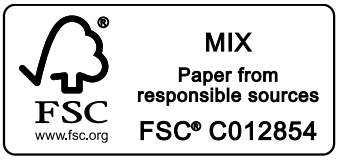
Generalized Krylov methods for large-scale matrix problems



This work is part of the research program “Innovative methods for large matrix problems”, which is (partly) financed by the Netherlands Organization for Scientific Research (NWO).

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-4251-2

Generalized Krylov methods for large-scale matrix problems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op maandag 26 juni 2017 om 14:00 uur

door

Ian Nathan Zwaan

geboren te Virginia, Verenigde Staten van Amerika

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecomissie is als volgt:

voorzitter:	prof.dr. M.G.J. van den Brand
promotor:	prof.dr.ir. B. Koren
copromotor(en):	dr. M.E. Hochstenbach
leden:	prof.dr. K.J. Batenburg (Universiteit Leiden, CWI)
	dr.ir. M.B. van Gijzen (Technische Universiteit Delft)
	prof.dr. P.C. Hansen (Technical University of Denmark)
	prof.dr. K. Meerbergen (University of Leuven)
	prof.dr. S. Weiland

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

Chapter 1

Introduction

This dissertation concerns the development of new Krylov subspace methods for two classes of well-known problems encountered in numerical mathematics. The first class consists of standard matrix eigenvalue problems (cf., e.g., [3, 78]), where the goal is to find scalars λ and nonzero vectors $\boldsymbol{x}$ such that

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{1.1}$$

for a given matrix A. The second class consists of ill-posed problems (cf., e.g., [28]), where the objective is to reconstruct an unknown vector $\boldsymbol{x}$ from

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1.2}$$

where A is an ill-conditioned matrix and the right-hand side $\boldsymbol{b}$ is contaminated by noise. In both cases, we will assume that A is too large for the practical use of direct methods, but is sufficiently structured to facilitate fast matrix-vector products. For example, A may be a sparse matrix with dimensions ranging from the order a few thousands up to a few billions.

Eigenvalue problems are important in many fields, for instance, chemistry, control theory, dynamic systems, geology, mechanics, pattern recognition, quantum mechanics, signal processing, statistics, vibration analysis, etc. Ill-posed problems are found in astronomy, computed tomography, computer vision, other fields related to image analysis and restoration, geophysics, signal processing, statistics, etc.

Krylov subspace methods are popular for solving large-scale problems of the form (1.1) and (1.2), when using direct methods is infeasible and fast matrix-vector products are available. The central idea of subspace methods is to project the problem onto a lower dimensional search space, and to extract an approximate solution by solving a small-scale problem instead of the original large-scale problem. However, challenging problems remain; and four of those problems are outlined below, as well as the contributions of this dissertation toward solutions.

1.1 Eigenvalue inclusion regions

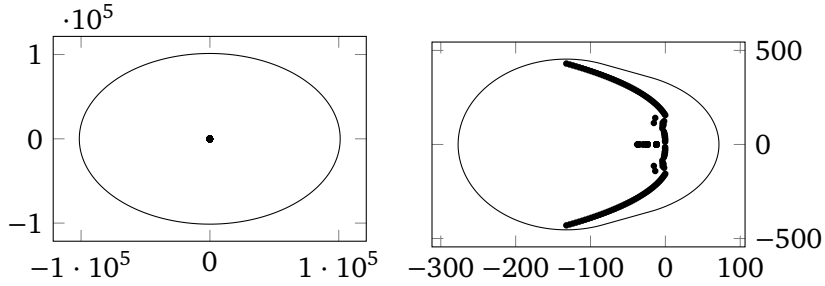


Figure 1.1: The boundary of the field of values (solid line) in relation to the eigenvalues (dots) of the Tolosa matrix [2] of dimension 340; before (left) and after (right) balancing.

Computing eigenvalues accurately can be computationally expensive, even with state-of-the-art iterative methods; thus, it may be desirable to have a fast alternative for initial and exploratory phases. For example, sometimes it suffices to have regions in the complex plane which are guaranteed to contain the (desired) eigenvalues, without knowing exactly where the eigenvalues reside in those regions. An attractive eigenvalue inclusion region is the field of values given by

$$W(A) = \{\boldsymbol{x}^* A \boldsymbol{x} \colon \boldsymbol{x} \in \mathbb{C}^n, \|\boldsymbol{x}\| = 1\}.$$

The region defined by $W(A)$ is convex and guaranteed to contain all eigenvalues; furthermore, its boundary can be approximated efficiently and is often tight around the eigenvalues. However, occasionally the numerical radius $r(A)$ of A is much larger than the spectral radius $\rho(A)$ of A, where $r(A)$ and $\rho(A)$ are defined as

$$r(A) = \max_{z \in W(A)} |z| \qquad \text{and} \qquad \rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda|,$$

making $W(A)$ meaningless as an inclusion region.

In Chapter 2 we show that the quality of the field of values as an inclusion region may often be improved by balancing the matrix A when $r(A)/\rho(A) \gg 1$; see Figure 1.1 for an example. Balancing is an existing technique designed to decrease the disparity between row and column norms through a carefully constructed diagonal similarity transform. Several interesting connections with the nonnormality of matrices are investigated and emphasized. Moreover, we propose a new, simple, and fast balancing methodology for computing spectral inclusion regions, where the Hessenberg matrix resulting from the Arnoldi process

is balanced and used to approximate the field of values. The effectiveness of the method is demonstrated with numerical experiments.

1.2 Eigenvalue sensitivity

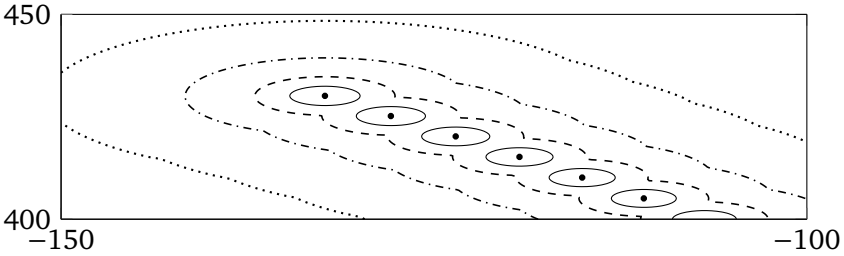


Figure 1.2: Pseudospectra level curves for the 340×340 Tolosa matrix, for $\varepsilon = 10^{-1.1}$ (dotted), $10^{-1.4}$ (dash dotted), $10^{-1.7}$ (dashed), and 10^{-2} (solid).

Although inclusion regions can be appropriate in a preliminary stage of eigenvalue computations, they only provide limited information. Relevant information, apart from the eigenvalues and eigenvectors themselves, often includes the behavior of eigenvalues under perturbations. Information of this kind can be provided, for instance, by asymptotic error bounds and pseudospectra. An important and intuitively straightforward asymptotic error bound is given by

$$|\widetilde{\lambda} - \lambda| \lesssim \kappa(\lambda)\|E\| \qquad \text{with} \qquad \kappa(\lambda) = \frac{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|}{|\boldsymbol{x}^*\boldsymbol{y}|},$$

where $(\lambda, \boldsymbol{x})$ is a simple eigenpair of A with corresponding left eigenvector $\boldsymbol{y}$, $\widetilde{\lambda}$ is some eigenvalue of $A + E$, and $\boldsymbol{x}^*$ denotes the conjugate transpose of $\boldsymbol{x}$. Accordingly, the eigenvalue condition number $\kappa(\lambda)$ provides an indication of the worst-case sensitivity of λ to perturbations of A. The challenge is to approximate $\kappa(\lambda)$, and therefore $\boldsymbol{x}$ and $\boldsymbol{y}$, efficiently and accurately for nonnormal matrices. One-sided methods may unappealingly require two runs to compute acceptable approximations to both the left and right eigenvectors; while current two-sided methods simultaneously approximate the left and right eigenvectors, and thus eigenvalue condition numbers, but also face inherent difficulties with restarts, numerical stability, and error analysis.

Additional insight may be gained from pseudospectra; indeed, one possible definition of the ε-pseudospectrum of A is

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \colon z \in \Lambda(A + E) \text{ for some } E \text{ with } \|E\| < \varepsilon\}.$$

A visual representation of the boundary of Λ_ε reveals the extent with which eigenvalues can "move" under perturbations; see, for example, Figure 1.2. Computing pseudospectra is computationally demanding; hence, one-sided subspace methods are sometimes used to approximate parts of pseudospectra and to speedup the process. However, the choice for one-sided methods is arbitrary to some extent.

We present an extension of the Krylov–Schur restarting method to the two-sided Arnoldi method for large-scale nonnormal matrices in Chapter 3. This extension allows for the simultaneous approximation of left and right eigenvectors, and thus eigenvalue condition numbers, while working exclusively with orthonormal bases. Specifically, two-sided Krylov–Schur maintains orthonormal bases for a separate left and right Krylov subspace, and applies only orthonormal transformations to these bases during the restarts. Therefore, we may expect, with a careful implementation, better numerical stability compared to unsymmetric Lanczos. We derive algorithms for both standard Ritz extraction and harmonic Ritz extraction, and present a quantitative and qualitative error analysis. We complete the chapter with numerical examples where we compute the least sensitive eigenvalues and use the left and right shift-invariant bases to approximate pseudospectra.

1.3 Tikhonov regularization

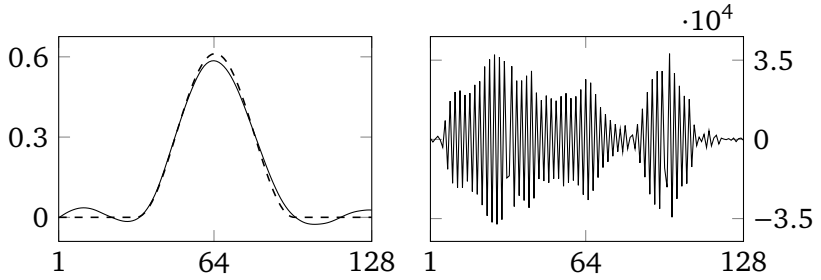


Figure 1.3: The exact solution (dashed) and reconstructed solution (solid) of the `phillips` test problem from Regularization Tools [27], with regularization (left) and without regularization (right).

