, D \in \mathbb{R}^{m\times n}$ have full rank. Additionally, let $A \in \mathbb{R}^{n\times n}$ and $BA^{-T}D^T \in \mathbb{R}^{m\times m}$ be nonsingular. Then, the projector $N = D^T(BA^{-T}D^T)^{-1}BA^{-T}$ has eigenvalues $\lambda = 0, 1$. The eigenspace of N associated with $\lambda = 1$ has dimension m and is spanned by the columns of $V = [\, v_1, \ldots, v_m \,]$ if and only if $V = D^T R^T$, where $R \in \mathbb{R}^{m\times m}$ is nonsingular. The $n - m$ eigenvectors v_i, $i = m + 1, \ldots, n$ associated with $\lambda = 0$ lie in $A^T \operatorname{null}(B)$, where $\operatorname{null}(B)$ is the nullspace of B.*

Proof. It is easily verified that N is a projector and so has eigenvalues $\lambda = 0, 1$. Eigenvectors v_i, $i = 1, \ldots, m$ corresponding to $\lambda = 1$ lie in $\operatorname{range}(N)$, the range of N, while eigenvectors v_i, $i = m + 1, \ldots, n$ corresponding to $\lambda = 0$ lie in $\operatorname{null}(N)$. Comparison of $N = D^T(BA^{-T}D^T)^{-1}BA^{-T}$ with (1.66) in [118] shows that N projects on to $\operatorname{range}(D^T)$ orthogonal to $A^{-1}B^T$. It follows from [118, Section 1.12.1] that $\operatorname{range}(N) = \operatorname{range}(D^T)$, which has dimension m since D has full rank. Since $v_1, \ldots, v_m$ must be linearly independent and lie in $\operatorname{range}(D^T)$, $V = [\, v_1, \ldots, v_m \,]$ must be of the form $V = D^T R^T$, where $R \in \mathbb{R}^{m\times m}$ is nonsingular.

Also from [118, Section 1.12.1], $\operatorname{null}(N) = \operatorname{range}(A^{-1}B^T)^{\perp} = \operatorname{null}(BA^{-T})$, which has dimension $n - m$, since B has full rank. Since A is nonsingular, any vector in $\operatorname{null}(BA^{-T})$ is of the form $A^T n$, where $n \in \operatorname{null}(B)$.

□

Now we require some results regarding the solutions of Sylvester equations

$$EM = XF \tag{4.30}$$

for the unknown real symmetric matrices $M \in \mathbb{R}^{m\times m}$ and $X \in \mathbb{R}^{n\times n}$, where $E, F \in \mathbb{R}^{n\times m}$ are known.

Definition 4.6. [77, Definition 3.1] A real $m \times m$ matrix M will be called **admissible** if $(F^T E)M = M(F^T E)^T$, $M^T = M$ and $\operatorname{null}(F) \subseteq \operatorname{null}(EM)$.

Lemma 4.7. [77, Theorem 4.2] *If M is an admissible matrix and the rank condition*

$$\operatorname{rank} F^T E = \operatorname{rank} E$$

holds, the general real symmetric solution X of (4.30) is given by

$$X = EM(F^T EM)^{\#} ME^T + X_0, \tag{4.31}$$

where $(F^T EM)^{\#}$ denotes any (1,2)-inverse of $F^T EM$ (see Appendix A.6), and X_0 is any real symmetric matrix for which $X_0 F = 0$.

For positive definite solutions Y we require further restrictions.

Lemma 4.8. [77, Theorem 7.1] *The equation* (4.30) *has a solution pair* ($M > 0$, $X > 0$) *if and only if* $F^T E$ *is simple with all eigenvalues nonnegative, and*

$$\operatorname{rank} EF^T = \operatorname{rank} E = \operatorname{rank} F = \operatorname{rank} F^T E.$$

These solutions X are characterized by the following lemma.

Lemma 4.9. (adapted from [77, Lemma 7.1 and Theorem 7.2]) *Let the rank conditions in Lemma* 4.8 *hold. Then for each positive definite admissible* M*, the matrix*

$$X = EM(F^T EM)^\dagger ME^T + X_0,$$

where $(F^T EM)^\dagger$ *is the Moore-Penrose pseudoinverse of* $F^T EM$ *and* $X_0 F = 0$*, determines a solution pair* ($M > 0$, $X > 0$) *if and only if* X_0 *is real and symmetric positive definite.*

We are now in a position to determine the symmetric bilinear forms with respect to which the Murphy-Golub-Wathen block diagonal preconditioned saddle point matrix is self-adjoint.

Theorem 4.10. *Let the nonsymmetric saddle point matrix*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ D & 0 \end{bmatrix},$$

where $n \geq m$*,* $A \in \mathbb{R}^{n\times n}$ *is nonsingular,* $B, D \in \mathbb{R}^{m\times n}$ *have full rank and the Schur complement* $DA^{-1}B^T$ *is nonsingular, be left preconditioned by the Murphy-Golub-Wathen block diagonal preconditioner*

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & DA^{-1}B^T \end{bmatrix}.$$

Moreover, let

$$N = D^T(BA^{-T}D^T)^{-1}BA^{-T} \tag{4.32}$$

have diagonalization $N = Z\Lambda Z^{-1}$ *and* R *be any nonsingular matrix that satisfies* $RDA^{-1}B^T = BA^{-T}D^T R^T$ *or the zero matrix. Then, all symmetric bilinear forms* $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ *with respect to which* $\mathcal{P}^{-1}\mathcal{A}$ *is self-adjoint are defined by*

$$\mathcal{W} = \begin{bmatrix} W_{11} & W_{12}^T \\ W_{12} & W_{22} \end{bmatrix},$$

with $W_{11} = ZZ^T$*,* $W_{12} = B^T R$ *and* $W_{22} = \widehat{W}_{22} - BA^{-T}D^T R^T$ *where, if* $(NW_{11})^\#$ *is a* (1,2)*-inverse of* NW_{11}*,* $\widehat{W}_{22} = BA^{-T}W_{11}(NW_{11})^\# W_{11}A^{-1}B^T$*. This symmetric bilinear form is an inner product when* $W_{22} > W_{12}W_{11}W_{12}^T$*.*

Proof. Conditions (4.28) in Lemma 4.4 for the block diagonal preconditioner are

$$W_{12}^T(DA^{-1}B^T)^{-1}D = D^T(BA^{-T}D^T)^{-1}W_{12} \tag{4.33}$$

$$W_{12}A^{-1}B^T = BA^{-T}W_{12}^T \tag{4.34}$$

$$W_{11}A^{-1}B^T = W_{12}^T + D^T(BA^{-T}D^T)^{-1}W_{22}. \tag{4.35}$$

Postmultiplying (4.33) by $A^{-1}B^T$ and using (4.34) gives that

$$W_{12}^T = D^T(BA^{-T}D^T)^{-1}W_{12}A^{-1}B^T = D^T(BA^{-T}D^T)^{-1}BA^{-T}W_{12}^T \tag{4.36}$$

or, using (4.32), that

$$NW_{12}^T = 1W_{12}^T. \tag{4.37}$$

Thus, either $W_{12} = 0$ or the columns of W_{12}^T are eigenvectors of N associated with $\lambda = 1$. In the latter case, we infer from Lemma 4.5 that $W_{12}^T = D^T R^T$ where R is nonsingular. Accordingly, (4.37) is satisfied by $W_{12} = RD$, where R is either nonsingular or the zero matrix.

Substituting $W_{12} = RD$ into (4.34) and (4.35) shows that we additionally require that

$$RDA^{-1}B^T = BA^{-T}D^T R^T, \tag{4.38}$$

i.e., that $RDA^{-1}B^T$ is symmetric, and that

$$\begin{aligned} W_{11}A^{-1}B^T &= (RD)^T + D^T(BA^{-T}D^T)^{-1}W_{22} \\ &= D^T(BA^{-T}D^T)^{-1}(BA^{-T}D^T R^T + W_{22}) \\ &= D^T(BA^{-T}D^T)^{-1}\widehat{W}_{22}. \end{aligned} \tag{4.39}$$

Symmetry of W_{22} and $BA^{-T}D^T R^T$ (by (4.38)) ensures that $\widehat{W}_{22} = BA^{-T}D^T R^T + W_{22}$ is symmetric.

To solve (4.39) we postmultiply both sides by $(DA^{-1}B^T)^{-1}D$ to obtain that

$$W_{11}N^T = D^T(BA^{-T}D^T)^{-1}\widehat{W}_{22}(DA^{-1}B^T)^{-1}D.$$

The right-hand side of this equation is symmetric with W_{11} symmetric, which is exactly the condition (2.13) for self-adjointness of N^T with respect to $\langle\cdot,\cdot\rangle_{W_{11}}$. By Lemma 4.5, $N = Z\Lambda Z^{-1}$ and $N^T = Z^{-T}\Lambda Z^{-1}$ are diagonalizable with real eigenvalues and it follows from Corollary 2.28 that $W_{11} = ZZ^T$ defines a real inner product with respect to which N is self-adjoint. Moreover, W_{11} is admissible (see Definition 4.6).

It remains to find W_{22}. Since

$$\operatorname{rank}(BA^{-T}) = \operatorname{rank}(N) = \operatorname{rank}(N^T) = \operatorname{rank}((DA^{-1}B^T)^{-1}D) = m,$$

we can apply Lemma 4.7 to the transpose of (4.39), i.e., to

$$(BA^{-T})W_{11} = \widehat{W}_{22}((DA^{-1}B^T)^{-1}D)$$

to obtain that

$$\widehat{W}_{22} = BA^{-T}W_{11}(NW_{11})^{\#}W_{11}A^{-1}B^T, \tag{4.40}$$

where $X_0 = 0$ in (4.31) since $\operatorname{rank}((DA^{-1}B^T)^{-1}D) = m$. Indeed, if the (1,2)-inverse is the Moore-Penrose pseudoinverse we can be sure that $\widehat{W}_{22}$ is positive definite by Lemma 4.9. The condition for positive definiteness of W_{22} follows from [75, Theorem 7.7.6]. □

Remark 4.11. When $\mathcal{A}$ is symmetric with A symmetric positive definite and B full rank, Theorem 4.10 simplifies dramatically. Under these assumptions, $\mathcal{P}$ is symmetric positive definite and the conditions of the theorem are satisfied by $\mathcal{W} = \mathcal{P}$. In other words,

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}^{-1} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$$

is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}.$$

The choice of eigenvector matrix Z of N in (4.32) that leads to this inner product is $Z = A^{\frac{1}{2}}U$, where U is an orthogonal matrix of eigenvectors for the symmetric matrix $A^{-\frac{1}{2}}NA^{\frac{1}{2}}$.

Symmetric bilinear forms with respect to which Murphy-Golub-Wathen block triangular preconditioned saddle point matrix is self-adjoint can be obtained in the same way.

Theorem 4.12. *Let $\mathcal{A}$, N and Z be as described in Theorem* 4.10*, with $n \geq m$, $A \in \mathbb{R}^{n\times n}$ and $DA^{-1}B^T \in \mathbb{R}^{m\times m}$ nonsingular and $B, D \in \mathbb{R}^{m\times n}$ full rank. Then, if $\mathcal{A}$ is preconditioned by the Murphy-Golub-Wathen block triangular preconditioner*

$$\mathcal{P} = \begin{bmatrix} A & B^T \\ 0 & DA^{-1}B^T \end{bmatrix},$$

any symmetric bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint that is defined by the block diagonal matrix

$$\mathcal{W} = \begin{bmatrix} W_{11} & 0 \\ 0 & W_{22} \end{bmatrix}$$

has blocks of the form $W_{11} = ZZ^T$ and $W_{22} = BA^{-T}W_{11}(NW_{11})^{\#}A^{-1}B^T$. If we choose the Moore-Penrose pseudoinverse then $\mathcal{W}$ is positive definite and defines an inner product.

Proof. The conditions (4.28) in Lemma 4.4 for self-adjointness are that

$$W_{11}A^{-1}B^T(DA^{-1}B^T)^{-1}D = W_{11}N^T$$

is symmetric and that

$$W_{11}A^{-1}B^T = D^T(BA^{-T}D^T)^{-1}W_{22}.$$

The latter is identical in form to (4.39) in Theorem 4.10 above and can be solved similarly to the block diagonal case above. By so doing, we find that if N has diagonalization $N = Z\Lambda Z^{-1}$ then $W_{11} = ZZ^T$ and $W_{22} = BA^{-T}W_{11}(NW_{11})^{\#}A^{-1}B^T$. Additionally, if the Moore-Penrose pseudoinverse is used, Lemma 4.9 guarantees that W_{22} is positive. The first condition for self-adjointness is automatically satisfied by this choice of W_{11}. □

Remark 4.13. When $\mathcal{A}$ is symmetric, A is symmetric positive definite and B has full rank the Murphy-Golub-Wathen block diagonal preconditioner is symmetric positive definite and defines an inner product with respect to which the block triangular preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint. i.e.,

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} A & B^T \\ 0 & BA^{-1}B^T \end{bmatrix}^{-1} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$$

is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \begin{bmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}.$$

It is interesting that the same inner product renders both the block diagonal and block triangular preconditioned saddle point matrices self-adjoint in this case.

These characterizations of symmetric bilinear forms, although for ideal preconditioners, show how one may find symmetric bilinear forms with respect to which a more practical preconditioner is self-adjoint. In this spirit, we conclude this section with a simple extension to Theorem 4.10, along with an example of its application to a PDE problem.

Corollary 4.14. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$ has full rank and $n \geq m$, be left preconditioned by

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix},$$

where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are invertible but may be nonsymmetric. If $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to the symmetric bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \begin{bmatrix} W_{11} & 0 \\ 0 & W_{22} \end{bmatrix},$$

$W_{11} \in \mathbb{R}^{n\times n}$, $W_{22} \in \mathbb{R}^{m\times m}$ are symmetric and invertible, then $A_0^{-1}A$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{W_{11}}$ and $S_0^{-1}BA_0^{-1}B^T$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{W_{22}}$.

Proof. Setting $W_{12} = 0$ in Lemma 4.4 shows that the self-adjointness conditions are

$$W_{11}A_0^{-1}A = A^T A_0^{-T} W_{11} \tag{4.41}$$

$$W_{11}A_0^{-1}B^T = B^T S_0^{-T} W_{22}. \tag{4.42}$$

Equation (4.41) is a self-adjointness relationship for $A_0^{-1}A$ (see 2.13), i.e., we require that $A_0^{-1}A$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{W_{11}}$. Premultiplying (4.42) by BA_0^{-T}, as in the proof of Theorem 4.10, gives that

$$BA_0^{-T}W_{11}A_0^{-1}B^T = BA_0^{-T}B^T S_0^{-T}W_{22}.$$

The left-hand side of this equation is symmetric so W_{22} must enforce symmetry of

$$BA_0^{-T}B^T S_0^{-T}W_{22}.$$

However, from (2.13) it is clear that this is equivalent to requiring that $S_0^{-1}BA_0^{-1}B^T$ is W_{22}-self-adjoint. □

Consider now the approximate block diagonal preconditioner (4.24) with $D = B$, $A = E - F$ with E and $BE^{-1}B^T$ nonsingular, so that

$$\mathcal{P} = \begin{bmatrix} E & 0 \\ 0 & BE^{-1}B^T \end{bmatrix}.$$

If $A_0 = E$, $A_0^{-1}A$ in Corollary 4.14 is

$$A_0^{-1}A = E^{-1}A = E^{-1}(E - F) = I - E^{-1}F,$$

which implies that $A_0^{-1}A$ is W_{11}-self-adjoint if and only if $E^{-1}F$ is W_{11}-self-adjoint. Additionally, since $S_0 = BE^{-1}B^T$,

$$S_0^{-1}BA_0^{-1}B^T = (BE^{-1}B^T)^{-1}(BE^{-1}B^T) = I.$$

From Corollary 4.14 we infer that any nonsingular symmetric matrix is a valid choice for W_{22} and the only restriction on $\mathcal{W}$ is that W_{11} define a symmetric bilinear form with respect to which $E^{-1}F$ is self-adjoint.

As an example of how this structure might be applied, we consider the Navier-Stokes equations

$$\begin{aligned} -\nu\nabla^2 u + u \cdot \nabla u + \nabla p &= f \\ \nabla \cdot u &= 0 \end{aligned}$$

on a two-dimensional domain $\Omega = [0, 1] \times [0, 1]$ with Dirichlet boundary conditions. We first discretize the domain to obtain the discrete points $x_k = kh_x$, $k = 1, \ldots, n_x$, and $y_j = jh_y$, $j = 1, \ldots n_y$. We use Picard iteration to solve the Navier-Stokes equations, so that at each iteration we must solve an Oseen problem of the form

$$\begin{aligned} -\nu\nabla^2 u + w \cdot \nabla u + \nabla p &= f \\ \nabla \cdot u &= 0, \end{aligned}$$

where w is known. Assume that at the current step w is constant throughout the domain, i.e., that $w(x, y) = (w_x, w_y)^T$ for constants w_x, w_y. In this case, discretizing the Oseen equations by centred finite differences leads to the nonsymmetric saddle point system

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

where A is a finite difference approximation of the convection-diffusion operator of the sort in Example 6.2, i.e., $A = I_{n_y} \otimes A_x + A_y \otimes I_{n_x}$, where

$$A_i = \begin{bmatrix} 2\beta_i & \alpha_i - \beta_i & & \\ -\alpha_i - \beta_i & \ddots & \ddots & \\ & \ddots & \ddots & \alpha_i - \beta_i \\ & & -\alpha_i - \beta_i & 2\beta_i \end{bmatrix},$$

with

$$\alpha_i = w_i \frac{1}{h_i}, \quad \beta_i = \frac{2\nu}{h_i^2}, \ i = x, y.$$

We assume that the Peclet numbers $\frac{w_i h_i}{\nu} < 2$. The matrix B is a central difference approximation of the divergence operator.

We seek a matrix splitting $A = E - F$, E nonsingular, and an inner product $\langle \cdot, \cdot \rangle_{W_{11}}$ such that $E^{-1}F$ is W_{11}-self-adjoint. We begin by finding such splittings for each of A_x and A_y. These are then combined, using Kronecker products, to obtain the desired matrices E, F and W_{11}.

Both A_x and A_y are tridiagonal Toeplitz matrices and so their eigenvalues are [18, Theorem 2.4]

$$\lambda_j^{(i)} = 2\beta_i + 2\sqrt{\beta_i^2 - \alpha_i^2} \cos \frac{\pi j}{n+1},$$

$i = x, y$. Now, $\lambda_j^{(i)}$ is real if $\beta_i^2 \geq \alpha_i^2$ or $2\nu > h_i$, $i = x, y$. However, since we assume that the mesh Peclet numbers are less than 2, this condition is satisfied. Consequently, A_x and A_y have real eigenvalues and are both self-adjoint with respect to inner products. These inner products, which can be obtained similarly to that for the one-dimensional finite difference approximation in Section 2.3 (see Theorem 2.33), are defined by

$$W_i = \frac{2}{n+1} \Theta_i,$$

$i = x, y$, where

$$\Theta_i = \operatorname{diag}\left(\left(\frac{\beta_i - \alpha_i}{\beta_i + \alpha_i} \right), \left(\frac{\beta_i - \alpha_i}{\beta_i + \alpha_i} \right)^2, \ldots, \left(\frac{\beta_i - \alpha_i}{\beta_i + \alpha_i} \right)^n \right).$$

We choose the matrix splittings $A_i = E_i - F_i$, $i = x, y$, where $E_i = \operatorname{diag}(A_i) = 2\beta_i I$.

Since E_i and W_i are diagonal, they commute and it follows that

$$\begin{aligned} W_i A_i &= A_i^T W_i \\ W_i(E_i - F_i) &= (E_i^T - F_i^T)W_i \\ 2\beta_i W_i - W_i F_i &= 2\beta_i W_i - F_i^T W_i \\ W_i F_i &= F_i^T W_i \\ (W_i E_i)E_i^{-1} F_i &= (E_i^{-1} F_i)^T (W_i E_i). \end{aligned}$$

Consequently, $E_i^{-1}F_i$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{W_i E_i}$ and F_i is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{W_i}$. Both $(W_i E_i)$ and W_i are diagonal matrices with positive entries and, therefore, define inner products.

Now let us see how E_i, F_i and W_i can be combined to give E, F and W_{11} using the Kronecker product properties in Appendix A.2. We have that

$$\begin{aligned} A &= I_{n_y} \otimes A_x + A_y \otimes I_{n_x} \\ &= I_{n_y} \otimes (E_x - F_x) + (E_y - F_y) \otimes I_{n_x} \\ &= \left(I_{n_y} \otimes E_x + E_y \otimes I_{n_x}\right) - \left(I_{n_y} \otimes F_x + F_y \otimes I_{n_x}\right) \\ &= E - F, \end{aligned}$$

where $E = I_{n_y} \otimes E_x + E_y \otimes I_{n_x}$ and $F = I_{n_y} \otimes F_x + F_y \otimes I_{n_x}$. This gives us the matrix splitting $A = E - F$. We choose $W = W_y \otimes W_x$, which is symmetric since W_x and W_y are each symmetric. By substituting for W and E we find that

$$\begin{aligned} WE &= (W_y \otimes W_x)\left(I_{n_y} \otimes E_x + E_y \otimes I_{n_x}\right) \\ &= W_y \otimes W_x E_x + W_y E_y \otimes W_x \\ &= 2\beta_x W_y \otimes W_x + 2\beta_y W_y \otimes W_x. \end{aligned}$$

Thus,

$$WE = [2(\beta_x + \beta_y)W_y] \otimes W_x. \tag{4.43}$$

By using Lemma A.2 and the transpose property in Appendix A.2 it can be shown that WE is symmetric positive definite.

Additionally, since F_i is W_i-self-adjoint, $i = x, y$, we have that

$$\begin{aligned} WEE^{-1}F = WF &= (W_y \otimes W_x)\left(I_{n_y} \otimes F_x + F_y \otimes I_{n_x}\right) \\ &= W_y \otimes W_x F_x + W_y F_y \otimes W_x \\ &= W_y \otimes F_x^T W_x + F_y^T W_y \otimes W_x \\ &= \left(I_{n_y} \otimes F_x^T + F_y^T \otimes I_{n_x}\right)(W_y \otimes W_x) \\ &= F^T W, \end{aligned}$$

where we have used the transpose property in Appendix A.2. Thus,

$$WEE^{-1}F = F^TW = F^TE^{-T}E^TW = \left(E^{-1}F\right)^T(WE),$$

since WE and W are symmetric. Finally, we find that one choice for the matrix $\mathcal{W}$ is $\mathcal{W} = \text{diag}(WE, I)$, where WE is given by (4.43). Since $\mathcal{W}$ is a Kronecker sum of diagonal matrices, inner products with $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ are cheap to compute.

This example shows that even for nonsymmetric saddle point problems it may be possible to obtain inner products with respect to which the preconditioned coefficient matrix is self-adjoint. The availability of an inner product means that for this nonsymmetric problem it possible to apply $\mathcal{W}$-PMINRES, with its short-term recurrences and minimization property, rather than a standard method for nonsymmetric problems, provided $\mathcal{W}$ is not too ill-conditioned. Although the PDE considered here is quite simple, and assumes constant flow, it shows that practical Krylov subspace methods in nonstandard inner products are not necessarily restricted to symmetric saddle point problems.

4.3 The SZ$^+$ preconditioner

Let us return to symmetric saddle point problems. As mentioned in Section 4.1, scaling A_0 and S_0 to ensure positive definiteness of $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ can be expensive but is often required for the SZ preconditioner (4.8). Similarly to the BP$^+$ preconditioner (4.6), however, the SZ preconditioner can be modified, to form what we call the "Schöberl-Zulehner$^+$" (SZ$^+$) preconditioner for which $\mathcal{W}$ always defines an inner product. It is always possible to apply a $\mathcal{W}$-PMINRES algorithm to the resulting SZ$^+$-preconditioned system and it is this robustness that makes the preconditioner particularly attractive. We have derived this preconditioner, by considering the self-adjointness conditions in Lemma 4.4, independently of [85, Table 3]. Throughout this section we assume $C = 0$.

Specifically, take

$$\mathcal{P} = \begin{bmatrix} A_0 & -B^T \\ -B & BA_0^{-1}B^T + S_0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ -BA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & -A_0^{-1}B^T \\ 0 & I \end{bmatrix} \tag{4.44}$$

which is a Krzyżanowski preconditioner with $c = d = -1$. The preconditioned coefficient matrix is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where

$$\mathcal{W} = \begin{bmatrix} A + A_0 & \\ & BA_0^{-1}B^T + S_0 \end{bmatrix}. \tag{4.45}$$

Whenever $A + A_0 > 0$ and $BA_0^{-1}B^T + S_0 > 0$, $\mathcal{W}$ defines an inner product so that a sufficient condition for positive definiteness is that A_0 and S_0 are themselves positive definite.

Although $\mathcal{W}$ always defines an inner product, $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to this inner product, similarly to the BP^+ preconditioner.

Theorem 4.15. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$, A is symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full rank and $n \geq m$, be left preconditioned by the SZ^+ preconditioner (4.44)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite. Then $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where $\mathcal{W}$ is given by* (4.45)*.*

Proof. Positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ is equivalent to positive definiteness of $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ with respect to the Euclidean inner product (see Lemma 2.31). To ascertain whether $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite, we first factor $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ as

$$\begin{aligned}
\mathcal{W}\mathcal{P}^{-1}\mathcal{A} &= \begin{bmatrix} A + A_0 & \\ & BA_0^{-1}B^T + S_0 \end{bmatrix} \begin{bmatrix} A_0^{-1} & A_0^{-1}B^T S_0^{-1} \\ & S_0^{-1} \end{bmatrix} \begin{bmatrix} I & \\ BA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \\
&= \begin{bmatrix} M & (A + A_0)VB^T \\ BV(A + A_0) & BVB^T \end{bmatrix} \\
&= \begin{bmatrix} I & \\ BV(A + A_0)M^{-1} & I \end{bmatrix} \begin{bmatrix} M & \\ & T \end{bmatrix} \begin{bmatrix} I & M^{-1}(A + A_0)VB^T \\ & I \end{bmatrix}, \qquad (4.46)
\end{aligned}$$

where

$$M = A + AA_0^{-1}A + (A + A_0)A_0^{-1}B^T S_0^{-1}BA_0^{-1}(A + A_0), \qquad (4.47)$$

$$T = BV(V^{-1} - (A + A_0)M^{-1}(A + A_0))VB^T \qquad (4.48)$$

and

$$V = A_0^{-1} + A_0^{-1}B^T S_0^{-1}BA_0^{-1}. \qquad (4.49)$$

The congruence transform (4.46) shows that $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ whenever $M > 0$ and $T > 0$. Now, A_0 and S_0 are positive definite and B has full rank so M and V, given by (4.47) and (4.49), are each the sum of a positive definite matrix and a positive semidefinite matrix and are themselves positive definite. However, the matrix T in (4.48) is negative definite as we now show. Negative definiteness of T is equivalent to

$$V^{-1} < (A + A_0)M^{-1}(A + A_0)$$

or, equivalently, since M and V are positive definite (see Corollary A.16), that

$$V > (A + A_0)^{-1} M (A + A_0)^{-1}.$$

Substituting for V and M gives that

$$A_0^{-1} + A_0^{-1} B^T S_0^{-1} B A_0^{-1} > A_0^{-1} A (A + A_0)^{-1} + A_0^{-1} B^T S_0^{-1} B A_0^{-1}$$

or that

$$0 > A_0^{-1} A (A + A_0)^{-1} - A_0^{-1}. \tag{4.50}$$

Now,

$$A_0^{-1} A (A + A_0)^{-1} = A_0^{-1} \left[A (A + A_0)^{-1} A_0 \right] A_0^{-1} = A_0^{-1} \left(A^{-1} + A_0^{-1} \right)^{-1} A_0^{-1}$$

and so the inequality (4.50) can be expressed as

$$0 > A_0^{-1} ((A^{-1} + A_0^{-1})^{-1} - A_0) A_0^{-1}.$$

This holds if and only if

$$A_0 > (A^{-1} + A_0^{-1})^{-1}$$

or

$$A_0^{-1} < A_0^{-1} + A^{-1}.$$

Since A^{-1} is positive definite this is satisfied and T is negative definite. Considering (4.46) we see that, since M is positive definite and T is negative definite, $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ is indefinite by Sylvester's law of inertia (see Theorem A.14 in Appendix A.4). It follows from Lemma 2.31 that $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. □

The indefiniteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\mathcal{W}$ means that a $\mathcal{W}$-PCG method cannot be reliably applied to this problem. However, it is always possible to apply $\mathcal{W}$-PMINRES, the convergence of which, as Corollary 3.15 indicates, is determined essentially by the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$. Although it is of course possible to apply I-MINRES to the original saddle point problem (4.2), the SZ$^+$ preconditioner aims to achieve a nicer eigenvalue distribution, that should accelerate convergence.

Bounds on the eigenvalues of the preconditioned coefficient matrix allow the speed of convergence of $\mathcal{W}$-PMINRES to be ascertained a priori. In order to obtain such bounds, suppose we can find δ, Δ and Φ such that

$$1 < \delta \leq \psi = \frac{t^T \tilde{A} t}{t^T A t} \leq \Delta \tag{4.51}$$

and

$$0 \leq \omega = \frac{t^T B^T \tilde{S}^{-1} B t}{t^T A t} \leq \Phi \tag{4.52}$$

for all nonzero $t \in \mathbb{R}^n$, where $\tilde{A} = A + A_0$ and $\tilde{S} = BA_0^{-1}B^T + S_0$. When $A = A_0$ and $S = S_0$, i.e., when the approximations are exact, δ and Δ in (4.51) are equal to two and Φ in (4.52) is equal to $\frac{1}{2}$. Bounds on the eigenvalues of the SZ$^+$ preconditioned saddle point matrix that utilize (4.51) and (4.52) can be combined with Corollary 3.15 in Chapter 3 to provide a priori information about $\mathcal{W}$-PMINRES convergence.

Lemma 4.16. *Let*

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

where $A \in \mathbb{R}^{n\times n}$, A is symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full rank and $n \geq m$, be left preconditioned by the SZ$^+$ preconditioner (4.44)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite. Then the eigenvalues λ of $\mathcal{P}^{-1}\mathcal{A}$ are real and, letting*

$$\mu = \frac{\lambda}{1+\lambda},$$

satisfy either

$$\frac{1 - \sqrt{1 + 4\Phi\delta}}{2\delta} \leq \mu < 0 \tag{4.53}$$

or

$$\frac{1}{\Delta} \leq \mu \leq \frac{1 + \sqrt{1 + 4\Phi\delta}}{2\delta}, \tag{4.54}$$

where δ, Δ and Φ are given by (4.51) *and* (4.52)*.*

Proof. The SZ$^+$ preconditioner is a Krzyżanowski preconditioner with $c = -1$. From Theorem 4.2 we know that λ is real and that $\lambda \neq 0, -1$. To progress further we re-write the SZ$^+$ generalized eigenvalue problem $\mathcal{A}x = \lambda\mathcal{P}x$ using the approach in [122]. Expressly, we have that

$$\begin{aligned} \mathcal{A}x &= \lambda\mathcal{P}x \\ (1+\lambda)\mathcal{A}x &= \lambda(\mathcal{P} + \mathcal{A})x \\ \mathcal{A}x &= \mu\mathcal{W}x, \end{aligned}$$

where

$$\mathcal{W} = \mathcal{A} + \mathcal{P} = \begin{bmatrix} A + A_0 & 0 \\ 0 & BA_0^{-1}B^T + S_0 \end{bmatrix}$$

defines the SZ$^+$ inner product (see 4.45) and

$$\mu = \frac{\lambda}{\lambda + 1}.$$

Since $\lambda \neq 0, -1$, it follows that $\mu \neq 0, \infty$.

The matrix $\mathcal{W}$ has the structure of a Krzyżanowski preconditioner with $c = d = 0$. This allows us to use Theorem 4.2 provided we replace A_0 and S_0 by $A + A_0$ and $S_0 + BA_0^{-1}B^T$, respectively.

From Theorem 4.2, when $Bu = 0$

$$\lambda = \frac{u^T A u}{u^T \tilde{A} u}$$

and so

$$0 < \frac{1}{\Delta} \leq \mu = \frac{u^T A u}{u^T \tilde{A} u} \leq \frac{1}{\delta} < 1. \tag{4.55}$$

If $Bu \neq 0$, Theorem 4.2 gives that

$$\mu = \frac{1 \pm \sqrt{1 + 4\omega\psi}}{2\psi}, \tag{4.56}$$

where ω and ψ are the generalized Rayleigh quotients (4.51) and (4.52). The eigenvalues λ are real since $4\omega\psi \geq 0$, with the larger positive and the smaller negative. We now bound these eigenvalues using (4.51) and (4.52).

Let us consider the larger root first. Now,

$$\mu = \frac{1 + \sqrt{1 + 4\omega\psi}}{2\psi} \leq \frac{1 + \sqrt{1 + 4\Phi\psi}}{2\psi} \leq \frac{1 + \sqrt{1 + 4\Phi\delta}}{2\delta}.$$

Meanwhile, the lower bounds $0 \leq \omega$ and $1/\Delta \leq \psi$ show that

$$\mu = \frac{1 + \sqrt{1 + 4\omega\psi}}{2\psi} \geq \frac{1}{\psi} \geq \frac{1}{\Delta}$$

so that if μ satisfies (4.56) then

$$\frac{1}{\Delta} \leq \mu \leq \frac{1 + \sqrt{1 + 4\Phi\delta}}{2\delta}.$$

Since $4\omega\delta \geq 0$, (4.55) is subsumed by this bound.

The negative root of (4.56) and the lower bound in (4.52) imply that

$$\mu = \frac{1 - \sqrt{1 + 4\omega\psi}}{2\psi} \leq \frac{1 - \sqrt{1 + 0}}{2\psi} \leq 0.$$

Since $\mu \neq 0$, this simply states that μ is negative.

Substituting the upper bound in (4.52) into the negative root in (4.56) gives that

$$\mu = \frac{1 - \sqrt{1 + 4\omega\psi}}{2\psi} \geq \frac{1 - \sqrt{1 + 4\Phi\psi}}{2\psi} \tag{4.57}$$

the minimum of which, with respect to ψ, can be found by differentiating. We find that

$$\frac{1}{2}\frac{d}{d\psi}\left[(1-(1+4\Phi\psi)^{\frac{1}{2}})\psi^{-1}\right] = \frac{1+2\Phi\psi-\sqrt{1+4\Phi\omega}}{2\psi^2\sqrt{1+4\Phi\psi}}. \tag{4.58}$$

The numerator and denominator of the right-hand side of (4.58) are positive since $\sqrt{(1+2\Phi\psi)^2} = \sqrt{1+4\Phi\psi+4\Phi^2\psi^2} > \sqrt{1+4\Phi\psi}$. Thus,

$$\frac{1}{2}\frac{d}{d\psi}\left[(1-(1+4\Phi\psi)^{\frac{1}{2}})\psi^{-1}\right] > 0 \tag{4.59}$$

and

$$\frac{1-\sqrt{1+4\Phi\psi}}{2\psi}$$

is a monotonically increasing function of ψ. The choice $\psi = \delta$ minimizes (4.57) and

$$\mu = \frac{1-\sqrt{1+4\omega\psi}}{2\psi} \geq \frac{1-\sqrt{1+4\Phi\delta}}{2\delta}. \tag{4.60}$$

Combining (4.59) and (4.60) shows us that if μ is negative it must be bounded by

$$\frac{1-\sqrt{1+4\Phi\delta}}{2\psi} \leq \mu < 0.$$

□

We illustrate the bounds in Lemma 4.16 for a Stokes flow example (Example 4.2 below with $h = 2^{-4}$). We take A_0 to be the incomplete Cholesky factorization with zero-fill computed by the Matlab function `ichol` and S_0 to be the pressure mass matrix computed by Ifiss [125].

The eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ lie in $(-0.26,-0.09)\cup(0.34,5.6)$ and so $\mu \in (-0.36,-0.10)\cup(0.25,0.85)$. Additionally,

$$1.8 < \frac{x^T\tilde{A}x}{x^TAx} < 6.2 \quad \text{and} \quad \frac{y^TB^T\tilde{S}^{-1}By}{y^TAy} < 0.77.$$

Therefore, Lemma 4.16 gives the bounds

$$-0.43 \leq \mu < 0 \quad \text{or} \quad 0.16 \leq \mu \leq 0.98.$$

These clearly bound μ and are shown, along with the exact values of μ, in Figure 4.1. We note however, that the smallest negative eigenvalue is bounded away from the origin, which is not highlighted by the bound.

The lack of a useful upper bound on negative eigenvalues is unfortunate, and stems from (4.52). However, for certain classes of problems it may be possible to improve this bound on negative eigenvalues; i.e., to bound eigenvalues away from the origin as in [126] for block diagonal preconditioners of Stokes problems.

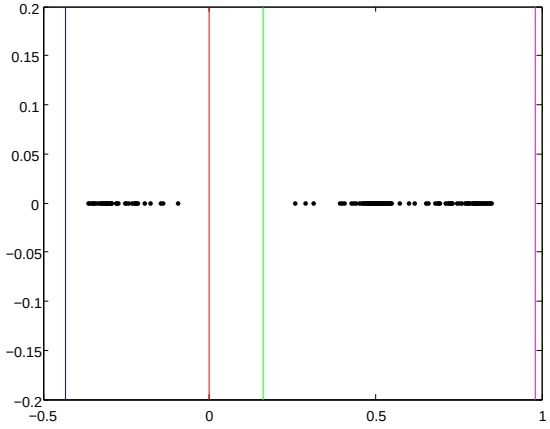


Figure 4.1: Bounds on μ for the SZ$^+$ preconditioned saddle point matrix for the channel flow example (Example 4.2) with $h = 2^{-4}$.

4.4 Numerical experiments

In this section we compare our SZ$^+$ preconditioner to more established preconditioners for symmetric saddle point problems: the block diagonal, BP, BP$^+$ and SZ preconditioners described in Section 4.1. Both I-PMINRES and $\mathcal{W}$-PMINRES are applied to the block diagonal preconditioned system, $\mathcal{W}$-PCG and $\mathcal{W}$-PMINRES to the BP and SZ preconditioned systems and $\mathcal{W}$-PMINRES to the BP$^+$ and SZ$^+$ preconditioned systems (see Table 4.2 for a summary). In all cases, the stopping criterion is a reduction in the relative preconditioned residual, measured in the I-norm, of 10^{-6} and the maximum number of iterations is 200. If convergence is not achieved by this maximum we say that the method has failed, which we denote by '–' in the following tables. For clarity, quantities associated with each preconditioner are differentiated by abbreviations. Thus BDI and BDW distinguish between block diagonal preconditioned I-MINRES and $\mathcal{W}$-MINRES while BP-CG and BP-MR, and SZ-CG and SZ-MR discriminate between $\mathcal{W}$-PCG and $\mathcal{W}$-PMINRES for the BP and SZ preconditioners.

4.4.1 Optimization problems

Example 4.1. Our first test problems are matrices from the CUTEr test set [60] and SIT100 from the GHS_Indef test set in the University of Florida Sparse Matrix

Method	Preconditioner	Inner Product	Krylov subspace method
BDI	Block diagonal (4.10)	I	Preconditioned MINRES ([62, p122])
BDW	Block diagonal (4.10)	$\mathcal{W}_{BDW}$ (4.11)	$\mathcal{W}$-PMINRES (Section 3.2.1)
BP-MR	BP (4.4)	$\mathcal{W}_{BP}$ (4.5)	$\mathcal{W}$-PMINRES (Section 3.2.1)
BP-CG	BP (4.4)	$\mathcal{W}_{BP}$ (4.5)	$\mathcal{W}$-PCG (Algorithm 3.3)
BP$^+$	BP$^+$ (4.6)	$\mathcal{W}_{BP^+}$ (4.7)	$\mathcal{W}$-PMINRES (Section 3.2.1)
SZ-MR	SZ (4.8)	$\mathcal{W}_{SZ}$ (4.9)	$\mathcal{W}$-PMINRES (Section 3.2.1)
SZ-CG	SZ (4.8)	$\mathcal{W}_{SZ}$ (4.9)	$\mathcal{W}$-PCG (Algorithm 3.3)
SZ$^+$	SZ$^+$ (4.44)	$\mathcal{W}_{SZ^+}$ (4.45)	$\mathcal{W}$-PMINRES (Section 3.2.1)

Table 4.2: A summary of the preconditioners, inner products and methods used in the saddle point numerical experiments.

Problem	n	m	$nnz(A)$	$nnz(B)$	γ
BLOCKQP1	2010	1001	2010	14010	1.1
HUES-MOD	10000	2	10000	20000	0
HUESTIS	10000	2	10000	20000	0
MOSARQP1	2500	700	2590	3422	0
MOSARQP2	2500	700	2590	3422	0
PRIMAL4	1489	75	1488	16031	0.0012
QGROW15	645	300	962	5620	0.8
QGROW22	946	440	1639	8252	0.8
SIT100	7142	3120	32478	14284	0

Table 4.3: Problem data for the optimization problems.

Collection [31]. We tested all preconditioners on a larger number of CUTEr problems but present only a representative set here. The SIT100 matrix has a constraint matrix B with rank $m-1$. Consequently, our theory does not apply although we observe that the SZ$^+$ preconditioner still performs well for this problem.

For some of the CUTEr matrices A in (4.2) is indefinite. This is remedied by replacing A by $A_{aug} = A+\gamma I$, where γ is a positive parameter that ensures that A_{aug} is positive definite. These values of γ are tabulated in Table 4.3. The right-hand side vectors were randomly generated, excepting that for SIT100 for which we set $g = 0$ to ensure that the system was consistent. The problem data for each system, including the number of nonzeros, nnz, is given in Table 4.3. The matrix A_0 is obtained by applying a symmetric approximate minimum degree reordering to A, obtained with the Matlab command `symamd`, before performing an incomplete Cholesky factorization with zero-fill with the Matlab function `ichol`. The Schur complement approximation is

$$S_0 = BA_0^{-1}B^T. \tag{4.61}$$

Table 4.4 highlights the robustness of the SZ$^+$ method since only it and the BP$^+$ method converge for every problem. Rather alarmingly, the choice (4.61) for S_0

Problem	BDI	BDW	BP-MR	BP-CG	BP$^+$	SZ-MR	SZ-CG	SZ$^+$
BLOCKQP1	3	3	2	5	3	–	–	3
HUES-MOD	–	3	2	–	3	–	1	3
HUESTIS	3	3	2	–	3	–	–	3
MOSARQP1	–	3	2	2	3	–	1	3
MOSARQP2	–	3	–	2	3	–	1	3
PRIMAL4	3	3	2	–	3	–	–	3
QGROW15	–	3	2	2	3	–	1	3
QGROW22	–	5	–	4	6	–	3	4
SIT100	16	–	–	–	28	–	–	13

Table 4.4: Iteration counts for the saddle point preconditioners for optimization problems.

makes the (2,2) block of $\mathcal{W}_{SZ}$, defined by (4.9), the zero matrix. Although there are some problems for which SZ-CG is fastest, there are others for which it does not converge while SZ-PMINRES is crippled by the absence of an inner product, failing for every problem. The unreliability of the BP-MINRES and BP-CG methods is likely due to indefiniteness of $A - A_0$, although we note that BP-CG occasionally converges even when $\mathcal{W}_{BP}$ is indefinite. In contrast, the BDW, SZ$^+$ and BP$^+$ preconditioners are reliable. This is a particularly appealing feature when little is known about the eigenvalues of A, A_0, S and S_0, i.e., when one of the other methods may fail. Moreover, the SZ$^+$ preconditioner is either more effective than, or as effective as, the BP$^+$ and BDW preconditioners. We observe that both BP$^+$ and SZ$^+$ solve the SIT100 saddle point problem, and that application of the SZ$^+$ preconditioner leads to particularly fast convergence, but we cannot offer an explanation. GMRES could be applied instead in this case. However, the preconditioner may not be as effective for GMRES, the convergence of which can depend on factors other than the eigenvalues of the preconditioned system.

4.4.2 Stokes problems

Our remaining examples are finite element approximations of four different Stokes problems. Each comes from [43] and satisfies the coupled partial differential equations

$$-\nabla^2 u + \nabla p = 0 \tag{4.62}$$

$$\nabla \cdot u = 0 \tag{4.63}$$

on Ω, where $u = \left[u_x, u_y \right]^T$ is the velocity and p the scalar pressure. However, the domains and boundary conditions differ, as we now describe.

Example 4.2. (Channel flow) On the domain $\Omega = (-1,1) \times (-1,1)$ a Stokes solution is

$$u_x = 1 - y^2,\ u_y = 0,\ p = -2x + \text{constant}. \tag{4.64}$$

This velocity solution defines a Dirichlet condition on the inflow boundary $x = -1$ while at $y = \pm 1$, the characteristic boundary, $u_x = u_y = 0$. The stress-free Neumann condition

$$\frac{\partial u_x}{\partial x} - p = 0 \quad \text{and} \quad \frac{\partial u_y}{\partial x} = 0$$

is imposed at the outflow boundary ($x = 1, -1 < y < 1$), which has the effect of uniquely specifying p.

Example 4.3. (Backward step) The domain Ω is L-shaped and its boundary is defined by $-1 < x < 5$ with $0 < y < 1$ when $x < 0$ and $-1 < y < 1$ when $x > 0$. A Poiseuille flow profile (similar to (4.64)) is specified on the inflow boundary ($x = -1, 0 \leq y \leq 1$) and a no-flow condition imposed on the walls. The mean outflow pressure is zero because of the stress-free Neumann condition

$$\frac{\partial u_x}{\partial x} - p = 0 \quad \text{and} \quad \frac{\partial u_y}{\partial x} = 0$$

which is applied at the outflow boundary ($x = 5, -1 \leq y \leq 1$).

Example 4.4. (Regularized cavity) The third problem is defined on the square domain $\Omega = (-1,1) \times (-1,1)$. It is a regularized cavity flow problem with a lid that moves from left to right with velocity $u_x = 1 - x^4$ on $y = 1,\ -1 \leq x \leq 1$. No-flow conditions are imposed at the walls.

Example 4.5. (Colliding flow) The final problem is an example of a colliding flow on the domain $\Omega = (-1,1) \times (-1,1)$, for which

$$u_x = 200xy^3,\ u_y = 5x^4 - 5y^4,\ p = 60x^2y - 20y^3 + \text{ constant}$$

with compatible Dirichlet boundary conditions.

To arrive at the finite element approximation, each is cast as a variational problem and discretized using stable Taylor-Hood (Q_2-Q_1) finite elements by the Matlab package IFISS [125]. The resulting symmetric saddle point problem is

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

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{m \times n}$. The matrix dimensions for each example and different mesh sizes are given in Table 4.5. The approximation A_0 is an incomplete

Problem	h	n	m	$nnz(A)$	$nnz(B)$
Channel flow	2^{-3}	162	25	1338	552
	2^{-4}	578	81	6698	2444
	2^{-5}	2178	289	29546	10768
Backward step	2^{-3}	418	61	3834	1603
	2^{-4}	1538	209	18842	7140
	2^{-5}	5890	769	82582	30483
Regularized cavity	2^{-3}	162	25	1122	494
	2^{-4}	578	81	6178	2318
	2^{-5}	2178	289	28418	10460
Colliding flow	2^{-3}	162	25	1122	494
	2^{-4}	578	81	6178	2318
	2^{-5}	2178	289	28418	10460

Table 4.5: Problem data for the Stokes flow problems.

Problem	h	BDI	BDW	BP-MR	BP-CG	BP$^+$	SZ-MR	SZ-CG	SZ$^+$
Channel flow	2^{-3}	38	38	–	22	41	–	19	40
	2^{-4}	55	57	–	36	59	–	28	63
	2^{-5}	91	95	–	69	95	–	51	109
Backward step	2^{-3}	54	55	–	32	57	–	26	67
	2^{-4}	83	83	–	59	88	–	47	98
	2^{-5}	147	149	–	121	147	–	88	169
Regularized cavity	2^{-3}	28	32	–	19	34	–	17	37
	2^{-4}	43	48	–	35	52	–	25	60
	2^{-5}	68	81	–	66	88	–	47	99
Colliding flow	2^{-3}	26	28	–	19	28	–	15	33
	2^{-4}	36	41	–	32	46	–	25	50
	2^{-5}	65	71	–	61	72	–	43	81

Table 4.6: Iteration counts for the saddle point preconditioners for the Stokes problems.

Cholesky factorization of A with zero-fill, computed by the Matlab function `ichol` and S_0 is the pressure mass matrix provided by Ifiss. It would be interesting to investigate the effect of replacing A_0 by a spectrally equivalent approximation of A, for which mesh-independent convergence could be observed. We note that Examples 4.4 and 4.5 are rank deficient since $\text{rank}(B) = m - 1$ but the systems are consistent and pose no difficulty for the solvers. However, as for SIT100 in the previous example, our theory does not apply in this case.

Table 4.6 shows the iteration counts for the different preconditioners. The BP- and SZ-CG methods are the most efficient despite $A - A_0$ being indefinite. This was previously noted in [34, 135] for the BP method applied to Stokes problems. We stress, however, that the unpredictability of these methods, exemplified by the results for Example 4.1, make them less attractive for general problems. Even for these Stokes problems scaling A_0 by σ, with $\sigma < 0.3$, is enough to cause stagnation of

the SZ-CG method for all problems except Examples 4.2, 4.4 and 4.5 with $h = 2^{-3}$, i.e., on the smallest grid. The block diagonal preconditioner performs well for these problems, probably because A_0 and S_0 are sufficiently good approximations of A and S for Stokes flow problems (see [43, Chapters 5 and 6]), although the block diagonal preconditioner may be less effective on other problems such as those in Example 4.1.

4.5 Conclusions

In this chapter we found explicit expressions for the eigenvalues of the Krzyżanowski preconditioned saddle point matrix. We additionally characterized the symmetric bilinear forms with respect to which a nonsymmetric saddle point problem, preconditioned by a Krzyżanowski-type preconditioner, is self-adjoint. This was then used to determine the symmetric bilinear forms with respect to which two Murphy-Golub-Wathen preconditioned saddle point matrices, and a second block diagonal preconditioned matrix, are self-adjoint. We derived eigenvalue bounds for the SZ^+ preconditioner and showed that $\mathcal{W}_{SZ^+}$ always defines an inner product, although it is one with respect to which $\mathcal{P}_{SZ^+}^{-1}\mathcal{A}$ is indefinite. The importance of this inner product cannot be understated; it means that a $\mathcal{W}$-PMINRES method can be reliably applied to SZ^+ preconditioned problems without the additional computations required to scale the preconditioner or the possibility of failure. Numerical results confirm the robustness of the the SZ^+ preconditioner for a range of different test problems for which the block diagonal, BP and SZ preconditioners can fail. Furthermore, we found that for the optimization problems considered the SZ^+ preconditioner is more effective than, or as effective as, the BP^+ and BDW preconditioners. Additional theory that explains the good convergence of the SZ^+ preconditioner for the SIT100 matrix in Example 4.1, for which the constraint matrix B is rank deficient, would be welcome. In the next chapter we shall see how combining some of the preconditioners described in this chapter, including the SZ^+ preconditioner, leads to superior methods.

Chapter 5

Combination preconditioning for saddle point systems

In the previous chapter we introduced a number of preconditioners for symmetric saddle point matrices

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}$$

for which the preconditioned saddle point matrix is self-adjoint with respect to an inner product. Two such preconditioners can be combined by a technique Stoll and Wathen [135] call **combination preconditioning** to form a new preconditioner such that the preconditioned saddle point matrix is self-adjoint with respect to a symmetric bilinear form or inner product. The combination is controlled by two parameters, for certain choices of which the combination preconditioner is more effective than either of the original preconditioners. Furthermore, its use requires the same number of solves with A_0 and S_0 as the more expensive of the two. The process of combination preconditioning is described in the following lemma.

Lemma 5.1. [135, Lemma 3.5] *If $\mathcal{P}_1$ and $\mathcal{P}_2$ are left preconditioners for the symmetric matrix $\mathcal{A}$ for which symmetric matrices $\mathcal{W}_1$ and $\mathcal{W}_2$ exist with $\mathcal{P}_1^{-1}\mathcal{A}$ self-adjoint in $\langle\cdot,\cdot\rangle_{\mathcal{W}_1}$ and $\mathcal{P}_2^{-1}\mathcal{A}$ self-adjoint in $\langle\cdot,\cdot\rangle_{\mathcal{W}_2}$ and if*

$$\alpha\mathcal{P}_1^{-T}\mathcal{W}_1 + \beta\mathcal{P}_2^{-T}\mathcal{W}_2 = \mathcal{P}_3^{-T}\mathcal{W}_3 \tag{5.1}$$

for some matrix $\mathcal{P}_3$ and some symmetric matrix $\mathcal{W}_3$, then $\mathcal{P}_3^{-1}\mathcal{A}$ is self-adjoint in $\langle\cdot,\cdot\rangle_{\mathcal{W}_3}$.

Remark 5.2. An alternative proof to that in [135] results from premultiplying (5.1) by $\mathcal{A}$. Since $\mathcal{P}_i^{-1}\mathcal{A}$ is $\mathcal{W}_i$-self-adjoint, $i = 1, 2$, $\mathcal{A}\mathcal{P}_i^{-T}\mathcal{W}_i$ is symmetric (see (2.13)).

Provided $\mathcal{P}_3$ and $\mathcal{W}_3$ can be factored as in (5.1), with $\mathcal{W}_3$ symmetric (and nonsingular), $\mathcal{A}\mathcal{P}_3^{-T}\mathcal{W}_3$ is symmetric and, using (2.13) again, we find that $\mathcal{P}_3^{-1}\mathcal{A}$ is self-adjoint with respect to the symmetric bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{W}_3}$.

Lemma 5.1 shows precisely how two preconditioners may be combined. To construct combinations of the preconditioners in the previous chapter, we first apply Lemma 5.1 to two different Krzyżanowski preconditioners (4.12) with bilinear forms (4.14), namely

$$\mathcal{P}_1 = \begin{bmatrix} I & 0 \\ c_1 B A_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & d_1 A_0^{-1} B^T \\ 0 & I \end{bmatrix} \tag{5.2}$$

with its bilinear form defined by

$$\mathcal{W}_1 = \delta_1 \begin{bmatrix} A_0 - c_1 A & 0 \\ 0 & S_0 + c_1 d_1 B A_0^{-1} B^T + d_1 C \end{bmatrix} \tag{5.3}$$

and

$$\mathcal{P}_2 = \begin{bmatrix} I & 0 \\ c_2 B A_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & d_2 A_0^{-1} B^T \\ 0 & I \end{bmatrix} \tag{5.4}$$

with

$$\mathcal{W}_2 = \delta_2 \begin{bmatrix} A_0 - c_2 A & 0 \\ 0 & S_0 + c_2 d_2 B A_0^{-1} B^T + d_2 C \end{bmatrix}. \tag{5.5}$$

Lemma 5.1 gives that

$$\mathcal{P}_3^{-T}\mathcal{W}_3 = \alpha \mathcal{P}_1^{-T}\mathcal{W}_1 + \beta \mathcal{P}_2^{-T}\mathcal{W}_2 \tag{5.6}$$

$$= \alpha\delta_1 \begin{bmatrix} G_{11}^{(1)} & G_{12}^{(1)} \\ G_{21}^{(1)} & G_{22}^{(1)} \end{bmatrix} + \beta\delta_2 \begin{bmatrix} G_{11}^{(2)} & G_{12}^{(2)} \\ G_{21}^{(2)} & G_{22}^{(2)} \end{bmatrix} \tag{5.7}$$

where, for $i = 1, 2$,

$$\begin{aligned} G_{11}^{(i)} &= (A_0^{-1} + c_i d_i A_0^{-1} B^T S_0^{-1} B A_0^{-1})(A_0 - c_i A), \\ G_{12}^{(i)} &= -c_i A_0^{-1} B^T S_0^{-1} (S_0 + c_i d_i B A_0^{-1} B^T + d_i C), \\ G_{21}^{(i)} &= -d_i S_0^{-1} B A_0^{-1} (A_0 - c_i A), \\ G_{22}^{(i)} &= S_0^{-1} (S_0 + c_i d_i B A_0^{-1} B^T + d_i C). \end{aligned} \tag{5.8}$$

It appears difficult, in general, to extract from the additive combination (5.6) the preconditioner $\mathcal{P}_3$ and symmetric bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{W}_3}$. However, for certain choices of c_1, c_2, d_1 and d_2 this can be achieved. The resulting combination preconditioner also retains the structure of a Krzyżanowski preconditioner.

Theorem 5.3. *Let $\mathcal{P}_1$, $\mathcal{W}_1$, $\mathcal{P}_2$ and $\mathcal{W}_2$ be defined by* (5.2), (5.3), (5.4) *and* (5.5), *respectively, where $A, A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite, $C \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite, $B \in \mathbb{R}^{m\times n}$ has full rank and $n \geq m$. Then, if $c_1 = c_2 = c$, $\mathcal{P}_3^{-T}\mathcal{W}_3$ in* (5.6) *is given by*

$$\mathcal{P}_3 = \begin{bmatrix} I & 0 \\ cBA_0^{-1} & I \end{bmatrix} \begin{bmatrix} \frac{1}{\alpha\delta_1+\beta\delta_2}A_0 & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} I & (\alpha\delta_1 d_1 + \beta\delta_2 d_2)A_0^{-1}B^T \\ 0 & I \end{bmatrix}$$

and

$$\mathcal{W}_3 = \begin{bmatrix} (A_0 - cA) & 0 \\ 0 & (\alpha\delta_1 + \beta\delta_2)S_0 + (\alpha\delta_1 d_1 + \beta\delta_2 d_2)(cBA_0^{-1}B^T + C) \end{bmatrix}.$$

Alternatively, if $d_1 = d_2 = 0$, we obtain that

$$\mathcal{P}_3 = \begin{bmatrix} I & 0 \\ (\alpha\delta_1 c_1 + \beta\delta_2 c_2)BA_0^{-1} & I \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & \frac{1}{\alpha\delta_1+\beta\delta_2}S_0 \end{bmatrix}$$

and

$$\mathcal{W}_3 = \begin{bmatrix} (\alpha\delta_1 + \beta\delta_2)A_0 - (\alpha\delta_1 c_1 + \beta\delta_3 c_2)A & 0 \\ 0 & S_0 \end{bmatrix}.$$

Proof. When $c_1 = c_2 = c$, it is straightforward to show that (5.7) becomes

$$\mathcal{P}_3^{-T}\mathcal{W} = \begin{bmatrix} P_{11}W_{11} & P_{12}W_{22} \\ P_{21}W_{11} & P_{22}W_{22} \end{bmatrix}, \tag{5.9}$$

where

$$\begin{aligned} P_{11} &= (\alpha\delta_1 + \beta\delta_2)A_0^{-1} + (\alpha\delta_1 d_1 + \beta\delta_2 d_2)cA_0^{-1}B^T S_0^{-1}BA_0^{-1}, \\ P_{12} &= -cA_0^{-1}B^T S_0^{-1}, \\ P_{21} &= -(\alpha\delta_1 d_1 + \beta\delta_2 d_2)S_0^{-1}BA_0^{-1}, \\ P_{22} &= S_0^{-1}, \end{aligned}$$

$$W_{11} = A_0 - cA \quad \text{and} \quad W_{22} = (\alpha\delta_1 + \beta\delta_2)S_0 + (\alpha\delta_1 d_1 + \beta\delta_2 d_2)(cBA_0^{-1}B^T + C).$$

Thus, a natural choice is

$$\begin{aligned} \mathcal{P}_3^{-1} &= \begin{bmatrix} P_{11}^T & P_{21}^T \\ P_{12}^T & P_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} I & -(\alpha\delta_1 d_1 + \beta\delta_2 d_2)A_0^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} (\alpha\delta_1 + \beta\delta_2)A_0^{-1} & 0 \\ 0 & S_0^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -cBA_0^{-1} & I \end{bmatrix} \end{aligned}$$

so that

$$\mathcal{W}_3 = \begin{bmatrix} W_{11} & 0 \\ 0 & W_{22} \end{bmatrix}$$

and $\mathcal{P}_3$ is as stated in the theorem.

Similarly, when $d_1 = d_2 = 0$ we find from (5.7) and (5.8) that

$$\mathcal{P}_3^{-T}\mathcal{W}_3 = \begin{bmatrix} A_0^{-1}((\alpha\delta_1 + \beta\delta_2)A_0 - (\alpha\delta_1 c_1 + \beta\delta_2 c_2)A) & -(\alpha\delta_1 c_1 + \beta\delta_2 c_2)A_0^{-1}B^T \\ 0 & (\alpha\delta_1 + \beta\delta_2)I \end{bmatrix}$$

and so we can take

$$\mathcal{P}_3^{-1} = \begin{bmatrix} A_0^{-1} & 0 \\ 0 & (\alpha\delta_1 + \beta\delta_2)S_0^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\frac{\alpha\delta_1 c_1 + \beta\delta_2 c_2}{\alpha\delta_1 + \beta\delta_2} B A_0^{-1} & I \end{bmatrix}$$

and

$$\mathcal{W}_3 = \begin{bmatrix} (\alpha\delta_1 + \beta\delta_2)A_0 - (\alpha\delta_1 c_1 + \beta\delta_3 c_2)A & 0 \\ 0 & S_0 \end{bmatrix}.$$

□

Remark 5.4. The combination preconditioners in Theorem 5.3 look somewhat unwieldy. However, similarly to Theorem 4.2, simplifications can be made for specific choices of $\mathcal{P}_1$ and $\mathcal{P}_2$ as we shall see in the remainder of this chapter.

Theorem 5.3 provides several options for creating new preconditioners. Stoll and Wathen [135] proposed a BP-BP$^+$ combination and a BP-SZ combination. In Section 5.1 we construct the BP$^+$-BDW combination preconditioner and in Section 5.2 the BP-BDW combination preconditioner. The combination preconditioner we propose in Section 5.3 is a different BP-SZ combination preconditioner than that in [135]. We feel that it is advantageous that the symmetric bilinear form for our BP-SZ combination preconditioner involves α and β, since tuning these to ensure positive definiteness of the symmetric bilinear form can be easier than scaling A_0 and S_0. Additionally, the two constituent preconditioners are more easily discerned from the new combination. Both $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG require only three-term recurrences and possess an optimality property. Furthermore, their convergence is reasonably well understood, i.e., it depends primarily on the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ (see Corollaries 3.5 and 3.15). These are desirable properties and they make a characterization of the conditions under which $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG can be applied valuable. Thus, for each of the combination preconditioners we derive conditions for the associated symmetric bilinear form to be an inner product and for positive definiteness of the preconditioned coefficient matrix with respect to this inner product. We also compare the effectiveness of the combination preconditioner to that of certain saddle point preconditioners in Chapter 4. A summary of our findings is provided in Section 5.4.

One of the attractive properties of the BP preconditioner defined in Section 4.1 is that it transforms the original indefinite saddle point matrix $\mathcal{A}$ into a matrix that is positive definite with respect to an inner product. Unfortunately, the BP$^+$ and BDW preconditioners are each equipped with an inner product with respect to which the preconditioned linear system is indefinite. Surprisingly, by combining the BP$^+$ and BDW preconditioners as we propose in Section 5.1 the preconditioned coefficient matrix can be made positive definite with respect to an inner product. This enables the solution of the preconditioned linear system by a less expensive conjugate gradient method. The positive definiteness of the BP$^+$-BDW preconditioned matrix also highlights the power of combination preconditioning, that makes it *possible* to achieve definiteness even when the constituent preconditioners do not.

A BP$^+$-SZ$^+$ combination preconditioner can also be formed using Theorem 5.3. However, although the BP$^+$-SZ$^+$ combination preconditioner can outperform each of the BP$^+$ and SZ$^+$ preconditioners it is not as effective as the other preconditioners presented here. Additionally, the BP$^+$-SZ$^+$ is never positive definite with respect to an inner product. For these reasons we reserve analysis of the BP$^+$-SZ$^+$ preconditioner for Appendix C.

Throughout this chapter we assume that $A > 0$, $C \geq 0$, $A_0 > 0$ and $S_0 > 0$ and that B has full rank. As in Chapter 4, we denote by $\mathcal{A}$ the symmetric saddle point matrix in (4.2), by $\mathcal{P}$ a left preconditioner and by $\mathcal{W}$ the symmetric bilinear form with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint. When we need to distinguish between different preconditioners, we will again use subscripts: BP, BP^+, SZ and SZ^+ are associated with the BP, BP$^+$, SZ and SZ$^+$ preconditioners so that, for example, the BP preconditioner will be denoted by $\mathcal{P}_{BP}$. We differentiate between the block diagonal preconditioner with $\mathcal{W}$-MINRES and preconditioned MINRES in the Euclidean inner product by BDW and BDI, respectively; and *comb* represents the relevant combination preconditioner.

In numerical experiments, the termination criteria for $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG are that the preconditioned residual is reduced by a factor of 10^{-6} or that 200 iterations is exceeded. Failure is denote by '–'. For all problems the approximation A_0 is a no-fill incomplete Cholesky factorization, which may be scaled by $\sigma \in \mathbb{R}$, $\sigma > 0$ to ensure positive definiteness of $A - A_0$. These scaling factors are given in Table 5.1. The Schur complement approximation is the pressure mass matrix computed by Ifiss. We perform $\mathcal{W}$-PMINRES or $\mathcal{W}$-PCG for values of α and β between -2 and 2 in steps of 0.1, neglecting those that cause $\mathcal{P}$ to be singular. The $(\alpha, \beta) \in [-2, 2] \times [-2, 2]$ pairs

Problem	h	σ	Problem	h	σ
Channel flow	2^{-3}	0.53	Regularized cavity	2^{-3}	0.59
	2^{-4}	0.19		2^{-4}	0.25
	2^{-5}	0.047		2^{-5}	0.072
Backward step	2^{-3}	0.43	Colliding flow	2^{-3}	0.59
	2^{-4}	0.15		2^{-4}	0.25
	2^{-5}	0.038		2^{-5}	0.072

Table 5.1: Scaling factors for A_0

shown in tables in this chapter are those that give the lowest number of iterations, possibly subject to $\mathcal{W}_{comb}$ defining and inner product and, when a conjugate gradient method is applied, to $\mathcal{P}_{comb}^{-1}\mathcal{A}$ being positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}_{comb}}$.

5.1 The BP$^+$-BDW combination preconditioner

We ask here whether combination preconditioning can be employed to obtain a $\mathcal{W}$-positive definite preconditioned saddle point matrix from two indefinite configurations. One of the advantages of positive definiteness is that $\mathcal{W}$-PCG may be reliably applied instead of $\mathcal{W}$-PMINRES. Even for $\mathcal{W}$-PMINRES, the eigenvalues of a $\mathcal{W}$-positive definite preconditioned saddle point matrix lie on the positive real line and, if clustered, might lead to faster convergence than can be achieved for an indefinite system.

The preconditioners we consider are the BP$^+$ preconditioner (4.6) and the BDW preconditioner (4.10), both of which are described in Section 4.1. From Theorem 5.3, since $c_1 = -1$, $c_2 = 0$, $d_1 = d_2 = 0$ and $\delta_1 = \delta_2 = 1$, we have that

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ -\frac{\alpha}{\alpha+\beta}B & \frac{1}{\alpha+\beta}S_0 \end{bmatrix} \tag{5.10}$$

and

$$\mathcal{W} = \mathcal{W}_3 = \begin{bmatrix} \alpha(A+A_0)+\beta A_0 & 0 \\ 0 & S_0 \end{bmatrix}. \tag{5.11}$$

As we shall see, for this choice of $\mathcal{P}$ and $\mathcal{W}$ there *do* exist α and β for which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to an inner product.

Theorem 5.5. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full rank, $C \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite and $n \geq m$. Let $\mathcal{A}$ be left preconditioned by the BP$^+$-BDW combination preconditioner (5.10)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite, so that $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where $\mathcal{W}$ is defined by* (5.11)*.*

When

I. *$\alpha > 0$ and $\alpha+\beta < 0$, $\mathcal{W}$ defines an inner product, with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite, if and only if*

$$A_0 < -\frac{\alpha}{\alpha+\beta}A;$$

II. *$\alpha > 0$ and $\alpha+\beta > 0$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is indefinite;*

III. *$\alpha < 0$ and $\alpha+\beta > 0$, $\mathcal{W}$ defines an inner product if and only if*

$$A_0 > -\frac{\alpha}{\alpha+\beta}A$$

but $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to this inner product;

IV. *$\alpha < 0$ and $\alpha+\beta < 0$, $\mathcal{W}$ does not define an inner product.*

Proof. Recall from Lemma 2.31 that $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite if and only if $\mathcal{W}\mathcal{P}^{-1}\mathcal{A} > 0$, where

$$\begin{aligned}
\mathcal{W}\mathcal{P}^{-1}\mathcal{A} &= \begin{bmatrix}\alpha(A+A_0)+\beta A_0 & 0\\ 0 & S_0\end{bmatrix}\begin{bmatrix}A_0^{-1} & 0\\ \alpha S_0^{-1}BA_0^{-1} & (\alpha+\beta)S_0^{-1}\end{bmatrix}\begin{bmatrix}A & B^T\\ B & -C\end{bmatrix}\\
&= \begin{bmatrix}\alpha AA_0^{-1}(A+A_0)+\beta A & \alpha AA_0^{-1}B^T+(\alpha+\beta)B^T\\ \alpha BA_0^{-1}(A+A_0)+\beta B & \alpha BA_0^{-1}B^T-(\alpha+\beta)C\end{bmatrix}\\
&= \begin{bmatrix}M & MA^{-1}B^T\\ BA^{-1}M & T\end{bmatrix}\\
&= \begin{bmatrix}I & 0\\ BA^{-1} & I\end{bmatrix}\begin{bmatrix}M & 0\\ 0 & T-BA^{-1}MA^{-1}B^T\end{bmatrix}\begin{bmatrix}I & A^{-1}B^T\\ 0 & I\end{bmatrix},
\end{aligned} \tag{5.12}$$

with

$$M = \alpha AA_0^{-1}A+(\alpha+\beta)A = A(\alpha A_0^{-1}+(\alpha+\beta)A^{-1})A \tag{5.13}$$

and

$$T = \alpha BA_0^{-1}B^T-(\alpha+\beta)C.$$

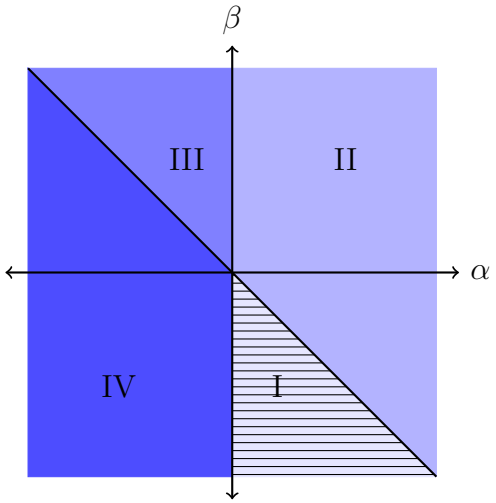


Figure 5.1: Regions (distinguished by different colours) that are considered in Theorem 5.5. The horizontal lines show where $\mathcal{P}^{-1}\mathcal{A}$ can be made positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ by choosing α and β appropriately or by scaling A_0 and S_0.

From (5.12) and Sylvester's law of inertia (see Theorem A.14 in Appendix A.4) it follows that $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite when $M > 0$ and $T > BA^{-1}MA^{-1}B^T$. The latter condition is equivalent to requiring that

$$\alpha BA_0^{-1}B^T - (\alpha+\beta)C > \alpha BA_0^{-1}B^T + (\alpha+\beta)BA^{-1}B^T$$

or that

$$-(\alpha+\beta)C > (\alpha+\beta)BA^{-1}B^T. \tag{5.14}$$

The conditions for positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ are, therefore, that $M > 0$ and (5.14) is satisfied. To ascertain whether these conditions are fulfilled we consider separately each case in Figure 5.1.

I. If $\alpha+\beta < 0$ and $\alpha > 0$, $\mathcal{W}$ is positive definite if and only if

$$A > -\frac{\alpha+\beta}{\alpha}A_0.$$

Corollary A.16 shows that this is also the condition for positive definiteness of M in (5.13). Now, (5.14) is equivalent to $C > -BA^{-1}B^T$. Since the left-hand side of this inequality is symmetric positive semidefinite and the right-hand side is negative definite, it is always satisfied. We have, therefore, that when $\alpha+\beta < 0$, $\alpha > 0$ and $\mathcal{W}$ defines an inner product, $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite.

II. When $\alpha > 0$ and $\alpha+\beta > 0$, $\mathcal{W}$ and M are positive definite. However, since C is positive semidefinite, A is positive definite and B has full rank, (5.14) cannot hold. Indeed, $T - BA^{-1}MA^{-1}B^T < 0$ and $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$.

III. When $\alpha < 0$ with $\alpha + \beta > 0$, $\mathcal{W}$ defines an inner product if and only if

$$\frac{1}{\alpha+\beta}A < -\frac{1}{\alpha}A_0.$$

Furthermore, M in (5.13) is positive definite if and only if $\alpha A_0^{-1} + (\alpha+\beta)A^{-1} > 0$ which, since A_0 and A are positive definite, is equivalent to

$$\frac{1}{\alpha+\beta}A < -\frac{1}{\alpha}A_0$$

by Corollary A.16. It follows that $M > 0$ whenever $\mathcal{W} > 0$. However, as in Case II, (5.14) is not satisfied and $T - BA^{-1}MA^{-1}B^T < 0$. Thus, when $\alpha + \beta > 0$ and $\mathcal{W}$ defines an inner product it is one with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is indefinite.

IV. If $\alpha < 0$ and $\alpha + \beta < 0$, $\mathcal{W}$ is always indefinite and does not define an inner product. We find by calculations similar to those above that M in (5.13) is negative definite. The left-hand side of (5.14) is positive semi-definite while the right-hand side is negative definite. Thus, $T - BA^{-1}MA^{-1}B^T$ is positive definite. □

The pivotal result of Theorem 5.5 is that when $\alpha + \beta < 0$ with $\alpha > 0$ it *is* possible to obtain from the BP$^+$ and BDW preconditioners—for each of which the preconditioned saddle point matrix is indefinite with respect to an inner product—a combination preconditioner for which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint and positive definite with respect to a nonstandard inner product. As previously mentioned, a conjugate gradient method can be used to solve the preconditioned linear system in this circumstance. We shall see below that for these parameters convergence of the combination preconditioner is rapid.