The (pseudo)inverse of an ill-conditioned matrix tends to amplify high-frequency components in the right-hand side when used to solve least-squares problems of the form (1.2). Hence, if the measured data $\boldsymbol{b}$ is contaminated by noise, the noise will be amplified and dominate the solution; see, for example, Figure 1.3. Dampening the high-frequency components may often improve the quality of the solution, and can be achieved with regularization methods. A well-

known method is, for instance, standard form Tikhonov regularization, where the solution of

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|\boldsymbol{x}\|^2$$

for some $\mu > 0$ is used as a solution for (1.2). This minimization problem can be solved efficiently for large-scale problems by projecting it onto a Krylov subspace generated with Golub–Kahan–Lanczos bidiagonalization or the Arnoldi method if A is square. For certain problems it may be more appropriate to use general form Tikhonov regularization, and solve

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|L\boldsymbol{x}\|^2,$$

for a suitably chosen L. When $L \neq I$, it is no longer obvious that standard Krylov subspaces are satisfactory search spaces, and generalized Krylov subspaces may be considered instead. This concern is further exacerbated in multiparameter Tikhonov regularization:

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \sum_{i=1}^{\ell} \mu_i \|L_i\boldsymbol{x}\|^2,$$

which faces the additional problem of parameter selection. Numerous methods exist for selecting a sensible μ in standard and general form Tikhonov regularization; however, selecting "good" μ_i in multiparameter regularization is more complicated.

We introduce a new method for large-scale multiparameter Tikhonov regularization with general regularization operators in Chapter 4. The method works by repeatedly extending the search space in multiple directions, similar to generalized Krylov, and subsequently removing the less promising directions to ensure moderate growth of the search space. Moreover, we propose a discrepancy principle based parameter selection strategy related to perturbation results. Numerical experiments are performed to test the algorithms.

1.4 The generalized singular value decomposition

Consider general form Tikhonov regularization and suppose, for example, that

$$A \in \mathbb{R}^{m\times n}, \qquad L \in \mathbb{R}^{p\times n}, \qquad m \geq n \geq p, \qquad \mathcal{N}(A) \cap \mathcal{N}(L) = \{\boldsymbol{0}\};$$

then the generalized singular value decomposition (GSVD) of the matrix pair (A, L) is given by

$$\begin{aligned} A &= UCX^{-1}, \qquad U^TU = I, \qquad \operatorname{diag}(c_1, \ldots, c_n) \in [0, 1]^{n\times n}, \\ L &= VSX^{-1}, \qquad V^TV = I, \qquad \operatorname{diag}(s_1, \ldots, s_p) \in [0, 1]^{p\times n}, \end{aligned}$$

where X is nonsingular and $C^TC + S^TS = I$. The regularized solution can now be written as

$$\begin{aligned} \boldsymbol{x}_{\boldsymbol{\mu}} &= (A^TA + \mu L^TL)^{-1}A^T\boldsymbol{b} \\ &= X(C^TC + \mu S^TS)^{-1}CU^T\boldsymbol{b} = \sum_{i=1}^{n} \frac{c_i}{c_i^2 + \mu s_i^2}\, \boldsymbol{x}_i \boldsymbol{u}_i^T\boldsymbol{b}, \end{aligned}$$

where $s_i = 0$ for $i = p+1, \ldots, n$. This formula indicates that solutions for different μ or multiple right-hand sides $\boldsymbol{b}$ can be obtained efficiently once the GSVD has been computed. It also motivates the truncated GSVD solution, which is obtained by setting $\mu = 0$ and only summing the terms corresponding to the k largest c_i. Typically, $k \ll n$ for ill-conditioned A, and the intention is to exclude the terms where $1/c_i$ becomes too large and amplify unwanted components excessively. Besides regularization problems, the GSVD is also useful for solving eigenvalue problems of the form

$$s_i^2 A^TA\boldsymbol{x}_i = c_i^2 L^TL\boldsymbol{x}_i,$$

without using the products A^TA and L^TL and potentially losing information. Unfortunately, computing the GSVD using direct methods is only feasible for moderately sized matrix pairs.

In Chapter 5 we derive two new algorithms for computing of a few of the extremal generalized singular values and their corresponding generalized singular vectors. The context and connections with existing methods are stated, convergence behavior is investigated, and error analysis is provided. The chapter ends with numerical experiments demonstrating the competitiveness of the methods, and illustrating their suitability for the approximation of the truncated GSVD of matrix pairs with rapidly decaying generalized singular values.

1.5 Outline

The structure of this thesis follows the structure of the previous sections and is summarized below.

Chapter 2 is dedicated to matrix balancing for field of value based spectral inclusion regions with connections to nonnormality of matrices, together with a new and efficient methodology for the approximation of these inclusion regions.

An extension of the Krylov–Schur restarting method to two-sided Arnoldi is given in Chapter 3, along with an extensive error analysis. Suggestions for a robust implementation and possible applications are included.

In Chapter 4 a multidirectional subspace expansion technique is considered for large-scale multiparameter Tikhonov regularization. Furthermore, a selection strategy for multiple parameters is proposed.

A generalized Davidson algorithm and an alternative multidirectional version are derived and analyzed in Chapter 5. Both methods depart from previous iterative methods for the GSVD, and depend on restarts with multiple vectors instead of inner-outer iterations.

Chapter 2 and Chapter 5 have been submitted for publication [38, 97], Chapter 3 has been accepted for publication in SIAM J. Matrix Anal. Appl. [98], and Chapter 4 has appeared in J. Sci. Comput. [99]. All articles have been edited for this dissertation and have minor editorial changes and differences.

1.6 Notation

We use the following notation, unless stated otherwise. Regular capital letters are used for matrices and calligraphic letters for subspaces; for example, $\mathcal{N}(A)$ and $\mathcal{R}(A)$ denote the nullspace and range of A, respectively. Bold lowercase letters denote vectors, while regular lowercase Roman and Greek letters denote scalars. Specifically, we use I for the identity matrix, D for diagonal matrices, E and F for error matrices, and $\boldsymbol{e}_i$ for the i-th standard basis vector; m, n, p, k, l, and ℓ for dimensions and sizes; i, j, k, l, and ℓ for indices; κ for condition numbers, λ for eigenvalues, μ for regularization parameters, and σ for (generalized) singular values. The quantities $\sigma_{\max}(G)$ and $\sigma_{\min}(G)$ are defined as the largest and smallest generalized singular values, respectively, of a general matrix G. $\mathbb{R}$ and $\mathbb{C}$ signify the sets of real and complex numbers, respectively. Different flavors of the same letter are usually related; for instance, the elements of a diagonal matrix D are d_i, the elements of a general matrix A are a_{ij}, the columns $\boldsymbol{v}_i$ of V form a basis of the subspace $\mathcal{V}$. The transpose of A is A^T, and the Hermitian transpose is A^*. Finally, we use the notation $\|\cdot\| = \|\cdot\|_2$ for the Euclidean norm and $\|\cdot\|_F$ for the Frobenius norm.

Chapter 2

Matrix balancing for field of value type inclusion regions

Abstract. The field of values may be an excellent tool for generating a spectral inclusion region: it is easy to approximate numerically, and for many matrices this convex region fits relatively tightly around the eigenvalues. However, for some matrices the field of values may be a poor eigenvalue inclusion region (which happens, more precisely, if the numerical radius is much larger than the spectral radius). In this chapter, we show that balancing the matrix, also known as scaling, may be very helpful for generating a quality inclusion region based on the field of values. We review some known balancing techniques, present an implementation for the balancing of sparse matrices, and introduce a new scaling method by scaling the Hessenberg matrix resulting from a Krylov process. Moreover, several interesting connections with nonnormality of matrices are pointed out. We show that a combination of balancing and a projected field of values may render excellent approximate spectral inclusion regions.

Key words. Field of values, numerical range, (approximate) spectral inclusion region, eigenvalue inclusion region, eigenvalue localization, matrix balancing, matrix scaling, Krylov scaling, nonnormal matrix, (relative) measure of nonnormality, large sparse matrix, eigenvalue problem, strongly connected components.

AMS subject classification. 47A12, 65F10, 65F15, 65F30, 65F35, 65F50.

2.1 Introduction

Let A be a large sparse real or complex $n \times n$ matrix and let $\|\cdot\|$ denote the 2-norm. The field of values (or numerical range)

$$W(A) = \{\boldsymbol{x}^* A \boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{C}^n,\ \|\boldsymbol{x}\| = 1\}$$

may be an attractive spectral inclusion region for two main reasons. First, it is a convex and compact set which is guaranteed to contain all eigenvalues (cf. [44]). Second, it can be efficiently approximated by the method proposed by Johnson [45], who pointed out that $W(A)$ can be efficiently approximated by computing

the maximal and minimal eigenvalues of the Hermitian part of $e^{i\alpha}A$:

$$\mathcal{H}(e^{i\alpha}A) = \tfrac{1}{2}\,(e^{i\alpha}A + (e^{i\alpha}A)^*)$$

for a number of angles $\alpha \in [0, \pi)$. Therefore, we use

$$\begin{aligned} \max_{z\in W(A)} \operatorname{Re}(e^{-i\alpha}z) &= \tfrac{1}{2}\,\lambda_{\max}(e^{i\alpha}A + (e^{i\alpha}A)^*), \\ \min_{z\in W(A)} \operatorname{Re}(e^{-i\alpha}z) &= \tfrac{1}{2}\,\lambda_{\min}(e^{i\alpha}A + (e^{i\alpha}A)^*), \end{aligned} \tag{2.1}$$

for every angle α, where Re denotes the real part of a complex number, and $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalue of a Hermitian matrix.