Regardless of whether $\mathcal{W}$-PMINRES or $\mathcal{W}$-PCG is applicable, convergence for the BP$^+$-BDW preconditioned system depends heavily on the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$. These can be bounded when $C = 0$, as the following theorem shows.

Theorem 5.6. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full rank, $C \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite and $n \geq m$. Let $\mathcal{A}$ be left preconditioned by the BP$^+$-BDW combination preconditioner (5.10)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite and let $\mathcal{W}$ in* (5.11) *be positive definite, so that $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to an inner product. Additionally, let*

$$0 < \delta \leq \psi = \frac{u^T A_0 u}{u^T A u} \leq \Delta \tag{5.15}$$

and

$$0 \leq \omega = \frac{u^T B^T S_0^{-1} B u}{u^T A u} \leq \Phi. \tag{5.16}$$

Let $\left[\, u^T,\, v^T \,\right]^T$ *be an eigenvector of* $\mathcal{P}^{-1}\mathcal{A}$*. Then, if* $Bu = 0$*, the corresponding eigenvalue of* $\mathcal{P}^{-1}\mathcal{A}$*,* λ^+ *is positive and satisfies*

$$\frac{1}{\Delta} \leq \lambda^+ \leq \frac{1}{\delta}. \tag{5.17}$$

Otherwise, when

I. $\alpha > 0$ *but* $\alpha + \beta < 0$ *the remaining eigenvalues* λ^+*, which are all positive, satisfy*

$$\frac{1}{\Delta} \leq \lambda^+ \leq \frac{(1+\alpha\Phi) + \sqrt{(1+\alpha\Phi)^2 + 4\delta\Phi(\alpha+\beta)}}{2\delta} \tag{5.18}$$

or

$$0 < \lambda^+ \leq \frac{(1+\alpha\Phi) - \sqrt{(1+\alpha\Phi)^2 + 4\delta\Phi(\alpha+\beta)}}{2\delta}; \tag{5.19}$$

II and III. $\alpha + \beta > 0$*, the remaining positive eigenvalues* λ^+ *of* $\mathcal{P}^{-1}\mathcal{A}$ *satisfy*

$$\frac{1}{\Delta} \leq \lambda^+ \leq \frac{(1+\alpha\Phi) + \sqrt{(1+\alpha\Phi)^2 + 4\delta\Phi(\alpha+\beta)}}{2\delta}$$

while negative eigenvalues λ^- *are bounded by*

$$\frac{(1+\alpha\Phi) - \sqrt{(1+\alpha\Phi)^2 + 4\delta\Phi(\alpha+\beta)}}{2\delta} \leq \lambda^- < 0. \tag{5.20}$$

Proof. The eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are those of the generalized eigenvalue problem

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} A_0 & 0 \\ -\frac{\alpha}{\alpha+\beta} B & \frac{1}{\alpha+\beta} S_0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}. \tag{5.21}$$

The BP$^+$-BDW combination preconditioner is a Krzyżanowski preconditioner with $c = -\frac{\alpha}{\alpha+\beta}$, $d = 0$, S_0 replaced by $\frac{1}{\alpha+\beta} S_0$ and, consequently, ω replaced by $(\alpha+\beta)\omega$ in Theorem 4.2. Thus, by Theorem 4.2, $\lambda \neq 0, -\frac{\alpha+\beta}{\alpha}$ and $u \neq 0$. Additionally, when $Bu = 0$,

$$\lambda = \frac{u^T A u}{u^T A_0 u},$$

and (5.15) then implies that

$$\frac{1}{\Delta} \leq \lambda \leq \frac{1}{\delta}.$$

If $Bu \neq 0$ then , by Theorem 4.2,

$$\lambda = \frac{(1+\alpha\omega) \pm \sqrt{(1+\alpha\omega)^2 + 4\psi\omega(\alpha+\beta)}}{2\psi}. \tag{5.22}$$

These roots must be real since $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to an inner product (see Theorem 5.5 and Lemma 2.32). However, it is possible to verify this using the condition that $\alpha(A + A_0) + \beta A_0 > 0$. Cases I–III in Figure 5.1 require separate examination to determine bounds on the eigenvalues. We present the proof of one case here because the proofs for the remaining two are similar.

I. Let $\alpha > 0$ and $\alpha + \beta < 0$. Both roots in (5.22) are positive, since $\alpha + \beta$ is negative, ψ is positive and ω is nonnegative. The larger root is given by

$$\lambda = \frac{(1 + \alpha\omega) + \sqrt{(1 + \alpha\omega)^2 + 4\psi\omega(\alpha + \beta)}}{2\psi} \tag{5.23}$$

which we see, by considering (5.15) and (5.16), is minimized when $\omega = 0$ and $\psi = \Delta$ and so $\lambda \geq \frac{1}{\Delta}$. Conversely, (5.23) is maximized when $\omega = \Phi$ and $\psi = \delta$ which gives that

$$\frac{1}{\Delta} \leq \lambda \leq \frac{(1 + \alpha\Phi) + \sqrt{(1 + \alpha\Phi)^2 + 4\delta\Phi(\alpha + \beta)}}{2\delta}.$$

The smaller root of (5.22) is

$$\lambda = \frac{(1 + \alpha\omega) - \sqrt{(1 + \alpha\omega)^2 + 4\psi\omega(\alpha + \beta)}}{2\psi}, \tag{5.24}$$

which is maximized when $\omega = 0$ and $\psi = \Delta$. This implies that $\lambda \geq 0$. However, we concluded previously that $\lambda \neq 0$ and we must have that $\lambda > 0$. Additionally, (5.24) is minimized by choosing $\psi = \delta$ and $\omega = \Phi$ which gives that

$$\lambda \geq \frac{(1 + \alpha\Phi) - \sqrt{(1 + \alpha\Phi)^2 + 4\delta\Phi(\alpha + \beta)}}{2\delta}.$$

□

Remark 5.7. As for the SZ$^+$ preconditioner in Section 4.3, neither (5.19) nor (5.20) bound the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ away from the origin. The difficulty is caused by (5.16). However, it may be possible to obtain tighter bounds for certain applications.

Remark 5.8. Theorem 5.6 accords with Lemma 2.32 since the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are real and positive only when $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint and positive definite with respect to the inner product defined by $\mathcal{W}$. When $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is indefinite, the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are real but lie on both sides of the imaginary axis.

Remark 5.9. When $A_0 = A$, (5.17) shows that $\mathcal{P}^{-1}\mathcal{A}$ has eigenvalues at 1. Since the dimension of null(B) is $n - m$, there are $n - m$ such eigenvalues. The remaining eigenvalues can be bounded by (5.18) and (5.19). Alternatively, by expressing (5.21) in component form it can be shown that

$$\frac{\lambda(\lambda - 1)}{\alpha + \beta + \alpha\lambda} Au = B^T S_0^{-1} Bu$$

from which it is clear that the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ depend nonlinearly on the eigenvalues of $A^{-1}B^T S_0^{-1} B$. These, in turn, depend on how well S_0 approximates the Schur complement $S = BA^{-1}B^T$.

If it is possible to obtain bounds of the form (5.15) and (5.16) without too great an expense we can ascertain a priori the eigenvalue distribution of $\mathcal{P}^{-1}\mathcal{A}$. This largely determines the convergence of both $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG. Moreover, it may be possible to use these bounds to choose α and β (and scale A_0 and S_0 if necessary) to obtain good eigenvalues.

Table 4.6 in Chapter 4 shows that the BP$^+$ and BDW preconditioners are effective when applied to the four Stokes problems 4.2–4.5. We now see how much is gained by instead using the combination preconditioner. Since $\mathcal{P}_{comb}^{-1}\mathcal{A}$ can be made positive definite with respect to an inner product, we apply both $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG to the combination preconditioned system, while $\mathcal{W}$-PMINRES is used to solve the BP$^+$ and BDW preconditioned systems. Recall that we perform $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG for values of α and β between -2 and 2, with a step size of 0.1.

We have tabulated in Tables 5.2 and 5.3 the lowest iteration count for the combination preconditioner for which $\mathcal{W}_{comb}$ defines an inner product and, for the conjugate gradient method, for which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to this inner product. For most problems this count is obtained by multiple choices of α and β. More generally, provided $\mathcal{W}_{comb}$ defines an inner product the number of iterations is only mildly sensitive to changes in α and β.

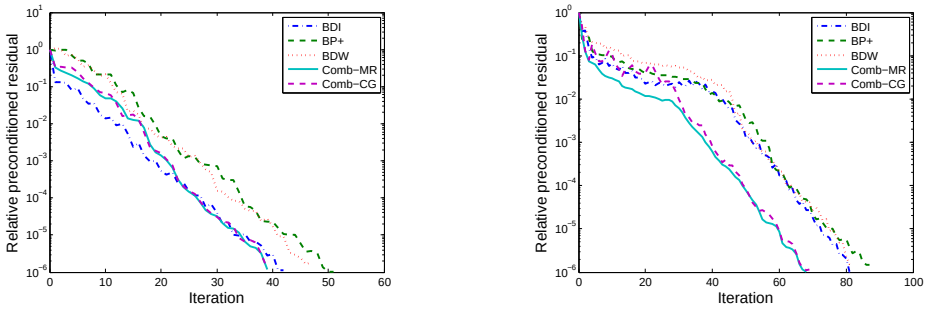
It is clear that the combination preconditioner offers superior performance to either the BP$^+$ preconditioner or the BDW preconditioner. The relative reduction in the number of iterations required by the combination preconditioner, in comparison to the better performing of the BDW and BP$^+$ preconditioners, is 19.8% on average for $\mathcal{W}$-PMINRES and 20.3% on average for $\mathcal{W}$-PCG. When compared with the best of the robust methods, BDI preconditioned I-MINRES, the combination preconditioner is 14.2% better on average when $\mathcal{W}$-PMINRES is applied and 14.9% better when

Problem	h	BDI	BP$^+$	BDW	Comb (α, β)	% BP$^+$/BDW	% BDI
Channel flow	2^{-3}	38	41	38	27 (1.3,-2)	29	29
	2^{-4}	55	59	57	43 (1.7,-2)	25	22
	2^{-5*}	91 (93)	95 (94)	95 (92)	86 (0.7,-0.6)	9	5
Backward step	2^{-3}	54	57	55	41 (1.4,-2)	25	24
	2^{-4}	83	88	83	69 (1.4,-1.6)	17	17
	$2^{-5\dagger}$	147 (148)	147 (145)	148 (155)	140 (1.2,-1)	5	5
Regularized cavity	2^{-3}	28	34	32	21 (1.1,-1.8)	34	25
	2^{-4}	43	52	48	40 (1.2,-1.5)	17	7
	2^{-5}	68	88	81	73 (1.9,-2)	10	-7
Colliding flow	2^{-3}	26	28	28	20 (1.1,-1.8)	29	23
	2^{-4}	36	46	41	34 (0.8,-1)	17	6
	2^{-5}	65	72	71	56 (1.4,-1.5)	21	14

* For positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ it was necessary to scale A_0 by 0.9. Thus, the iteration counts for BDI, BP$^+$ and BDW in parentheses are obtained with $A_0 = 0.9LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab.

† For positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ it was necessary to scale A_0 by 0.7. Thus, the iteration counts for BDI, BP$^+$ and BDW in parentheses are obtained with $A_0 = 0.7LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab.

Table 5.2: Iteration counts for the BP$^+$ preconditioner, the BDW preconditioner, the BDI preconditioner and the best possible BP$^+$-BDW combination preconditioner for $\mathcal{W}$-PMINRES, as well as the relative reduction in the number of iterations required by BP$^+$-BDW combination-MINRES when compared with the better performing of BP$^+$- and BDW-MINRES and compared with block diagonal preconditioned MINRES in the Euclidean inner product. The $(\alpha, \beta) \in [-2, 2] \times [-2, 2]$ pairs shown are those that give the lowest number of iterations and for which $\mathcal{W}_{comb}$ defines an inner product.



(a) backward step

(b) regularized cavity

Figure 5.2: Convergence plots for the BP$^+$-BDW combination preconditioner with $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG for (a) the backward step flow (Example 4.3) with $\alpha = 1.4$ and $\beta = -1.6$ and (b) for regularized cavity flow (Example 4.4) with $\alpha = 1.2$ and $\beta = -1.5$.

$\mathcal{W}$-PCG is used. We remark that in many cases the optimal choices of α and β for $\mathcal{W}$-PMINRES and $\mathcal{W}$-PCG coincide, suggesting that $\mathcal{W}$-PMINRES performs best when $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite with respect to the inner product $\langle\cdot,\cdot\rangle_{\mathcal{W}_{comb}}$. Additionally, the number of iterations for each method is close, and so it appears that for these problems the cheaper $\mathcal{W}$-PCG method is preferable to $\mathcal{W}$-PMINRES. Convergence plots for the values of α and β listed in Tables 5.2 and 5.3 are shown in Figure 5.2. We observe that the $\mathcal{W}$-PCG and $\mathcal{W}$-PMINRES curves are very similar and decrease more rapidly than those of BP$^+$ and BDW preconditioned $\mathcal{W}$-MINRES.

Although optimal $\mathcal{W}_{comb}$-PMINRES convergence occurs when $\mathcal{W}_{comb}$ defines an inner product, faster convergence can be observed for the conjugate gradient method when $\mathcal{W}_{comb}$ is not positive definite, as can be seen from Table 5.3 and Figure 5.2. The average relative reduction when compared with the better of BDW- or BP$^+$-MINRES is 34.8% while the relative reduction when compared with the BDI preconditioner method is 29.9%. However, robustness is not guaranteed when $\langle\cdot,\cdot\rangle_{\mathcal{W}_{comb}}$ is not an inner product and we caution that the BP$^+$-BDW combination-CG method may fail for other approximations of A_0 or S_0 or for different saddle point problems. Note also that the convergence is far from monotone.

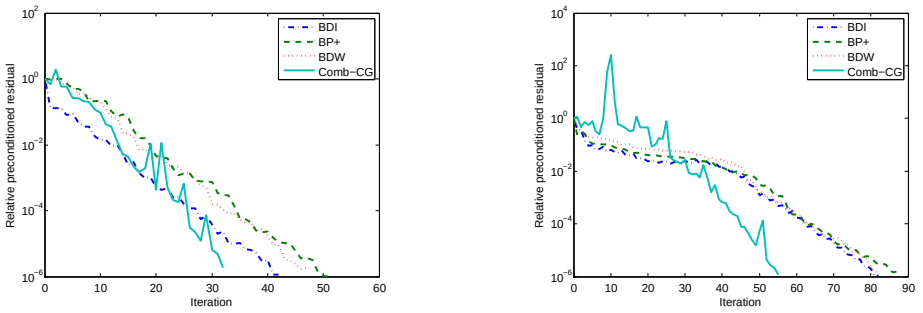
We additionally examined the effect of scaling the BP$^+$ and BDW preconditioners separately, by setting β or α to zero and varying the other parameter. Both preconditioners were sensitive to scaling, although the optimal scaling appeared to be problem dependent.

		$\mathcal{W}_{comb} > 0$			$\mathcal{W}_{comb} \not> 0$		
Problem	h	Comb (α, β)	% BP$^+$/BDW	% BDI	Comb (α, β)	% BP$^+$/BDW	% BDI
Channel flow	2^{-3}	27 (1.3,-2)	29	29	21 (1,-1.9)	45	45
	2^{-4}	43 (1.6,-1.9)	25	22	34 (1.3,-2)	40	38
	2^{-5}*	80 (1.9,-2)	13	14	58 (1.7,-2)	39	36
Backward step	2^{-3}	41 (1.4,-2)	25	24	31 (0.8,-1.7)	44	43
	2^{-4}	70 (1.4,-1.6)	16	16	56 (1.2,-1.9)	33	33
	2^{-5}†	118 (1.7,-1.8)	19	20	99 (1.7,-1.9)	33	33
Regularized cavity	2^{-3}	22 (1.3,-2)	31	21	19 (1,-2)	41	32
	2^{-4}	40 (1.2,-1.5)	17	7	33 (1,-1.9)	31	23
	2^{-5}	73 (1.4,-1.5)	10	-7	59 (1.3,-1.9)	27	13
Colliding flow	2^{-3}	21 (1.3,-1.9)	25	19	19 (1,-2)	32	27
	2^{-4}	35 (1.2,-1.5)	15	3	31 (1.2,-2)	24	14
	2^{-5}	58 (1.5,-1.6)	18	11	51 (1.3,-1.9)	28	22

* For positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ it was necessary to scale A_0 by 0.9. Thus, the iteration count with $\mathcal{W} > 0$ is obtained with $A_0 = 0.9LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab. The iteration count with $\mathcal{W} \not> 0$ is obtained with $A_0 = LL^T$.

† For positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ it was necessary to scale A_0 by 0.7. Thus, the iteration count with $\mathcal{W} > 0$ is obtained with $A_0 = 0.7LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab. The iteration count with $\mathcal{W} \not> 0$ is obtained with $A_0 = LL^T$.

Table 5.3: Iteration counts for the best possible BP$^+$-BDW combination preconditioner for which $\mathcal{W} > 0$ and for which $\mathcal{W} \not> 0$, as well as the relative reduction in the number of iterations required by BP$^+$-BDW combination-MINRES when compared with the better performing of BP$^+$- and BDW-MINRES and compared with block diagonal preconditioned MINRES in the Euclidean inner product. Iteration counts for BP$^+$- and BDW-MINRES and for block diagonal preconditioned MINRES in the Euclidean inner product are given in Table 5.2. The $(\alpha, \beta) \in [-2, 2] \times [-2, 2]$ pairs shown are those that give the lowest number of iterations. For the left half of the table, we restrict (α, β) pairs to those for which $\mathcal{W}_{comb}$ defines an inner product with respect to which $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite.



(a) backward step (b) regularized cavity

Figure 5.3: $\mathcal{W}$-PCG Convergence plots for (a) the backward step flow (Example 4.3) with $\alpha = 1.2$ and $\beta = -1.9$ and (b) the regularized cavity flow (Example 4.4) with $\alpha = 1$ and $\beta = -1.9$. $\mathcal{W}$ is not positive definite for these parameter choices.

In summary, the BP$^+$-BDW combination preconditioner is very effective when used with either a conjugate gradient method or a MINRES algorithm. For BP$^+$-BDW combination-MINRES, the preconditioner is most effective when $\mathcal{W}_{comb}$ defines an inner product with respect to which $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite. Positive definiteness of both $\mathcal{W}_{comb}$ and of the preconditioned saddle point matrix with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ is characterized by Theorem 5.5. Combination preconditioned $\mathcal{W}$-CG performs very well when $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is self-adjoint and positive definite with respect to a nonstandard inner product as the theory supports and good convergence is observed, but remains unexplained, when $\mathcal{W}$ is indefinite.

5.2 The BP-BDW combination preconditioner

We saw in the previous section that it was possible to combine the BP$^+$ and BDW preconditioners. A natural extension is to construct a BP-BDW combination preconditioner. However, the approximation to the Schur complement in the BP preconditioner differs from that in the BDW preconditioner and so Lemma 5.1, rather than Theorem 5.3, is used to obtain a combination.

For the BP preconditioner (4.4), with its symmetric bilinear form defined by (4.5), and the BDW preconditioner (4.10), with its symmetric bilinear form defined by (4.11), Lemma 5.1 gives that

$$\mathcal{P}_3^{-T}\mathcal{W}_3 = \begin{bmatrix} A_0^{-1} & \alpha A_0^{-1}B^T S_0^{-1} \\ 0 & (\beta-\alpha)S_0^{-1} \end{bmatrix} \begin{bmatrix} \alpha A + (\beta-\alpha)A_0 & 0 \\ 0 & S_0 \end{bmatrix}.$$

Accordingly, we take

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ \frac{\alpha}{\alpha-\beta}B & \frac{1}{\beta-\alpha}S_0 \end{bmatrix} \tag{5.25}$$

and

$$\mathcal{W} = \begin{bmatrix} \alpha A + (\beta-\alpha)A_0 & 0 \\ 0 & S_0 \end{bmatrix}. \tag{5.26}$$

Recall that $\mathcal{W}$-PMINRES can be applied to the BP$^+$-BDW combination preconditioned system when $\mathcal{W}$ in (4.9) is positive definite and that the application of a $\mathcal{W}$-PCG method additionally requires that $\mathcal{P}^{-1}\mathcal{A}$ be positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$. The next theorem shows that $\mathcal{W}$ defines an inner product for many choices of α and β and that for certain of these $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to this inner product.

Theorem 5.10. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full rank, $C \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite and $n \geq m$. Let $\mathcal{A}$ be left preconditioned by the BP-BDW combination preconditioner (5.25)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite, so that $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where $\mathcal{W}$ is defined by* (5.26)*.*

When

- **I.** *$\alpha > 0$ and $\alpha > \beta$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite when $\alpha A > (\alpha - \beta)A_0$;*
- **II.** *$\beta > \alpha > 0$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is indefinite;*
- **III.** *$\alpha < 0$ and $\beta > \alpha$, $\mathcal{W}$ defines an inner product when $(\beta - \alpha)A_0 > -\alpha A$ but $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$;*
- **IV.** *$\beta < \alpha < 0$, $\mathcal{W}$ does not define an inner product.*

Proof. The proof differs only slightly from that for Theorem 5.5 and so we only briefly sketch the details here. Again, $\mathcal{W}$-positive definiteness of $\mathcal{P}^{-1}\mathcal{A}$ depends on the positive definiteness of

$$\mathcal{W}\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & T - BA^{-1}MA^{-1}B^T \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix},$$

where

$$M = A(\alpha A_0^{-1} + (\beta - \alpha)A^{-1})A \tag{5.27}$$

and

$$T = \alpha B A_0^{-1} B^T + (\alpha - \beta)C. \tag{5.28}$$

By Sylvester's law of inertia in Appendix A.4, $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ is positive definite when $M > 0$ and $T > BA^{-1}MA^{-1}B^T$. Substituting for M and T shows that the second condition is satisfied if and only if

$$(\alpha - \beta)C > (\beta - \alpha)BA^{-1}B^T. \tag{5.29}$$

Each case in Figure 5.4 is considered separately to determine whether $M > 0$ and (5.29) holds.

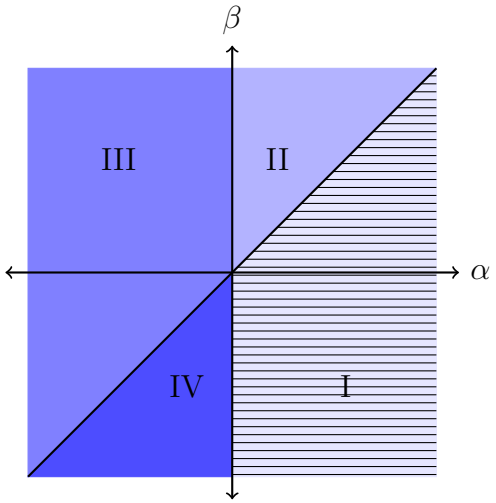


Figure 5.4: Regions (distinguished by different colours) that are considered in Theorem 5.10. The horizontal lines show where $\mathcal{P}^{-1}\mathcal{A}$ can be made positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ by choosing α and β appropriately or by scaling A_0 and S_0.

I. If $\alpha > \beta$ and $\alpha > 0$ the condition for self-adjointness of $\mathcal{W}$ is immediately evident and it is easily shown that when it is satisfied $M > 0$. Additionally, since $(\alpha - \beta)C$ is positive semidefinite and $(\beta - \alpha)BA^{-1}B^T$ is negative definite, (5.29) is satisfied. Thus, when $\alpha > \beta$, $\alpha > 0$ and $\mathcal{W}$ defines an inner product, it is one with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite.

II. With $\beta > \alpha > 0$, $\mathcal{W}$ and M are positive definite. However, the left-hand side of (5.29) is negative semidefinite while the right-hand side is positive definite. Thus, $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$.

III. Case III corresponds to $\alpha < 0$ and $\beta > \alpha$. The condition for positive definiteness of $\mathcal{W}$ is readily verified and simple algebra shows that $M > 0$ when $\mathcal{W} > 0$. However, as in Case II the left-hand side of (5.29) is negative semidefinite and the right-hand side is definite. It follows that $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$.

IV. When $\beta < \alpha < 0$, $\mathcal{W}$ is indefinite. We find that $M < 0$ but that (5.29) is satisfied. □

From Theorem 5.10 we conclude that $\mathcal{W}$-PMINRES is applicable for a wide range of α and β and that, furthermore, for certain values of $\alpha > 0$ with $\alpha > \beta$, a $\mathcal{W}$-PCG method can be applied.

We report now on the performance of the BP-BDW combination-CG method, comparing it to BP-CG and BDW-MINRES. The convergence of BP-BDW combination-MINRES was also investigated but this method was found to be less effective and

				$\mathcal{W}_{comb} > 0$		$\mathcal{W}_{comb} \not> 0$	
Problem	h	BP-CG	BDW	Comb (α, β)	% BP/BDW	Comb (α, β)	% BP/BDW
Channel flow	2^{-3}	28	36	27 (1.3,0)	4	20 (1.7,-2)	29
	2^{-4}	43	56	43 (1,0)	0	35 (0.8,-2)	19
	2^{-5}	77	91	73 (1.1,-0.1)	5	59 (0.3,-2)	23
Backward step	2^{-3}	43	52	41 (1.3,0)	5	30 (1.1,-1.4)	30
	2^{-4}	71	90	65 (1.9,0)	8	56 (0.7,-1.5)	21
	2^{-5}	125	164	116 (1.4,0)	7	98 (0.7,-1.7)	22
Regularized cavity	2^{-3}	22	30	21 (1.4,-0.1)	5	18 (2,-1.2)	18
	2^{-4}	41	44	40 (0.9,0)	2	33 (1,-2)	20
	2^{-5}	74	70	70 (1.2,0)	0	62 (0.3,-2)	11
Colliding flow	2^{-3}	21	28	20 (1.4,-0.1)	5	18 (2,-1.5)	14
	2^{-4}	34	43	34 (0.9,0)	0	29 (1.8,-2)	15
	2^{-5}	61	66	54 (1.5,-0.1)	11	52 (1.6,-1.1)	15

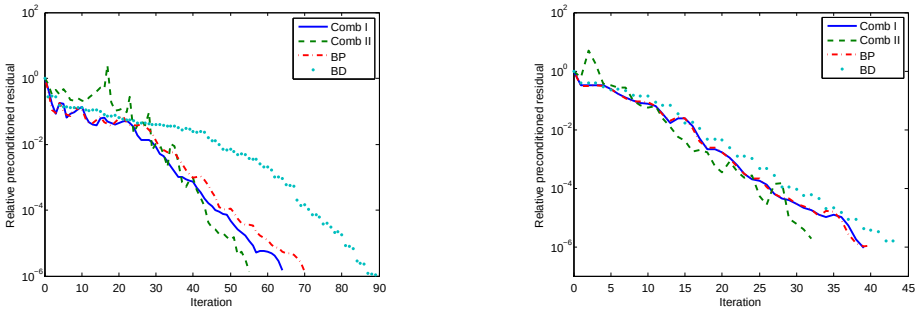
Table 5.4: Iteration counts for BP-CG, BDW-MINRES and the best possible BP-BDW combination-CG, as well as the relative reduction in the number of iterations required by BP-BDW combination-CG when compared with the better performing of BP-CG and BDW-MINRES. The $(\alpha, \beta) \in [-2, 2] \times [-2, 2]$ pairs shown are those that give the lowest number of iterations. For the left half of the table, we restrict (α, β) pairs to those for which $\mathcal{W}_{comb}$ defines an inner product with respect to which $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite.

so the results are omitted. To ensure that $A - A_0$ is positive definite, so that $\mathcal{W}_{BP}$ defines an inner product, A_0 is a no-fill incomplete Cholesky factorization of A, scaled by σ as in Table 5.1.

Table 5.4 shows that the combination preconditioner can be more effective than either the BP or BDW preconditioners. When $\mathcal{W}_{comb} > 0$ and $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}_{comb}}$, the BP-BDW combination preconditioner gives a percentage decrease in the number of iterations required of, on average, 4.3% and a maximum reduction of 11%, when compared with the better of the BP and BDW preconditioners. For some problems the best option is to scale the better performing of the BP or BDW preconditioners—which highlights that the BP and BDW preconditioners are sensitive to scaling—but the greatest improvement occurs when a good combination involving both preconditioners can be found.

Faster convergence is again observed when $\mathcal{W}_{comb}$ is not positive definite, with the average reduction in the number of iterations required jumping to 19.9% and the maximum reduction increasing to 30%. However, as mentioned for the BP$^+$-BDW combination preconditioner our theory cannot guarantee the convergence of BP-BDW combination-CG when $\mathcal{W}$ is indefinite.

We plot the relative preconditioned residuals for the flow over a backward step (Example 4.3) and the regularized cavity flow (Example 4.4) with $h = 2^{-4}$ in Figure 5.5, showing convergence plots when $\mathcal{W}_{comb} > 0$ and $\mathcal{W}_{comb} \not> 0$. These highlight that, particularly for Example 4.3, convergence is more rapid when the BP-BDW combina-



(a) backward step

(b) regularized cavity

Figure 5.5: Convergence plots for the BP-BDW combination preconditioner with $\mathcal{W}$-PCG for (a) the backward step flow (Example 4.3) with $\alpha = 1.9$ and $\beta = 0$ (Comb I) and $\alpha = 0.7$ and $\beta = -1.5$ (Comb II) and (b) the regularized cavity flow (Example 4.4) with $\alpha = 1.1$ and $\beta = 0$ (Comb I) and $\alpha = 1$ and $\beta = -2$ (Comb II).

tion preconditioner is used. We also see that convergence is more erratic when $\mathcal{W}_{comb}$ does not define an inner product.

5.3 The BP-SZ combination preconditioner

Section 4.4 showed that both the BP and SZ preconditioners in Section 4.1 can be effective. Stoll and Wathen [135] demonstrated that faster convergence than either of these methods attains can be achieved by combining them. Here a different BP-SZ combination preconditioner is analysed. We assume in this section that the condition, required by the SZ preconditioner, that $C = 0$ is met.

It is shown by Stoll and Wathen [135] that for the BP preconditioner (4.4), with its bilinear form defined by (4.5), and the SZ preconditioner (4.8), with its bilinear form defined by (4.9), Lemma 5.1 gives that

$$\begin{aligned}\mathcal{P}_3^{-T}\mathcal{W}_3 &= \alpha\mathcal{P}_1^{-T}\mathcal{W}_1 + \beta\mathcal{P}_2^{-T}\mathcal{W}_2 \\ &= \begin{bmatrix} I & -A_0^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} (\alpha-\beta)A_0^{-1} & 0 \\ -\beta S_0^{-1}BA_0^{-1} & T \end{bmatrix} \begin{bmatrix} A - A_0 & 0 \\ 0 & S_0 \end{bmatrix},\end{aligned}$$

where $T = (\beta - \alpha)S_0^{-1} - \beta S_0^{-1}BA_0^{-1}B^T S_0^{-1}$. They take the combination

$$\mathcal{P}_3^{-T} = \begin{bmatrix} I & -A_0^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} (\alpha-\beta)A_0^{-1} & 0 \\ -\beta S_0^{-1}BA_0^{-1} & (\beta-\alpha)S_0^{-1} - \beta S_0^{-1}BA_0^{-1}B^T S_0^{-1} \end{bmatrix}$$

and

$$\mathcal{W}_3 = \begin{bmatrix} A - A_0 & 0 \\ 0 & S_0 \end{bmatrix}. \tag{5.30}$$

However, the second term in $\mathcal{P}_3^{-T}$ can be factored as

$$\begin{aligned}&\begin{bmatrix}(\alpha-\beta)A_0^{-1} & 0\\ -\beta S_0^{-1}BA_0^{-1} & (\beta-\alpha)S_0^{-1}-\beta S_0^{-1}BA_0^{-1}B^TS_0^{-1}\end{bmatrix}\\ =&\begin{bmatrix}A_0^{-1} & 0\\ \frac{\beta}{\beta-\alpha}S_0^{-1}BA_0^{-1} & -S_0^{-1}\end{bmatrix}\begin{bmatrix}(\alpha-\beta)I & 0\\ 0 & (\alpha-\beta)I+\beta BA_0^{-1}B^TS_0^{-1}\end{bmatrix}\end{aligned}$$

and so another option is

$$\mathcal{P}_3^{-T}=\begin{bmatrix}I & -A_0^{-1}B^T\\ 0 & I\end{bmatrix}\begin{bmatrix}A_0^{-1} & 0\\ \frac{\beta}{\beta-\alpha}S_0^{-1}BA_0^{-1} & -S_0^{-1}\end{bmatrix},$$

so that

$$\mathcal{P}=\mathcal{P}_3=\begin{bmatrix}I & 0\\ BA_0^{-1} & I\end{bmatrix}\begin{bmatrix}A_0 & \frac{\beta}{\beta-\alpha}B^T\\ 0 & -S_0\end{bmatrix}=\begin{bmatrix}A_0 & \frac{\beta}{\beta-\alpha}B^T\\ B & \frac{\beta}{\beta-\alpha}BA_0^{-1}B^T-S_0\end{bmatrix} \tag{5.31}$$

and

$$\begin{aligned}\mathcal{W}=\mathcal{W}_3&\begin{bmatrix}(\alpha-\beta)I & 0\\ 0 & (\alpha-\beta)I+\beta BA_0^{-1}B^TS_0^{-1}\end{bmatrix}\begin{bmatrix}A-A_0 & 0\\ 0 & S_0\end{bmatrix}\\ &=\begin{bmatrix}(\alpha-\beta)(A-A_0) & 0\\ 0 & (\alpha-\beta)S_0+\beta BA_0^{-1}B^T\end{bmatrix}.\end{aligned} \tag{5.32}$$

By comparing (5.32) with (5.30) we see that the former can be scaled by adjusting α and β, which may make it easier to ensure that $\mathcal{W}$ is positive definite. Accordingly, in this thesis only the BP-SZ combination preconditioner (5.31) is considered. Additionally, from our combination preconditioner the BP and SZ preconditioners can be clearly identified: the BP preconditioner is recovered when $\alpha = 1$ and $\beta = 0$ while the SZ preconditioner is retrieved by setting $\alpha = 0$ and $\beta = 1$. From preliminary tests we observed that both the combination preconditioner proposed by Stoll and Wathen and (5.31) converged similarly. It would be interesting to carry out a thorough comparison of the performance of the two combination preconditioners.

We remark that the BP-SZ combination preconditioner (5.31) can alternatively be derived from Theorem 5.3, since, as is clear from Table 4.1, $c_1 = c_2 = 1$. If this approach is used the Schur complement approximation S_0 in Theorem 5.3 should be replaced by $-S_0$. To obtain (5.31) and (5.32) from the preconditioner and symmetric bilinear form in Theorem 5.3 scaling by $\beta - \alpha$ is required.

As previously discussed, knowledge of the conditions for positive definiteness of $\mathcal{W}$, and of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, is helpful when deciding which Krylov subspace method is most appropriate. To ascertain these conditions we begin with a technical lemma.

Lemma 5.11. *Let $A0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ be symmetric positive definite, $B \in \mathbb{R}^{n\times m}$ have full rank and $n \geq m$. If*

$$\lambda_{min} \leq \frac{x^T B A_0^{-1} B^T x}{x^T S_0 x} \leq \lambda_{max}$$

for all nonzero $x \in \mathbb{R}^m$ then

$$0 \leq \frac{y^T B^T S_0^{-1} B y}{y^T A_0 y} \leq \lambda_{max} \tag{5.33}$$

for all $y \in \mathbb{R}^n$.

Proof. Any eigenvalue λ of $S_0^{-1} B A_0^{-1} B^T$ satisfies the eigenvalue problem

$$\begin{aligned} S_0^{-1} B A_0^{-1} B^T x &= \lambda x \\ \Rightarrow A_0^{-1} B^T S_0^{-1} B (A_0^{-1} B^T x) &= \lambda (A_0^{-1} B^T x) \\ \Leftrightarrow B^T S_0^{-1} B y &= A_0 \lambda y, \end{aligned}$$

where $y = A_0^{-1} B^T x$ and $y = 0 \Leftrightarrow x = 0$. Thus, any eigenvalue of $S_0^{-1} B A_0^{-1} B^T \in \mathbb{R}^{m\times m}$ is an eigenvalue of $A_0^{-1} B^T S_0^{-1} B \in \mathbb{R}^{n\times n}$. To determine the remaining $n - m$ eigenvalues of $A_0^{-1} B^T S_0^{-1} B$, note that this matrix is similar to $A_0^{-1/2} B^T S_0^{-1} B A_0^{-1/2}$ which, in turn, is congruent to the rank m matrix $B^T S_0^{-1} B \in \mathbb{R}^{n\times n}$. Thus, Sylvester's law of inertia (see Appendix A.4) indicates that $A_0^{-1} B^T S_0^{-1} B$, like $B^T S_0^{-1} B$, has $n-m$ zero eigenvalues.

If λ_{min} and λ_{max} are the minimum and maximum eigenvalues of $S_0^{-1} B A_0^{-1} B^T$ they bound the generalized Rayleigh quotient

$$\lambda_{min} \leq \frac{x^T B A_0^{-1} B^T x}{x^T S_0 x} \leq \lambda_{max}$$

and it follows that

$$0 \leq \frac{y^T B^T S_0^{-1} B y}{y^T A_0 y} \leq \lambda_{max}.$$

□

We now state the requirements for positive definiteness of $\mathcal{W}$ and of $\mathcal{P}^{-1}\mathcal{A}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ for the BP-SZ combination preconditioner.

Theorem 5.12. *Let*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite $B \in \mathbb{R}^{m\times n}$ has full rank and $n \geq m$. Let $\mathcal{A}$ be left preconditioned by the BP-SZ combination preconditioner (5.31)*, where $A_0 \in \mathbb{R}^{n\times n}$ and $S_0 \in \mathbb{R}^{m\times m}$ are symmetric positive definite, so that $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where $\mathcal{W}$ is defined by* (5.32)*.*

When

I. *$\alpha > \beta > 0$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint if and only if $A > A_0$;*

II. *$\alpha < \beta$ and $\beta > 0$, $\mathcal{W}$ defines an inner product if and only if*

$$A < A_0 \quad \text{and} \quad S_0 < \frac{\beta}{\beta-\alpha} B A_0^{-1} B^T$$

and $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to this inner product if and only if

$$T = BV\left(V^{-1} - (A-A_0)M^{-1}(A-A_0)\right)VB^T > 0, \tag{5.34}$$

where

$$M = (\alpha-\beta)(A-A_0)A_0^{-1}A + \beta(A-A_0)A_0^{-1}B^T S_0^{-1} B A_0^{-1}(A-A_0) \tag{5.35}$$

and

$$V = (\alpha-\beta)A_0^{-1} + \beta A_0^{-1}B^T S_0^{-1} B A_0^{-1}; \tag{5.36}$$

III. *$\alpha < \beta < 0$, $\mathcal{W}$ does not define an inner product;*

IV. *$\beta < 0$ and $\alpha > \beta$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint and positive definite if and only if*

$$A > A_0 \quad \text{and} \quad S_0 > \frac{\beta}{\beta-\alpha} B A_0^{-1} B^T.$$

Proof. Since $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite if and only if $\mathcal{W}\mathcal{P}^{-1}\mathcal{A} > 0$ let us investigate the positive definiteness of $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$. Now,

$$\begin{aligned}\mathcal{W}\mathcal{P}^{-1}\mathcal{A} &= \begin{bmatrix} M & (A-A_0)VB^T \\ BV(A-A_0) & BVB^T \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ BV(A-A_0)M^{-1} & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} I & M^{-1}(A-A_0)VB^T \\ 0 & I \end{bmatrix},\end{aligned} \tag{5.37}$$

where T, M and V are given by (5.34), (5.35) and (5.36), respectively. Since (5.37) is a congruence transform, by Sylvester's law of inertia (see Appendix A.4), $\mathcal{P}^{-1}\mathcal{A}$ is

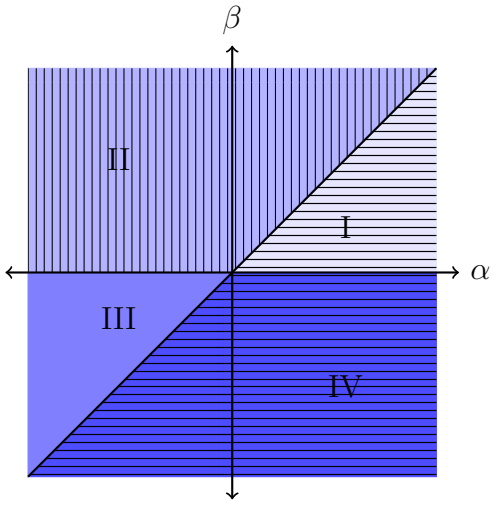


Figure 5.6: Regions (distinguished by different colours) that are considered in Theorem 5.12. The horizontal lines depict the region in which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ while the vertical lines show the region in which positive definiteness depends on T, given by (5.34).

positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ if and only if $M > 0$ and $T > 0$. The positive definiteness of M and T depends on α and β, with four distinct cases requiring consideration. These are depicted in Figure 5.6.

I. When $\alpha > \beta > 0$, $\mathcal{W}$ is positive definite if and only if $A > A_0$. By observation, $V > 0$ and $M > 0$. Now $T > 0$ if and only if

$$V^{-1} > (A - A_0)M^{-1}(A - A_0)$$

or, equivalently,

$$V < (A - A_0)^{-1}M(A - A_0)^{-1}$$

since M and V are positive definite. By substituting for M and V we find that this requirement is equivalent to

$$A_0 < (A_0^{-1} - A^{-1})^{-1}. \tag{5.38}$$

We must have, since $A > A_0$, that $A_0^{-1} > A^{-1}$ (see Corollary A.16) or $A_0^{-1} - A^{-1} > 0$ and so (5.38) holds if and only if

$$A_0^{-1} > A_0^{-1} - A^{-1}$$

which is always true. Thus, when $\alpha > \beta > 0$, $\mathcal{W}$ defines an inner product with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is positive definite if and only if $A > A_0$.

II. Next we consider Case II, for which $\beta > \alpha$ with $\beta > 0$. By observation, $\mathcal{W}$ defines an inner product if and only if

$$A < A_0 \quad \text{and} \quad S_0 < \frac{\beta}{\beta - \alpha} B A_0^{-1} B^T.$$

If $M > 0$,

$$\frac{\beta}{\beta - \alpha} B^T S_0^{-1} B > (A_0^{-1} - A^{-1})^{-1}$$

which, since $(A_0^{-1} - A^{-1})^{-1} < 0$ if and only if $A < A_0$, is satisfied if and only if $\mathcal{W} > 0$. If $V > 0$,

$$\frac{\beta}{\beta - \alpha} B^T S_0^{-1} B > A_0.$$

However,

$$\frac{\beta}{\beta - \alpha} B^T S_0^{-1} B$$

is positive semidefinite while A_0 is positive definite and this inequality cannot be satisfied. It follows that V is never positive definite and that it is not possible to significantly simplify the condition for positive definiteness of T; one can only state that $M > 0$ and that when $T > 0$, $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite.

III. Case III in Figure 5.6 corresponds to $0 > \beta > \alpha$. However, if $A_0 > 0$ and $S_0 > 0$ then $(\alpha - \beta) S_0 + \beta B A_0^{-1} B^T < 0$ and $\mathcal{W}$ is not positive definite. Similar calculations to those above show that $M < 0$ and $V < 0$ and that positive definiteness of T depends on the positive definiteness of $A - A_0$.

IV. The range $\alpha > \beta$ and $\beta < 0$ corresponds to case IV in Figure 5.6. For $\mathcal{W}$ to be positive definite we require that $A > A_0$ and

$$S_0 > \frac{\beta}{\beta - \alpha} B A_0^{-1} B^T.$$

To determine whether V is positive definite we note that if

$$S_0 > \frac{\beta}{\beta - \alpha} B A_0^{-1} B^T > 0$$

then

$$\frac{\beta - \alpha}{\beta} > \lambda_{max} \geq \frac{x^T B A_0^{-1} B^T x}{x^T S_0 x} \geq \lambda_{min} > 0.$$

From (5.33) in Lemma 5.11 we have that

$$\frac{\beta - \alpha}{\beta} > \lambda_{max} \geq \frac{y^T B^T S_0^{-1} B y}{y^T A_0 y} \geq 0$$

or that

$$V = (\alpha - \beta)A_0^{-1} + \beta A_0^{-1} B^T S_0^{-1} B A_0^{-1} > 0.$$

Thus, $V > 0$ when $\mathcal{W} > 0$.

For M to be positive definite a similar argument to that used for Case I shows that we require

$$(A_0^{-1} - A^{-1})^{-1} > \frac{\beta}{\beta - \alpha} B^T S_0^{-1} B$$

which, since $A_0 < (A_0^{-1} - A^{-1})^{-1}$ is satisfied whenever $V > 0$.

If $T > 0$ then, as in Case I,

$$A_0^{-1} > A_0^{-1} - A^{-1},$$

which is always true. Thus $\mathcal{P}^{-1}\mathcal{A}$ is positive definite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ when $\alpha > \beta$ with $\beta < 0$ if and only if $A > A_0$ and $S_0 > \frac{\beta}{\beta - \alpha} B A_0^{-1} B^T$. □

Regardless of whether $A > A_0$ or $A_0 > A$, Theorem 5.12 indicates that for many choices of α and β, $\mathcal{W}$ defines an inner product and $\mathcal{W}$-PMINRES may be reliably used to solve the symmetric saddle point system (4.2). Under additional assumptions described in Theorem 5.12, $\mathcal{W}$-PCG is robust.

We now demonstrate the effectiveness of the BP-SZ combination preconditioner for Examples 4.2–4.5 in Section 4.4. To all preconditioned systems we apply $\mathcal{W}$-PCG and scale the incomplete Cholesky factorization of A by σ to ensure that $A - A_0$ is positive definite; this requires a lower bound on the smallest eigenvalue of $A_0^{-1}A$. As noted in Chapter 4, this approximation may be costly to compute. The values of σ are given in Table 5.1.

Table 5.5 shows that when $\mathcal{W}_{comb} > 0$ the convergence of combination preconditioned $\mathcal{W}$-PCG can be significantly faster than BP- or SZ-CG. Moreover, although only one parameter choice for the combination preconditioner is listed, multiple choices gave the same iteration count for most problems. The relative reduction in the number of iterations, compared with the better preforming of the BP and SZ preconditioners, was up to 33% and was 20.2% on average.

Even faster convergence was achieved for these problems when $\mathcal{W}_{comb}$ was indefinite (and did not define an inner product). In this case the percentage decrease was up to 33% and was 26.7% on average. We note that the values of α and β that lead to this rapid convergence are near values for which $\langle \cdot, \cdot \rangle_{\mathcal{W}_{comb}}$ is an inner product (see Table 5.5). However, we have no theory to explain the effectiveness of the BP-SZ combination-CG method when $\langle \cdot, \cdot \rangle_{\mathcal{W}_{comb}}$ is not an inner product.

				$\mathcal{W}_{comb} > 0$		$\mathcal{W}_{comb} \not> 0$	
Problem	h	BP-CG	SZ-CG	Comb (α, β)	% BP/SZ	Comb (α, β)	% BP/SZ
Channel flow	2^{-3}	28	33	21 (1.7,-1)	25	20 (1.8,-2)	29
	2^{-4}	43	–	35 (1.4,-0.2)	19	31 (2,-0.4)	28
	2^{-5}	77	–	77 (1,0)	0	63 (2,-0.1)	18
Backward step	2^{-3}	43	69	36 (1.9,-0.8)	16	31 (1.7,-1.3)	28
	2^{-4}	71	–	59 (1.8,-0.2)	17	54 (1.7,-0.3)	24
	2^{-5}	125	–	125 (1,0)	0	111 (1.9,-0.1)	11
Regularized cavity	2^{-3}	22	25	16 (2,-2)	27	16 (1.5,-2)	27
	2^{-4}	41	–	30 (1.9,-0.5)	27	29 (2,-0.7)	29
	2^{-5}	74	–	54 (1.6,-0.1)	27	51 (1.3,-0.1)	31
Colliding flow	2^{-3}	21	23	14 (2,-2)	33	14 (1.6,-2)	33
	2^{-4}	34	–	25 (1.9,-0.5)	26	23 (2,-0.6)	32
	2^{-5}	61	–	46 (1.6,-0.1)	25	43 (1.3,-0.1)	30

Table 5.5: Iteration counts for the BP preconditioner, the SZ preconditioner and the best possible BP-SZ combination preconditioner for $\mathcal{W}$-PCG, as well as the relative reduction in the number of iterations required by BP-SZ combination-CG when compared with the better performing of BP- and SZ-CG. The $(\alpha, \beta) \in [-2, 2] \times [-2, 2]$ pairs shown are those that give the lowest number of iterations. For the left half of the table, we restrict (α, β) pairs to those for which $\mathcal{W}_{comb}$ defines an inner product with respect to which $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite.

Convergence plots for the flow over a backward step and the regularized cavity, both with $h = 2^{-4}$, are shown in Figure 5.7. The values of α and β used are given in Table 5.5. Both curves for the combination preconditioner decrease similarly, and rapidly, indicating that convergence is almost as rapid when $\langle \cdot, \cdot \rangle_{\mathcal{W}_{comb}}$ defines an inner product. Additionally, we are guaranteed convergence when $\mathcal{W}_{comb} > 0$.

We also tested the combination preconditioner on the same Stokes examples with $\mathcal{W}$-PMINRES. Scaling of A_0 and S_0 was found to be critical, as for the BP and SZ preconditioners (see Table 4.6) and $\mathcal{W}$-PCG tended to outperform $\mathcal{W}$-PMINRES. However, when scaled so that $\mathcal{W}_{comb}$ defined an inner product, BP-SZ combination-MINRES could outperform BP-MINRES, SZ-MINRES and either of the block diagonal preconditioned MINRES methods (BDW and BDI). The BP-SZ combination preconditioner was observed to be most effective for these examples, unsurprisingly, when $\mathcal{P}_{comb}^{-1}\mathcal{A}$ had nicely distributed eigenvalues. This implies that, for these Stokes problems, Corollary 3.15 appears to be descriptive.

In summary, we have shown that the BP-SZ combination preconditioner improves on the good performance of its constituent preconditioners for Examples 4.2 to 4.5. For a wide range of α and β values, $\langle \cdot, \cdot \rangle_{\mathcal{W}_{comb}}$ is an inner product and for a subset of these $\mathcal{P}_{comb}^{-1}\mathcal{A}$ is positive definite with respect to this inner product so that a conjugate gradient method can be reliably applied.

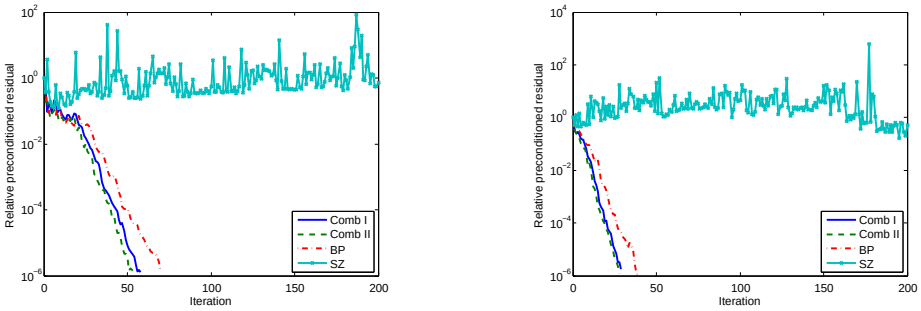


Figure 5.7: Comparison of the BP-SZ combination preconditioner for (a) the flow over a backward step (Example 4.3) with $\alpha = 1.8$ and $\beta = -0.2$ (Comb I) and $\alpha = 1.7$ and $\beta = -0.3$ (Comb II) and (b) the regularized cavity flow (Example 4.4) with $\alpha = 1.9$ and $\beta = -0.5$ (Comb I) and $\alpha = 2$ and $\beta = -0.7$ (Comb II).

5.4 Conclusions

This chapter examined a new formulation of the BP-SZ combination preconditioner and the novel BP$^+$-BDW and BP-BDW combination preconditioners. For each, we derived conditions for $\mathcal{W}_{comb}$ to define an inner product and found that these are satisfied for many choices of α and β, i.e. for many specific combinations. Particularly for PMINRES, we observed from numerical experimentation that the absence of an inner product led to poor convergence, making conditions for positive definiteness important.

Requirements for positive definiteness of the BP-BDW combination preconditioned saddle point matrix and of the BP-SZ combination preconditioned matrix were given. Significantly, we also showed that the BP$^+$-BDW combination preconditioned saddle point matrix can be made positive definite with respect to an inner product, despite neither the BP$^+$ preconditioned saddle point matrix nor the BDW preconditioned matrix being positive definite with respect to its inner product. This allows a $\mathcal{W}$-conjugate gradient method to be reliably used to solve the BP$^+$-BDW combination preconditioned system and, perhaps more importantly, it shows the potential for using combination preconditioning to change the inertia of a saddle point matrix. There may certainly be other combinations of preconditioners for which this is true and we feel that for this reason the development of indefinite preconditioners for saddle point problems, such as the SZ$^+$ preconditioner in Section 4.3, can be fruitful even when one seeks a preconditioner for which $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint and positive definite with

respect to an inner product. However, the difficulty lies, in general, in obtaining the preconditioner and inner product from Lemma 5.1 or Theorem 5.3.

Problem	h	BDI-MR	BP-CG*	BP-SZ*	BP$^+$-BDW-CG	BP$^+$-BDW-MR	BP-BDW*
Channel flow	2^{-3}	38	28	21	27	27	27
	2^{-4}	55	43	35	43	43	43
	$2^{-5\dagger}$	91 (93)	77	77	80	86	73
Backward step	2^{-3}	54	43	36	41	55	41
	2^{-4}	83	71	59	70	69	65
	$2^{-5\ddagger}$	147 (148)	125	125	118	116	
Regularized cavity	2^{-3}	28	22	16	22	21	21
	2^{-4}	43	41	30	40	40	40
	2^{-5}	68	74	54	73	73	70
Colliding flow	2^{-3}	26	21	14	21	20	20
	2^{-4}	36	34	25	35	34	34
	2^{-5}	65	61	46	58	56	54

* A_0 is scaled for positive definiteness of $A - A_0$ with scaling factors given in Table 5.1.

† For positive definiteness of the BP$^+$-BDW combination preconditioned saddle point matrix with respect to its inner product it was necessary to scale A_0 by 0.9. Thus, the iteration count for BDI in parentheses, and the iteration count for BP$^+$-BDW-CG, are obtained with $A_0 = 0.9LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab.

‡ For positive definiteness of the BP$^+$-BDW combination preconditioned saddle point matrix with respect to its inner product it was necessary to scale A_0 by 0.7. Thus, iteration count for BDI in parentheses, and the iteration count for BP$^+$-BDW-CG, are obtained with $A_0 = 0.7LL^T$, where L is the incomplete Cholesky factor computed by `ichol` in Matlab.

Table 5.6: Comparison of combination preconditioners for the Stokes problems.

Our computational results confirm that for the Stokes problems examined each combination preconditioner can be more effective than either of the constituent preconditioners, often significantly so. Overall, Table 5.6 shows that the BP-SZ combination preconditioner performs best. However, both the BP-SZ and BP-BDW combination preconditioners require that $A - A_0$ be scaled (as do the individual BP and SZ preconditioners) which we found to be costly for the Stokes problems with fine grids. We observed from our numerical experiments that for the BP$^+$-BDW preconditioner scaling of A_0 was not required for the reliable application of $\mathcal{W}$-PMINRES, and in many cases for $\mathcal{W}$-PCG. Additionally, both BP$^+$-BDW combination-MINRES and BP$^+$-BDW combination-MINRES outperformed, by a considerable margin, the best method for the Stokes problems for which scaling is not required: block diagonal preconditioned MINRES in the Euclidean inner product. This indicates that when scaling $A - A_0$ is costly, the BP$^+$-BDW combination preconditioner provides an attractive alternative.

The faster performance of $\mathcal{W}$-PCG observed for each combination preconditioner when $\mathcal{W}_{comb}$ was not positive definite remains unexplained. The same behaviour has been previously observed [34, 135] and analysed [115], although the analysis appears difficult to generalize to our preconditioners.

Chapter 6

Preconditioners for nonsymmetric matrices

As discussed in the introduction, a lack of understanding of the factors that determine GMRES convergence for general nonsymmetric (nonnormal) matrices makes it difficult to deduce mathematical criteria for effective preconditioners. If the preconditioned coefficient matrix is close to normal and has nicely distributed eigenvalues, either the standard eigenvalue bound (3.27) or standard pseudospectral bound (3.28) imply that convergence will be rapid [62]. However, it is difficult to know exactly which characteristics of a preconditioner improve the normality of the preconditioned coefficient matrix. In the absence of mathematical criteria, preconditioners are generally chosen heuristically.

Broadly speaking, these preconditioners are either designed to be widely applicable or are constructed for a particular type of problem[1]. General purpose preconditioners typically approximate either the coefficient matrix or its inverse, one of the most common types being incomplete factorization preconditioners, introduced in [99]. Another is the set of sparse approximate inverse preconditioners. The original sparse approximate inverse preconditioners [14] solve or approximate

$$\min_{P \in \mathcal{L}} \|I - BP\|_F,$$

where $\mathcal{L}$ is a set of sparse matrices and $\|\cdot\|_F$ is the Frobenius norm, although factored sparse approximate inverse preconditioners have since been proposed. For a survey of these general purpose preconditioners see [15].

[1]There is, inevitably, some overlap. For example, it may be possible to optimize a general purpose preconditioner by considering a characteristic particular to the coefficient matrix of interest [15].

Applicable to partial differential equations are a number of additional preconditioners. Algebraic [21, 116, 117] and geometric multigrid methods (see [68] for an early reference) and domain decomposition methods (see [142] for an overview) were originally developed as stand-alone techniques for solving linear systems. However, they can be powerful preconditioners. More specialized preconditioners also exist, such as the diffusion synthetic acceleration preconditioner for the transport equation [4, 86, 96] and the least-squares commutator and pressure convection-diffusion preconditioners for Newton or Picard iteration for the Navier-Stokes equations [43, Chapter 8]. A preconditioner for the Euler equations that takes into account the eigenvectors of the coefficient matrix and thus the symmetric bilinear form with respect to which B is self-adjoint, is given by Darmofal [30]. For Toeplitz matrices (2.18) (that are not necessarily related to PDEs) circulant preconditioners have proved popular [29, 136, 140, 143].

Often these heuristically chosen preconditioners are effective. However, it can be difficult to determine reasons for this and the characteristics common to good preconditioners. This chapter appraises some known preconditioners in view of self-adjointness and normality with respect to nonstandard symmetric bilinear forms. In Section 6.1 we examine the relationship between the eigenvectors of the preconditioner P and those of the coefficient matrix B and discuss how this relates to the symmetric bilinear forms with respect to which P and B are self-adjoint. Although eigenvectors are a useful analytical tool, their accurate computation can be extremely difficult. Pseudospectra, in contrast, are far less sensitive to perturbations [145, Theorem 52.4]. Accordingly, in Section 6.1 we also relate the nearness of eigenvectors to similarities between the pseudospectra of P and B. Section 6.2 investigates preconditioners that are self-adjoint with respect to the same symmetric bilinear form as the coefficient matrix. The case that the preconditioner is self-adjoint with respect to a symmetric bilinear form that is nearby to a symmetric bilinear form with respect to which B is self-adjoint is addressed in Section 6.3. A summary of the chapter can be found in Section 6.4.

Unless stated otherwise, matrices are assumed to be diagonalizable. Throughout, a diagonalizable matrix B will have a matrix of eigenvectors Z while a matrix of eigenvectors of a diagonalizable preconditioner P will be denoted by $\mathcal{Z}$. All pseudospectral plots are computed by Tom Wright's EigTool package [154] for Matlab. I-MR residuals are measured in the I-norm, while for any nonstandard inner product $\langle \cdot, \cdot \rangle_M$ the residuals of the corresponding M-MR method are measured in the M-norm.

6.1 Nearby eigenvectors

In this section preconditioners with eigenvectors that are close to those of the coefficient matrix are considered. Relationships between the nearness of these eigenvectors and the nearness of the symmetric bilinear forms with respect to which B and P are self-adjoint—the focus of Sections 6.2 and 6.3—are discussed. We also see that the nearness of the eigenvectors of B and P is reflected in their pseudospectra.

Let us begin by examining the effect of nearby eigenvectors on $P^{-1}B$. If $P = \mathcal{Z}\Lambda_P\mathcal{Z}^{-1}$ and $B = Z\Lambda_B Z^{-1}$ then

$$\mathcal{Z}^{-1}B\mathcal{Z} = \Lambda_B + C, \tag{6.1}$$

where we would expect C to be small when $\mathcal{Z}$ is close to an eigenvector matrix of B. It follows that

$$\mathcal{Z}^{-1}P^{-1}B\mathcal{Z} = (\mathcal{Z}^{-1}P\mathcal{Z})(\mathcal{Z}^{-1}B\mathcal{Z}) = \Lambda_P^{-1}(\Lambda_B + C) \tag{6.2}$$

so that

$$P^{-1}B = \mathcal{Z}\Lambda_P^{-1}\Lambda_B\mathcal{Z}^{-1} + \mathcal{Z}\Lambda_P^{-1}C\mathcal{Z}^{-1}. \tag{6.3}$$

If C were a small perturbation we would expect the first term to dominate. Under the additional assumption that Λ_P is a good approximation of Λ_B, $\Lambda_P^{-1}\Lambda_B$ is close to the identity, as is $P^{-1}B$. Moreover, the preconditioned coefficient matrix is nearly normal.

We now study the symmetric bilinear forms with respect to which the preconditioner and coefficient matrix are self-adjoint. If B and P have the same eigenvectors, the former is self-adjoint with respect to the symmetric bilinear form defined by $W = Z^{-*}YZ^{-1}$ and the latter with respect to the symmetric bilinear form defined by $\mathcal{W} = Z^{-*}\mathcal{Y}Z^{-1}$. These symmetric bilinear forms are equal when $Y = \mathcal{Y}$, i.e., when the number of real eigenvalues of P matches that of B.

If the difference in the cardinality of real eigenvalues of B and P is small and the eigenvectors of P and B are close, then we might expect that P and B will be self-adjoint with respect to nearby symmetric bilinear forms. In particular,

$$W - \mathcal{W} = Z^{-*}YZ^{-1} - \mathcal{Z}^{-*}\mathcal{Y}\mathcal{Z}^{-1} = \mathcal{Z}^{-*}\left(\mathcal{Z}^*Z^{-*}YZ^{-1}\mathcal{Z} - \mathcal{Y}\right)\mathcal{Z}^{-1}$$

so that, if $Z^{-1}\mathcal{Z} = I + F$, then

$$\mathcal{Z}^*(W - \mathcal{W})\mathcal{Z} = (I + F)^*Y(I + F) - \mathcal{Y}. \tag{6.4}$$

If F is small we would expect $(I+F)^*Y(I+F) \approx Y$. The difference between Y and $\mathcal{Y}$ is related to how well P captures the number of real and complex conjugate eigenvalues of B. When this is well approximated by P, we would expect $W - \mathcal{W}$ to be small. In other words, when P approximates the eigenvalues and eigenvectors of B well, P and B are self-adjoint with respect to nearby symmetric bilinear forms. It is more difficult to make concrete statements about F when $W - \mathcal{W}$ is small.

To summarize, when B and P have nearby eigenvectors the symmetric bilinear forms with respect to which each is self-adjoint are nearby, in the sense that the difference between the matrices defining these symmetric bilinear forms is small. When P and B have nearby eigenvectors and P approximates well the eigenvalues of B, $P^{-1}B$ is close to normal. The difficulty lies, in general, in computing the eigenvector matrix of P that is closest to a given eigenvector matrix of B, since the number of possible orderings of columns increases combinatorially with the dimension of B. Although normalizing each eigenvector removes the possibility that eigenvectors may be scaled differently, the sign of the eigenvector must also be chosen.

6.1.1 Pseudospectra of P and B

Often the coefficient matrix and preconditioner exhibit pseudospectra that are remarkably similar. Some examples are given in Figures 6.5, 6.7, 6.10 and 6.13. We use the notion of nearby eigenvectors and nonstandard norms to provide some explanation of this phenomenon.

The following theorem describes a condition for the pseudospectra of P and B to be identical.

Theorem 6.1. *The matrices $P \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ have identical I-norm pseudospectra if P and B have diagonalizations $P = \mathcal{Z}\Theta\mathcal{Z}^{-1}$ and $B = Z\Theta Z^{-1}$, $\Theta, Z, \mathcal{Z} \in \mathbb{C}^{n\times n}$, where $Z^*Z = \mathcal{Z}^*\mathcal{Z}$. Under these conditions, P and B have the same singular values.*

Proof. From Definition A.4 in Appendix A.3 the I-norm ϵ-pseudospectra of B is

$$\Lambda_\epsilon(B) = \{z | z \in \Lambda(B+E), \|E\|_I < \epsilon\},$$

where $\Lambda(B+E)$ is the spectrum of $B+E$. Now,

$$\Lambda(B+E) = \Lambda(Z\Theta Z^{-1} + E) = \Lambda(\Theta + Z^{-1}EZ) = \Lambda(\Theta + G),$$

where $E = ZGZ^{-1}$. It follows that $\|E\|_I = \|ZGZ^{-1}\|_I = \|G\|_T$, where

$$Z^*Z = \mathcal{Z}^*\mathcal{Z} = T \tag{6.5}$$

and we have used (2.2) to relate the I-norm to the T-norm.

Thus, $\Lambda_\epsilon(B) = \{z | z \in \Lambda(\Theta + G), \|G\|_T < \epsilon\}$.

The ϵ-pseudospectra of P is analogously defined as

$$\Lambda_\epsilon(P) = \{z | z \in \Lambda(P + E), \|E\|_I < \epsilon\}$$

which, by employing the same arguments as were used for B and (6.5), gives that

$$\Lambda_\epsilon(P) = \{z | z \in \Lambda(\Theta + G), \|G\|_T < \epsilon\}.$$

Comparison with the pseudospectra of B confirms that $\Lambda_\epsilon(B) = \Lambda_\epsilon(P)$.

To show that P and B have identical singular values, we recall that the singular values of B are the square roots of the eigenvalues of B^*B. Since

$$B^*B = Z^{-*}\Theta^*Z^*Z\Theta Z^{-1} = Z^{-*}\Theta^*T\Theta Z^{-1}$$

which is similar to $Z^{-1}Z^{-*}\Theta^*T\Theta = T^{-1}\Theta^*T\Theta$, the singular values of B are the square roots of the eigenvalues of $T^{-1}\Theta^*T\Theta$. The same process applied to P shows that its singular values are also the square roots of the eigenvalues of $T^{-1}\Theta^*T\Theta$. Thus, the singular values of P and B are the same. □

Remark 6.2. More generally, the pseudospectra of $P = \mathcal{Z}\Theta\mathcal{Z}^{-1}$ and $B = Z\Theta Z^{-1}$, with respect to the X-norm, are identical when $\mathcal{Z}^*X\mathcal{Z} = Z^*XZ$. The X-singular values (see Definition 2.13) of P and B are the same in this case.

Remark 6.3. Under the conditions of Theorem 6.1, $\|p_k(B)\|_I = \|p_k(P)\|_I$ for any polynomial $p_k \in \Pi_k$. This is proved similarly to the equality of the pseudospectra:

$$\|p_k(B)\|_I = \|Zp_k(\Theta)Z^{-1}\|_I = \|p_k(\Theta)\|_T = \|\mathcal{Z}p_k(\Theta)\mathcal{Z}^{-1}\|_I = \|p_k(P)\|_I.$$

In particular, P and B have the same ideal GMRES polynomials (see page 51).

Remark 6.4. The condition that $Z^*Z = \mathcal{Z}^*\mathcal{Z}$ in Theorem 6.1 above is sufficient for the pseudospectra of B and P to be identical but is not necessary. Merely scaling the eigenvectors of B shows this. Specifically, if $B = Z\Theta Z^{-1}$ and $P = \mathcal{Z}\Theta\mathcal{Z}^{-1}$ with $\mathcal{Z} = \alpha Z$ for $\alpha \in \mathbb{R}$, $\alpha \neq 0, 1$ then, since $P = (\alpha Z)\Theta(\alpha Z)^{-1} = B$ the pseudospectra of P and B are the same. However, $\mathcal{Z}^T\mathcal{Z} = \alpha^2 Z^TZ \neq Z^TZ$. It would be interesting to ascertain sufficient conditions for the pseudospectra of P and B to be identical.

$\kappa_I(B)$	$\kappa_I(P_{nearby})$	$\kappa_I(Z)$	$\kappa_I(\mathcal{Z}_{nearby})$	$\|T - \mathcal{Z}_{chol}^T \mathcal{Z}_{chol}\|_I$	$\|T - \mathcal{Z}_{sqrt}^T \mathcal{Z}_{sqrt}\|_I$
9.2×10^5	8.3×10^5	303	267	6.1×10^{-14}	2.7×10^{-12}

Table 6.1: Problem data for the preconditioners with identical or nearby pseudospectra.

Remark 6.5. A related result to Theorem 6.1 is given in Theorem 1 in [3] where it is shown that under the more restrictive assumption that P and B have super-identical pseudospectra[2] they are similar.

Theorem 6.1 shows that if P exactly captures the eigenvalues of B and has eigenvectors that are not necessarily identical to P but for which $\mathcal{Z}^*\mathcal{Z} = Z^*Z$, the pseudospectra will reflect this structure. This does not necessarily mean convergence of GMRES for $P^{-1}B$ will be fast; it is entirely possible that the eigenvectors are nothing like those of B in which case convergence might be poor. We illustrate this with a small example.

We solve $Bx = b$, where $B = Z\Lambda Z^{-1} \in \mathbb{R}^{100\times 100}$ and Z, b and the diagonal entries of Λ have random normal entries. We denote by T the matrix $Z^{-T}Z^{-1}$. The spectrum and pseudospectra of B are shown in Figure 6.1. The eigenvalues are represented by black dots (on the real line), while the contours represent the boundaries of ϵ-pseudospectra for different ϵ in $[10^{-3.5}, 10^{-1}]$. To this linear system we apply two preconditioners that have the same spectra and pseudospectra as B: P_{chol} and P_{sqrt}. The eigenvectors of $P_{chol} = \mathcal{Z}_{chol}\Lambda\mathcal{Z}_{chol}^{-1}$ are given by the Cholesky decomposition $T = \mathcal{Z}_{chol}^T\mathcal{Z}_{chol}$, which is computed by the standard Matlab function. The matrix square root computed in the Matrix Function Toolbox [72], which is based on the real Schur form algorithm in [73, Algorithm 6.7], is used to obtain the eigenvectors of $P_{sqrt} = \mathcal{Z}_{sqrt}\Lambda\mathcal{Z}_{sqrt}^{-1}$. Thus,

$$T = Z^TZ = \mathcal{Z}_{chol}^T\mathcal{Z}_{chol} = \mathcal{Z}_{sqrt}^T\mathcal{Z}_{sqrt} \tag{6.6}$$

and B, P_{chol} and P_{sqrt} have identical pseudospectra by Theorem 6.1.

From Theorem 6.1 we know that P_{chol}, P_{sqrt} and B have the same singular values and, therefore, the same condition number. It is also clear from (6.6) that Z, $\mathcal{Z}_{sqrt}$ and $\mathcal{Z}_{chol}$ have the same condition number. These are given in Table 6.1, as are the backward errors for $\mathcal{Z}_{chol}$ and $\mathcal{Z}_{sqrt}$. A third preconditioner, $P_{nearby} = \mathcal{Z}_{nearby}\Lambda\mathcal{Z}_{nearby}^{-1}$ is also constructed that has the same eigenvalues as B and eigenvectors that are

[2]The pseudospectra of $P, B \in \mathbb{C}^{n\times n}$ are super-identical if the singular values of $zI - P$ equal those of $zI - B$ for every $z \in \mathbb{C}$ [19]. This is a more restrictive condition than that the I-norm pseudospectra of P and B are equal for which only equality, for all $z \in \mathbb{C}$, of the smallest singular value of $zI - B$ and $zI - P$ is required.

nearby, but not identical, to those of B. Specifically, $\mathcal{Z}_{nearby} = Z+$`0.01randn(100)`, where `randn` represents the Matlab function of the same name. The pseudospectra of P_{nearby} are similar to those of B, as also shown in Figure 6.1. Again, the black dots denote the eigenvalues of P_{nearby} (that are the same as those of B), while the contours represent the boundaries of ϵ-pseudospectra for $\epsilon \in [10^{-3.5}, 10^{-1}]$. However, since the pseudospectral curves are slightly closer to the eigenvalues than those of B, P is slightly closer to normal. (Lemma A.8 states that the ϵ-pseudospectrum of an I-normal matrix consists of open ϵ-balls around the eigenvalues of the matrix). The condition number $\mathcal{Z}_{nearby}$, given in Table 6.1, reflects this, since it is smaller than that of Z. The relative preconditioned residuals in Figure 6.1 show that neither preconditioner with identical pseudospectra to that of B is effective but that the preconditioner with nearby eigenvectors accelerates convergence significantly.

Remark 6.3 shows that B, P_{chol} and P_{sqrt} have the same ideal GMRES polynomials. From the discussion on page 51, the ideal GMRES polynomial provides an upper bound on worst-case GMRES behaviour, which can be sharp. We might expect, therefore, that convergence of GMRES applied to $Bx = b$, $P_{chol}x_c = b$ and $P_{sqrt}x_s = b$, for the same $b \in \mathbb{R}^n$, is similar. Figure 6.1 (d) shows that this is true, at least for the random right-hand side vector used in Figure 6.1 (c). Convergence for $P_{nearby}x_n = b$ is also very similar to that for $Bx = b$. This example shows that preconditioners P that satisfy the conditions of Theorem 6.1 have the same pseudospectra as B and often exhibit similar GMRES convergence curves but may not be good preconditioners. The preconditioner that approximates the eigenvectors of B in this example is effective, although we note that for this example, although P and B are somewhat nonnormal (since the boundaries of the ϵ-pseudospectra in Figure 6.1 are further than a distance ϵ from the eigenvalues), both Z and $\mathcal{Z}_{nearby}$ are well-conditioned.