For large sparse matrices, it is not necessary to compute all eigenvalues of the Hermitian parts $\frac{1}{2}(e^{i\alpha}A + (e^{i\alpha}A)^*)$ of the matrices $e^{i\alpha}A$; instead, we may use the Lanczos method (see, e.g., [90]) to approximate the largest and smallest eigenvalue of the Hermitian parts. The Lanczos method generates a low-dimensional Krylov subspace to approximate eigenpairs of large (sparse) matrices. It generally approximates the extremal eigenvalues well, particularly the largest and smallest eigenvalue. We may run a new Lanczos process for every α in a selected discrete set (cf. [11]); in this case, the largest eigenpair will be approximated using a different Krylov subspace for each angle α, generated by a different matrix of the form $e^{i\alpha}A + (e^{i\alpha}A)^*$ and an initial vector, for instance a random vector, or the approximate eigenvector for a previous value of α. The resulting set will be a subset of $W(A)$, and will often be a good approximation to this set. As $W(A)$ contains all eigenvalues, such an approximating subset of $W(A)$ will usually also be an eigenvalue inclusion region.

We can also generate a cruder approximation to $W(A)$ that is computationally cheaper and therefore more attractive, by using one single Krylov subspace for all angles α. We first carry out an Arnoldi process (see, e.g., [90]) on A and an initial vector $\boldsymbol{w}_1$ of unit length (for instance, a random vector). Let

$$\mathcal{V}_k = \mathcal{K}_k(A, \boldsymbol{w}_1) = \operatorname{span}(\boldsymbol{w}_1, A\boldsymbol{w}_1, \ldots, A^{k-1}\boldsymbol{w}_1)$$

be the Krylov space of dimension k generated by A and $\boldsymbol{w}_1$, where we make the common assumption that $\mathcal{V}_k$ has dimension k. Performing k steps of the Arnoldi process yields the decomposition

$$AV_k = V_k H_k + h_{k+1,k}\boldsymbol{w}_{k+1}\boldsymbol{e}_k^*, \tag{2.2}$$

where the columns of V_k form an orthonormal basis for $\mathcal{V}_k$ with $\boldsymbol{w}_1$ as its first column, H_k is an upper Hessenberg matrix, and $\boldsymbol{e}_k$ is the kth canonical basis

vector. One can now approximate $W(A)$ by $W(A) \approx W(V_k^* A V_k) = W(H_k)$, as was originally suggested by Manteuffel [58, 59]; see also [36, 37]. This approximation has the following attractive monotonic inclusion property (see [58]).

Proposition 2.1.

$$W(H_k) \subseteq W(H_{k+1}) \subseteq W(A).$$

In this proposition the convex set $W(H_k) = W(V_k^* A V_k)$ may be interpreted as the field of values of A restricted to the Krylov subspace $\mathcal{V}_k$. In particular, we know that after k steps $W(H_k)$, and therefore also $W(A)$, contains the convex hull of the eigenvalues of H_k, which are the Ritz values of A with respect to $\mathcal{V}_k$. Since $k \ll n$ in a subspace method, determining $W(H_k)$ is computationally very attractive. We note that the interior approximation $W(H_k)$ to $W(A)$ will usually result in a smaller region than a region where a separate Krylov space per angle α is used; however, the difference may be small, cf. [36, Ex. 2.2]. The resulting field of values $W(H_k)$ is often a good approximate inclusion region for the spectrum.

However, in some cases $W(H_k)$ may be much too small, or much too large. A simple extreme example of the former case is the 10×10 matrix A with as its only elements $a_{j+1,j} = 10^{j-1}$, for $j = 1, \ldots, 9$, on its subdiagonal, where the starting vector is $\boldsymbol{w}_1 = \boldsymbol{e}_1$. It is easy to see that the k-dimensional Krylov space generated by A and $\boldsymbol{w}_1$ is span($[\boldsymbol{e}_1 \ \ldots \ \boldsymbol{e}_k]$), and that the logarithmic norm of H_k (the largest real part; cf. the remainder of this section) increases exponentially with k. However, since one usually chooses a random initial vector, this extreme behavior is rare.

The main motivation for this chapter is the opposite case: unfortunately, for some matrices $W(A)$ and $W(H_k)$ may be poor spectral inclusion regions because they are much *larger* than the convex hull of the spectrum. We give an example for the 4000×4000 tols4000 matrix [2]; see Figure 2.1.
For this matrix, the spectral radius

$$\rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda| \tag{2.3}$$

is $\rho(A) \approx 4.84 \cdot 10^3$, where $\Lambda(A)$ denotes the spectrum of A. The numerical radius

$$r(A) = \max_{z \in W(A)} |z| \tag{2.4}$$

is significantly larger: $r(A) \approx 1.17 \cdot 10^7$, so that the ratio

$$\frac{r(A)}{\rho(A)}, \tag{2.5}$$

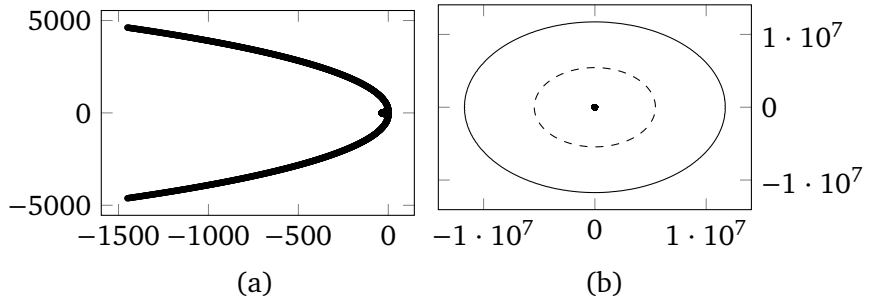


Figure 2.1: (a): Spectrum; (b): spectrum, fields of values $W(A)$ (solid) and $W(H_{20})$ (dash) for tols4000. Note that the spectrum is barely visible (as a dot) because of the large scale.

which ideally should be close to one for a tight spectral inclusion region, is approximately $2.42 \cdot 10^3$. Ratio (2.5) will be of interest throughout the rest of this chapter.

In this example, the field of values $W(H_k)$ is also far too large for a useful approximate spectral inclusion region: for $k = 20$, we have the numerical radius $r(H_{20}) \approx 5.38 \cdot 10^6$, so that $r(H_{20})/\rho(A) \approx 1.11 \cdot 10^3$. Therefore, both $W(A)$ and $W(H_{20})$ are poor spectral inclusion regions. A striking property of the tols4000 matrix is that it is very badly scaled: the ratio of the largest and smallest column norms is $\mathcal{O}(10^7)$, and similarly for the row norms. Therefore, we investigate balancing (or scaling, two terms that are generally used interchangeably) of the matrix to improve the (approximate) eigenvalue inclusion regions based on the field of values.

Eigenvalue inclusion regions are useful in many applications, for instance, to get a quick estimate of the spectrum, or to determine a suitable region for the computation of pseudospectra [12, Thm. 2.1]. As a historical note, Bendixson already showed in 1902 [5, Thm. II] that $\min \operatorname{Re}(W(A)) \leq \operatorname{Re}(\lambda) \leq \max \operatorname{Re}(W(A))$ for all eigenvalues λ of A; in fact, both the real and imaginary parts can be bounded in this way; see, for instance [92]. We note that for some applications, not the spectrum is important but the field of values itself. One example is the quantity $\frac{d}{dt}\|e^{tA}\|_{|t=0}$, which is the logarithmic norm (or numerical abscissa), equal to $\max_{z \in W(A)} \operatorname{Re}(z) = \frac{1}{2}\lambda_{\max}(A + A^*)$; see, e.g., [89]. In these cases, the techniques of this chapter, which aim at getting good spectral inclusion regions by modifying the field of values, may be less relevant.

The rest of this chapter is organized as follows. Section 2.2.1 explores why scaling may be a good idea to generate high-quality approximate eigenvalue inclusion regions. In Section 2.3 we review existing scaling methods for the

matrix to generate better spectral inclusion regions based on the field of values. We also present a new implementation of a sparse balancing routine spbalance. Section 2.4 introduces a simple new Krylov scaling approach. We end with some numerical experiments and conclusions in Sections 2.5 and 2.6.

2.2 Scaling

2.2.1 Matrix scaling for the field of values

A simple but key observation of this chapter is given in the following proposition.

Proposition 2.2. *Let $D \in \mathbb{R}^{n\times n}$ be a nonsingular matrix. Then $W(D^{-1}AD)$ is a spectral inclusion region for A. Moreover, $\rho(D^{-1}AD) = \rho(A)$.*

Proof. This follows easily from the observation that A and $D^{-1}AD$ have the same eigenvalues. □

In the context of matrix scaling, D is usually restricted to be either a diagonal matrix with positive elements, or a permutation thereof; this explains the choice of the notation. Also, several scaling methods restrict the diagonal elements to powers of two in order to avoid roundoff errors.

The next ingredient that we need is the fact that the matrix two-norm tightly squeezes the numerical radius, as quantified by the following proposition; see [39, p. 331].

Proposition 2.3.

$$\tfrac{1}{2}\,\|A\| \le r(A) \le \|A\|.$$

We may attempt to choose suitable scaling matrices D such that $W(D^{-1}AD)$ is a tighter inclusion region than $W(A)$; in particular, we hope that $r(D^{-1}AD) \ll r(A)$. Then also (cf. (2.5))

$$\frac{r(D^{-1}AD)}{\rho(D^{-1}AD)} = \frac{r(D^{-1}AD)}{\rho(A)} \ll \frac{r(A)}{\rho(A)}. \tag{2.6}$$

In view of Propositions 2.2 and 2.3, it is a good idea to try to reduce $\|A\|$ to reach this goal. Proposition 2.3 leads to the following relatively straightforward key corollary.

Corollary 2.4. *If the nonsingular diagonal matrix D is such that $\|D^{-1}AD\| < \frac{1}{2}\,\|A\|$ then $r(D^{-1}AD) < r(A)$.*

Scaling tends to decrease the matrix norm; the following example illustrates this idea.

Example 2.1. Let $A = \left[\begin{smallmatrix} 0 & 4 \\ 1 & 0 \end{smallmatrix}\right]$, then $\|A\| = 4$, while the norm of the scaled matrix $D^{-1}AD$ using $D = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ is 2. It may also be checked that $W(A)$ has numerical radius $\rho(A) = \frac{5}{2}$, while $W(D^{-1}AD)$ is the interval $[-2, 2]$ and therefore $\rho(D^{-1}AD) = 2$.

Since

$$\tfrac{1}{\sqrt{n}} \max\{ \|A\|_1,\ \|A\|_\infty \} \le \|A\| \le \sqrt{n} \min\{ \|A\|_1,\ \|A\|_\infty \},$$

having approximately equal row and column sums may decrease $\|A\|$. This norm decrease is guaranteed if scaling A decreases $\|A\|_1$ or $\|A\|_\infty$ by at least a factor n.

2.2.2 Scaling and nonnormality

Another viewpoint on scaling is the following. The optimal convex eigenvalue inclusion region would be the convex hull of the spectrum. Therefore, the field of values for a normal matrix with the same eigenvalues would provide this optimal inclusion region. Hence, we can view balancing as an attempt to transform A into a matrix with the same eigenvalues that is closer to normal. (Note that this is a different question as the one studied in [77], where the closest normal matrix is sought with no conditions on the spectrum.)

We will therefore consider some measures of nonnormality; some are mentioned in [19]. Let

$$A = Q(\Lambda + N)Q^* \tag{2.7}$$

be a Schur decomposition of A. Here, Λ is a diagonal matrix containing the eigenvalues, while the quantities $\|N\|$ (minimized over all possible Schur forms) and $\|N\|_F$ are possible measures of nonnormality introduced by Henrici [29]; see $\widetilde{\mu}_3$ and μ_3 in [19]. Here, $\|\cdot\|_F$ denotes the Frobenius norm. These quantities is invariant under shifts of the matrix ($A \to A + \tau I$), but not under scalar multiplication $A \to \gamma A$. In contrast, the measures $\|N\|/\|A\|$ and $\|N\|/\|\Lambda\|$ (minimized over all Schur forms) are invariant under scalar multiplication but not under shifts. Following [9, 10], we will call measures of nonnormality that are invariant under scalar multiplication *relative* measures. In [9, 10], two such measures are given: $\|N\|_F/\|\Lambda\|_F$ and $\|AA^* - A^*A\|_F/\|A^2\|_F$, where the latter would perhaps be more natural.

Denote the singular values of A by $\sigma_1 \ge \cdots \ge \sigma_n$, and arrange the eigenvalues according to their moduli: $|\lambda_1| \ge \cdots \ge |\lambda_n|$. In view of Proposition 2.3,

ratio (2.5) is closely connected to a comparison of the largest singular value ($\sigma_1 = \|A\|$) and the largest eigenvalue in modulus. The measure of nonnormality μ_4 in [19], introduced by Ruhe [75, Thm. 1], is

$$\max_i \ |\sigma_i - |\lambda_i| \,|.$$

Restricting to the case $i = 1$, we might define a relative version of this quantity as

$$\|A\| \,/\, \rho(A).$$

Both ratios $r(A)/\rho(A)$ and $\|A\|/\rho(A)$ can be seen as relative measures of nonnormality, as will become clear in the next proposition, in which some connections between the various measures are given. When an inequality involves $\|N\|$, it holds for all possible N in a Schur form (2.7).

In the following proposition we give some properties of ratio (2.5) in terms of other measures of nonnormality.

Proposition 2.5. *If $\rho(A) > 0$; then*

$$\text{(a)} \quad \max\left\{1, \ \frac{1}{2}\,\frac{\|A\|}{\rho(A)}\right\} \le \frac{r(A)}{\rho(A)} \le \min\left\{1 + \frac{\|N\|}{\rho(A)}, \ \frac{r(A)}{|\,\|A\| - \|N\|\,|}\right\},$$

$$\text{(b)} \quad 1 - \frac{\rho(A)}{r(A)} \le 1 - \frac{\rho(A)}{\|A\|} \le \frac{\|N\|}{\|A\|} \le 1 + \frac{\rho(A)}{\|A\|} \le 1 + \frac{\rho(A)}{r(A)},$$

and if A is diagonalizable with $A = X\Lambda X^{-1}$, then

$$\text{(c)} \quad \frac{\|A\|}{\rho(A)} \le \kappa(X).$$

Proof. The first two inequalities in (a) follow from Proposition 2.3 and the fact that $W(A)$ contains all eigenvalues. The last inequality follows from Proposition 2.3 and some triangular inequalities: note that $\|\Lambda\| = \rho(A)$, and since $Q^*AQ = N + \Lambda$, we have $\|A\| - \rho(A) \le \|N\| \le \|A\| + \rho(A)$, and $|\,\|A\| - \|N\|\,| \le \rho(A) \le \|A\|$. We can use similar arguments for (b).

Part (c) was noted in [11], where it was also concluded that if (2.5) is large then any eigenvector basis of A is ill conditioned, which means that A is far from normal. □

In the following example we illustrate the various quantities using a well-known matrix with real eigenvalues. We will also show that ratio (2.5) can become arbitrarily large by scaling a symmetric matrix.

Example 2.2. Let A be the $n \times n$ tridiagonal matrix with stencil $[1, 0, 1]$. It is well known that $\lambda_{\min} \approx -2$ and $\lambda_{\max} \approx 2$. Let $D_\alpha = \operatorname{diag}(1, \alpha, \ldots, \alpha^{n-1})$ be a diagonal scaling matrix, and define $A_\alpha = D_\alpha^{-1} A D_\alpha$. Then A_α has stencil $[\alpha^{-1}, 0, \alpha]$ and, it may be checked that, for $\alpha \to \infty$,

$$r(A_\alpha) \sim \alpha, \quad \frac{r(A_\alpha)}{\rho(A_\alpha)} \sim \tfrac{1}{2}\alpha, \quad \frac{\|A_\alpha\|}{\rho(A_\alpha)} \sim \tfrac{1}{2}\alpha, \quad \frac{\|N_\alpha\|}{\|\Lambda_\alpha\|} \sim \tfrac{1}{2}\alpha, \quad \frac{\|N_\alpha\|}{\|A_\alpha\|} \to 1.$$

In particular, we see that ratio (2.5) can get arbitrarily large by scaling. However, in this case balancing the matrix may improve the situation. For instance, for $n = 10$ and $\alpha = 10$, we have $r(A_{10})/\rho(A_{10}) \approx 5.05$. After applying the spbalance routine (see next section), this ratio reduces slightly to 4.38. (Note that in this case [69] and Matlab's balance give the same balancing as spbalance since no permutations are carried out.) For $\alpha = 100$ the scaling gives a clearer improvement. In this case, $r(A_{100})/\rho(A_{100}) \approx 50.0$, while spbalance gives a reduction to 4.01.

During our experiments we encountered the following interesting simple, but unfortunate situation.

Example 2.3. Consider the 2×2 matrix

$$A_\varepsilon = \begin{bmatrix} 1 & 1 \\ -1 & -1+\varepsilon \end{bmatrix}.$$

Then, for $\varepsilon \to 0$, it may be checked that $\rho(A_\varepsilon) \sim \sqrt{\varepsilon}$, while $r(A_\varepsilon) \to 1$. Therefore, the ratio (2.5) can also get arbitrarily large in this case, but now scaling will yield no improvement, since the norms of the first row and column, and the norms of the second row and column are equal.

2.2.3 Scaling and the field of values in the D^2-inner product

Let B be a symmetric positive definite matrix. Instead of the standard inner product $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}^* \boldsymbol{x}$, we now consider the B-inner product $(\boldsymbol{x}, \boldsymbol{y})_B = \boldsymbol{y}^* B \boldsymbol{x}$. In particular, we look at the inner product induced by D^2, where D is the diagonal scaling matrix with positive diagonal entries.

Interestingly enough, the next result gives a connection between the field of values of the balanced matrix, and the field of values of the original matrix in the D^2-inner product. This gives another elegant interpretation of the field of values of the scaled matrix.

Definition 2.6. We define the field of values $W_B(A)$ with respect to the B-inner product to be the set

$$\left\{ \frac{\boldsymbol{x}^* B A \boldsymbol{x}}{\boldsymbol{x}^* B \boldsymbol{x}} \mid \boldsymbol{x} \in \mathbb{C}^n, \ \|\boldsymbol{x}\| = 1 \right\}.$$

Proposition 2.7. *Let D be a symmetric positive definite matrix. The field of values $W_{D^{-2}}(A)$ with respect to the D^{-2}-inner product is identical to the standard field of values of the scaled matrix $D^{-1}AD$.*

Proof. This follows directly from the fact that, with $\boldsymbol{y} = D\boldsymbol{x}$,

$$\frac{\boldsymbol{x}^* D^{-2} A \boldsymbol{x}}{\boldsymbol{x}^* D^{-2} \boldsymbol{x}} = \frac{\boldsymbol{y}^* D^{-1} A D \boldsymbol{y}}{\boldsymbol{y}^* \boldsymbol{y}}$$

for all nonzero $\boldsymbol{x}$. □

Similarly, the $W_{D^2}(A)$-field of values obtained by the D^2-inner product is equal to the standard field of values of the scaled matrix DAD^{-1}.

After these mainly theoretical properties, we now switch our focus to practical matrix balancing techniques.

2.3 Existing scaling methods and a new implementation

We will review several known scaling methods and present a new implementation of spbalance.