In practice, we might identify a preconditioner P that has eigenvectors that are alike but not identical to those of B, such as P_{nearby} in the example above, and whose eigenvalues approximate those of B. The next theorem shows that we can still expect that the pseudospectra of B and P will be nearby if the eigenvalues and eigenvectors are captured well enough.

Theorem 6.6. *Let $P \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ have diagonalizations $P = \mathcal{Z}\Theta_P\mathcal{Z}^{-1}$ and $B = Z\Theta_B Z^{-1}$, $\Theta_P, \Theta_B, \mathcal{Z}, Z \in \mathbb{C}^{n\times n}$. Additionally, let $T = Z^*Z$ and $\mathcal{T} = \mathcal{Z}^*\mathcal{Z}$. Then the ϵ-pseudospectra of P and B are given by*

$$\Lambda_\epsilon(B) = \{z | z \in \Lambda(\Theta_B + G), \|G\|_T < \epsilon\} \tag{6.7}$$

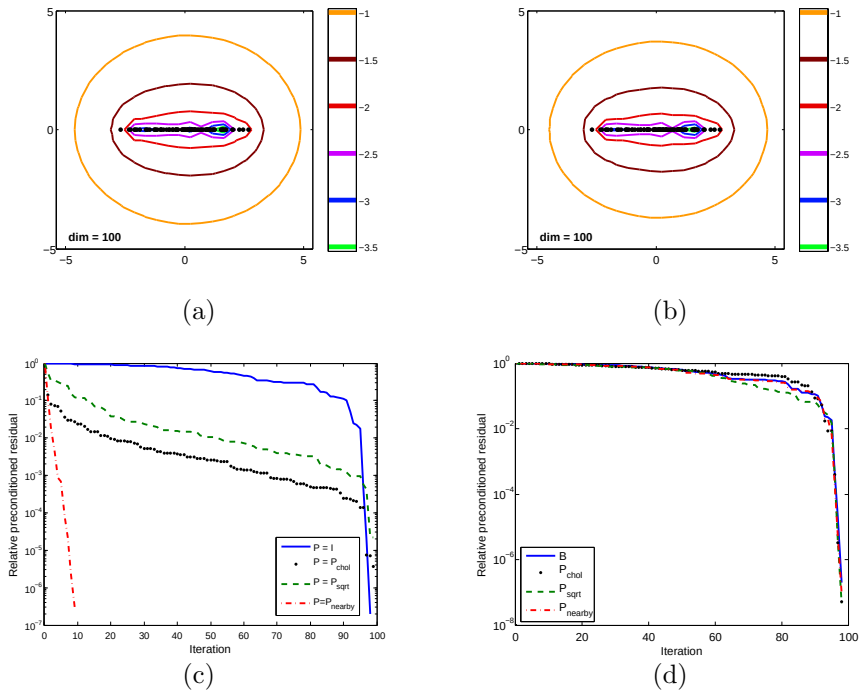


Figure 6.1: The pseudospectra of (a) B (and P_{chol} and P_{sqrt}) and (b) P_{nearby} as well as (c) the relative preconditioned residuals for GMRES applied to the preconditioned systems and (d) the relative residuals for GMRES (without preconditioning) applied to each of B, P_{chol}, P_{sqrt} and P_{nearby}. On each pseudospectral plot the black dots represent the eigenvalues of the matrix, while each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-3.5}, 10^{-1}]$.

and

$$\Lambda_\epsilon(P) = \{z | z \in \Lambda(\Theta_P + G), \|G\|_{\mathcal{T}} < \epsilon\}, \tag{6.8}$$

where

$$\frac{1}{\kappa_I(Z\mathcal{Z}^{-1})}\|G\|_{\mathcal{T}} \leq \|G\|_T \leq \kappa_I(Z\mathcal{Z}^{-1})\|G\|_{\mathcal{T}}$$

and $\kappa_I(\cdot)$ *is the* I*-norm condition number.*

Proof. As in Theorem 6.1, the I-norm ϵ-pseudospectrum of B is given by

$$\Lambda_\epsilon(B) = \{z | z \in \Lambda(\Theta_B + G), \|G\|_T < \epsilon\}.$$

Meanwhile, the ϵ-pseudospectra of P is the set

$$\Lambda_\epsilon(P) = \{z | z \in \Lambda(P + E), \|E\|_I < \epsilon\}.$$

The spectrum of $P + E$ is

$$\Lambda(P + E) = \Lambda(\Theta_P + K),$$

where $K = \mathcal{Z}^{-1}E\mathcal{Z}$. Taking norms gives that $\|E\|_I = \|\mathcal{Z}K\mathcal{Z}^{-1}\|_I = \|K\|_{\mathcal{T}}$ and so

$$\Lambda_\epsilon(P) = \{z | z \in \Lambda(\Theta_P + F), \|G\|_{\mathcal{T}} < \epsilon\},$$

where we have written G instead of K to aid comparison of the pseudospectra of B and P.

To relate $\|G\|_T$ and $\|G\|_{\mathcal{T}}$, we use (2.2) to show that

$$\|G\|_T = \|ZGZ^{-1}\|_I = \|Z\mathcal{Z}^{-1}\mathcal{Z}G\mathcal{Z}^{-1}\mathcal{Z}Z^{-1}\|_I \leq \|Z\mathcal{Z}^{-1}\|_I\|\mathcal{Z}Z^{-1}\|_I\|G\|_{\mathcal{T}},$$

which is the upper bound. The lower bound follows from swapping the roles of $\mathcal{Z}$ and Z (and, consequently, of $\mathcal{T}$ and T). □

Remark 6.7. If the eigenvectors of P and B are nearby, so that the condition number of $Z\mathcal{Z}^{-1}$ is close to 1, then $\|G\|_T \approx \|G\|_{\mathcal{T}}$ and there will be a large overlap in the set of perturbation matrices in (6.7) and in (6.8). If, additionally, the eigenvalues of P are good approximations of those of B we would expect the pseudospectra of P and B to be nearby. Theorem 6.6 shows that the pseudospectra highlight more than the spectrum alone.

6.2 Preconditioners with the same symmetric bilinear form

Preconditioners P that are self-adjoint with respect to the same symmetric bilinear form as the coefficient matrix B are the focus of this section. A sufficient condition for this is that P and B have the same eigenvectors and the same eigenvalue structure (so that $Y = \mathcal{Y}$), as discussed in Section 6.1.

Since P is self-adjoint with respect to the same symmetric bilinear form (see Lemma 2.18), under the assumption that B and P are both W-self-adjoint, WP and $P^{-1}B$ satisfy the self-adjointness relation (2.13), i.e.,

$$(WP)(P^{-1}B) = WB = B^T W = B^T P^{-T} P^T W = (P^{-1}B)^T (WP). \tag{6.9}$$

This shows that $P^{-1}B$ is WP-self-adjoint. If $P^{-1}B$ is additionally diagonalizable, there exists a diagonalization $P^{-1}B = S\Lambda S^{-1}$ such that $WP = S^{-*}YS^{-1}$, $WP \in \mathbb{R}^{n\times n}$ (see Corollary 2.28) and B is M-normal, where $M = S^{-*}S^{-1}$, $M \in \mathbb{R}^{n\times n}$ (see Lemma 2.29 and Corollary 2.28). Note that it is possible to avoid complex arithmetic by using the real Jordan form instead (see Theorem 2.20) but we find it more convenient to use the complex diagonalization (which is easily computed by Matlab) than the real Jordan form. By Corollary 3.15, the preconditioned M-MR residuals, $P^{-1}r_k^M$, satisfy that

$$\frac{\|P^{-1}r_k^M\|_M}{\|P^{-1}r_0\|_M} \leq \min_{p\in\Pi_k,\, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)| \tag{6.10}$$

and that

$$\frac{\|P^{-1}r_k\|_M}{\|P^{-1}r_0\|_M} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|,$$

where Γ_ϵ encloses $\Lambda(P^{-1}B) + \triangle_\epsilon$, the set $\triangle_\epsilon$ is described in Appendix A.3, and L_ϵ is the arc length of Γ_ϵ. Thus, since $P^{-1}B$ is M-normal its eigenvalues largely dictate the convergence of M-MR methods, at least when measured by the M-norm.

Applying M-MR involves inner products with M, which may be costly to compute even if this matrix is known. However, the analysis of Section 3.2.4 allows us to relate the convergence of any M-MR method to that of any I-MR method. In particular, Corollary 3.20 gives that

$$\frac{\|P^{-1}r_k^I\|_I}{\|P^{-1}r_0\|_I} \leq \frac{\sigma_{max}(S^{-1})}{\sigma_{min}(S^{-1}V_{k+1}^I)} \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)|$$

and

$$\frac{\|P^{-1}r_k^I\|_I}{\|P^{-1}r_0\|_I} \leq \frac{\sigma_{max}(S^{-1})}{\sigma_{min}(S^{-1}V_{k+1})}\frac{L_\epsilon}{2\pi\epsilon}\min_{p\in\Pi_k, p(0)=1}\max_{z\in\Gamma_\epsilon}|p_k(z)|,$$

where the columns of V_k form an I-orthonormal basis of $\mathcal{K}_k\left(P^{-1}B, P^{-1}r_0\right)$.When

$$\frac{\sigma_{max}(S^{-1})}{\sigma_{min}(S^{-1}V_{k+1})}$$

is small compared with $\kappa_I(S)$, this may give a more accurate picture of convergence than the standard eigenvalue bound (3.27).

If we instead apply a W-self-adjoint matrix P as a right preconditioner then, by Lemma 2.18, B^{-1} and P^{-1} are also W-self-adjoint and

$$(WB^{-1})BP^{-1} = WP^{-1} = P^{-T}W = P^{-T}B^TB^{-T}W = (BP^{-1})^T(WB^{-1})$$

so that BP^{-1} is WB^{-1}-self-adjoint[3]. We find that $P^{-1}B = L\Lambda L^{-1}$, where $WB^{-1} = L^{-*}YL^{-1}$ for some Y of the form described in Theorem 2.27. Additionally, $P^{-1}B$ is X-normal, where $X = L^{-*}L^{-1}$ and so, similarly to the left preconditioned case, Corollary 3.16 implies that relative residuals of any minimum residual method in $\langle\cdot,\cdot\rangle_X$ satisfy that

$$\frac{\|r_k^X\|_X}{\|r_0\|_X} \leq \min_{p\in\Pi_k,\, p(0)=1}\max_{\lambda\in\Lambda(BP^{-1})}|p(\lambda)|$$

and

$$\frac{\|r_k^X\|_X}{\|r_0\|_X} \leq \frac{L_\epsilon}{2\pi\epsilon}\min_{p\in\Pi_k, p(0)=1}\max_{z\in\Gamma_\epsilon}|p_k(z)|,$$

where Γ_ϵ encloses $\Lambda(BP^{-1}) + \triangle_\epsilon = \Lambda(P^{-1}B) + \triangle_\epsilon$ and L_ϵ is the arc length of Γ_ϵ. Again, as a result of the X-normality of $P^{-1}B$, eigenvalues influence the convergence of this X-MR method and a good preconditioner will improve the eigenvalues of B. As for left preconditioning, convergence bounds for I-MR methods are obtained by utilizing the relationships in Section 3.2.4.

A symmetric bilinear form with respect to which a persymmetric matrix (see page 19) is self-adjoint is defined by the reverse identity matrix (2.19). We now examine three preconditioned linear systems for which P and B are both persymmetric, or constructed from persymmetric matrices, and so are self-adjoint with respect to the same symmetric bilinear forms.

[3]Clearly, one would not form WB^{-1} but these results could be related to I-MR through the relationships in Section 3.2.4.

6.2.1 Toeplitz matrices

As noted at the start of Section 2.2, a Toeplitz matrix

$$B = \begin{bmatrix} b_0 & b_{-1} & b_{-2} & \dots & \dots & b_{-n+1} \\ b_1 & b_0 & b_{-1} & \ddots & & \vdots \\ b_2 & b_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_{-1} & b_{-2} \\ \vdots & & \ddots & b_1 & b_0 & b_{-1} \\ b_{n-1} & \dots & \dots & b_2 & b_1 & b_0 \end{bmatrix}$$

is self-adjoint with respect to the symmetric bilinear form $\langle \cdot, \cdot \rangle_{\hat{Y}}$ defined by the reverse identity matrix

$$\hat{Y} = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}.$$

A particular Toeplitz matrix is the circulant matrix

$$C = \begin{bmatrix} c_0 & c_{n-1} & \dots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \dots & c_1 & c_0 \end{bmatrix}$$

which is (I-)normal, with unitary eigenvectors that are related to the Fourier matrix

$$F = (f_{jk}) \quad \text{with} \quad f_{jk} = e^{-\frac{2(j-1)(k-1)\pi \mathrm{i}}{n}}, \quad \text{for} \quad j, k = 1, \dots, n.$$

Specifically, if $U = \frac{1}{\sqrt{n}} F$, a diagonalization of C is

$$C = U^* diag(Fc) U, \tag{6.11}$$

where c is the first column of C.

Although Toeplitz matrices can be highly nonnormal [145, Chapter 7] they are often effectively preconditioned by circulant matrices [26, 27, 28, 29, 106, 123, 136, 137, 140, 143]. Part of the appeal of circulants is that, because of their relation to the Fourier matrix, linear systems with C can be solved in $O(n \log n)$ operations by utilizing the fast Fourier transform [136, 29]. The original circulant preconditioner P for B, proposed by Strang [136], has as its first column

$$[P_n]_{k,0} = \begin{cases} b_k, & 0 \le k \le \lfloor \frac{n}{2} \rfloor, \\ b_{n-k}, & \lfloor \frac{n}{2} \rfloor < k < n. \end{cases}$$

Different circulant preconditioners have since been suggested, some of which minimize a function of the Frobenius norm over the set of circulant matrices [29, 140].

We now appraise the effectiveness of the Strang preconditioner for a highly nonnormal Toeplitz matrix.

Example 6.1. The linear system we consider involves the Grcar matrix of dimension 100,

$$B = \begin{bmatrix} 1 & 1 & 1 & & & \\ -1 & 1 & 1 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & 1 \\ & & & \ddots & \ddots & 1 \\ & & & & -1 & 1 \end{bmatrix},$$

and a random right-hand side. Although B, the Strang preconditioner P and $P^{-1}B$ are well-conditioned, the eigenvector matrices of B and $P^{-1}B$ are not, as can be seen from Table 6.2.

To the unpreconditioned linear system $Bx = b$, where B has diagonalization $B = Z\Lambda Z^{-1}$, we sought to apply M-MR, where $M = Z^{-*}Z^{-1}$. However, the matrix of eigenvectors Z was too ill-conditioned for computations with M to be feasible. To $P^{-1}Bx = P^{-1}b$ we applied M-MR with $M = S^{-*}S^{-1}$, as in (6.10), where the inner product $\langle \cdot, \cdot \rangle_M$ is related to the symmetric bilinear form $\langle \cdot, \cdot \rangle_{WP}$. To simplify the computations, we chose $S = \mathcal{Z}$ to be the matrix of eigenvectors of P computed by Matlab's `eig` function. Thus,

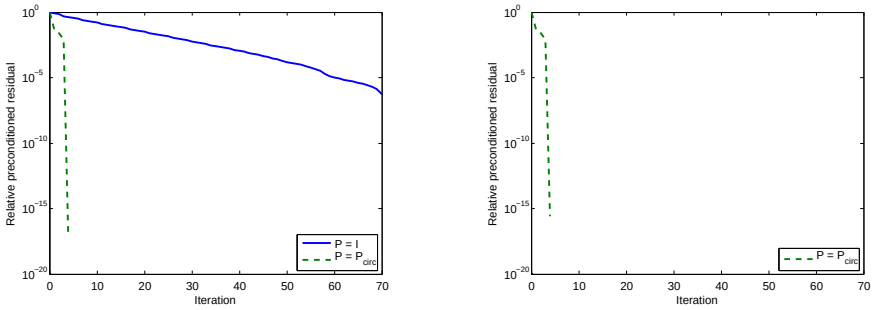
$$WP = S^{-*}YS^{-1} \tag{6.12}$$

for some Y of the form prescribed in (2.21). (Some of the signs ϵ_i, $i = 1, \ldots, j$ may be -1, since S is prescribed.) Since S and Y are known, a diagonalization of P that would lead to this factorization can be determined if desired.

From Figure 6.2 it is clear that the unpreconditioned system converges slowly but that the preconditioner is very effective. Both M-MR and I-MR for the preconditioned system converge similarly, as might be expected from Theorem 3.23, given that M has I-norm condition number 1.9. However, the convergence of I-MR is faster than that predicted from the standard eigenvalue GMRES bound (3.27) since the eigenvectors of $P^{-1}B$ are so poorly conditioned (see Table 6.2). We note that if we were to use the Fourier matrix U (c.f. (6.11)) rather than the eigenvectors determined by Matlab

	$\kappa_I(B)$	$\kappa_I(P)$	$\kappa_I(P^{-1}B)$	$\kappa_I(\mathcal{Z}_B)$	$\kappa_I(\mathcal{Z}_P)$	$\kappa_I(\mathcal{Z}_{P^{-1}B})$
Grcar	3.59	3.60	2.69	5.1×10^{17}	1.4	1.2×10^{8}

Table 6.2: Condition numbers of B, P and $P^{-1}B$, and of their eigenvectors, for the Grcar matrix in Example 6.1.



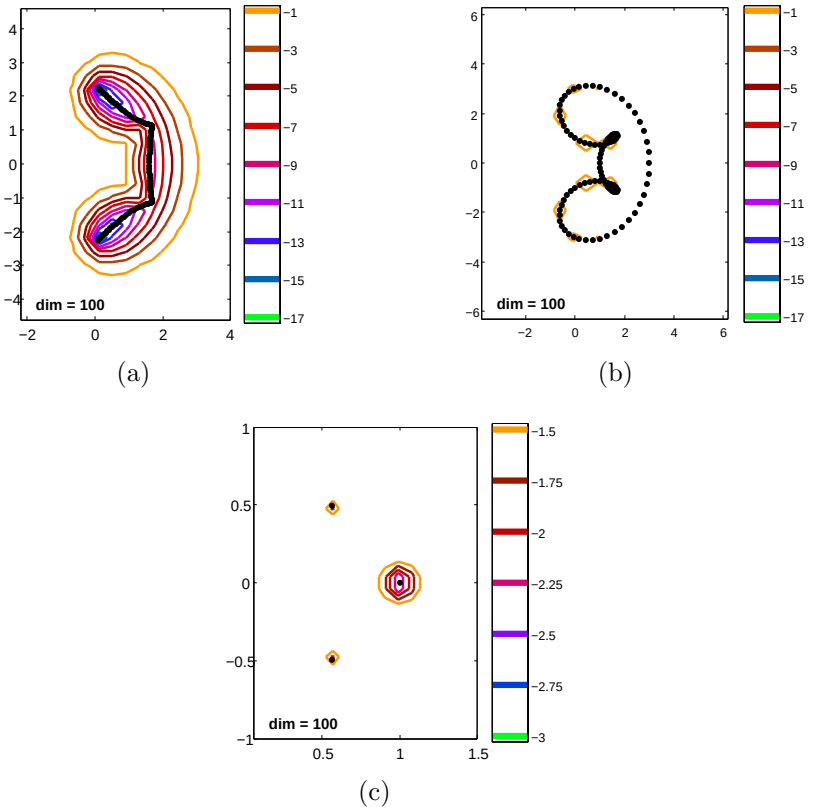
(a) I:MR (b) M-MR

Figure 6.2: Relative preconditioned residuals for (a) I-MR and (b) M-MR convergence of the preconditioned linear system for the Grcar matrix in Example 6.1.

to compute M we would have that $M = U^{-*}U^{-1} = I$, i.e. M-MR and I-MR would be identical methods. In this case the bound (3.59) in Theorem 2.20 would be much more descriptive of I-MR convergence than the standard eigenvalue bound (3.27)!

Much is known about the pseudospectra of Toeplitz and circulant matrices, both of which are related to the symbol of the matrix (see [145, Chapter 7]). Pseudospectral plots of B, P and $P^{-1}B$ are shown in Figure 6.3. Since each ϵ-pseudospectral curve of B is much further than a distance ϵ away from its eigenvalues (see Figure 6.3), this matrix is nonnormal. (Recall from Lemma A.8 that the ϵ-pseudospectrum of a normal matrix comprises ϵ-balls around the eigenvalues.) Although P is normal its eigenvalues, shown in Figure 6.3, follow the profile of the pseudospectra[4] of B; the circulant preconditioner captures information about B that is different from the eigenvalues. Additionally, the preconditioned coefficient matrix appears to have few distinct eigenvalues and is much closer to normal than B, since its ϵ-pseudospectra in Figure 6.3 are almost open ϵ-balls around the eigenvalues. This normality and clustering of eigenvalues explains the fast GMRES convergence for the preconditioned system.

[4]The eigenvalues of P are completely described by Theorem 7.1 in [145].



(a) (b) (c)

Figure 6.3: Pseudospectra of (a) B, (b) P and (c) $P^{-1}B$ for the Grcar matrix in Example 6.1. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-17}, 10^{-1}]$ for P and B and for selected ϵ in $[10^{-3}, 10^{-1.5}]$ for $P^{-1}B$.

6.2.2 Kronecker products of Toeplitz matrices

Certain finite difference approximations of, for example, convection-diffusion equations in two (or more) dimensions can be represented in terms of Kronecker products of Toeplitz matrices. We consider an approximation of the form

$$B = I_{n_y} \otimes B_x + B_y \otimes I_{n_x},$$

where B_x and B_y are Toeplitz matrices of dimensions n_x and n_y, respectively. The matrix B can be effectively preconditioned by a semi-circulant preconditioner,

$$P = I_{n_y} \otimes C_x + B_y \otimes I_{n_x},$$

where C_x is a circulant approximation of B_x [69]. As in the previous example, the fast Fourier transform makes P cheap to apply.

At the start of Section 2.2, and again in Section 6.2.1, we stated that Toeplitz (and circulant) matrices are self-adjoint with respect to symmetric bilinear forms defined by reverse identity matrices (2.19). Thus, B_x and C_x are $\hat{Y}_{n_x}$-self-adjoint and B_y is $\hat{Y}_{n_y}$-self-adjoint, where $\hat{Y}_{n_x}$ and $\hat{Y}_{n_y}$ are reverse identity matrices of dimension n_x and n_y, respectively. It is natural to ask whether B and P are each self-adjoint with respect to $\langle \cdot, \cdot \rangle_W$, where

$$W = \hat{Y}_{n_y} \otimes \hat{Y}_{n_x}.$$

By using the Kronecker product properties in Appendix A.2, similarly to that for the example on page 85, it is straightforward to show that $WB = B^T W$ which means that B is self-adjoint with respect to $\langle \cdot, \cdot \rangle_W$ (c.f. (2.13)). The preconditioner P has the same structure as B and so is also $\langle \cdot, \cdot \rangle_W$-self-adjoint. Thus, P and B are self-adjoint with respect to the same symmetric bilinear form. We now examine whether, in addition, the preconditioner is effective.

Example 6.2. The equation we consider is a two-dimensional convection-diffusion equation

$$-\epsilon \nabla^2 u + w \cdot \nabla u = f \text{ in } \Omega = (0,1) \times (0,1), \quad u = g \text{ on } \partial\Omega \tag{6.13}$$

with $w = (0,1)^T$ and boundary conditions

$$g(x,0) = x, \quad g(0,y) = -1, \quad g(1,y) = 1, \quad g(x,1) = 0.$$

Following [69] we let n_x and n_y be the number of grid points used in the discretization in the x and y directions, respectively, so that $x_k = kh_x$, $k = 1, \ldots, n_x$ and $y_j = jh_y$, $j = 1, \ldots n_y$, where

$$h_x = \frac{1}{n_x + 1} \quad \text{and} \quad h_y = \frac{1}{n_y + 1}.$$

The second-order centered finite difference[5] approximation of (6.13) leads to equations of the form

$$-\epsilon\left(\frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h_x^2}+\frac{u_{k,j+1}-2u_{k,j}+u_{k,j-1}}{h_y^2}\right)+\frac{u_{k,j+1}-u_{k,j-1}}{2h_y}=g(x_k,y_j), \quad (6.14)$$

where $k=1,\ldots,n_x, j=1,\ldots,n_y$. These are related to the one-dimensional finite difference approximation in Section (2.3), where an upwind finite difference approximation was used to discretize a one-dimensional convection-diffusion equation. Let

$$u=\begin{bmatrix}u_{1,1}, & u_{2,1}, & \ldots, & u_{n_x,1}, & u_{1,2}, & \ldots, & u_{n_x,n_y}\end{bmatrix}^T,$$

$$\alpha_x=0, \quad \alpha_y=\frac{1}{h_y}, \quad \beta_i=\frac{2\epsilon}{h_i^2}, \ i=x,y$$

and

$$B_i=\begin{bmatrix}2\beta_i & \alpha_i-\beta_i & & \\ -\alpha_i-\beta_i & \ddots & \ddots & \\ & \ddots & \ddots & \alpha_i-\beta_i \\ & & -\alpha_i-\beta_i & 2\beta_i\end{bmatrix},$$

for $i=x,y$. Then u is the solution of $Bx=b$, where $B=I_{n_y}\otimes B_x+B_y\otimes I_{n_x}$ and

$$b_{k,j}=2g(x_k,y_j), \ k=1,\ldots,n_x, \ j=1,\ldots,n_y.$$

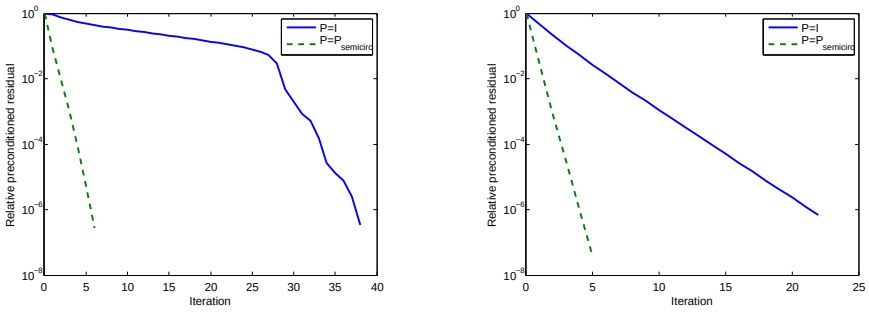
We choose $n_x=7$, $n_y=28$, and $\epsilon=0.0325$, for which the mesh Peclet number is $\frac{w_y h_y}{\epsilon}=1.06$.

The semi-circulant preconditioner proposed by Hemmingsson [69] is $P=I_{n_y}\otimes C_x+B_y\otimes I_{n_x}$, where

$$C_x=\begin{bmatrix}2\beta_1 & -\beta_1 & & -\beta_1 \\ -\beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & -\beta_1 \\ -\beta_1 & & -\beta_1 & 2\beta_1\end{bmatrix}.$$

Both M-MR and I-MR were used to solve this problem. To the unpreconditioned problem $Bx=b$, where B has diagonalization $B=Z\Lambda Z^{-1}$, we apply M-MR with $M=Z^{-*}Z^{-1}$ and Z computed by `eig` in Matlab. To the preconditioned system $P^{-1}Bx=P^{-1}b$ we apply M-MR with $M=S^{-*}S^{-1}$ where, as discussed for Toeplitz

[5] A centered finite difference approximation may be safely used because the mesh Peclet number is less than two (see [69]).



(a) I-MR (b) M-MR

Figure 6.4: Relative preconditioned residuals for the finite-difference discretization of the convection-diffusion problem, preconditioned by a semi-circulant preconditioner in Example 6.2 when convergence is measured in the (a) I-norm and (b) M-norm.

matrices in Section 6.2.1, S is the matrix of eigenvectors of P computed by Matlab's `eig` function.

The relative residuals are shown in Figure 6.4. Although there is some disparity between M- and I-norm convergence for the original problems, the preconditioned systems converge at a similar (and rapid) rate. The semi-circulant preconditioner is clearly effective for this problem. However, the upper bound in Theorem 3.23, that relates M-MR and I-MR convergence, is unduly pessimistic since, for the preconditioned system $\kappa_I(M) = 5.6 \times 10^{15}$. The standard I-MR eigenvalue bound (3.27) is also poor as a result of the ill-conditioning of the eigenvectors of $P^{-1}B$ (see Table 6.3).

The pseudospectra of P and B in Figure 6.5 show that both matrices are nonnormal. However, their pseudospectra are very similar, perhaps indicating that the eigenvectors and eigenvalues of P are nearby to those of B, as discussed in Theorem 6.6. The pseudospectra of $P^{-1}B$ in Figure 6.5 indicates that the preconditioned coefficient matrix is much closer to normal than the original and that the eigenvalues of $P^{-1}B$ are very clustered. (Lemma A.8 states that, for any ϵ, the ϵ-pseudospectra of a normal matrix is comprised of open ϵ-balls centred at the eigenvalues of the matrix). This is in line with the fast convergence we observe from Figure 6.4.

$\kappa_I(B)$	$\kappa_I(P)$	$\kappa_I(P^{-1}B)$	$\kappa_I(Z_B)$	$\kappa_I(Z_P)$	$\kappa_I(Z_{P^{-1}B})$
62.0	70.2	2.6	9.0×10^6	7.5×10^7	1.5×10^7

Table 6.3: Condition numbers of B, P and $P^{-1}B$, and of their eigenvectors, for the finite-difference discretization of the convection-diffusion problem preconditioned by a semi-circulant preconditioner in Example 6.2.

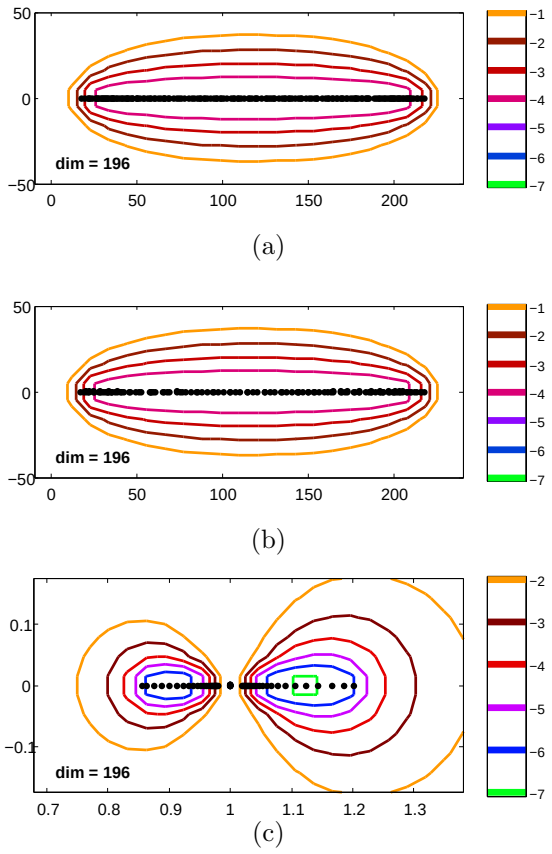


Figure 6.5: Pseudospectra of (a) B, (b) P and (c) $P^{-1}B$ for the finite-difference discretization of the convection-diffusion problem, preconditioned by a semi-circulant preconditioner in Example 6.2. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-7}, 10^{-1}]$ for P and B and for selected ϵ in $[10^{-7}, 10^{-2}]$ for $P^{-1}B$.

$\kappa_I(B)$	$\kappa_I(P)$	$\kappa_I(P^{-1}B)$	$\kappa_I(Z_B)$	$\kappa_I(\mathcal{Z}_P)$	$\kappa_I(Z_{P^{-1}B})$
89.1	89.1	1.2	1.7×10^8	1.6×10^8	4.5×10^{10}

Table 6.4: Condition numbers of B, P and $P^{-1}B$, and of the eigenvectors of P, B and $P^{-1}B$, for the SUPG matrix and finite difference preconditioner in Example 6.3.

6.2.3 A persymmetric finite element discretization

We examine in Example 6.3 below a streamline upwind Petrov-Galerkin (SUPG) finite element discretization of another convection-diffusion operator that is persymmetric but not Toeplitz. The geometric multigrid (GMG) preconditioner of Ramage [111] has been shown to be effective for SUPG convection-diffusion problems for which the viscosity is not too small. For the problem we consider in Example 6.3 below, the preconditioner is also persymmetric, and so P and B are self-adjoint with respect to the same symmetric bilinear form, defined by the reverse identity matrix (2.19).

Example 6.3. The test problem we consider is the third example in [43, Chapter 3]. The differential equation satisfies

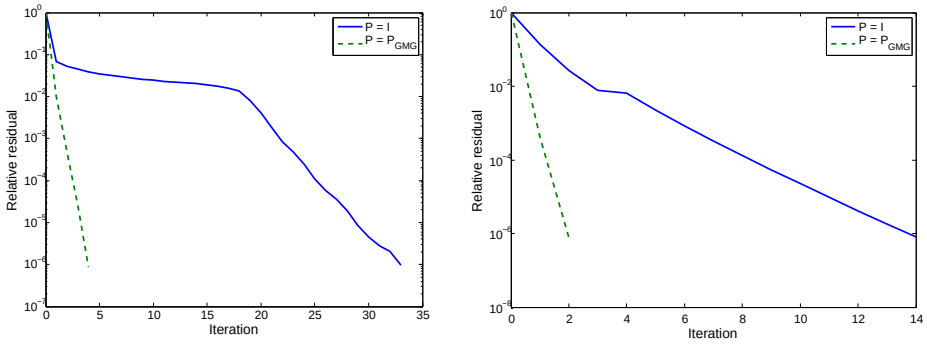
$$-\frac{1}{200}\nabla^2 u + w \cdot \nabla u = f \text{ in } \Omega = (-1,1) \times (-1,1), \quad u = g \text{ on } \partial\Omega$$

with a wind $w = (-\sin\frac{\pi}{6}, \cos\frac{\pi}{6})^T$ at a 30° angle to the left of vertical and boundary conditions

$$\begin{aligned} u(x,-1) &= 1, & 0 < x \leq 1 \\ u(1,y) &= 1, & -1 \leq y < 1 \\ u(x,y) &= 0, & \text{elsewhere on } \partial\Omega. \end{aligned}$$

The convection-diffusion equation is discretized using Ifiss 3.1 [125] with the default settings of Q1 finite elements on a 16×16 grid, and SUPG with the default stabilization parameter (defined in [43, Equation 3.44]). The one-dimensional mesh Peclet numbers in the x- and y-directions, that were used in Section 6.2.2, are $w_x h_x/\epsilon = 12.5$ and $w_y h_y/\epsilon = 21.7$. The element mesh Peclet number[6] [43, page 132] is 14.4. Before boundary conditions are enforced—by a process described for the Poisson equation in [43, Section 1.4.3]—the SUPG matrix is Toeplitz. Afterwards, it is merely persymmetric. The multigrid preconditioner is also computed by Ifiss with the default settings of line Gauss-Seidel smoothing in 2 directions with 1 pre-smoothing step and 1 post-smoothing step.

[6]The element mesh Peclet number is $\|w\|h/(2\epsilon)$ where, if $\theta = \arctan(|w_y/w_x|)$ at the centroid, $h = \min\{h_x/\cos\theta, h_y/\sin\theta\}$.



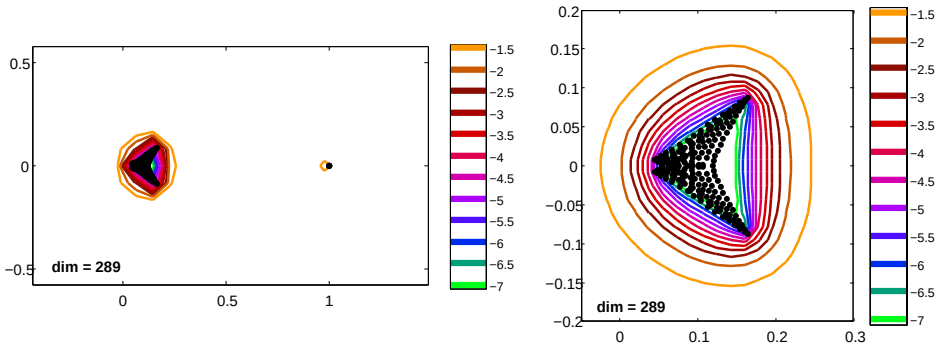
(a) I-MR (b) M-MR

Figure 6.6: Relative preconditioned residuals for I-MR (left column) and M-MR convergence for the SUPG matrix and multigrid preconditioner in Example 6.3.