2.3.1 Matlab's balance

First, we consider a well-established scaling technique implemented in the LAPACK routine xGEBAL and Matlab function balance. It goes back to work by Osborne [63] and Parlett and Reinsch [69]. The idea is to find a matrix D, a permutation of a diagonal matrix, that scales the rows and columns of A in a given norm, for instance the 1-norm or 2-norm, such that

$$\|D^{-1}AD\,\boldsymbol{e}_i\| \approx \|\boldsymbol{e}_i^*\, D^{-1}AD\|, \qquad i = 1, \ldots, n.$$

We used Matlab's implementation balance, which claims to render a matrix that "has, as nearly as possible, approximately equal row and column norms". We found in experiments that this is certainly not always the case.[1] However, the balanced matrix may have a (sometimes much) smaller norm (in norms such as the 1-norm, 2-norm, ∞-norm, or the Frobenius norm).

In Figure 2.2(a), we plot the spectrum and field of values of the balanced matrix B for $A =$ tols4000. We conclude that $W(B)$ is an immensely improved eigenvalue inclusion region; cf. Figure 2.1(b).

[1] Note that for the balanced tols4000 matrix, the ratio of the norms of the rows and corresponding columns could still be as large as ≈ 340, both in the 1-norm and 2-norm!

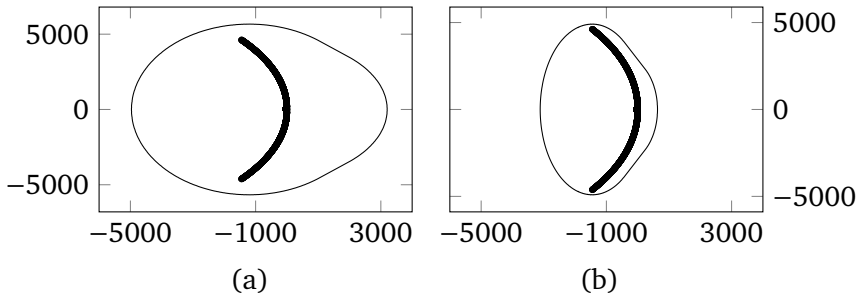


Figure 2.2: (a) Spectrum and fields of values $W(B)$ (solid) for the scaled matrix B for tols4000 by Matlab's balance. (b) Idem for our implementation spbalance for sparse matrices.

However, Matlab's function balance is currently available for dense matrices only. This was a motivation for Chen and Demmel [16] to develop several balancing methods for large sparse matrices. We will consider these and other balancing methods in the next subsections.

2.3.2 Sparse balancing

The function spbalance for the balancing of sparse matrices improves on Matlab's balance: spbalance finds the strongly connected components of a directed graph whose adjacency matrix has the same structure as A and then sorts the components using a topological sort. Chen and Demmel provide an implementation[2] of spbalance. We made a memory-efficient implementation ourselves, which is also suitable for 64-bit systems. The function spbalance seems to improve on Matlab's balance since the permutation algorithm used in balance is a special case of the more general strongly connected components algorithm; see [16] for details.

In Figure 2.2(b), we see that we get a tighter spectral inclusion region with a different scaled matrix. In Section 2.5 we will see that our spbalance implementation gives excellent results for field of values type inclusion regions. Note that Figure 2.2 plots the fields of values for the large balanced matrix; in the numerical experiments in Section 2.5 we consider the fields of values of the projection of the (balanced) matrices onto a Krylov space. This makes the generation of a spectral inclusion region computationally even more attractive (cf. Section 2.1).

[2]http://www.cs.pomona.edu/~tzuyi/Research/Balancing/

2.3.3 Chen and Demmel's Krylov balancing

In this and the next subsection we review three Krylov balancing methods introduced by Chen and Demmel [16]. Let $|A|$ be the matrix with entries $|a_{ij}|$, and assume that $|A|$ is irreducible. Then, since $|A|$ is an irreducible nonnegative matrix, it has a unique maximal eigenvalue which is real and positive, called the Perron eigenvalue. The corresponding right and left eigenvectors, which we will denote by $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, are called the Perron vectors. It can be shown that the scaling

$$D = \operatorname{diag}(\boldsymbol{x}_i) = \operatorname{diag}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \tag{2.8}$$

achieves the lower bound on $\|D^{-1}AD\|_\infty$, which is $\rho(|A|)$ [16].

For large sparse matrices, approximating the Perron vectors of $|A|$ may not be feasible if A is not given explicitly but instead by a routine carrying out the matrix-vector product. Therefore, Chen and Demmel [16] introduced scaling methods based on the approximation of the Perron vectors by a power-type method on A instead of on $|A|$. The name Krylov balancing was chosen since the method uses the results of some matrix-vector products of A with vectors with elements ± 1. This approach was motivated by several statistical observations in [16]. Algorithm 2.1 generates a scaling matrix by approximating the right Perron vector $\boldsymbol{x}$.

Algorithm 2.1 (Krylov balancing (function KrylovAz))**.**
Input: A, $m \in \mathbb{N}$ (number of steps, default: 5).
Output: A scaling matrix D such that $D^{-1}AD$ is (hopefully) better scaled than A.
1. $D = I$
2. **for** $k = 1, \ldots, m$ **do**
3. $\quad \boldsymbol{z} =$ vector of random ± 1s
4. $\quad \boldsymbol{p} = D^{-1}AD\,\boldsymbol{z}$
5. $\quad D = D \cdot \operatorname{diag}(|\boldsymbol{p}_i|)$

We note that Algorithm 2.1 performs relatively poorly in the numerical experiments in Section 2.5. One of the possible reasons being that the elements of A are only accessed via a matrix-vector product, which is also an advantage of the method; see Section 2.5 for a further comparison and discussion. Also, in spite of the name, we remark that this method does not use a Krylov space, but some matrix-vector products with A, although we may still project the scaled matrix onto a Krylov space; see Section 2.5.

2.3.4 Chen and Demmel's two-sided Krylov balancing

Chen and Demmel [16] also introduce a scaling method based on a two-sided Krylov method, exploiting matrix-vector products with both A and A^T. Similar to Algorithm 2.1, this method is matrix free (it does not need A explicitly but only the action of A and A^T applied to a vector); in contrast to Algorithm 2.1, this method needs actions with the transpose. Chen and Demmel expect that this method will generally give better results than the one-sided method of Section 2.3.3, since it uses more information.

They show that with the scaling

$$D = \operatorname{diag}(\sqrt{\boldsymbol{x}_1/\boldsymbol{y}_1}, \ldots, \sqrt{\boldsymbol{x}_n/\boldsymbol{y}_n}) \tag{2.9}$$

$B = D^{-1}AD$ is balanced in the weighted sense, that is, $B\boldsymbol{w} = B^T\boldsymbol{w}$ for $\boldsymbol{w} = D^{-1}\boldsymbol{x}$. Moreover, the Perron eigenvalue is perfectly scaled (has eigenvalue condition number equal to 1), since the right and left Perron vectors coincide for the scaled matrix.

Proposition 2.8. *With the choice* (2.9)*, B is balanced in the weighted sense.*

Proof. [16]; see also [7]. □

Algorithm 2.2 generates a scaling matrix by approximating both $\boldsymbol{x}$ and $\boldsymbol{y}$ by using matrix-vector products with A and A^T instead of $|A|$ and $|A|^T$, which may not be available. Still, in Section 2.5 we will also carry out experiments where $|A|$ and/or $|A|^T$ are used to approximate the Perron vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Algorithm 2.2 (Two-sided Krylov balancing (function KrylovATz))**.**
Input: A, $m \in \mathbb{N}$ (number of steps, default: 5).
Output: A scaling matrix D such that $D^{-1}AD$ is (hopefully) better scaled than A.

1. $D = I$
2. **for** $k = 1, \ldots, m$ **do**
3. $\quad \boldsymbol{z} =$ vector of random ±1s
4. $\quad \boldsymbol{p} = D^{-1}AD\,\boldsymbol{z}$
5. $\quad \boldsymbol{q} = DA^TD^{-1}\,\boldsymbol{z}$
6. $\quad D = D \cdot \operatorname{diag}(\sqrt{|\boldsymbol{p}_i/\boldsymbol{q}_i|})$

Finally, Chen and Demmel [16] propose to add a *cutoff* value to Algorithm 2.2: in line 5, if $|\boldsymbol{p}_i|$ or $|\boldsymbol{q}_i|$ is smaller than the cutoff value, then $\sqrt{|\boldsymbol{p}_i/\boldsymbol{q}_i|}$ is replaced by 1. The chosen default cutoff value is 10^{-8}. We will see in Section 2.5 that this Cutoff approach is superior to KrylovAz and KrylovATz for our purposes. Still, spbalance and the method of the next section are even better.

2.4 A new Krylov balancing approach

We now propose a new Krylov scaling approach, which is simple yet seems to be powerful in numerical experiments. The key idea is to first carry out a modest number of Krylov steps, giving the Arnoldi decomposition of the type (2.2). Subsequently, we scale the Hessenberg matrix H_k instead of the original matrix A. We note that this balancing of the Hessenberg matrix does not imply a scaling of the original matrix.

In our numerical examples, we use our implementation spbalance to scale the matrix. In fact, in the numerical experiments of Section 2.5 it turns out that often the Hessenberg matrices are well scaled already, so that scaling is performed only in a limited number of cases. Moreover, in the examples this balancing is often "modest": the maximal condition number of the scaling matrix D in the experiments is equal to 8, except for tols4000, for which it is very large ($\mathcal{O}(10^9)$).

Another option for the scaling of the Hessenberg matrix would be to use the weighted Perron scaling of Section 2.3.4. However, in the numerical experiments this approach sometimes performs poorly because the norm of the skew-Hermitian part $\frac{1}{2}(H_k - H_k^*)$ increases a lot, resulting in a field of values that is much too large; see also Section 2.5.

We present pseudocode for the new approach in Algorithm 2.3.