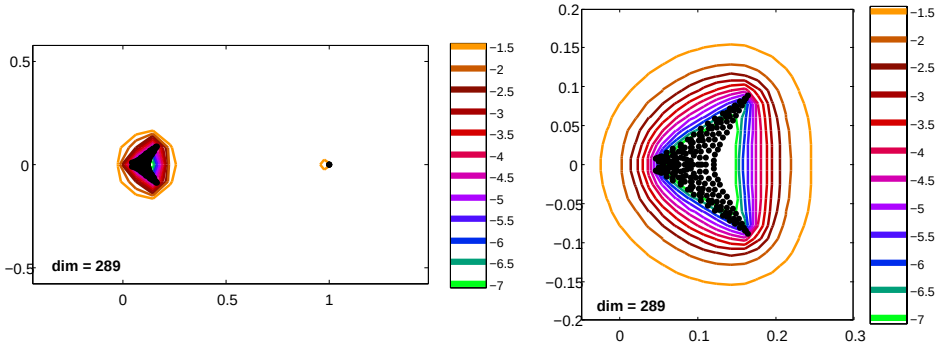
As in the previous examples, we apply both M-MR and I-MR. If B has diagonalization $B = Z\Lambda Z^{-1}$ we use M-MR, with $M = Z^{-*}Z^{-1}$ and Z computed by `eig` in Matlab, on the unpreconditioned problem $Bx = b$. To the preconditioned system $P^{-1}Bx = P^{-1}b$ we again use M-MR with $M = S^{-*}S^{-1}$ where we choose $S = \mathcal{Z}$ to be the matrix of eigenvectors of P computed by Matlab's `eig` function, as discussed in Section 6.2.1 for Toeplitz matrices.

The first observation we make is that preconditioning makes the condition number of the eigenvectors worse, as is clear from Table 6.4. Although the preconditioner is effective, as is evident from Figure 6.6, the standard eigenvalue I-GMRES bound (3.27) is incapable of predicting the fast convergence. Theorem 3.23, which relates I-MR and M-MR convergence is also poor since, for the preconditioned system, $\kappa_I(M) = 1.7 \times 10^6$.

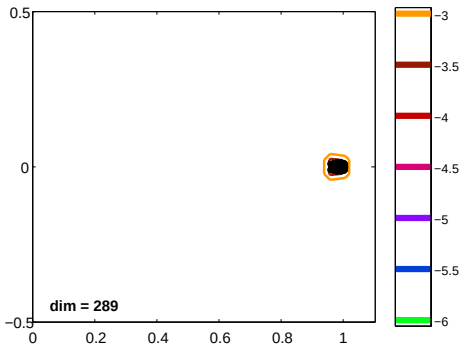
Figure 6.7 shows that P and B are similarly nonnormal and it is not surprising that convergence for the unpreconditioned problem, shown in Figure 6.6, is slow. Since these matrices have nearly identical pseudospectra, their eigenvectors may be close, as in Theorem 6.6. The preconditioned coefficient matrix has pseudospectral curves that are small balls around its clustered eigenvalues, indicating that $P^{-1}B$ is close to normal. As a result, fast GMRES convergence for the preconditioned system is to be expected.



(a) (b)



(c) (d)



(e)

Figure 6.7: Pseudospectra of (a) and (b) B, (c) and (d) P and (e) $P^{-1}B$ for the SUPG matrix and multigrid preconditioner in Example 6.3. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-7}, 10^{-1.5}]$ for P and B and for selected ϵ in $[10^{-6}, 10^{-3}]$ for $P^{-1}B$.

6.3 Preconditioners with a nearby symmetric bilinear form

In the last section we discussed preconditioners that were self-adjoint with respect to the same symmetric bilinear form as B. Unfortunately, for many applications a W-self-adjoint preconditioner is not known (except when $W = I$, so that B and P are symmetric, or when W is the reverse identity matrix, so that B and P are persymmetric). However, if a preconditioner is self-adjoint with respect to a symmetric bilinear form that is close to W, say $\mathcal{W}$, we might expect that we can find a matrix $\mathcal{S}$ such that $\mathcal{W}P = \mathcal{S}^{-*}Y\mathcal{S}^{-1}$ and $P^{-1}B = (\mathcal{S} - E)\Lambda(\mathcal{S} - E)^{-1}$ is a diagonalization of $P^{-1}B$, with E a small perturbation. In this case, if $\mathcal{M} = \mathcal{S}^{-*}\mathcal{S}^{-1}$, $P^{-1}B$ is nearly $\mathcal{M}$-normal, in the sense that $(P^{-1}B)(P^{-1}B)^\star - (P^{-1}B)^\star(P^{-1}B)$ is small.

Lemma 6.8. *Let $P, B \in \mathbb{R}^{n\times n}$ and let $P^{-1}B = (\mathcal{S} - E)\Lambda(\mathcal{S} - E)^{-1}$, $\mathcal{M} = \mathcal{S}^{-*}\mathcal{S}^{-1}$ and $G = I - \mathcal{S}^{-1}E$, where $\mathcal{S}, \Lambda, E \in \mathbb{C}^{n\times n}$. Then, if $(P^{-1}B)^\star$ is the $\mathcal{M}$-adjoint of $P^{-1}B$,*

$$(P^{-1}B)(P^{-1}B)^\star - (P^{-1}B)^\star(P^{-1}B) = \mathcal{S}\left[G\Lambda G^{-1}G^{-*}\Lambda^* G^* - G^{-*}\Lambda^* G^* G\Lambda G^{-1}\right]\mathcal{S}^{-1}.$$

Proof.

$$\begin{aligned}
&(P^{-1}B)(P^{-1}B)^\star - (P^{-1}B)^\star(P^{-1}B)\\
=&(\mathcal{S} - E)\Lambda(\mathcal{S} - E)^{-1}\mathcal{S}\mathcal{S}^*(\mathcal{S} - E)^{-*}\Lambda^*(\mathcal{S} - E)^*\mathcal{S}^{-*}\mathcal{S}^{-1}\\
&- \mathcal{S}\mathcal{S}^*(\mathcal{S} - E)^{-*}\Lambda^*(\mathcal{S} - E)^*\mathcal{S}^{-*}\mathcal{S}^{-1}(\mathcal{S} - E)\Lambda(\mathcal{S} - E)^{-1}
\end{aligned}$$

from which the result follows. □

Remark 6.9. Provided $\mathcal{S}^{-1}E$ is small enough, G is close to the identity and the two terms in brackets in Lemma 6.8 will be nearly identical. In this case $P^{-1}B$ is nearly $\mathcal{M}$-normal and $P^{-1}B$ has eigenvectors that are close to $\mathcal{M}$-unitary. If the eigenvectors were $\mathcal{M}$-unitary then, similarly to the I-normal case, the relative preconditioned residuals of $\mathcal{M}$-MR applied to $P^{-1}Bx = P^{-1}b$, measured in the $\mathcal{M}$-norm, could be bounded by polynomials evaluated at the eigenvalues of $P^{-1}B$ (c.f., (6.10)). If $P^{-1}B$ is nearly $\mathcal{M}$-normal then we might expect that these relative preconditioned residuals would be bounded by polynomials evaluated at the eigenvalues of $P^{-1}B$ up to some small constant (greater than 1).

Furthermore, the convergence of $\mathcal{M}$-MR is still governed chiefly by eigenvalues, as the following theorem illustrates. As discussed in Section 6.2, the relationships in Section 3.2.4, particularly Corollary 3.20 and Theorem 3.23, allow these results to be translated to convergence bounds for I-MR.

Theorem 6.10. *Let $P, B \in \mathbb{R}^{n\times n}$, and let*

$$\mathcal{M} = \mathcal{S}^{-*}\mathcal{S}^{-1}, \tag{6.15}$$

$\mathcal{S} \in \mathbb{C}^{n\times n}$. Suppose $P^{-1}B$ has diagonalization

$$P^{-1}B = (\mathcal{S} - E)\Lambda(\mathcal{S} - E)^{-1}, \tag{6.16}$$

where $\Lambda, E \in \mathbb{C}^{n\times n}$. Then the preconditioned residual $P^{-1}r_k^{\mathcal{M}}$ of any minimum residual method in $\langle\cdot,\cdot\rangle_{\mathcal{M}}$ applied to $P^{-1}Bx = P^{-1}b$, $b \in \mathbb{R}^n$, satisfies

$$\frac{\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \kappa_I(I - \mathcal{S}^{-1}E) \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)| \tag{6.17}$$

and

$$\frac{\|P^{-1}r_k\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \frac{L_\epsilon}{2\pi\epsilon} \min_{p\in\Pi_k, p(0)=1} \max_{z\in\Gamma_\epsilon} |p_k(z)|, \tag{6.18}$$

where Γ_ϵ is any curve or union of curves that encloses the ϵ-pseudospectrum of the matrix $\mathcal{S}^{-1}P^{-1}B\mathcal{S}$ and L_ϵ is the arc length of Γ_ϵ.

If, moreover, $\|\mathcal{S}^{-1}E\|_I < 1$ then

$$\frac{\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \frac{1 + \|\mathcal{S}^{-1}E\|_I}{1 - \|\mathcal{S}^{-1}E\|_I} \min_{p\in\Pi_k,\ p(0)=1} \max_{\lambda\in\Lambda(P^{-1}B)} |p(\lambda)|. \tag{6.19}$$

Proof. We start from the polynomial representation of the residual,

$$P^{-1}r_k^{\mathcal{M}} = p_k(P^{-1}B)P^{-1}r_0.$$

Upon taking $\mathcal{M}$-norms we obtain

$$\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}} = \|p_k(P^{-1}B)P^{-1}r_0\|_{\mathcal{M}}.$$

From (2.2) and (6.16) it follows that

$$\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}} \leq \|\mathcal{S}^{-1}(\mathcal{S} - E)p_k(\Lambda)(\mathcal{S} - E)^{-1}\mathcal{S}\|_I \|P^{-1}r_0\|_{\mathcal{M}}.$$

Thus,

$$\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}} \leq \|I - \mathcal{S}^{-1}E\|_I \|(I - \mathcal{S}^{-1}E)^{-1}\|_I \|p_k(\Lambda)\|_I \|P^{-1}r_0\|_{\mathcal{M}}.$$

Since $\mathcal{M}$-MR minimizes the residual and Λ is diagonal, we have (6.17).

If $\|\mathcal{S}^{-1}E\|_I < 1$ then

$$\|(I - \mathcal{S}^{-1}E)^{-1}\|_I \leq \frac{1}{1 - \|\mathcal{S}^{-1}E\|_I}$$

(see, for example, [75, page 301]) and

$$\frac{\|P^{-1}r_k^{\mathcal{M}}\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \frac{1 + \|\mathcal{S}^{-1}E\|_I}{1 - \|\mathcal{S}^{-1}E\|_I} \min_{p \in \Pi_k,\, p(0)=1} \max_{\lambda \in \Lambda(P^{-1}B)} |p(\lambda)|.$$

which is (6.19). The pseudospectral bound (6.18) is obtained from Theorem 3.13 by choosing $\mathcal{M} = \mathcal{S}^{-*}\mathcal{S}^{-1}$. □

Remark 6.11. The pseudospectral bound depends on the normality of $\mathcal{S}^{-1}P^{-1}B\mathcal{S}$. This matrix should be close to diagonal, in the sense that the off-diagonal entries are small in magnitude, and therefore close to normal, when P and B are self-adjoint with respect to nearby symmetric bilinear forms.

There appears to be a link between the nearness of symmetric bilinear forms discussed in this section and the nearness of eigenvectors discussed in Section 6.1. It is difficult to make firm connections, but we briefly comment on the similarity between the role of E in Theorem 6.10 and of C in (6.1) and F in (6.4).

From (6.3) we have that

$$P^{-1}B = \mathcal{Z}\Lambda_P^{-1}\Lambda_B\mathcal{Z}^{-1} + \mathcal{Z}\Lambda_P^{-1}C\mathcal{Z}^{-1} \tag{6.20}$$

while, since F in (6.4) is given by $Z^{-1}\mathcal{Z} = I + F$,

$$\begin{aligned} P^{-1}B &= \mathcal{Z}\Lambda_P^{-1}(I+F)^{-1}\Lambda_B(I+F)\mathcal{Z}^{-1} \\ &= \mathcal{Z}\Lambda_P^{-1}(I+F)^{-1}\Lambda_B\mathcal{Z}^{-1} + \mathcal{Z}\Lambda_P^{-1}(I+F)^{-1}\Lambda_B F\mathcal{Z}^{-1}. \end{aligned} \tag{6.21}$$

On the other hand, (6.17) in Theorem 6.10 gives that

$$\begin{aligned} P^{-1}B &= (\mathcal{Z} - E)\Lambda_{P^{-1}B}(\mathcal{Z} - E)^{-1} \\ &= (\mathcal{Z} - E)\Lambda_{P^{-1}B}(I - \mathcal{Z}^{-1}E)^{-1}\mathcal{Z}^{-1} \\ &= \mathcal{Z}\Lambda_{P^{-1}B}(I - (\mathcal{Z}^{-1}E))^{-1}\mathcal{Z}^{-1} - \mathcal{Z}(\mathcal{Z}^{-1}E)\Lambda_{P^{-1}B}(I - (\mathcal{Z}^{-1}E))^{-1}\mathcal{Z}^{-1}, \end{aligned}$$

which is not dissimilar in form to (6.20) or (6.21).

6.3.1 Capturing real and complex eigenvalues

We saw in Section 6.3 that when P and B are self-adjoint with respect to nearby symmetric bilinear forms, convergence of a Krylov method in a nonstandard inner product applied to the preconditioned system depends primarily on eigenvalues. We analyse here the effect of differences between Y_P and Y_B, that appear in the symmetric bilinear forms for P and B (see Theorem 2.20), on convergence for $P^{-1}B$ of a minimal residual method in a nonstandard inner product. Accordingly, we assume that P and B have the same eigenvectors. In this section it is convenient to work in real arithmetic and so we use the real Jordan form in preference to the possibly complex diagonalization.

We let B and P have as their real Jordan forms (see Appendix A.1)

$$B = Z\mathcal{J}_B Z^{-1} \tag{6.22}$$

and

$$P = Z\mathcal{J}_P Z^{-1}. \tag{6.23}$$

From Theorem 2.20, B is self-adjoint with respect to $\langle\cdot,\cdot\rangle_W$ and P is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, where $W = ZY_BZ^{-1}$ and $\mathcal{W} = ZY_PZ^{-1}$, and any disparity between W and $\mathcal{W}$ is due to Y_B and Y_P. Since P is given by (6.23), we choose $\mathcal{M} = Z^{-T}Z^{-1}$ so that $\mathcal{M}$ is akin to that in Theorem 6.10 but depends on the real Jordan form of P rather than the complex diagonalization. To determine the convergence of $\mathcal{M}$-MR we consider three cases, namely, that B has fewer, an equal number of, or more real eigenvalues than P.

When the number of real eigenvalues is the same P and B are self-adjoint with respect to the same symmetric bilinear form and convergence of $\mathcal{M}$-MR is described in Section 6.2. This leaves us with the possibility that the number of real eigenvalues of B differs from that of P. We consider in detail the effect on the convergence of $\mathcal{M}$-MR when B has the greater number of real eigenvalues by first expressing the real Jordan form of $P^{-1}B$ in terms of Z and then applying the convergence result in Corollary 3.12.

Let

$$\mathcal{J}_B = \operatorname{diag}\left(\Lambda_1^B, \ldots, \Lambda_{\gamma_B}^B, \mathcal{J}_{\gamma_B+1}^B, \ldots, \mathcal{J}_\delta^B\right),$$

with

$$\Lambda_1^B = \begin{cases} \lambda_1, & n \text{ odd}, \\ \begin{bmatrix} \lambda_1^{(a)} & 0 \\ 0 & \lambda_1^{(b)} \end{bmatrix}, & n \text{ even}, \end{cases}$$

$$\Lambda_i^B = \begin{bmatrix} \lambda_i^{(a)} & 0 \\ 0 & \lambda_i^{(b)} \end{bmatrix}, i = 2, \ldots, \gamma_B \quad \text{and} \quad \mathcal{J}_i^B = \begin{bmatrix} \alpha_i & \beta_i \\ -\beta_i & \alpha_i \end{bmatrix}, i = \gamma_B + 1, \ldots, \delta.$$

Let $\mathcal{J}_P$ be conformally partitioned so that

$$\mathcal{J}_P = \text{diag}\left(\Lambda_1^P, \ldots, \Lambda_{\gamma_P}^P, \mathcal{J}_{\gamma_P+1}^P, \ldots, \mathcal{J}_\delta^P\right),$$

where

$$\Lambda_1^P = \begin{cases} \phi_1, & n \text{ odd}, \\ \begin{bmatrix} \phi_1^{(a)} & 0 \\ 0 & \phi_1^{(b)} \end{bmatrix}, & n \text{ even}, \end{cases}$$

$$\Lambda_i^P = \begin{bmatrix} \phi_i^{(a)} & 0 \\ 0 & \phi_i^{(b)} \end{bmatrix}, i = 2, \ldots, \gamma_P \quad \text{and} \quad J_i^P = \begin{bmatrix} \mu_i & \nu_i \\ -\nu_i & \mu_i \end{bmatrix}, i = \gamma_P + 1, \ldots, \delta.$$

Then, from (6.22) and (6.23) we have that

$$P^{-1}B = Z\mathcal{J}_P^{-1}\mathcal{J}_B Z^{-1} = ZJZ^{-1}. \tag{6.24}$$

where

$$J = \text{diag}\left(\mathcal{J}_1, \ldots, \mathcal{J}_{\gamma_P}, J_{\gamma_P+1}, \ldots, J_{\gamma_B}, \mathcal{J}_{\gamma_B+1}, \ldots \mathcal{J}_\delta\right) \tag{6.25}$$

with

$$\mathcal{J}_1 = \begin{cases} \frac{\lambda_1}{\mu_1}, & n \text{ odd }, \\ \begin{bmatrix} \frac{\lambda_1^{(a)}}{\mu_1^{(a)}} & 0 \\ 0 & \frac{\lambda_1^{(b)}}{\mu_1^{(b)}} \end{bmatrix}, & n \text{ even}, \end{cases}$$

$$\mathcal{J}_i = \begin{cases} \begin{bmatrix} \frac{\lambda_i^{(a)}}{\mu_1^{(a)}} & 0 \\ 0 & \frac{\lambda_i^{(b)}}{\mu_1^{(b)}} \end{bmatrix}, & i = 2, \ldots, \gamma_P, \\ \frac{1}{\mu_i^2+\nu_i^2} \begin{bmatrix} \mu_i\alpha_i + \nu_i\beta_i & \mu_i\beta_i - \nu_i\alpha_i \\ -(\mu_i\beta_i - \nu_i\alpha_i) & \mu_i\alpha + \nu_i\beta_i \end{bmatrix}, & i = \gamma_B + 1, \ldots, \delta \end{cases}$$

and

$$J_i = \frac{1}{\mu_i^2 + \nu_i^2} \begin{bmatrix} \mu_i\lambda_1^{(a)} & -\nu_i\lambda_i^{(b)} \\ \nu_i\lambda_i^{(a)} & \mu_i\lambda_1^{(b)} \end{bmatrix}, \ i = \gamma_P + 1, \ldots, \gamma_B.$$

It is clear that the blocks $\mathcal{J}_i, \ i = 1, \ldots, \gamma_P, \gamma_B + 1, \ldots, \delta$ are already in real Jordan form. Let us now find a real Jordan form for the remaining blocks, $J_i, \ i = \gamma_{P+1}, \ldots \gamma_B$. Each has as its diagonalization

$$J_i = \frac{1}{\mu_i^2 + \nu_i^2} \begin{bmatrix} \mu_i\lambda_1^{(a)} & -\nu_i\lambda_i^{(b)} \\ \nu_i\lambda_i^{(a)} & \mu_i\lambda_1^{(b)} \end{bmatrix} = R_i\Theta_i R_i^{-1},$$

where

$$R_i = \frac{1}{2\tau_i}\begin{bmatrix} \tau_i + \mu_i(\lambda_i^{(a)} - \lambda_i^{(b)}) & \tau_i - \mu_i(\lambda_i^{(a)} - \lambda_i^{(b)}) \\ 2\nu_i\lambda_i^{(a)} & -2\nu_i\lambda_i^{(a)} \end{bmatrix} \tag{6.26}$$

and

$$\Theta_i = \frac{1}{\mu_i^2 + \nu_i^2}\begin{bmatrix} \mu_i(\lambda_i^{(a)} + \lambda_i^{(b)}) + \tau_i & 0 \\ 0 & \mu_i(\lambda_i^{(a)} + \lambda_i^{(b)}) - \tau_i \end{bmatrix}, \tag{6.27}$$

with $\tau_i^2 = \mu_i^2(\lambda_i^{(a)} - \lambda_i^{(b)})^2 - 4\nu_i^2\lambda_i^{(a)}\lambda_i^{(b)}$.

When $\tau_i^2 \geq 0$, so that τ_i is real, this is a real diagonalization of J_i and we have found the real Jordan form. Otherwise, we must convert the complex diagonalization of J_i into the real Jordan form

$$\Theta_i = \frac{1}{\mu_i^2 + \nu_i^2}\begin{bmatrix} \mu_i(\lambda_i^{(a)} + \lambda_i^{(b)}) & \tau_i i \\ -\tau_i i & \mu_i(\lambda_i^{(a)} + \lambda_i^{(b)}) \end{bmatrix} \tag{6.28}$$

and

$$R_i = \frac{1}{2\tau_i^2}\begin{bmatrix} \tau_i^2 & \tau_i\mu_i(\lambda_i^{(a)} - \lambda_i^{(b)})i \\ 0 & 2\nu_i\lambda_i^{(a)}\tau_i i \end{bmatrix}. \tag{6.29}$$

By combining (6.24), (6.25), (6.26), (6.27), (6.28) and (6.29) we obtain that

$$P^{-1}B = ZS\mathcal{J}S^{-1}Z^{-1},$$

where

$$\begin{aligned}\mathcal{J} &= \operatorname{diag}((\Lambda_1^P)^{-1}\Lambda_1^B, \ldots, (\Lambda_{\gamma_P}^P)^{-1}\Lambda_{\gamma_P}^B, \Theta_{\gamma_P+1}, \ldots, \Theta_{\gamma_B}, (\mathcal{J}_{\gamma_B+1}^P)^{-1}\mathcal{J}_{\gamma_B+1}^B, \ldots, (\mathcal{J}_\delta^P)^{-1}\mathcal{J}_\delta^B) \\ &= \operatorname{diag}(\mathcal{J}_1, \ldots, \mathcal{J}_{\gamma_P}, \mathcal{J}_{\gamma_P+1}, \ldots, \mathcal{J}_{\gamma_B}, \mathcal{J}_{\gamma_B+1}, \ldots, \mathcal{J}_\delta)\end{aligned}$$

and

$$S = \operatorname{diag}(I, \ldots, I, R_{\gamma_P+1}, \ldots, R_{\gamma_B}, I, \ldots, I)$$

with Θ_i, $i = \gamma_P \ldots, \gamma_B$ given by (6.27) or (6.28) and the blocks R_i, $i = \gamma_P + 1, \ldots, \gamma_B$ given by (6.26) or (6.29), depending on whether τ_i is real.

Application of Corollary 3.12 to this real Jordan form of $P^{-1}B$ allows us to bound the $\mathcal{M}$-MR preconditioned residuals by

$$\frac{\|P^{-1}r_k\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \kappa_I(S) \min_{p\in\Pi_k, p(0)=1} \max_{\ell=1,\ldots,\delta} |p(\mathcal{J}_\ell)|.$$

When the number of real eigenvalues possessed by B and P is identical, $\kappa_I(S)$ achieves its minimum value of 1. If S has condition number greater than 1, $\mathcal{M}$-MR convergence may be worse than if P and B have an equal number of real eigenvalues. The condition

number of S depends, in turn, on the conditioning of the individual blocks R_i given by (6.26) or (6.29) and the number of blocks in S that differ from the identity matrix is dictated by the disparity between the number of real eigenvalues of P and B.

When P has more real eigenvalues than B, i.e., when $\gamma_B < \gamma_P$, $P^{-1}B$ can again be expressed as $P^{-1}B = ZS\mathcal{J}S^{-1}Z$ using (6.22) and (6.23). Now, however,

$$\begin{aligned}\mathcal{J} &= \operatorname{diag}((\Lambda_1^P)^{-1}\Lambda_1^B, \ldots, (\Lambda_{\gamma_B}^P)^{-1}\Lambda_{\gamma_B}^B), \Theta_{\gamma_B+1}, \ldots, \Theta_{\gamma_B}, (\mathcal{J}_{\gamma_P+1}^P)^{-1}\mathcal{J}_{\gamma_P+1}^B, \ldots, (\mathcal{J}_\delta^P)^{-1}\mathcal{J}_\delta^B) \\ &= \operatorname{diag}(\mathcal{J}_1, \ldots, \mathcal{J}_{\gamma_P}, \mathcal{J}_{\gamma_P+1}, \ldots, \mathcal{J}_{\gamma_B}, \mathcal{J}_{\gamma_B+1}, \ldots, \mathcal{J}_\delta)\end{aligned}$$

where, if

$$\tau_i = \alpha_i^2(\phi_i^{(a)} - \phi_i^{(b)})^2 - 4\beta_i\phi_i^{(a)}\phi_i^{(b)} \tag{6.30}$$

is nonnegative,

$$\Theta_i = \frac{1}{2\phi_i^{(a)}\phi_i^{(b)}}\begin{bmatrix}\alpha_i(\phi_i^{(a)} + \phi_i^{(b)}) + \tau_i & 0 \\ 0 & \alpha_i(\phi_i^{(a)} + \phi_i^{(b)}) - \tau_i\end{bmatrix}.$$

Otherwise,

$$\Theta_i = \frac{1}{2\phi_i^{(a)}\phi_i^{(b)}}\begin{bmatrix}\alpha_i(\phi_i^{(a)} + \phi_i^{(b)}) & \tau_i i \\ -\tau_i i & \alpha_i(\phi_i^{(a)} + \phi_i^{(b)})\end{bmatrix}.$$

The matrix S, which is again the key to determining $\mathcal{M}$-MR convergence, is still conformally partitioned with $\mathcal{J}_B$ and $\mathcal{J}_P$ but is of the form

$$S = \operatorname{diag}(I, \ldots, I, R_{\gamma_B+1}, \ldots, R_{\gamma_P}, I, \ldots, I)$$

where, if τ_i in (6.30) is nonnegative

$$R_i = \frac{1}{2\tau_i}\begin{bmatrix}\tau_i + \alpha_i(\phi_i^{(a)} - \phi_i^{(b)}) & \tau_i - \alpha_i(\phi_i^{(a)} - \phi_i^{(b)}) \\ -2\beta_i\phi_i^{(a)} & 2\beta_i\phi_i^{(a)}\end{bmatrix} \tag{6.31}$$

while if $\tau_i < 0$,

$$R_i = \frac{1}{2\tau_i^2}\begin{bmatrix}\tau_i^2 & \alpha_i(\phi_i^{(a)} - \phi_i^{(b)})\tau_i i \\ 0 & -2\beta_i\phi_i^{(a)}\tau_i i\end{bmatrix}. \tag{6.32}$$

By again using Corollary 3.12 we find that $\mathcal{M}$-MR has preconditioned relative residuals that are are bounded above by

$$\frac{\|P^{-1}r_k\|_{\mathcal{M}}}{\|P^{-1}r_0\|_{\mathcal{M}}} \leq \kappa_I(S) \min_{p\in\Pi_k, p(0)=1} \max_{\ell=1,\ldots,\delta} |p(\mathcal{J}_\ell)|,$$

where $\kappa_I(S)$ depends on the conditioning of R_i given by (6.31) and (6.32). Again, $\kappa_I(S)$ takes its minimum value of 1 when when P and B have an equal number of real eigenvalues and the bound in Corollary 3.12 for the $\mathcal{M}$-MR method cannot be smaller when there is a mismatched number of real eigenvalues.

	$\kappa_I(B)$	$\kappa_I(P)$	$\kappa_I(P^{-1}B)$	$\kappa_I(Z_B)$	$\kappa_I(\mathcal{Z}_P)$	$\kappa_I(Z_{P^{-1}B})$
SUPG $h = 2^{-3}$	29.8	29.8	1.0	1.75×10^{10}	1.76×10^{10}	7.9×10^{9}
SUPG $h = 2^{-4}$	95.6	95.6	1.0	9.2×10^{17}	7.3×10^{17}	4.3×10^{11}
olm100	1.5×10^{4}	3.1×10^{4}	6.0×10^{4}	91.7	3.0×10^{5}	8.3×10^{18}
spectral	8.5×10^{4}	8.2×10^{4}	1.6×10^{5}	1.2×10^{4}	2.6×10^{5}	1.1×10^{6}

Table 6.5: Condition numbers of B, P and $P^{-1}B$, and of their eigenvectors, for the recirculating wind SUPG matrix in Example 6.4, olm100 in 6.5 and the exponential equation in Example 6.6.

6.3.2 Examples

We now examine GMRES convergence for preconditioners that are self-adjoint with respect to symmetric bilinear forms nearby to those with respect to which the coefficient matrices are self-adjoint. In each case, we apply minimum residual methods in both the Euclidean inner product and nonstandard inner products. As in the examples in Section 6.2, to unpreconditioned systems $Bx = b$, where B has diagonalization $B = Z\Lambda Z^{-1}$, we apply M-MR with $M = Z^{-*}Z^{-1}$ and Z computed by `eig` in Matlab. To preconditioned systems $P^{-1}Bx = P^{-1}b$ we apply $\mathcal{M}$-MR with $\mathcal{M} = \mathcal{S}^{-*}\mathcal{S}^{-1}$, as in (6.15), with $\mathcal{S} = \mathcal{Z}$ the matrix of eigenvectors of P computed by Matlab's `eig` function.

Example 6.4. We saw in Example 6.3 that the Ramage multigrid method is effective for SUPG finite element approximations of convection-diffusion problems. Here we examine GMRES convergence and the symmetric bilinear forms with respect to which B and P are self-adjoint for Example 2 in [43, Chapter 3]. The differential equation is again of the form

$$-\epsilon\nabla^2 u + w \cdot \nabla u = f \text{ in } \Omega = (-1,1) \times (-1,1), \quad u = g \text{ on } \partial\Omega$$

but now has a vertical wind

$$w = (0, 1 + (x+1)^2/4)^T,$$

that increases in strength as x increases. The velocity u is zero on the inflow boundary. It decreases quadratically to zero on the right wall and cubically to zero on the left wall. A nonzero Neumann condition is enforced on the outflow boundary. These conditions are shown in Figure 6.8.

We used the same discretization as in Example 6.3 and applied the Ramage multigrid preconditioner with the same settings. Here, however, we use two different meshes and viscosity parameters. When $h = 2^{-3}$, we take $\epsilon = 1/200$ so that the element Peclet

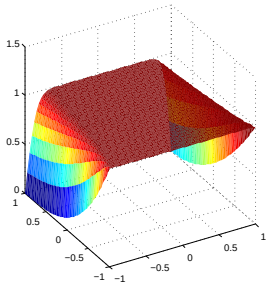


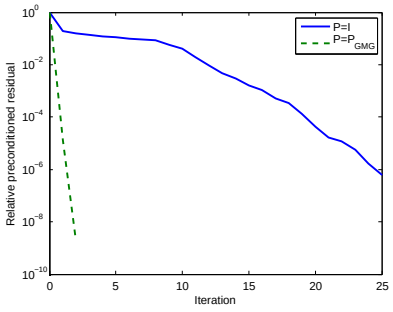
Figure 6.8: Boundary conditions for Example 6.4.

number (see Example 6.3) varies between 25.1 and 47.0, and we choose $\epsilon = 1/500$ when $h = 2^{-4}$, so that the element Peclet number varies between 31.3 and 60.6.

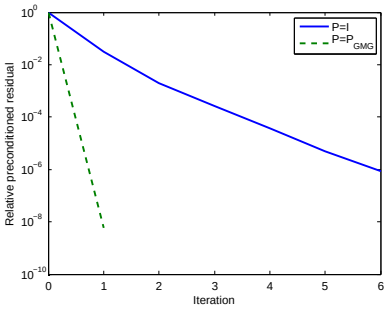
Table 6.5 shows that the condition number of the eigenvectors of $P^{-1}B$ is very large. Consequently, the standard eigenvalue GMRES bound (3.27) is incapable of predicting the fast convergence for both I-MR and $\mathcal{M}$-MR that is evident in Figure 6.9. The convergence curves for $\mathcal{M}$-MR and I-MR are very similar and, as we shall see below, P and B are self-adjoint with respect to nearby symmetric bilinear forms. As a result, Theorem 6.10 indicates that $\mathcal{M}$-MR convergence is determined largely by eigenvalues, while the similarity of the I-MR and $\mathcal{M}$-MR convergence curves indicate that I-MR convergence is also governed by the eigenvalues of $P^{-1}B$. However, for the preconditioned system, $\kappa_I(\mathcal{M}) = 3.1 \times 10^{20}$ when $h = 2^{-3}$ and 6.9×10^{36} when $h = 2^{-4}$, and consequently Theorem 3.23, that relates I-MR and $\mathcal{M}$-MR convergence, does not accurately relate the convergence of these methods for this example. Unfortunately, the iteration-dependent bounds in Section 3.2.4, although smaller, do not reflect the similarity in the $\mathcal{M}$-MR and I-MR convergence curves either for this matrix.

The pseudospectral plots in Figure 6.10 highlight the nonnormality of B and it is not surprising that convergence for the unpreconditioned problem, shown in Figure 6.9, is slow. However, P and B have similar pseudospectral plots, which indicates that these matrices are alike in some way; perhaps their eigenvectors are close, as described in Theorem 6.6. Application of the preconditioner, as for previous examples, results in a matrix that has clustered eigenvalues and that, since its ϵ-pseudospectra are more like open ϵ-balls about the eigenvalues, is also closer to normal. This is reflected in the rapid convergence of GMRES for the preconditioned problem.

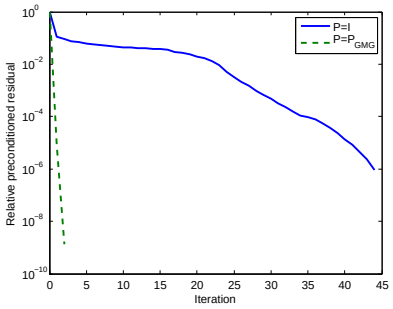
We saw from the pseudospectral plots of B and P that these matrices seem alike. We can glean some information by examining the symmetric bilinear forms with respect to which B and P are self-adjoint. For the coarser grid these can be computed by



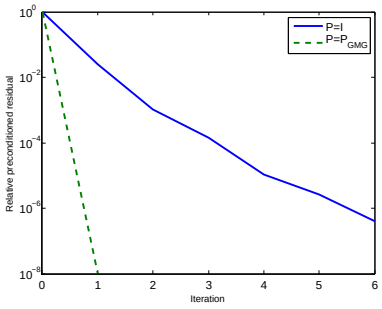
(a) $h = 2^{-3}$: I-MR



(b) $h = 2^{-3}$: M-MR/$\mathcal{M}$-MR

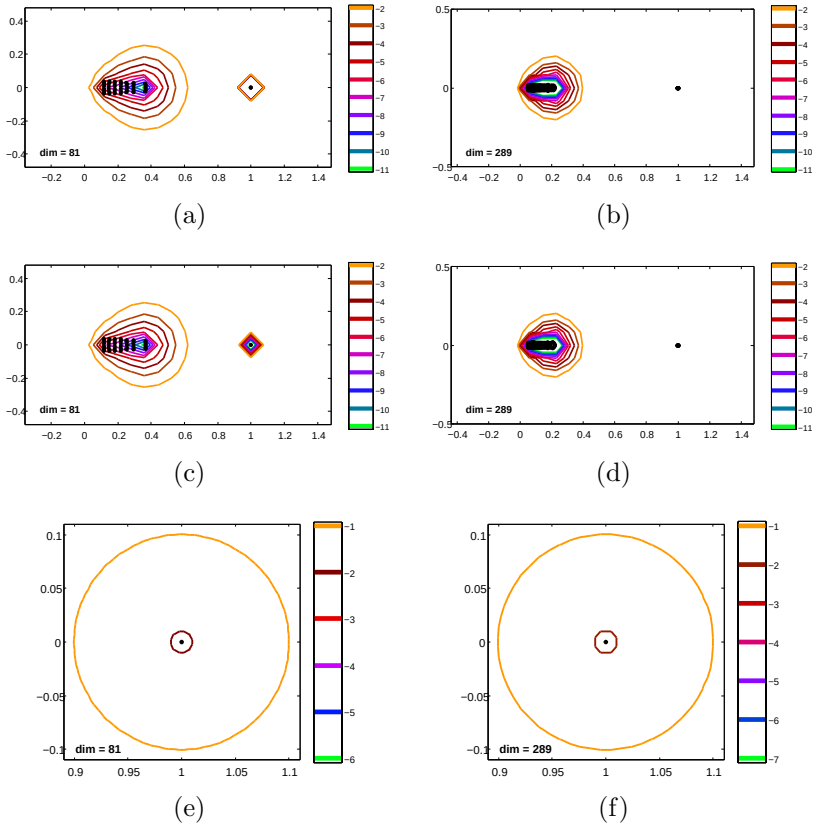


(c) $h = 2^{-4}$: I-MR



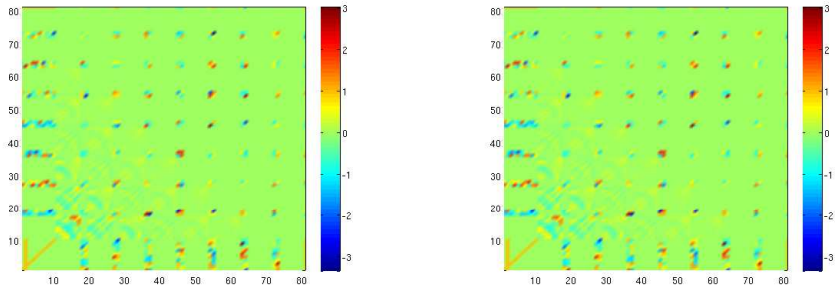
(d) $h = 2^{-4}$: M-MR/$\mathcal{M}$-MR

Figure 6.9: Relative preconditioned residuals for I-MR (left column) and M-MR convergence (for unpreconditioned systems) and $\mathcal{M}$-MR convergence (for preconditioned systems) for the SUPG matrix and multigrid preconditioner in Example 6.4.



(a) (b) (c) (d) (e) (f)

Figure 6.10: Pseudospectra of (a) B, (c) P and (e) $P^{-1}B$ when $h = 2^{-3}$ and (b) B, (d) P and (f) $P^{-1}B$ when $h = 2^{-4}$ for the SUPG matrix and multigrid preconditioner in Example 6.4. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-11}, 10^{-2}]$ for P and B and for selected ϵ in $[10^{-7}, 10^{-1}]$ for $P^{-1}B$.



(a) (b)

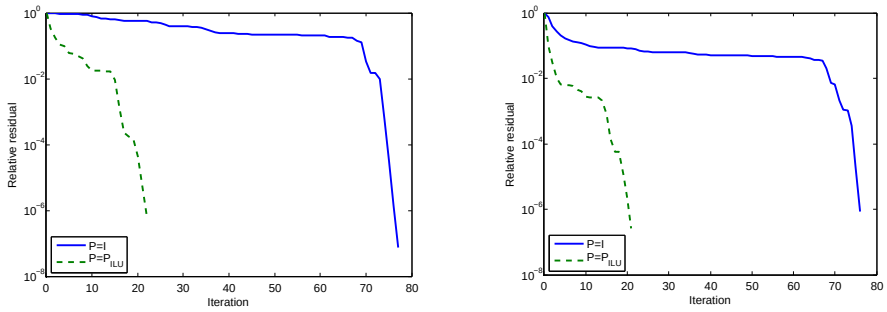
Figure 6.11: Comparison of (a) W and (b) $\mathcal{W}$ for the recirculating wind SUPG matrix and multigrid preconditioner in Example 6.4

the nullspace method in Section 2.2.2 and we find that $\|W - \mathcal{W}\|_I = 0.0092$ for the symmetric bilinear forms in Figure 6.11. By ordering the eigenvectors of P obtained by Matlab by hand, and then scaling them by 1 or -1 so that the resulting eigenvector matrix of P most closely matched the eigenvector matrix of B obtained by Matlab, we found that $\|\mathcal{Z} - Z\|_I = 0.1231$. This indicates that the eigenvectors of P are indeed good approximations to those of B. Figure 6.10 shows that the eigenvalues of P and B are close and so, in light of Theorem 6.6, it is not surprising that the pseudospectra of these matrices look so similar. It is possible that the difference between $\mathcal{Z}$ and Z could be made smaller by better scaling or ordering.

Example 6.5. Incomplete LU (ILU) factorizations constitute one of the most commonly used classes of general purpose preconditioners. It is known that ILU preconditioners tend to cluster eigenvalues but for nonnormal matrices clustered eigenvalues do not necessarily lead to fast GMRES convergence. We examine ILU(0) preconditioning [99], as implemented in Matlab, for the olm100 matrix from the University of Florida Sparse Matrix Collection [31].

Table 6.5 shows that the conditioning of the eigenvectors of $P^{-1}B$ is worse than the conditioning of the eigenvectors of B. Consequently, the standard eigenvalue convergence bound (3.27) is unlikely to describe accurately the fast convergence of the preconditioned system shown in Figure 6.12. The bounds relating $\mathcal{M}$-MR and I-MR convergence in Theorem 3.23 are also pessimistic, since $\kappa_I(\mathcal{M}) = 8.8 \times 10^{10}$ for the preconditioned system.

The pseudospectral plots in Figure 6.13 confirm that P in particular is nonnormal so that here, in contrast to Example 6.1, we apply a nonnormal preconditioner to a fairly normal coefficient matrix. However, there are certainly similarities in the



(a) I-MR (b) M-MR/$\mathcal{M}$-MR

Figure 6.12: Relative preconditioned residuals for (left column) I-MR and (right column) M-MR convergence (for unpreconditioned systems) and $\mathcal{M}$-MR convergence (for preconditioned systems) for olm100 and the ILU(0) preconditioner in Example 6.5.

pseudospectra of B and P, that might indicate that the eigenvectors of these matrices are nearby (see Theorem 6.6). Although the eigenvalues of $P^{-1}B$ are quite clustered, as is typical for ILU(0) preconditioning, the pseudospectra indicate that even this preconditioned coefficient matrix is somewhat nonnormal.

The symmetric bilinear forms with respect to which B and P are self-adjoint can, as in Example 6.4, be obtained by the nullspace method described in Section 2.2.2. We find that $\|W - \mathcal{W}\|_I = 0.012$ for the symmetric bilinear forms defined by the matrices whose entries are plotted in Figure 6.14.

Example 6.6. A third example comes from the spectral method applied to

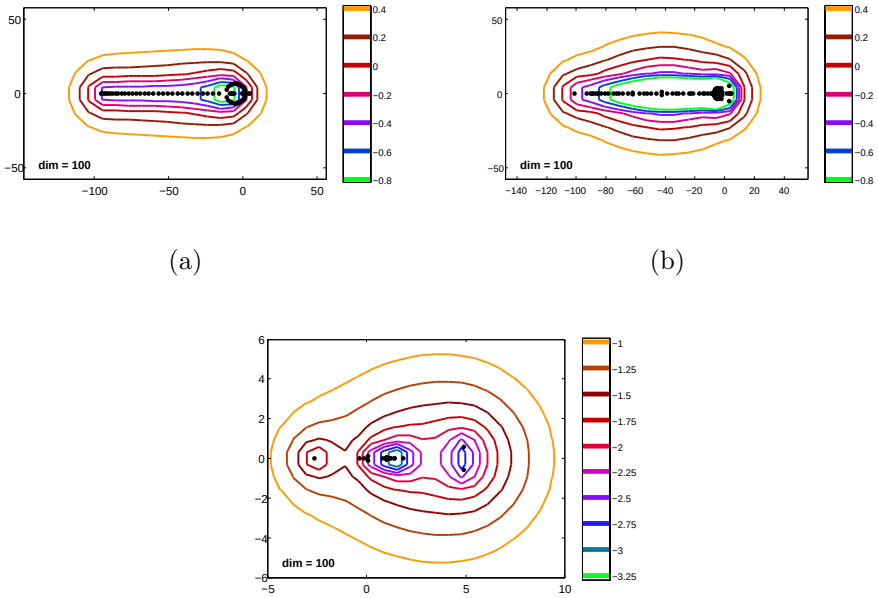
$$u' - u = 0 \quad \text{for} \quad x \in (0, 3)$$

with

$$u(0) = 1.$$

The discretization is performed by Chebfun [146], which creates a spectral matrix over $n-1$ Chebyshev points of the first kind $\{\xi_i\}_{i=1,\ldots,n-1}$. The boundary condition is enforced by appending $[\,1,\,0,\,\ldots,\,0\,]$ to the bottom of the spectral matrix (see Figure 6.15) and replacing the bottom row of the right-hand side vector by 1. The solution vector u is an approximation of u at Chebyshev points of the second kind, $\{x_i\}_{i=1,\ldots,n}$.

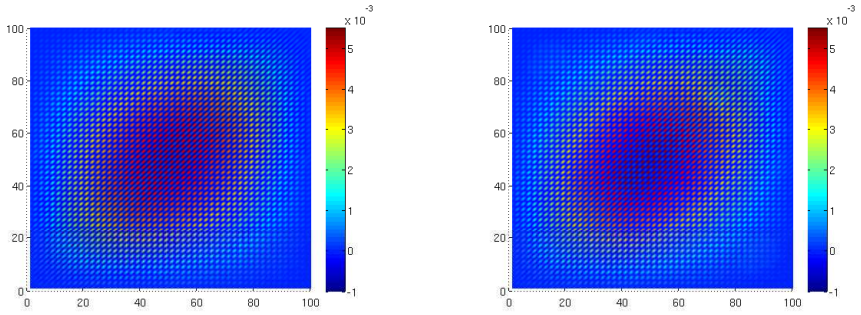
The preconditioner is formed from the finite difference matrix $F^{(1)}$ described by Funaro [55], that approximates the derivative at $n-1$ Chebyshev points of the first



(a) (b)

(c)

Figure 6.13: Pseudospectra of (a) B, (b) P and (c) $P^{-1}B$ for olm100 and the ILU(0) preconditioner in Example 6.5. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-0.8}, 10^{0.4}]$ for P and B and for selected ϵ in $[10^{-3.25}, 10^{-1}]$ for $P^{-1}B$.



(a) (b)

Figure 6.14: Comparison of (a) W and (b) $\mathcal{W}$ for olm100 and the ILU(0) preconditioner of Example 6.5.

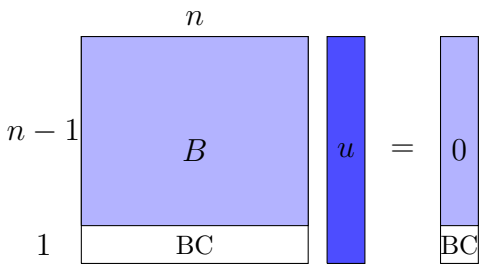


Figure 6.15: Structure of the spectral collocation matrix in Example 6.6. Light blue regions indicate where Chebyshev points of the first kind are used. Dark blue regions show where Chebyshev points of the second kind are used.

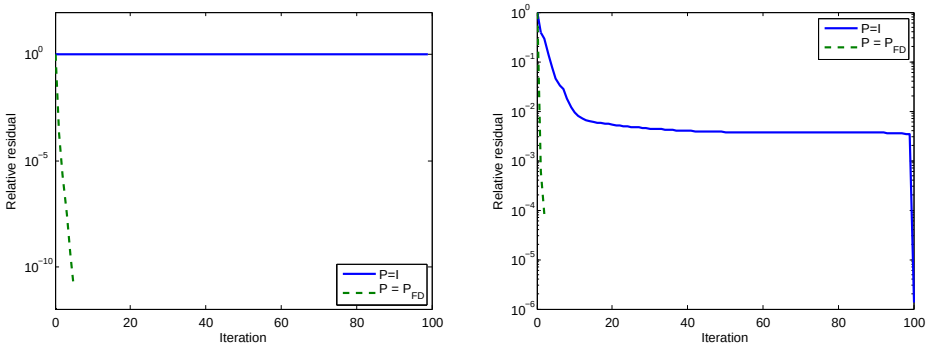
kind, and a matrix $F^{(0)}$, with entries [61]

$$f_{i,j}^{(0)} = \begin{cases} \frac{x_i - \xi_i}{x_i - x_{i-1}} & i = j \\ \frac{\xi_i - x_{i-1}}{x_i - x_{i-1}} & i + 1 = j \\ 0 & \text{otherwise} \end{cases}$$

that approximates u at Chebyshev points of the first kind. Specifically, $\widehat{P} = F^{(0)} + F^{(1)}$ and the preconditioner P is obtained from $\widehat{P}$ by replacing the bottom row by the boundary condition, as for the coefficient matrix B. The method is described in detail in [61].

The finite difference preconditioner is very effective, as is clear from Figure 6.16. However, as in previous examples the ill-conditioning of the eigenvectors of $P^{-1}B$ and P, evident from Table 6.5 make the standard eigenvalue bound (3.27) and the bound in Theorem 3.23 are poor predictors of convergence. Interestingly, none of P, B or $P^{-1}B$ is particularly close to normal. Indeed, the nonnormality of $P^{-1}B$ appears to manifest itself in the GMRES convergence for the preconditioned system, which shows a slower second phase of convergence. Despite this, application of the preconditioner clearly improves convergence dramatically.

The finite difference approximation, though low-order, appears to be an accurate enough representation of the initial value problem to be an effective preconditioner of the spectral discretization. We see from Figure 6.18 that P and B are also self-adjoint with respect to nearby symmetric bilinear forms. These are computed by the nullspace method in Section 2.2.2 and their difference is $\|W - \mathcal{W}\|_I = 0.019$.

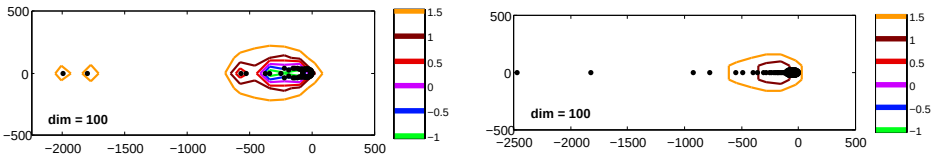


(a) I-MR (b) M-MR

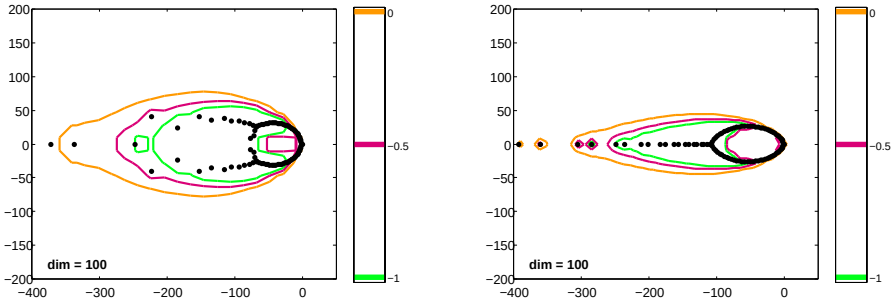
Figure 6.16: (a) I-MR and (b) M-MR convergence (for unpreconditioned systems) and $\mathcal{M}$-MR convergence (for preconditioned systems) for the exponential equation spectral discretization preconditioned by a finite difference preconditioner in Example 6.6.

6.4 Conclusions

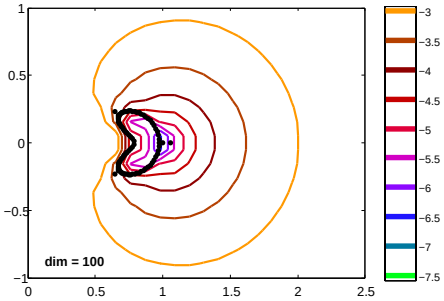
In this chapter we have found links between the nearness of the eigenvectors of P to those of B, the nearness of the symmetric bilinear forms with respect to which P and B are self-adjoint and the nearness of the pseudospectra of P and B. It was shown that an effective preconditioner for a Toeplitz matrix, and powerful preconditioners for two different two-dimensional discretizations of a convection-diffusion equation are self-adjoint with respect to the same symmetric bilinear form as the coefficient matrices. Other good preconditioners were shown to be self-adjoint with respect to nearby symmetric bilinear forms, indicating that it may be enough for these symmetric bilinear forms to be close.



(a) (b)

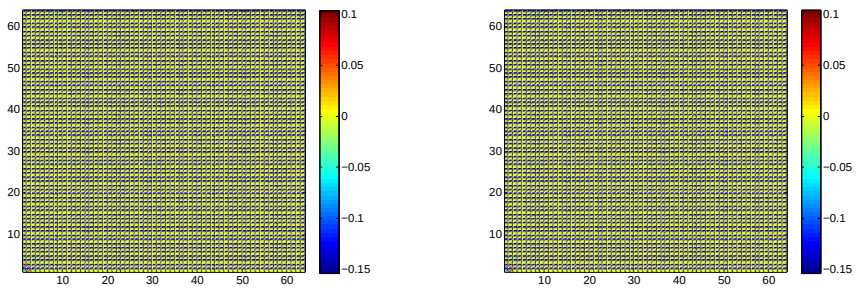


(c) (d)



(e)

Figure 6.17: Pseudospectral plots of (a) and (c) B, (b) and (d) P and (e) $P^{-1}B$ for the exponential equation spectral discretization preconditioned with a finite difference preconditioner in Example 6.6. On each pseudospectral plot the black dots represent the eigenvalues of the matrix. Each contour shows the boundary of an ϵ-pseudospectrum for selected ϵ in $[10^{-1}, 10^{1.5}]$ for P and B and for selected ϵ in $[10^{-7.5}, 10^{-3}]$ for $P^{-1}B$.



(a) (b)

Figure 6.18: Comparison of (a) W and (b) $\mathcal{W}$ for the exponential equation spectral discretization preconditioned with a finite difference preconditioner in Example 6.6.

Chapter 7

Conclusions

This thesis aimed to investigate ways in which nonstandard inner products can be used to develop new and examine existing effective preconditioned Krylov subspace methods. We focused on two subsets of real, nonsingular, diagonalizable matrices: saddle point matrices and general nonsymmetric (nonnormal) matrices.

For symmetric saddle point matrices, preconditioned Krylov subspace methods in nonstandard inner products are already used in practice, as discussed in Chapter 4. However, some of the most effective preconditioners, such as the Bramble-Pasciak and Schöberl-Zulehner preconditioners, require scaling before they may be robustly applied as part of a $\mathcal{W}$-PCG (or $\mathcal{W}$-PMINRES) method. Depending on the saddle point matrix of interest and the approximations A_0 and S_0 this scaling may be costly. Block diagonal preconditioners are robust and can be very effective when the blocks are good approximations of A and the (negative) Schur complement. However, we observed that they may perform poorly otherwise. The recently introduced Bramble-Pasciak$^+$ (BP$^+$) preconditioner is also robust and can be more effective than block diagonal preconditioners. Our contribution is to use our characterization of symmetric bilinear forms to modify the Schöberl-Zulehner preconditioner similarly. The result is the robust Schöberl-Zulehner$^+$ (SZ$^+$) preconditioner. We derived bounds on the eigenvalues of the SZ$^+$ preconditioned saddle point matrix that can be used to gauge, or improve, convergence of a Krylov subspace method in a nonstandard inner product. Importantly, however, even in the absence of such information we can guarantee convergence under the mild assumptions of positive definiteness of A, A_0 and S_0 and semi-positive definiteness of C. We observed that the SZ$^+$ preconditioner was as effective as, or more effective than, the BP$^+$ preconditioner on a set of problems from optimization applications.

When the saddle point matrix is nonsymmetric, common practice is to apply a preconditioner and solve the resulting system by a nonsymmetric Krylov subspace method. However, no method of this sort combines both short-term recurrences and optimality (with respect to an inner product that is independent of the initial iterate). We derived conditions for certain block-diagonal and block-triangular preconditioned saddle point matrices to be self-adjoint with respect to a symmetric bilinear form[1] or inner product. When this symmetric bilinear form is positive definite, a conjugate gradient or MINRES algorithm can be applied that *does* posses an optimality property and three-term recurrences. Whether this inner product is practical to use depends on the application; we presented one, simple example for which the inner product was block diagonal and, therefore, cheap to apply.

The focus of Chapter 5 was on combining preconditioners for symmetric saddle point matrices that are effective in their own right to achieve superior methods. We proved that certain Krzyżanowski preconditioners can be combined and used this result to create specific preconditioners. These included a different formulation of the Bramble-Pasciak-Schöberl-Zulehner (BP-SZ) combination preconditioner than [135], for which convergence could be made significantly faster than that for the BP or SZ preconditioners individually. Combination of the Bramble-Pasciak and block diagonal preconditioners gave rise to the Bramble-Pasciak-block diagonal (BP-BDW) combination preconditioner that was more effective than either the BP or BD preconditioners, although it was less efficient than the BP-SZ combination preconditioner.

The preconditioner that we feel best highlights the power of combination preconditioning, however, is the Bramble-Pasciak$^+$-block diagonal (BP$^+$-BDW) combination preconditioner, which combines two preconditioners for which the preconditioned coefficient matrices are self-adjoint but indefinite with respect to an inner product. The resulting combination preconditioner is *positive definite* and self-adjoint with respect to an inner product, for certain parameter choices that we fully characterized. We derived bounds on the eigenvalues of the BP$^+$-BDW-preconditioned saddle point matrix and observed fast convergence of $\mathcal{W}$-PCG and $\mathcal{W}$-PMINRES when this preconditioned matrix was self-adjoint and positive definite with respect to the inner product.

Nonstandard inner products and symmetric bilinear forms also proved informative for nonsymmetric matrices. In Chapter 3 we used the framework of nonstandard inner products to obtain an iteration-dependent convergence bound that can be a

[1] In this case, one may apply SQMR to the preconditioned system (see page 36).

much better predictor of convergence than the standard eigenvalue-based bound for (I-)GMRES. We also observed that this bound was capable of detecting superlinear convergence for a finite difference discretization of a convection-diffusion operator.

In Chapter 6 we first considered preconditioners P whose eigenvectors were nearby to those of the coefficient matrix B. A condition for the pseudospectra of P and B to be identical was discovered and the pseudospectra of these matrices when this condition was not met exactly was characterized. We next showed that effective preconditioners for persymmetric matrices, i.e., matrices that are self-adjoint with respect to the symmetric bilinear form defined by the reverse identity matrix $\langle \cdot, \cdot \rangle_{\widehat{Y}}$, are also $\widehat{Y}$-self-adjoint. A matrix formed from a Kronecker sum of persymmetric matrices and an effective preconditioner were also found to be self-adjoint with respect to the same symmetric bilinear form (that was defined by a Kronecker sum of reverse identity matrices). Good preconditioners for certain other nonsymmetric matrices were found to be self-adjoint with respect to symmetric bilinear forms that were nearby to those with respect to which the coefficient matrix was self-adjoint.

This thesis shows that examining preconditioned Krylov subspace methods in nonstandard inner products can aid the development of effective preconditioned Krylov subspace methods. Further discoveries using this approach may certainly be possible. The use of nonstandard inner products for preconditioned nonsymmetric saddle point matrices has received relatively little attention and this might make it a fruitful area for research. Perhaps the characterization of Lemma 4.4 in Chapter 4 provides the means of developing such preconditioned Krylov methods. In contrast to symmetric saddle point matrices, however, where symmetry can be used to devise rather general preconditioners and inner products, for nonsymmetric matrices this is a more difficult task and a problem-dependent approach might be necessary.

New combination preconditioners, either for symmetric or nonsymmetric matrices, could also be created. In particular it would be interesting to explore whether other combination preconditioners, for which positive definiteness can be achieved, can be constructed from preconditioners for which the preconditioned saddle point matrix is indefinite with respect to an inner product.

One area on which we have not focused is the prediction of α and β values that lead to fast convergence. It would be advantageous to develop methods or heuristics for choosing these parameter values. Additionally, although it appears difficult, a theoretical explanation of the performance of the combination preconditioned saddle point systems when the associated symmetric bilinear form is indefinite would be

valuable. It would also be interesting to apply the combination preconditioners, and the SZ^+ preconditioner, to other saddle point problems, such as those from PDE-constrained optimization—an area that has received much attention of late.

The work on nonsymmetric matrices could also be continued. It would be beneficial to be able to find or approximate the singular values a priori in our iteration-dependent I-GMRES bound. Additional links between the symmetric bilinear forms with respect to which the preconditioner and coefficient matrix are self-adjoint and fast I-GMRES convergence would also be of use. In particular, it would be beneficial to develop criteria, that can be cheaply checked, that indicate when a certain preconditioner will work well. An extension in a different direction could use the framework developed here to design new preconditioners for nonsymmetric matrices. This would require some further development of our results but would certainly be constructive.

Appendix A

Linear algebra results

This appendix contains selected definitions and results from linear algebra.

A.1 The real Jordan form

Let the Jordan canonical form of a matrix $B \in \mathbb{R}^{n\times n}$ be $B = ZJZ^{-1}$, $J, Z \in \mathbb{C}^{n\times n}$, where

$$J = \text{diag}(J(\lambda_1), \ldots, J(\lambda_j), J(\lambda_{j+1}), J(\overline{\lambda}_{j+1}), \ldots, J(\lambda_k), J(\overline{\lambda}_k)),$$

and for any scalar $\lambda \in \mathbb{C}$,

$$J(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}.$$

The corresponding real Jordan form is defined as follows.

Definition A.1. Real Jordan form [57, Appendix A.2] Let $\lambda_1, \ldots, \lambda_j$ be the real eigenvalues of B with eigenvectors $z_1, \ldots, z_j$ and let $\lambda_{j+1} = \mu_j + i\nu_j, \overline{\lambda}_{j+1} = \mu_{j+1} - i\nu_{j+1}, \ldots, \lambda_k = \mu_k + i\nu_k, \overline{\lambda}_k = \mu_k - i\nu_k$ be the $2(k-j)$ complex eigenvalues with eigenvectors $z_{j+1}, \overline{z}_{j+1}, \ldots, z_k, \overline{z}_k$. Then if $B = ZJZ^{-1}$ is the Jordan canonical form of B, where

$$Z = \left[z_1, \ldots, z_j, z_{j+1}, \overline{z}_{j+1}, \ldots, z_k, \overline{z}_k\right]$$

and

the corresponding real Jordan form of B is given by $B = R\mathcal{J}R^{-1}$ where $R \in \mathbb{R}^{n\times n}$ and $\mathcal{J} \in \mathbb{R}^{n\times n}$, with

$$R = \left[z_1, \ldots, z_j, \Re(z_{j+1}), \Im(z_{j+1}), \ldots, \Re(z_k), \Im(z_k)\right]$$

and

$$\mathcal{J} = \operatorname{diag}(J(\lambda_1), \ldots, J(\lambda_j), \mathcal{J}(\mu_{j+1} \pm i\nu_{j+1}), \ldots, \mathcal{J}(\mu_k \pm i\nu_k)).$$

where, if $J(\lambda_i)$ and $J(\overline{\lambda}_i)$ each have dimension m, $\mathcal{J}(\mu_{j+1} \pm i\nu_{j+1})$ has dimension $2m$ and

$$\mathcal{J}(\mu \pm i\nu) = \begin{bmatrix} \mu & \nu & 1 & 0 & & & \\ -\nu & \mu & 0 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \mu & \nu & 1 & 0 \\ & & & -\nu & \mu & 0 & 1 \\ & & & & & \mu & \nu \\ & & & & & -\nu & \mu \end{bmatrix}.$$

A.2 Kronecker products and vectorization

The (left) **Kronecker product** of $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

The Kronecker product satisfies that

$$\begin{aligned} &A \otimes (B + C) = A \otimes B + A \otimes C \\ &(A + B) \otimes C = A \otimes C + B \otimes C \\ &(kA) \otimes B = A \otimes (kB) = k(A \otimes B) \\ &(A \otimes B) \otimes C = A \otimes (B \otimes C) \\ &(A \otimes B)^T = A^T \otimes B^T \\ &\operatorname{rank}(A \otimes B) = \operatorname{rank} A \operatorname{rank} B, \end{aligned}$$

for any matrices A, B and C and scalar k. Provided A and B are invertible,

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Also, when the A and C are conformable, and B and D are conformable, the **mixed product property** states that

$$(A \otimes B)(C \otimes D) = AC \otimes BD.$$

The **vectorization** of a matrix $B \in \mathbb{R}^{m\times n}$ is the vector $\text{vec}(B) \in \mathbb{R}^{mn}$ obtained by stacking the columns of B, i.e.,

$$vec(B) = \left[b_{11}, \ldots, b_{m1}, b_{12}, \ldots, b_{m2}, \ldots, b_{1n}, \ldots, b_{mn}\right]^T.$$

When $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, the eigenvalues of $A \otimes B$ depend on the eigenvalues of A and B.

Lemma A.2. *Let $A \in \mathbb{R}^{n\times n}$ have eigenvalues λ_i, $i = 1, \ldots, n$ and let $B \in \mathbb{R}^{m\times m}$ have eigenvalues μ_i, $i = 1 \ldots, m$. Then $A \otimes B$ has eigenvalues*

$$\nu_{i,j} = \lambda_i \mu_j, i = 1 \ldots, n, j = 1, \ldots, m.$$

A.3 Pseudospectra

Pseudospectra are an alternative tool to eigenvalues for analyzing nonnormal matrices. Pseudospectra in the I-norm can be computed using Eigtool [154], as can pseudospectra in any other norm after some preprocessing [145, Chapter 39]. Three equivalent definitions of the **ϵ-pseudospectrum** of a matrix B, $\Lambda_\epsilon(B)$, are:

Definition A.3. [145, Chapter 2] The ϵ-pseudospectrum of B is the set of $z \in \mathbb{C}$ such that

$$\|(zI - B)^{-1}\| > \epsilon^{-1}. \tag{A.1}$$

Definition A.4. [145, Chapter 2] The ϵ-pseudospectrum of B is the set of $z \in \mathbb{C}$ such that

$$z \in \Lambda(B + E) \tag{A.2}$$

for some $E \in \mathbb{C}^{n\times n}$ with $\|E\| < \epsilon$, where $\Lambda(B + E)$ is the spectrum of $B + E$.

Definition A.5. [145, Chapter 2] The ϵ-pseudospectrum of B is the set of $z \in \mathbb{C}$ such that

$$\|(z - B)v\| < \epsilon \tag{A.3}$$

for some $v \in \mathbb{C}^n$ with $\|v\| = 1$.

When the pseudospectra are defined with respect to the I-norm, a fourth definition in [145, Chapter 2] is

Definition A.6. For $\|\cdot\| = \|\cdot\|_I$, $\Lambda_\epsilon(B)$ is the set of $z \in \mathbb{C}$ such that

$$\sigma_{min}(z - B) \leq \epsilon,$$

where $\sigma_{min}(C)$ is the smallest singular value of C.

More generally, using the W-singular values defined by Definition 2.13 and Lemma 2.16, we have that

Definition A.7. The ϵ-pseudospectrum of B with respect to $\|\cdot\|_W$ is the set of $z \in \mathbb{C}$ such that

$$\sigma^W_{min}(z - B) \leq \epsilon,$$

where $\sigma^W_{min}(C)$ is the smallest W-singular value of C.

Let the **open ϵ-ball** be denoted

$$\triangle_\epsilon = \{z \in \mathbb{C} : \ |z| < \epsilon\}. \tag{A.4}$$

The following lemma shows that when B is normal its pseudospectra are nicely characterized by $\triangle_\epsilon$.

Lemma A.8. (from Theorem 2.2 in [145]) *If $B \in \mathbb{R}^{n\times n}$ is I-normal and $\|\cdot\| = \|\cdot\|_I$, then*

$$\Lambda^I_\epsilon(B) = \Lambda(B) + \triangle_\epsilon \ \ \forall \epsilon > 0. \tag{A.5}$$

Corollary 2.28 shows that whenever a real matrix is self-adjoint it is normal with respect to a related inner product. Lemma A.8 can be generalized to this case.

Lemma A.9. *Let B be diagonalizable with diagonalization $B = Z\Lambda Z^{-1}$ and let $M = Z^{-*}Z^{-1}$. Then the ϵ-pseudospectrum of B with respect to $\|\cdot\|_M$, $\Lambda^M_\epsilon(B)$, is*

$$\Lambda^M_\epsilon(B) = \Lambda(B) + \triangle_\epsilon \ \ \forall \epsilon > 0. \tag{A.6}$$

Proof. By Definition A.3, $\Lambda^M_\epsilon(B)$ is the set of $z \in \mathbb{C}$ for which

$$\|(z - B)^{-1}\|_M > \epsilon^{-1}.$$

Now, by (2.2),

$$\|(z - B)^{-1}\|_M = \|Z^{-1}(z - B)^{-1}Z\|_I = \|(z - \Lambda)^{-1}\|_I$$

from which the result follows. □

A.4 Eigenvalues and singular values

The **Courant-Fischer theorem** for symmetric matrices shows that stationary points of the Rayleigh quotient of a symmetric matrix are its eigenvalues.

Theorem A.10. [118, Theorem 1.11] *If $A \in \mathbb{R}^{n\times n}$ is symmetric then*

$$\lambda_k(A) = \min_{\dim(S)=n-k+1} \max_{x\in S, x\neq 0} \frac{x^T A x}{x^T x}$$

and

$$\lambda_k(A) = \max_{\dim(S)=k} \min_{x\in S, x\neq 0} \frac{x^T A x}{x^T x}.$$

A consequence is that

$$\lambda_{min}(A) \leq \frac{x^T A x}{x^T x} \leq \lambda_{max}(A). \tag{A.7}$$

From the Courant-Fischer theorem it is possible to obtain Cauchy's interlacing property (see Theorem 4.3.15 in [75]), from which the **Poincarè separation theorem** follows.

Corollary A.11. [75, Corollary 4.3.16] *Let $A \in \mathbb{R}^{n\times n}$ be Hermitian, let r be a given integer with $1 \leq r \leq n$, and let $u_1, \ldots, u_r \in \mathbb{R}^n$ be r given orthonormal vectors. Let $H_r = U_r^T A U_r$, where $U_r = [\,u_1, \ldots, u_r\,]$. If the eigenvalues of A and H_r are arranged in increasing order, we have*

$$\lambda_k(A) \leq \lambda_k(H_r) \leq \lambda_{k+n-r}(A), \quad k = 1, 2, \ldots, r.$$

Since the singular values of a matrix $B \in \mathbb{R}^{n\times n}$ are the square roots of its eigenvalues, i.e., $\sigma_i(B) = \sqrt{\lambda_i(B^T B)}$, a consequence of the Poincaré separation theorem is the following.

Corollary A.12. *Let $B \in \mathbb{R}^{n\times n}$ and let U_r be as defined in Corollary* A.11*. Then*

$$\sigma_k(B) \leq \sigma_k(BU_r) \leq \sigma_{k+n-r}(B), \quad k = 1, 2, \ldots, r.$$

When considering positive definiteness of symmetric matrices, the **inertia** of a matrix becomes important.

Definition A.13. [75, Definition 4.5.6] Let $B \in \mathbb{R}^{n\times n}$ be a symmetric matrix. The inertia of B is the ordered triple

$$i(B) = (i_+(B), i_-(B), i_0(B)),$$

where $i_+(B)$ is the number of positive eigenvalues of B, $i_-(B)$ is the number of negative eigenvalues of B and $i_0(B)$ is the number of zero eigenvalues of B, all counting multiplicity.

Sylvester's law of inertia is useful for determining the inertia of a symmetric matrix from that of another.

Theorem A.14. [75, Theorem 4.5.8] *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices. There is a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that $A = SBS^*$ if and only if A and B have the same inertia.*

A.5 The Löwner ordering

If $A, H \in \mathbb{R}^{n\times n}$ are symmetric matrices use the notation $A > H$ to denote that $A - H$ is symmetric positive definite. Similarly, $A \geq H$ means that $A - H$ is symmetric positive semidefinite.

Lemma A.15. [75, Observation 7.7.2] *If $A, H \in \mathbb{R}^{n\times n}$ are symmetric, then*

$$A \geq H \quad \textit{implies} \quad C^T AC \geq C^T HC$$

for all $C \in \mathbb{R}^{n\times m}$; if $m \leq n$, we also have that

$$A\, H \quad \textit{implies} \quad C^T AC > C^T HT$$

whenever $C \in \mathbb{R}^{n\times m}$ has rand m.

Corollary A.16. [75, Corollary 7.7.4] *If $A, H \in \mathbb{R}^{n\times n}$ are symmetric positive definite, then*

- *$A \geq H$ if and only if $H^{-1} \geq A^{-1}$;*
- *If $A \geq H$, then* detA $\geq$ detH *and* trA $\geq$ trH*; and*
- *More generally, if $A \geq H$, then $\lambda_k(A) \geq \lambda_k(H)$ for all $k = 1, 2, \ldots, n$ if the respective eigenvalues of A and H are arranged in the same (increasing or decreasing) order.*

A.6 Generalized inverses

Definition A.17. The **Moore-Penrose pseudoinverse** $B^\dagger$ of $B \in \mathbb{C}^{n\times n}$ satisfies

1. $BB^\dagger B = B$;
2. $B^\dagger BB^\dagger = B^\dagger$;
3. $B^\dagger B$ is symmetric;
4. $BB^\dagger$ is symmetric.

Definition A.18. A **(1)-inverse** or generalized inverse satisfies condition (1) in Definition A.17 above.

Definition A.19. A **(1,2)-inverse** or reflexive generalized inverse satisfies conditions (1) and (2) in Definition A.17 above.

Appendix B

Eigenvalue bounds for the BP$^+$ preconditioner

The following corollary provides bounds on the eigenvalues of the BP$^+$ preconditioned saddle point matrix.

Corollary B.1. *Let*

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$$

be left preconditioned by the BP$^+$ preconditioner

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ -B & S_0 \end{bmatrix}$$

and let

$$0 < \delta \leq \frac{u^T A_0 u}{u^T A u} \leq \Delta \tag{B.1}$$

and

$$0 \leq \frac{u^T B^T S_0^{-1} B u}{u^T A u} \leq \Phi. \tag{B.2}$$

for $\delta, \Delta, \Phi \in \mathbb{R}$.