Algorithm 2.3 (Krylov balancing for a field of values type (approximate) eigenvalue inclusion region)**.**
Input: A, starting vector $\boldsymbol{w}_1$ (default: random), k (subspace dimension, default: 20).
Output: An approximate eigenvalue inclusion region.
1. Compute Arnoldi decomposition (2.2).
2. Scale H_k giving $\widetilde{H}_k = D^{-1} H_k D$ (see text).
3. Compute $W(\widetilde{H}_k)$.

We note that in Step 2 any sensible scaling method can be used; in our experiments, we used our spbalance implementation. The new method has several advantages. It is very simple, and has only two parameters with sensible defaults values. Balancing is necessary only for a small Hessenberg matrix and therefore virtually for free. In many cases, the Hessenberg matrices are already well scaled. Although Watkins [91] discusses situations where scaling of Hessenberg matrices may be disadvantageous for computing eigenvalues, we will see in the numerical experiments in the next section that scaling may be a good idea to generate high-quality spectral inclusion regions based on the field of values.

We summarize some properties of the various scaling methods in Table 2.1.

Table 2.1: Properties of the different scaling methods.

Method	Scales	Matrix free	Transpose free
Az (Alg. 1)	A	✓	✓
ATz, Cutoff (Alg. 2)	A	✓	—
spbalance	A	—	—
Krylov balancing (Alg. 3)	H_k	✓	✓

2.5 Numerical experiments

In this section, we will give the results of some extensive numerical tests, which we hope are interesting for the community. We test a number of large sparse matrices from the Matrix Market [60]. As an indication of the quality of the field of values of the scaled matrices as eigenvalue inclusion regions we use the ratio (2.5), which in an ideal case is 1, which corresponds to the optimal situation that the radius of the eigenvalue inclusion region is as large as the spectral radius. The values of $r(A)$ used in the fourth column are approximated by runs of Matlab's eigs for 16 different angles (cf. (2.1)). The matrices have various ratios (2.5), including some equal to 1.

In Table 2.2, we give the results for three different scaling methods by Chen and Demmel: KrylovAz, KrylovATz, and its variant Cutoff [16] as described in Sections 2.3.3 and 2.3.4. We first scale the original large matrix A to a matrix B, and then approximate $W(B)$ by $W(H_{20})$, where H_{20} is the 20×20 Hessenberg matrix generated by an Arnoldi decomposition (2.2) on B: see the columns labeled "w/o" (meaning "without extra scaling of the Hessenberg matrix"). We also add a new idea, by considering a second scaling, of the resulting Hessenberg matrix, by spbalance. This yields a scaled matrix $\widetilde{H}_{20}$ and a corresponding field of values $W(\widetilde{H}_{20})$. Therefore, this involves a double scaling: a scaling of the original matrix, and a scaling of the generated Hessenberg matrix (columns labeled "with"). In all cases, a random starting vector is used to generate the Krylov spaces. The total time needed for all experiments is also given for each method. Note that the main computational costs consist of 5 matrix-vector products (MVs) for KrylovAz and 10 for KrylovATz and Cutoff for the scaling, and then 20 MVs to obtain the Hessenberg matrix. As we see, KrylovAz and KrylovATz fail in several cases or yield poor results, also for matrices with ratio (2.5) close to 1. There are two possible reasons for this. The first is that the methods break down since the diagonal matrix contains a zero or because of a division by zero (see Step 3 in Algorithm 2.1 and Steps 4–6 in Algorithm 2.2). In the second case a scaling

matrix D is rendered, but the ratio $r(H_{20})/\rho(A)$ is poor (> 10).

The Cutoff method is more reliable. Of all methods in Table 2.2, Cutoff with double scaling gives the best results.

We now move to Table 2.3. In the column labeled Perron, we use the Perron scaling (2.8), where we approximate the Perron vector $\boldsymbol{x}$ by 5 steps of the power method with $|A|$. Subsequently, 20 steps of Arnoldi are carried out, with or without an extra scaling of the Hessenberg matrix with spbalance. In the column labeled "2-Perron", the same is done, but now the Perron scaling (2.9) is exploited, where the Perron vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are approximated by 5 steps of the power method with $|A|$ and $|A|^T$, respectively.

The next column shows the results of our implementation of spbalance. For this method, additional scaling of the Hessenberg matrices turns out to be unnecessary. Finally we display the results of our new K+B approach: Krylov balancing by scaling of the Hessenberg matrix only. We first carry out an Arnoldi process followed by a (possible) balancing of the Hessenberg matrix. The main computational costs consist of 5 MVs for Perron and 10 MVs for 2-Perron for the scaling, and then 20 MVs to obtain the Hessenberg matrix. The complexity of spbalance is $\mathcal{O}(n + \text{nnz})$ for the balancing, where "nnz" stands for the number of nonzeros (see [16]). Again, we need an additional 20 MVs for the Arnoldi method. The K+B approach is the cheapest approach with only 20 MVs, as the scaling of the Hessenberg matrix is practically for free. We stress the important fact that spbalance is available only when A is given explicitly, all other scaling methods can also be used if A is given via a matrix-vector product.

As we see, the spbalance and K+B methods give near optimal results in most cases. The K+B approach has difficulties with the tolosa matrices, for which spbalance performs very well. We note that for tols2000, the K+B suffers a problem similar to the one mentioned in Example 2.3. K+B is slightly better than spbalance for the tridiagonal matrix. We also give the ratio $r(\widetilde{H}_{20})/\rho(\widetilde{H}_{20})$ in the last column of the table, where $\widetilde{H}_{20}$ is the scaled Hessenberg matrix of the K+B method. (Note that $\rho(\widetilde{H}_{20}) = \rho(H_{20})$; see Proposition 2.2.) If $r(\widetilde{H}_{20})$ is much larger than $\rho(\widetilde{H}_{20})$, such as for tols2000 and tols4000, this may be viewed as a hint that the field of values may not be an excellent spectral inclusion region. Therefore, the ratio $r(\widetilde{H}_{20})/\rho(\widetilde{H}_{20})$ may serve as an indicator of the reliability of the approximate inclusion region.

As an example, we plot the approximate inclusion regions of three of the methods for the matrix olm5000 in Figure 2.3. The region generated by the K+B method is clearly the best.

Some of the matrices in the test set are unsymmetric but have all real eigenvalues. As an interesting note, we remark that there is no easy test to check if the

Table 2.2: Ratios $r(H_{20})/\rho(A)$ for some large sparse matrices A, where H_{20} denotes the scaled and projected matrix A. Three different methods are considered: Chen and Demmel's KrylovAz, KrylovATz, and Cutoff methods. First the matrix is scaled, then $k = 20$ steps of Krylov are carried out. In the w/o columns, the field of values is computed without further scaling of the Hessenberg matrix, for the with columns, the Hessenberg matrix undergoes a further scaling with spbalance. By "—", failures are indicated (breakdown of scaling method, or ratio > 10; see text).

				Az		ATz		Cutoff	
Matrix	n	$\rho(A)$	$r(A)/\rho(A)$	**w/o**	**with**	**w/o**	**with**	**w/o**	**with**
af23560	23560	$2.63 \cdot 10^2$	1.55	—	1.20	1.08	1.06	1.29	1.15
cry10000	10000	$4.22 \cdot 10^4$	1.01	3.87	1.17	1.53	1.26	1.00	1.00
dw8192	8192	$1.10 \cdot 10^2$	1.00	4.64	1.19	1.03	1.01	1.12	1.03
grcar10000	10000	$2.97 \cdot 10^0$	1.09	3.88	1.24	—	4.39	2.52	1.38
grcar10000+ 5I	10000	$7.74 \cdot 10^0$	1.03	—	—	1.97	1.30	1.78	1.12
grcar10000+10I	10000	$1.27 \cdot 10^1$	1.02	3.48	1.21	1.29	1.13	1.06	1.08
memplus	17758	$1.49 \cdot 10^0$	1.01	3.36	1.05	—	9.66	1.00	1.00
olm5000	5000	$2.53 \cdot 10^5$	5.04	—	8.52	1.64	1.23	1.54	1.16
rw5151	5151	$1.00 \cdot 10^0$	1.09	—	—	9.79	2.55	2.22	1.17
sherman2	1080	$2.71 \cdot 10^8$	5.00	1.09	1.06	1.27	1.07	1.14	1.04
sherman3	5005	$6.76 \cdot 10^6$	1.00	1.06	1.04	1.00	1.00	1.00	1.00
sherman5	3312	$5.95 \cdot 10^0$	4.51	3.55	1.26	1.00	1.00	1.00	1.00
tols2000	2000	$2.44 \cdot 10^3$	$1.2 \cdot 10^3$	—	1.61	2.94	1.50	1.33	1.14
tols4000	4000	$4.84 \cdot 10^3$	$2.4 \cdot 10^3$	—	1.96	2.55	1.36	5.08	1.68
tridiag(0.1,0,10)	2000	$2.00 \cdot 10^0$	$5.0 \cdot 10^2$	5.00	4.98	5.00	4.98	5.00	4.98
utm5940	5940	$1.76 \cdot 10^0$	1.18	7.20	1.27	—	—	1.36	1.15
Time (sec)				1.1		5.9		1.0	

Table 2.3: Ratios $r(H_{20})/\rho(A)$ for some large sparse matrices A, where H_{20} denotes the scaled and projected matrix A. We display the results of four different methods: scaling with an approximation of the Perron vector $\boldsymbol{x}$ (Perron); scaling with approximations of the Perron vector $\boldsymbol{x}$ and $\boldsymbol{y}$ (2-Perron, two-sided Perron), both without or with additional scaling of the resulting Hessenberg matrix; our implementation of spbalance; and our new "Krylov and balance" (K+B) approach. By "—" failures are indicated (breakdown of scaling method or ratio > 10; see text). The last column displays the ratio $r(\widetilde{H}_{20})/\rho(\widetilde{H}_{20})$, where $\widetilde{H}_{20}$ is the scaled Hessenberg matrix of the K+B method.