Positive eigenvalues $\lambda+$ of $\mathcal{P}^{-1}\mathcal{A}$ satisfy that

$$0 < \frac{1}{\Delta} \leq \lambda^+ \leq \frac{1}{\delta}, \tag{B.3}$$

$$0 < \frac{1}{\Delta} \leq \lambda^+ \leq \frac{(1+\Phi) + \sqrt{(1+\Phi)^2 + 4\delta\Phi}}{2\delta} \tag{B.4}$$

while negative eigenvalues λ^- satisfy

$$\frac{(1+\Phi) - \sqrt{(1+\Phi)^2 + 4\delta\Phi}}{2\delta} \leq \lambda^- < 0, \tag{B.5}$$

Proof. The Bramble-Pasciak$^+$ preconditioner is a Krzyżanowski preconditioner for which $c = -1$ and $d = 0$. It follows from Theorem 4.2 that $\lambda \neq 0, -1$ and that $u \neq 0$. Additionally, when $Bu = 0$ Theorem 4.2 states that

$$\lambda = \frac{u^T A u}{u^T A_0 u}$$

by (4.15) and, using (B.1), we obtain that

$$\frac{1}{\Delta} \leq \lambda \leq \frac{1}{\delta}.$$

Now assume that $Bu \neq 0$. If

$$\psi = \frac{u^T A_0 u}{u^T A u} \quad \text{and} \quad \omega = \frac{u^T B^T S_0^{-1} B u}{u^T A u}.$$

then the roots (4.21) in Theorem 4.2 are

$$\lambda = \frac{(1+\omega) \pm \sqrt{(1+\omega)^2 + 4\psi\omega}}{2\psi}, \tag{B.6}$$

which has one real positive root and one real negative root. This is to be expected since $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint but indefinite with respect to an inner product.

The larger root is given by

$$\lambda = \frac{(1+\omega) + \sqrt{(1+\omega)^2 + 4\psi\omega}}{2\psi}, \tag{B.7}$$

which it is possible to bound using (B.1) and (B.2). Equation (B.7) is minimized when $\omega = 0$ and $\psi = \Delta$ and so

$$\lambda \geq \frac{1}{\Delta}.$$

Conversely, it is maximized when $\omega = \Phi$ and $\psi = \delta$ which, in combination with the lower bound gives that

$$\frac{1}{\Delta} \leq \lambda \leq \frac{(1+\Phi) + \sqrt{(1+\Phi)^2 + 4\delta\Phi}}{2\delta}.$$

On the other hand, the smaller root of (B.6) is

$$\lambda = \frac{(1+\omega) - \sqrt{(1+\omega)^2 + 4\psi\omega}}{2\psi}, \tag{B.8}$$

which is also maximized when $\omega = 0$ and $\psi = \Delta$, so that $\lambda \leq 0$. However, we concluded previously that $\lambda \neq 0$ and so $\lambda < 0$ is the upper bound. Equation (B.8) is minimized by choosing $\psi = \delta$ and $\omega = \Phi$ which gives that

$$\lambda = \frac{(1+\Phi) - \sqrt{(1+\Phi)^2 + 4\Phi\delta}}{2\delta}.$$

□

Appendix C

BP$^+$-SZ$^+$ combination preconditioner

A combination not considered in Chapter 5 is that of the BP$^+$ and SZ$^+$ preconditioners. Since $c_1 = c_2 = -1$, Theorem 5.3 can be used to obtain a combination which can outperform each of the BP$^+$ and SZ$^+$ preconditioners. However, the BP-SZ, BP$^+$-BDW and BP-BDW combination preconditioners in Chapter 5 perform better again. Additionally, the combination preconditioned coefficient matrices in Chapter 5 can be made positive definite by choosing the combination preconditioner parameters α and β judiciously and, if necessary, scaling A_0 and S_0. This is not possible for the BP$^+$-SZ$^+$ combination preconditioner since $\mathcal{P}^{-1}\mathcal{A}$ is always $\mathcal{W}$-indefinite. For completeness, we describe the preconditioner here.

The combination of the BP$^+$ preconditioner (4.6) and SZ$^+$ preconditioner (4.44) by Theorem 5.3 is

$$\mathcal{P} = \begin{bmatrix} \frac{1}{\alpha+\beta}A_0 & \frac{-\beta}{\alpha+\beta}B^T \\ -\frac{1}{\alpha+\beta}B & \frac{\beta}{\alpha+\beta}BA_0^{-1}B^T + S_0 \end{bmatrix} \tag{C.1}$$

and

$$\mathcal{W} = \begin{bmatrix} A + A_0 & 0 \\ 0 & (\alpha+\beta)S_0 + \beta BA_0^{-1}B^T \end{bmatrix}. \tag{C.2}$$

The following theorem shows that $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ can be made an inner product but it is one with respect to which $\mathcal{P}^{-1}\mathcal{A}$ is never positive definite.

Theorem C.1. *Let*

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

be left preconditioned by the BP$^+$-SZ$^+$ combination preconditioner (C.1) *so that* $\mathcal{P}^{-1}\mathcal{A}$ *is self-adjoint with respect to* $\langle \cdot, \cdot \rangle_{\mathcal{W}}$*, where* $\mathcal{W}$ *is given by* (C.2).

When

I. $\alpha+\beta>0$ *and* $\beta>0$, $\mathcal{W}$ *defines an inner product;*

II. $\alpha+\beta<0$ *and* $\beta>0$, $\mathcal{W}$ *defines an inner product when*

$$S_0 < -\frac{\beta}{\alpha+\beta} B A_0^{-1} B^T;$$

III. $\alpha+\beta<0$ *and* $\beta<0$ $\mathcal{W}$ *does not define an inner product;*

IV. $\alpha+\beta>0$ *and* $\beta<0$, $\mathcal{W}$ *defines an inner product when*

$$S_0 > -\frac{\beta}{\alpha+\beta} B A_0^{-1} B^T.$$

When $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ *is an inner product, it is one with respect to which the preconditioned saddle point matrix* $\mathcal{P}^{-1}\mathcal{A}$ *is not positive definite.*

Proof. The preconditioned coefficient matrix is given by

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} (\alpha+\beta)A_0^{-1}A + \beta A_0^{-1}B^T S_0^{-1} B A_0^{-1}(A+A_0) & VB^T \\ S_0^{-1}BA_0^{-1} + S_0^{-1}B & S_0^{-1}BA_0^{-1}B^T \end{bmatrix},$$

where

$$V = (\alpha+\beta)A_0^{-1} + \beta A_0^{-1}B^T S_0^{-1} B A_0^{-1}. \tag{C.3}$$

By Lemma 2.31, the preconditioned saddle point matrix $\mathcal{P}^{-1}\mathcal{A}$ is $\mathcal{W}$-positive definite if and only if $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}>0$, where

$$\begin{aligned} \mathcal{W}\mathcal{P}^{-1}\mathcal{A} &= \begin{bmatrix} M & (A+A_0)VB^T \\ BV(A+A_0) & BVB^T \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ BV(A+A_0)M^{-1} & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} I & M^{-1}(A+A_0)VB^T \\ 0 & I \end{bmatrix}, \end{aligned} \tag{C.4}$$

with

$$\begin{aligned} M &= (\alpha+\beta)(A+A_0)A_0^{-1}A + \beta(A+A_0)A_0^{-1}B^T S_0^{-1} B A_0^{-1}(A+A_0) \\ &= (\alpha+\beta)A(A^{-1}+A_0^{-1})A + \beta(A+A_0)A_0^{-1}B^T S_0^{-1} B A_0^{-1}(A+A_0) \end{aligned} \tag{C.5}$$

and

$$T = BV[V^{-1} - (A+A_0)M^{-1}(A+A_0)]VB^T. \tag{C.6}$$

The congruence transform (C.4) and Sylvester's law of inertia (in Appendix A.4) show that the inertia of $\mathcal{W}\mathcal{P}^{-1}\mathcal{A}$ is determined by the eigenvalues of M and T which, in

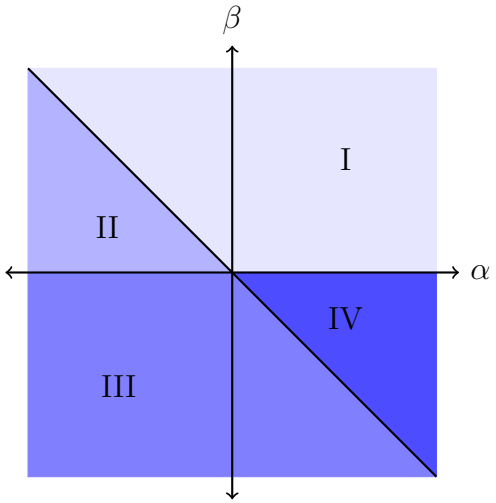


Figure C.1: Regions (distinguished by different colours) that are considered in Theorem C.1 for the BP$^+$-SZ$^+$ combination preconditioner.

turn, depend on α and β. We consider separately each of the four cases in Figure C.1.

I. Case I corresponds to taking $\alpha + \beta > 0$ with $\beta > 0$. By observation, $\mathcal{W}$, V and M are positive definite. However, T in (C.6) is negative definite. To show this, we first observe that $T < 0$ if and only if

$$V^{-1} < (A + A_0)M^{-1}(A + A_0)$$

or, since V and M are positive definite,

$$V > (A + A_0)^{-1}M(A + A_0)^{-1}.$$

It follows from substituting for V and M using (C.3) and (C.5), that T is negative definite if and only if

$$(\alpha + \beta)A_0^{-1} > (\alpha + \beta)A_0^{-1}A(A + A_0)^{-1}.$$

Premultiplying and postmultiplying both sides of this inequality by A_0 and dividing by the positive scalar $\alpha + \beta$ gives the equivalent condition

$$A_0 > A(A + A_0^{-1})^{-1}A_0 = (A_0^{-1} + A^{-1})^{-1}$$

which implies that

$$A_0^{-1} < A_0^{-1} + A^{-1}.$$

Since $A > 0$, this last inequality is true and T is negative definite. We conclude that $\mathcal{P}^{-1}\mathcal{A}$ is indefinite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, when $\alpha + \beta > 0$ and $\beta > 0$.

II. Consider now $\alpha+\beta < 0$ and $\beta > 0$ and let γ be the positive parameter $\gamma = -\frac{\beta}{\alpha+\beta}$. For these values of α and β, $(\alpha + \beta)S_0 + \beta BA_0^{-1}B^T > 0$ when $\gamma BA_0^{-1}B^T > S_0$.

The matrix M given by (C.5) is positive definite if and only if

$$(\alpha + \beta)(A + A_0)A_0^{-1}A + \beta(A + A_0)A_0^{-1}B^T S_0^{-1}BA_0^{-1}(A + A_0) > 0$$

or, upon premultiplying by $A_0(A + A_0)$ and postmultiplying by its transpose, if and only if

$$(\alpha + \beta)A(A + A_0)^{-1}A_0 + \beta B^T S_0 B > 0$$

or, equivalently, $(A^{-1} - A_0^{-1})^{-1} < \gamma B^T S_0^{-1}B$. However, since B has a nontrivial nullspace $\gamma B^T S_0^{-1}B$ is positive semidefinite. In contrast, $(A_0^{-1} + A^{-1})^{-1}$ is positive definite. Consequently, M is not positive definite and $\mathcal{P}^{-1}\mathcal{A}$ is not positive definite with respect to $\langle\cdot,\cdot\rangle_{\mathcal{W}}$.

III. The third case we investigate is $\alpha + \beta < 0$ and $\beta < 0$. For these choices of α and β, however, $(\alpha + \beta)S_0 + \beta BA_0^{-1}B^T < 0$ and $A + A_0 > 0$. This implies that $\mathcal{W}$ is indefinite. It is straightforward to show that, additionally, M and T, given by (C.5) and (C.6) are negative definite.

IV. We consider now Case IV, for which $\alpha + \beta > 0$ and $\beta < 0$ and introduce the quantity $\gamma = -\frac{\beta}{\alpha+\beta} > 0$. Now, $\mathcal{W} > 0$ when

$$(\alpha + \beta)S_0 + \beta BA_0^{-1}B^T > 0$$

or, if λ_{min} and λ_{max} are the smallest and largest eigenvalues of $S_0^{-1}BA_0^{-1}B^T$, respectively, when

$$\frac{1}{\gamma} \geq \lambda_{max} \geq \frac{x^T BA_0^{-1}B^T x}{x^T S_0 x} \geq \lambda_{min} > 0.$$

By (5.33),

$$\frac{1}{\gamma} > \lambda_{max} \geq \frac{y^T B^T S_0^{-1}By}{y^T A_0 y} \geq 0$$

and it follows that

$$A_0^{-1} > \gamma B^T S_0^{-1}B \tag{C.7}$$

is an equivalent condition for positive definiteness of $\mathcal{W}$.

By examining (C.3) we find that $V > 0$ if and only if

$$(\alpha + \beta)A_0^{-1} > -\beta A_0^{-1}B^T S_0^{-1}BA_0^{-1}$$

or, equivalently $A_0^{-1} > \gamma B^T S_0^{-1} B$, which is precisely the same as (C.7) and so is satisfied when $\mathcal{W}$ defines an inner product.

$M > 0$ if and only if

$$(\alpha + \beta)(A + A_0)A_0^{-1}A + \beta(A + A_0)A_0^{-1}B^T S_0^{-1} B A_0^{-1}(A + A_0) > 0.$$

Premultiplying by $(A + A_0)A_0^{-1}$ and postmultiplying by $A_0^{-1}(A + A_0)$ gives the equivalent condition

$$(\alpha + \beta)A(A + A_0)^{-1}A_0 > -\beta B^T S_0^{-1} B,$$

or $(A_0^{-1} + A^{-1})^{-1} > \gamma B^T S_0^{-1} B$.

If this is positive definite then using the same arguments as for the Case I above shows that $T < 0$. Thus, when $\alpha + \beta > 0$, $\beta < 0$, $\mathcal{P}^{-1}\mathcal{A}$ is not positive definite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. □

Theorem C.1 shows that it is not possible to apply a $\mathcal{W}$-PCG method to a BP$^+$-SZ$^+$ combination preconditioned saddle point problem. This is perhaps unsurprising since neither the BP$^+$ preconditioned system nor the SZ$^+$ preconditioned system is definite with respect to its inner product.

Bibliography

[1] M. Ainsworth and S. Sherwin. Domain decomposition preconditioners for p and hp finite element approximation of Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 175:243–266, 1999.

[2] M. Arioli, V. Pták and Z. Strakoš. Krylov sequences of maximal length and convergence of GMRES. *BIT*, 38:636–643, 1998.

[3] G. Armentia, J.-M. Gracia and F. E. Velasco. Identical pseudospectra of any geometric multiplicity. *Linear Algebra Appl.*, 2011. doi: 10.1016/j.laa.2011.01.014.

[4] S. F. Ashby, P. N. Brown, M. R. Dorr and A. C. Hindmarsh. A linear algebraic analysis of diffusion synthetic acceleration for the Boltzmann transport equation. *SIAM J. Numer. Anal.*, 32:128–178, 1995.

[5] S. F. Ashby, M. J. Holst, T. A. Manteuffel and P. E. Saylor. The role of the inner product in stopping criteria for conjugate gradient iterations. *BIT*, 41:26–52, 2001.

[6] S. F. Ashby, T. A. Manteuffel and P. E. Saylor. A taxonomy for conjugate gradient methods. *SIAM J. Numer. Anal.*, 27:1542–1568, 1990.

[7] O. Axelsson and M. Neytcheva. Preconditioning methods for linear systems arising in constrained optimization problems. *Numer. Linear Algebra Appl.*, 10:3–31, 2003.

[8] O. Axelsson and M. Neytcheva. Robust preconditioners for saddle point problems. In I. Dimov, I. Lirkov, S. Margenov and Z. Zlatev, editors, *Numerical Methods and Applications*, volume 2542 of *Lecture Notes in Computer Science*, pages 158–166. Springer-Verlag Berlin Heidelberg, 2003.

[9] R. E. Bank, B. D. Welfert and H. Yserentant. A class of iterative methods for solving saddle point problems. *Numer. Math.*, 56:645–666, 1990.

[10] T. Barth and T. Manteuffel. Variable metric conjugate gradient methods. In M. Natori and T. Nodera, editors, *PCG '94: Matrix Analysis and Parallel Computing*, pages 165–188. 1994.

[11] T. Barth and T. Manteuffel. Multiple recursion conjugate gradient algorithms Part I: Sufficient conditions. *SIAM J. Matrix Anal. Appl.*, 21:768–796, 2000.

[12] T. L. Barth and T. A. Manteuffel. Conjugate gradient algorithms using multiple recursions. In L. Adams and J. L. Nazareth, editors, *Proceedings of the AMS-IMS-SIAM Summer Research Conference held at the University of Washington, Seattle, WA, July 9–13, 1995*, pages 107–123. SIAM, Philadelphia, 1996.

[13] B. Beckermann, S. A. Goreinov and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM J. Matrix Anal. Appl.*, 27:772–778, July 2005. doi: 10.1137/040618849.

[14] M. W. Benson. *Iterative Solution of Large Scale Linear Systems.* Master's thesis, Lakehead University, Thunder Bay, Ontario, 1973.

[15] M. Benzi. Preconditioning techniques for large linear systems: A survey. *J. Comput. Phys.*, 182:418–477, 2002.

[16] M. Benzi, G. H. Golub and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.

[17] M. Benzi and A. J. Wathen. *Some Preconditioning Techniques for Saddle Point Problems*, volume 13 of *Mathematics in Industry.* Springer-Verlag Berlin Heidelberg, 2008.

[18] A. Böttcher and S. M. Grudsky. *Spectral Properties of Banded Toeplitz Matrices.* SIAM, 2005.

[19] M. F. Bourque and T. Ransford. Super-identical pseudospectra. *J. London Math. Soc.*, 2:511–528, 2009.

[20] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50:1–17, 1988.

[21] A. Brandt, S. F. McCormick and J. W. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In D. J. Evans, editor, *Sparsity and Its Applications.* Cambridge University Press, 1984.

[22] C. G. Broyden. A new taxonomy of conjugate gradient methods. *Comput. Math. Appl.*, 31:7–17, 1996.

[23] Z.-H. Cao. A note on spectrum distribution of constraint preconditioned generalized saddle point matrices. *Numer. Linear Algebra Appl.*, 16:503–516, 2009.

[24] P. Carrette. Reliable method for the evaluation of commuting matrices. Technical Report LiTH-ISY-R-2012, Department of Electrical Engineering, Linköping University, Sweden, 1998.

[25] C. Carstensen, M. Kuhn and U. Langer. Fast parallel solvers for symmetric boundary element domain decomposition equations. *Numer. Math.*, 79:321–347, 1998.

[26] R. H. Chan and X.-Q. Jin. *An Introduction to Iterative Toeplitz Solvers*. Fundamentals of Algorithms. SIAM, 2007.

[27] R. H. Chan, M. K. Ng and A. M. Yip. The best circulant preconditioners for Hermitian Toeplitz systems II: The multiple-zero case. *Numer. Math.*, 92:17–40, 2002.

[28] R. H. Chan and G. Strang. Toeplitz equations by conjugate gradients with circulant preconditioners. *SIAM J. Sci. Stat.Comput.*, 10:104–119, 1989.

[29] T. F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM J. Sci. Stat. Comput.*, 9:766–771, 1988.

[30] D. L. Darmofal and P. J. Schmid. The importance of eigenvectors for local preconditioning of the Euler equations. *J. Comput. Phys.*, 127:346–362, 1996.

[31] T. A. Davis and Y. Hu. The University of Florida Sparse Matrix Collection. *ACM Trans. Math. Softw.*, 38, 2011.

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

[33] C. R. Dohrmann and R. B. Lehoucq. A primal-based penalty preconditioner for elliptic saddle point systems. *SIAM J. Numer. Anal.*, 44:270–282, 2006.

[34] H. S. Dollar, N. I. M. Gould, M. Stoll and A. J. Wathen. Preconditioning saddle-point systems with applications in optimization. *SIAM J. Sci. Comput.*, 32(1):249–270, 2010.

[35] J. Drkošová, A. Grenbaum, M. Rozložník and Z. Strakoš. Numerical stability of GMRES. *BIT*, 35:309–330, 1995.

[36] J. Duintjer Tebbens, J. Liesen and Z. Strakos. Analysis of the second phase of the GMRES convergence for a convection-diffusion model problem, 2010. Submitted.

[37] J. Duintjer Tebbens and G. Meurant. Any Ritz value behavior is possible for Arnoldi and for GMRES with any convergence curve. Submitted.

[38] M. Eiermann. Fields of values and iterative methods. *Linear Algebra Appl.*, 180:167–197, 1993.

[39] M. Eiermann and O. G. Ernst. Geometric aspects in the theory of Krylov subspace methods. *Acta Numer.*, 10:251–312, 2001.

[40] H. Elman. Multigrid and Krylov subspace methods for the discrete Stokes equations. In N. D. Melson, T. A. Manteuffel, S. F. McCormick and C. C. Douglas, editors, *Seventh Copper Mountain Conference on Multigrid Methods*, volume CP 3339, pages 283–299. 1996.

[41] H. C. Elman. *Iterative methods for sparse, nonsymmetric systems of linear equations*. Ph.D. thesis, Yale University, Department of Computer Science., 1982.

[42] H. C. Elman and A. Ramage. Fourier analysis of multigrid for a model two-dimensional convection-diffusion equation. *BIT*, 46:283–306, 2006. 10.1007/s10543-006-0052-5.

[43] H. C. Elman, D. J. Silvester and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*. Oxford University Press, Oxford, 2005.

[44] M. Embree. How descriptive are GMRES convergence bounds? Technical report, Oxford University Computing Laboratory, 1999.

[45] A. Essai. Weighted FOM and GMRES for solving nonsymmetric linear systems. *Numer. Algorithms*, 18:277–292, 1998.

[46] V. Faber, W. Joubert, E. Knill and T. Manteuffel. Minimal residual method stronger than polynomial preconditioning. In *Proceedings of the Colerado Conference on Iterative Methods*. Breckenridge, Colorado, 1994.

[47] V. Faber, J. Liesen and P. Tichỳ. The Faber-Manteuffel theorem for linear operators. *SIAM J. Numer. Anal.*, 46:1323–1337, 2008.

[48] V. Faber, J. Liesen and P. Tichỳ. On orthogonal reduction to Hessenberg form with small bandwidth. *Numerical Algorithms*, 51:133–142, October 2009.

[49] V. Faber and T. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21:352–362, 1984.

[50] B. Fischer, A. Ramage, D. J. Silvester and A. J. Wathen. On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems. *Comput. Method Appl. M.*, 179, 1999.

[51] R. Fletcher. Conjugate gradient methods for indefinite systems. In G. A. Watson, editor, *Lecture Notes in Mathematics*, volume 506, pages 73–89. Springer-Verlag Berlin Heidelberg, New York, 1976.

[52] N. M. Freund, Roland W. and Nachtigal. Software for simplified Lanczos and QMR algorithms. *Appl. Numer. Math.*, 19:319–341, 1995.

[53] R. W. Freund and N. M. Nachtigal. QMR: A quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60:315–339, 1991.

[54] R. W. Freund and H. Zha. Simplfications of the nonsymmetric Lanczos process and a new algoithm for Hermitian indefinite linear systems. Technical report, AT&T Bell Labs, 1994.

[55] D. Funaro. A preconditioning matrix for the Chebyshev differencing operator. *SIAM J. Numer. Anal.*, 24:1024–1031, 1987.

[56] G. N. Gatica and N. Heuer. Conjugate gradient method for dual-dual mixed formulations. *Math. Comp.*, 71:1455–1472, 2002.

[57] I. Gohberg, P. Lancaster and L. Rodman. *Indefinite Linear Algebra and Applications*. Birkhäuser Verlag, Basel, 2005.

[58] G. H. Golub, C. Greif and J. M. Varah. An algebraic analysis of a block diagonal preconditioner for saddle point systems. *SIAM J. Matrix Anal. Appl.*, 27:779–792, 2006.

[59] G. H. Golub and C. F. van Loan. *Matrix Computations*. The John Hopkins University Press, Maryland, USA, third edition, 1996.

[60] N. I. M. Gould, D. Orban and P. Toint. CUTEr and SifDec: A constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Softw.*, 29:373–394, 2003.

[61] M. Green. *Finite Difference Preconditioners for Spectral Operators in Chebfun*. Master's thesis, University of Oxford, 2011.

[62] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, USA, 1997.

[63] A. Greenbaum and L. Gurvits. Max-min properties of matrix factor norms. *SIAM J. Sci. Comp.*, 15:348–358, 1994.

[64] A. Greenbaum, V. Pták and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17:465–469, 1996.

[65] A. Greenbaum and L. N. Trefethen. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM J. Sci. Comput.*, 15:359–368, 1994.

[66] M. H. Gutknecht. Changing the norm in conjugate gradient type algorithms. *SIAM J. Numer. Anal.*, 30:40–56, 1993.

[67] M. H. Gutknecht and M. Rozložník. A framework for generalized conjugate gradient methods-with special emphasis on contributions by Rüdiger Weiss. *Appl. Numer. Math.*, 41:7–22, 2002.

[68] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, 1985.

[69] L. Hemmingsson. A semi-circulant preconditioner for the convection-diffusion equation. *Numer. Math.*, 21:211–248, 1998.

[70] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31:2291–2317, 2010.

[71] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.

[72] N. J. Higham. The Matrix Function Toolbox. `http://www.ma.man.ac.uk/~higham/mftoolbox`.

[73] N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[74] N. J. Higham, C. Mehl and F. Tisseur. The canonical generalized polar decomposition. *SIAM J. Matrix Anal. Appl.*, 31:2163–2180, 2010.

[75] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990.

[76] I. C. F. Ipsen. Expressions and bounds for the GMRES residual. *BIT*, 40:524–535, 2000.

[77] A. Jameson, E. Kreindler and P. Lancaster. Symmetric, positive semidefinite, and positive definite real solutions of $AX = XA^T$ and $AX = YB$. *Linear Algebra Appl.*, 160:189–215, 1992.

[78] W. Joubert. On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems. *Numer. Linear Algebra Appl.*, 1:427–447, 1994.

[79] W. A. Joubert. A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems. *SIAM J. Sci. Comp.*, 15:427–439, 1994.

[80] A. Klawonn. Block-triangular preconditioners for saddle point problems with a penalty term. *SIAM J. Sci. Comput.*, 19:172–184, 1998.

[81] A. Klawonn. An optimal preconditioner for a class of saddle point problems with a penalty term. *SIAM J. Sci. Comp.*, 19:540–552, 1998.

[82] A. Klawonn and G. Starke. Block triangular preconditioners for nonsymmetric saddle point problems: Field-of-values analysis. *Numer. Math.*, 81:577–594, 1999.

[83] P. Konstantinos. *How Fast are Nonsymmetric Matrix Iterations? A 2010 Update.* Master's thesis, University of Oxford, 2010.

[84] P. Krzyżanowski. On block preconditioners for nonsymmetric saddle point problems. *SIAM J. Sci. Comput.*, 23:157–169, 2001.

[85] P. Krzyżanowski. On block preconditioner for saddle point problems with singular or indefinite (1,1) block. *Numer. Linear Algebra Appl.*, 18:123–140, 2011.

[86] E. W. Larsen. Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry discrete ordinates equations. Part I: Theory. *Nucl. Sci. Eng.*, 82:47–63, 1982.

[87] J. Liesen. Computable convergence bounds for GMRES. *SIAM J. Matrix Anal. Appl.*, 21:882–903, 2000.

[88] J. Liesen. When is the adjoint of a matrix a low degree rational function in the matrix? *SIAM J. Matrix Anal. Appl.*, 29:1171–1180, 2007.

[89] J. Liesen and B. N. Parlett. On nonsymmetric saddle point matrices that allow conjugate gradient iterations. *Numer. Math.*, 108:605–624, 2008.

[90] J. Liesen and P. E. Saylor. Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases. *SIAM J. Numer. Anal.*, 42:2148–2158, 2005.

[91] J. Liesen and Z. Strakoš. On optimal short recurrences for generating orthogonal Krylov subspace bases. *SIAM Rev.*, 50:485–503, 2008.

[92] J. Liesen and Z. Strakoš. GMRES convergence analysis for a convection-diffusion model problem. *SIAM J. Sci. Comput.*, 26:1989–2009, 2005.

[93] D. Loghin and A. J. Wathen. Analysis of preconditioners for saddle-point problems. *SIAM J. Sci. Comput.*, 25:2029–2049, 2004.

[94] K.-A. Mardal and R. Winther. Preconditioning discretizations of partial differential equations. *Numer. Linear Algebra Appl.*, 18:1–40, 2011.

[95] T. P. Mathew, M. Sarkis and C. E. Schaerer. Analysis of block matrix preconditioners for elliptic optimal control problems. *Numer. Linear Algebra Appl.*, 14(4):257–279, 2007.

[96] D. R. McCoy and E. W. Larsen. Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry discrete ordinates equations. Part II: Numerical results. *Nucl. Sci. Eng.*, 82:64–70, 1982.

[97] C. Mehl. Essential decomposition of polynomially normal matrices in real indefinite inner product spaces. *Electron. J. Linear Algebra*, 15:84–106, 2006.

[98] C. Mehl. On classification of normal matrices in indefinite inner product spaces. *Electron. J. Linear Algebra*, 15:50–83, 2006.

[99] J. A. Meijerink and H. A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comp.*, 31:148–162, 1977.

[100] G. Meurant. *Computer Solution of Large Linear Systems*, volume 28 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1999.

[101] G. Meurant and Z. Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer.*, 15:471–542, 2006.

[102] A. Meyer and T. Steidten. Improvements and experiments on the Bramble-Pasciak type CG for mixed problems in elasticity. Technical Report SFB393/01-13, TU Chemnitz, Chemnitz, Germany, 2001.

[103] K. W. Moreton. *Numerical Solution of Convection-Diffusion Problems*. Chapman and Hall, 1996.

[104] M. F. Murphy, G. H. Golub and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21:1969–1972, 1999.

[105] N. M. Nachtigal, S. C. Reddy and L. N. Trefethen. How fast are nonsymmetric matrix iterations? *SIAM J. Matrix Anal. Appl.*, 13:778–795, 1992.

[106] M. K. Ng. *Iterative Methods for Toeplitz Systems*. Numerical Mathematics and Scientific Computation. Oxford University Press, USA, 2004.

[107] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12:617–629, 1975.

[108] J. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE constrained optimization. Technical Report 1021, Mathematical Institute, University of Oxford, November 2010.

[109] J. Peters, V. Reichelt and A. Reusken. Fast iterative solvers for discrete Stokes equations. *SIAM J. Sci. Comput.*, 27:646–666, 2005.

[110] B. Philippe and L. Reichel. On the generation of Krylov subspace bases. *Appl. Numer. Math.*, 2011. doi: 10.1016/j.apnum.2010.12.009.

[111] A. Ramage. A multigrid preconditioner for stabilised discretisations of advection-diffusion problems. *J. Comput. Appl. Math.*, 110:187–203, 1999.

[112] T. Rees and M. Stoll. Block-triangular preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 17:977–996, 2010.

[113] T. Rees, M. Stoll and A. Wathen. All-at-once preconditioning in PDE-contrained optimization. *Kybernetika*, 46:341–360, 2010.

[114] H.-G. Roos, M. Stynes and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems*. Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, 1996.

[115] M. Rozložnìk and V. Simoncini. Krylov subspace methods for saddle point problems with indefinite preconditioning. *SIAM J. Matrix Anal. Appl.*, 24:368–391, 2002.

[116] J. W. Ruge. Algebraic multigrid (AMG) for geodetic survey problems. In *Prelimany Proc. Internat. Multigrid Conference*. 1983.

[117] J. W. Ruge and K. Stüben. Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG). In D. J. Paddon and H. Holstein, editors, *Multigrid Methods for Integral and Differential Equations*. The Institute of Mathematics and its Applications Conference Series, Clarendon Press, 1985.

[118] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003.

[119] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. and Stat. Comput.*, 7:856–869, 1986.

[120] H. Sadok and D. B. Szyld. A new look at CMRH and its relation to GMRES. Technical Report 10-04-25, Department of Mathematics, Temple University, April 2010.

[121] M. Sarkis and D. B. Szyld. Optimal left and right additive Schwarz preconditioning for minimal residual methods with Euclidean and energy norms. *Comput. Methods Ap. Mech. Engrg.*, 196:1612–1621, 2007.

[122] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29:752–773, 2007.

[123] S. Serra. Toeplitz preconditioners constructed from linear approximation processes. *SIAM J. Matrix Anal. Appl.*, 20:446–465, 1998.

[124] C. Siefert and E. de Sturler. Preconditioners for generalized saddle-point problems. *SIAM J. Numer. Anal.*, 44:1275–1296, 2006.

[125] D. Silvester, H. Elman and A. Ramage. Incompressible Flow and Iterative Solver Software (IFISS) version 3.1, January 2011. `http://www.manchester.ac.uk/ifiss/`.

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

[127] V. Simoncini. Block triangular preconditioners for symmetric saddle-point systems. *Appl. Numer. Math.*, 49:63–80, 2004.

[128] V. Simoncini and D. B. Szyld. The effect of non-optimal bases on the convergence of Krylov subspace methods. *Numer. Math.*, 100:711–733, 2005.

[129] G. L. G. Sleijpen and D. R. Fokkema. Bi-CGStab(ℓ) for linear equations involving unsymmetric matrices with complex spectrum. *Elec. Trans. Numer. Anal.*, 1:11–32, 1993.

[130] G. L. G. Sleijpen, H. A. van der Vorst and J. Modersitzki. Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems. *SIAM J. Matrix Anal. Appl.*, 22:726–751, 2000.

[131] P. Sonneveld. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 10(1):36–52, 1989.

[132] P. Sonneveld and M. B. van Gijzen. IDR(s): A family of simple and fast algorithms for solving large nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 31:1035–1062, 2008.

[133] G. Starke. Field-of-values analysis of preconditioned iterative methods for non-symmetric elliptic problems. *Numer. Math.*, 78:103–117, 1997.

[134] M. Stoll. *Solving Linear Systems using the Adjoint*. Ph.D. thesis, University of Oxford, 2008.

[135] M. Stoll and A. J. Wathen. Combination preconditioning and the Bramble-Pasciak$^+$ preconditioner. *SIAM J. Matrix Anal. Appl.*, 30:582–608, 2008.

[136] G. Strang. A proposal for Toeplitz matrix calculations. *Stud. Appl. Math.*, 74:171–176, 1986.

[137] V. V. Strela and E. E. Tyrtyshnikov. Which circulant preconditioner is better? *Math. Comp.*, 65:137–150, 1996.

[138] O. Taussky. The role of symmetric matrices in the study of general matrices. *Linear Algebra Appl.*, 5:147–154, 1972.

[139] O. Taussky and W. V. Parker. Problem 4846. *Amer. Math. Monthly*, 67:192–193, 1960.

[140] M. Tismenetsky. A decomposition of Toeplitz matrices and optimal circulant preconditioning. *Linear Algebra Appl.*, 154–156:105–121, 1991.

[141] K.-C. Toh. GMRES vs. ideal GMRES. *SIAM J. Matrix Anal. Appl.*, 18:30–36, 1997.

[142] A. Toselli and O. Widlund. *Domain Decomposition Methods — Algorithms and Theory*, volume 34 of *Springer Series in Compuational Mathematics*. Springer-Verlag Berlin Heidelberg, 2005.

[143] L. N. Trefethen. Approximation theory and numerical linear algebra. In J. C. Mason and M. G. Cox, editors, *Algorithms for Approximation II*. Chapman and Hall, 1990.

[144] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[145] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, 2005.

[146] L. N. Trefethen et al. *Chebfun Version 4.0.* The Chebfun Development Team, 2011. `http://www.maths.ox.ac.uk/chebfun/`.

[147] H. A. van der Vorst. Bi-CGStab: A fast and smoothly converging variant of BiCG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 13:631–644, 1992.

[148] H. A. van der Vorst and C. Vuik. The superlinear convergence behaviour of GMRES. *J. Comput. Appl. Math.*, 48:327–341, 1993.

[149] C. F. van Loan. Generalizing the singular value decomposition. *SIAM J. Numer. Anal.*, 13:76–83, 1976.

[150] C. F. van Loan. *Computational Frameworks for the Fast Fourier Transform.* SIAM, 1992.

[151] E. Vecharynski and J. Langou. Any admissible cycle-convergence behavior is possible for restarted GMRES at its initial cycles. *Numer. Linear Algebra Appl.*, 18:499–511, 2011.

[152] L. G. Vulkov, J. J. H. Miller and G. I. Shishkin, editors. *Analytical and Numerical Methods For Convection-Dominated and Singularly Perturbed Problems.* Nova Science Publishers, 2000.

[153] A. J. Wathen. Preconditioning and convergence in the right norm. *Int. J. Comput. Math.*, 84:1199–1209, 2007.

[154] T. G. Wright. Eigtool. `http://www.comlab.ox.ac.uk/pseudospectra/eigtool/`, 2002.

[155] J.-P. M. Zemke. *Krylov Subspace Methods in Finite Precision: A Unified Approach.* Ph.D. thesis, Hamburg-Harburg University of Technology, 2003.

[156] G.-F. Zhang, Z.-R. Ren and Y.-Y. Zhou. On HSS-based constraint preconditioners for generalized saddle-point problems. *Numer. Algor.*, 57:273–287, 2011.

[157] W. Zulehner. Analysis of iterative methods for saddle point problems: A unified approach. *Math. Comp.*, 71:479–505, 2002.

[158] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 32:536–560, 2011.