	Perron		2-Perron		spbal	K+B	$\frac{r(\widetilde{H}_{20})}{\rho(\widetilde{H}_{20})}$
Matrix	**w/o**	**with**	**w/o**	**with**			
af23560	7.53	1.34	1.10	1.12	1.01	1.02	1.02
cry10000	1.25	1.18	1.01	1.01	1.00	1.00	1.00
dw8192	2.27	1.73	1.00	1.00	1.00	1.00	1.00
grcar10000	—	2.85	—	7.05	1.08	1.08	1.10
grcar10000+ 5I	—	2.48	2.55	1.64	1.03	1.03	1.05
grcar10000+10I	9.45	1.72	1.50	1.10	1.02	1.02	1.03
memplus	1.30	1.04	1.00	1.00	1.00	1.00	1.00
olm5000	6.19	1.96	1.18	1.08	1.00	1.01	1.01
rw5151	—	2.68	1.40	1.29	0.98	0.98	1.06
sherman2	5.97	3.53	1.00	1.00	1.00	1.03	1.03
sherman3	1.39	1.16	1.00	1.00	1.00	1.00	1.00
sherman5	3.73	1.12	1.00	1.00	1.00	1.02	1.02
tols2000	1.40	1.24	1.02	1.03	1.01	—	14.7
tols4000	1.08	1.09	1.01	1.03	1.01	3.27	2.90
tridiag(0.1,0,10)	—	—	—	9.05	5.00	4.98	1.13
utm5940	9.98	1.15	0.98	1.08	0.95	0.94	1.02
Time (sec)	0.9		1.0		2.2	0.7	

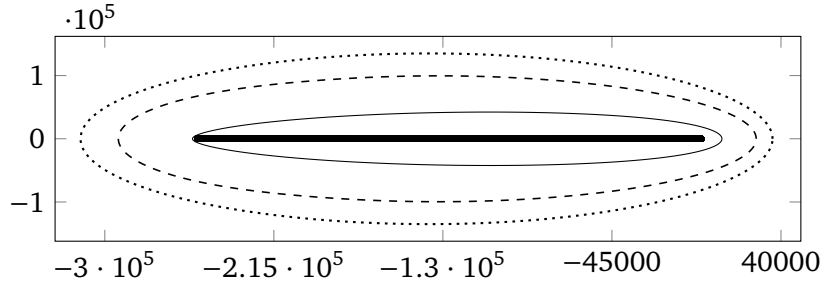


Figure 2.3: Spectrum and fields of values $W(\widetilde{H}_{20})$ (solid) for the new method: the scaled Hessenberg matrix $\widetilde{H}_{20}$, for the matrix olm5000. Also shown are the approximate inclusion regions obtained by the Cutoff (dash) and KrylovATz (dots) methods, where the Hessenberg matrices are also scaled.

spectrum is real. In particular, the field of values type inclusion regions cannot "feel" that the eigenvalues are all real. Indeed, for a matrix with real eigenvalues, the field of values may be arbitrarily large "in the direction of the imaginary axis".

For instance, for an upper triangular matrix with real eigenvalues, the skew-Hermitian part $\mathcal{S} = \frac{1}{2}(A - A^*)$ may have an arbitrarily large norm. This means that both

$$|\min\{\omega \mid i\omega \in W(A)\}| \quad \text{and} \quad |\max\{\omega \mid i\omega \in W(A)\}|$$

may be arbitrarily large. In Figure 2.3 for olm5000 (which is not upper triangular), we see examples of this phenomenon. The family of matrices A_α in Example 2.2 is another example: while A_α has the same spectrum independent of α, the field of values gets arbitrarily large for $\alpha \to \infty$. For matrices with real eigenvalues, we observe in numerical tests that Perron scaling of the Hessenberg matrix may significantly worsen the results, i.e., yield a spectral inclusion region that is much larger in the imaginary direction. This is one of the reasons that we opt for spbalance to scale the Hessenberg matrices.

2.6 Conclusion

For large sparse matrices, the field of values may be efficiently approximated by projection onto Krylov spaces. The resulting sets are approximate spectral inclusion regions and may often be of good quality. However, for some matrices this eigenvalue inclusion region may be much too large.

For these cases, we have shown that scaling may be a very helpful technique to generate tight spectral inclusion regions based on a field of values. We have reviewed various existing balancing methods of the original large matrix: the

KrylovAz, KrylovATz, and Cutoff methods. We have also considered two scaling methods by using approximate Perron vectors of the matrices $|A|$ and/or $|A|^T$. We are able to improve the results of these methods by scaling their resulting Hessenberg matrices from the Arnoldi decompositions. This therefore involves a double scaling: a scaling of the original matrix, and a scaling of the generated Hessenberg matrix. Of the mentioned approaches, the Cutoff method with extra Hessenberg scaling gives the best results.

Subsequently, we have proposed a new implementation of spbalance and a new promising scaling approach by only scaling the small Hessenberg matrix generated by an Arnoldi process. While the "Krylov and balance" (K+B) approach is simple and cheap, it provides equally good results compared with spbalance for almost all cases and better results than the other methods for most examples. We stress that this "Krylov and balance" technique involves a scaling of the Hessenberg matrix but does not imply a scaling of the original matrix.

Note that Matlab's balance and the spbalance method only are applicable if A is given in explicit form; the other scaling methods, including the new K+B approach, are also suitable for matrix-vector products given by functions. Moreover, besides being matrix-free, the K+B approach has the additional advantage of being transpose-free.

In fact, we believe that the combination of an Arnoldi decomposition, matrix balancing of the original matrix or the Hessenberg matrix, and the generation of the field of values of the (scaled) Hessenberg matrix yield an approximate eigenvalue inclusion region that may be very hard to beat both in quality and efficiency. We would like to stress the astonishing result that very good eigenvalue inclusion regions for large sparse matrices may be obtained with just a dozen of matrix-vector products. This is surprising since accurately finding just one eigenvalue may cost hundreds or even thousands of matrix-vector products. As an interesting illustration, we note that the eigenvalues of the 10000×10000 grcar matrix are so sensitive that state-of-the-art existing numerical methods, such as Matlab's eigs or Krylov–Schur [81], are unable to find these even modestly accurately. However, obtaining a high-quality inclusion region for *all* eigenvalues is very well possible!

Finally, the field of values of a scaled matrix is always convex, and it may therefore contain large regions with no eigenvalues. We remark that the method of this chapter may be combined with eigenvalue *exclusion* regions as described in [36]. The intersection of the field of values with one or more exclusion regions, which results in a non-convex inclusion region, may even provide (much) smaller spectral inclusion regions.

Chapter 3

Krylov–Schur-type restarts for the two-sided Arnoldi method

Abstract. We consider the two-sided Arnoldi method and propose a two-sided Krylov–Schur type restarting method. We discuss the restart for standard Rayleigh–Ritz extraction as well as harmonic Rayleigh–Ritz extraction. Additionally, we provide error bounds for Ritz values and Ritz vectors in the context of oblique projections and present generalizations of, e.g., the Bauer–Fike theorem and Saad's theorem. Applications of the two-sided Krylov–Schur method include the simultaneous computation of left and right eigenvectors and the computation of eigenvalue condition numbers. We demonstrate how the method can be used to find the least sensitive eigenvalues of a nonnormal matrix and how to approximate pseudospectra by using left and right shift-invariant subspaces. The results demonstrate that significant improvements in quality can be obtained over approximations with the (one-sided) Krylov–Schur method.

Key words. Two-sided Krylov–Schur, Krylov–Schur, two-sided Arnoldi, dual Arnoldi, implicit restart, harmonic two-sided extraction, eigenvalue condition number, pseudospectra, least sensitive eigenvalues.

AMS subject classification. 15A18, 15A23, 65F15, 65F50.

3.1 Introduction

The two-sided Lanczos algorithm (cf., e.g., [78, Sec. 6.4]) is an important alternative to the Arnoldi method (cf., e.g., [78, Sec. 6.2]) for nonnormal matrices. The former uses short three-term recurrences at the expense of double the number of matrix-vector multiplications. But if one wants the eigenvectors, then either the bases must be stored, or they must be computed in a second run. This means that either the storage needed for two-sided Lanczos becomes roughly twice that of Arnoldi, or the number of matrix-vector multiplications doubles again. Moreover, in practice re-biorthogonalization is often necessary because of the loss of biorthogonality in finite precision arithmetic. The accuracy and stability of the computed bases may be improved by using the two-sided Arnoldi method, proposed by Ruhe [76], to replace biorthonormal by orthonormal bases. In this

chapter, we propose an efficient restarting technique for two-sided Arnoldi, inspired by the Krylov–Schur algorithm [81, 84]. We also investigate perturbation and convergence properties using error bounds for Ritz values and Ritz vectors in the context of oblique projections.

There already are generalizations of the Krylov–Schur method, for example, for Hamiltonian matrices and the product eigenproblem by Kressner [50, 51], as well as a block method for symmetric matrices by Zhou and Saad [96], a version for unitary eigenproblems by Roden and Watkins [18], and a method for the truncated SVD by Stoll [86]. Jaimoukha and Kasenally [42] present a restarted two-sided Krylov method for model order reduction; however, their method uses projections to remove unstable elements without changing the initial vectors. Instead, we are interested in arbitrary exterior eigenvalues of general nonnormal matrices and allow our method to implicitly modify the initial vectors.

Applications that may benefit from two-sided Krylov–Schur include those where the condition number of eigenvalues is important and those where both the left and right eigenvectors are desired. In particular, we use two-sided Krylov–Schur to find eigenvalues with the lowest condition numbers and to approximate pseudospectra. The former may be useful to compute the least sensitive eigenvalues of parameterized matrices, or the most reliable eigenvalues of matrices containing uncertain data. The latter application can provide insight into the (worst-case) behavior of eigenvalues under perturbations. Our contribution is a new type of approximation using two shift-invariant subspaces.

The rest of this chapter is organized as follows. First we review Stewart's Krylov–Schur method in Section 3.2. Then we introduce a new two-sided Krylov–Schur method in Section 3.3 and its harmonic counterpart in Section 3.4. Section 3.5 explores the relation between two-sided Arnoldi and two-sided Lanczos. The focus of Section 3.6 is on perturbation and convergence theory, and that of Section 3.7 on distance properties. Finally, Sections 3.8 and 3.9 contain the numerical experiments and conclusions.

3.2 One-sided Krylov–Schur

The Krylov–Schur method by Stewart [81, 84] combines the Arnoldi method with a restarting mechanism based on the Schur decomposition. Let A be an $n \times n$ matrix and consider the Krylov subspace

$$\mathcal{V}_\ell = \mathcal{K}_\ell(A, \boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots, A^{\ell-1}\boldsymbol{v}\}, \tag{3.1}$$

where $\ell \ll n$. It is well known that the Arnoldi method creates a basis V_ℓ for $\mathcal{V}_\ell$ satisfying the decomposition

$$AV_\ell = V_\ell H_\ell + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* = V_{\ell+1}\underline{H}_\ell, \tag{3.2}$$

where $V_{k+1} = [V_\ell \ \boldsymbol{v}_{\ell+1}]$ has orthonormal columns and $\underline{H}_\ell = [H_\ell;\ \boldsymbol{h}_\ell^*]$ is upper-Hessenberg. When $\underline{H}_\ell$ is an arbitrary full-rank $(\ell+1)\times\ell$ matrix, it is nevertheless possible to transform the decomposition into the described upper-Hessenberg form [81, Thm 2.2]. To perform a restart, compute the Schur decomposition

$$H_\ell = QSQ^*,$$

where Q is unitary and S is upper triangular, and define $\widehat{V}_\ell = V_\ell Q$ and $\widehat{\boldsymbol{h}}_\ell = Q^*\boldsymbol{h}_\ell$; then

$$A\widehat{V}_\ell = \widehat{V}_\ell S + \boldsymbol{v}_{\ell+1}\widehat{\boldsymbol{h}}_\ell^*.$$

Partition the above decomposition as

$$A\begin{bmatrix}\widehat{V}_1 & \widehat{V}_2\end{bmatrix} = \begin{bmatrix}\widehat{V}_1 & \widehat{V}_2\end{bmatrix}\begin{bmatrix}S_{11} & S_{12}\\ 0 & S_{22}\end{bmatrix} + \boldsymbol{v}_{\ell+1}\begin{bmatrix}\widehat{\boldsymbol{h}}_1^* & \widehat{\boldsymbol{h}}_2^*\end{bmatrix},$$

where it may be assumed without loss of generality that the desired eigenvalues of H_ℓ are along the diagonal of S_{11}. Lastly, truncate to obtain

$$A\widehat{V}_1 = \widehat{V}_1 S_{11} + \boldsymbol{v}_{\ell+1}\widehat{\boldsymbol{h}}_1^*.$$

We summarize the one-sided Krylov–Schur method in Algorithm 3.1.

Algorithm 3.1 (One-sided Krylov–Schur [81])**.**
Input: $A \in \mathbb{C}^{n\times n}$, starting vector $\boldsymbol{v}_1$, minimum and maximum dimensions m and ℓ, tolerance tol.
Output: V_{m+1} and H_m such that $\|AV_m - V_m H_m\| \le$ tol.

1. **for** *number of restarts* **do**
2. Expand the Krylov decomposition to $AV_\ell = V_\ell H_\ell + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^*$.
3. Compute $H_\ell = QSQ^*$, and partition $Q = [Q_1\ Q_2]$ and $S = \left[\begin{smallmatrix}S_{11} & S_{12}\\ & S_{22}\end{smallmatrix}\right]$.
4. Set $V_m = V_\ell Q_1$, $H_m = S_{11}$, and $\boldsymbol{h}_m = Q_1^*\boldsymbol{h}_\ell$.
5. **if** $\|\boldsymbol{h}_m\| \le$ tol **then break**
6. **end**

The Krylov–Schur method extracts approximations to eigenvalues and eigenvectors using the standard Galerkin condition

$$AV_\ell\boldsymbol{c} - \theta V_\ell\boldsymbol{c} \perp \mathcal{V}_\ell.$$

However, it is also possible to extract eigenvalues by choosing a different test subspace $\mathcal{U}_\ell$ and imposing the modified Galerkin condition

$$AV_\ell \mathbf{c} - \theta V_\ell \mathbf{c} \perp \mathcal{U}_\ell.$$

In this case a Krylov–Schur type restart is more elaborate [84], but allows, for instance, restarts with harmonic Ritz value extraction. The following two sections show how the one-sided Krylov–Schur restart can be modified to restart either two-sided Arnoldi or harmonic two-sided Arnoldi.

3.3 Two-sided Krylov–Schur

In this section we derive the two-sided Krylov–Schur method. Assume A is a nonnormal $n \times n$ matrix, and consider the right Krylov subspace in (3.1) together with the left Krylov subspace

$$\mathcal{W}_\ell = \mathcal{K}_\ell(A^*, \boldsymbol{w}) = \operatorname{span}\{\boldsymbol{w}, A^*\boldsymbol{w}, (A^*)^2\boldsymbol{w}, \ldots, (A^*)^{\ell-1}\boldsymbol{w}\}.$$

The two-sided Arnoldi method proposed by Ruhe [76], and later as a block method by Cullum and Zhang [17], independently generates orthonormal bases for the right search space $\mathcal{V}_\ell$ and the left search space $\mathcal{W}_\ell$. This can be done by applying the (one-sided) Arnoldi method twice. Let the generated bases be denoted by V_ℓ and W_ℓ respectively; then the following relations are satisfied:

$$\begin{aligned} AV_\ell &= V_\ell H_\ell + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* = V_{\ell+1}\underline{H}_\ell, \\ A^*W_\ell &= W_\ell K_\ell + \boldsymbol{w}_{\ell+1}\boldsymbol{k}_\ell^* = W_{\ell+1}\underline{K}_\ell, \end{aligned} \tag{3.3}$$

where both $V_{\ell+1} = [V_\ell \ \boldsymbol{v}_{\ell+1}]$ and $W_{\ell+1} = [W_\ell \ \boldsymbol{w}_{\ell+1}]$ consist of orthonormal columns. The next step is to extract approximate eigenvectors and eigenvalues using the two-sided Rayleigh–Ritz method. For this purpose, the matrices H_ℓ and K_ℓ are modified to be Rayleigh quotients of A and A^*, respectively. The Rayleigh quotient of a matrix M is defined here as Y^*MX for a full column rank matrix X with left inverse Y^* (cf., e.g., [82, p. 252]). Assuming $W_\ell^*V_\ell$ is nonsingular, let

$$\begin{aligned} \widetilde{H}_\ell &= H_\ell + (W_\ell^*V_\ell)^{-1}W_\ell^*\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^*, \\ \widetilde{K}_\ell &= K_\ell + (V_\ell^*W_\ell)^{-1}V_\ell^*\boldsymbol{w}_{\ell+1}\boldsymbol{k}_\ell^* \end{aligned}$$

and

$$\begin{aligned} \widetilde{\boldsymbol{v}}_{\ell+1} &= (I - V_\ell(W_\ell^*V_\ell)^{-1}W_\ell^*)\boldsymbol{v}_{\ell+1}, \\ \widetilde{\boldsymbol{w}}_{\ell+1} &= (I - W_\ell(V_\ell^*W_\ell)^{-1}V_\ell^*)\boldsymbol{w}_{\ell+1}. \end{aligned}$$

Then it is possible to rewrite (3.3) as

$$\begin{aligned} AV_\ell &= V_\ell \widetilde{H}_\ell + \widetilde{\boldsymbol{v}}_{\ell+1} \boldsymbol{h}_\ell^*, \\ A^* W_\ell &= W_\ell \widetilde{K}_\ell + \widetilde{\boldsymbol{w}}_{\ell+1} \boldsymbol{k}_\ell^*. \end{aligned} \tag{3.4}$$

Since $\widetilde{\boldsymbol{v}}_{\ell+1}$ is orthogonal to W_ℓ and $\widetilde{\boldsymbol{w}}_{\ell+1}$ is orthogonal to V_ℓ, it follows that

$$\begin{aligned} \widetilde{H}_\ell &= (W_\ell^* V_\ell)^{-1} W_\ell^* A V_\ell, \\ \widetilde{K}_\ell &= (V_\ell^* W_\ell)^{-1} V_\ell^* A^* W_\ell. \end{aligned} \tag{3.5}$$

Because $(W_\ell^* V_\ell)^{-1} W_\ell^*$ and $(V_\ell^* W_\ell)^{-1} V_\ell^*$ are left inverses of V_ℓ and W_ℓ respectively, we recognize $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ as Rayleigh quotients of A and A^*. Furthermore, the eigenvalues of $\widetilde{H}_\ell$ and of $\widetilde{K}_\ell^*$ satisfy the following proposition due to Cullum and Zhang [17].

Proposition 3.1. *Using the previous definitions, $\widetilde{H}_\ell$ and $\widetilde{K}_\ell^*$ are similar if $W_\ell^* V_\ell$ is nonsingular.*

Proof. Since $(V_\ell^* W_\ell)^{-*} = (W_\ell^* V_\ell)^{-1}$, it is easy to deduce from (3.5) that

$$(W_\ell^* V_\ell) \widetilde{H}_\ell = W_\ell^* A V_\ell = \widetilde{K}_\ell^* (W_\ell^* V_\ell).$$

□

If $W_\ell^* V_\ell$ is singular, then one can perform additional steps of the Krylov process, or remove vectors until $W_j^* V_j$ is nonsingular for some j.

We are now ready to derive a new two-sided restarting approach inspired by (one-sided) harmonic Krylov–Schur restarts [84]. Consider the Schur decompositions

$$\widetilde{H}_\ell = QSQ^* \qquad \text{and} \qquad \widetilde{K}_\ell = ZTZ^*,$$

where the eigenvalues of $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ are ordered along the diagonals of S and T, respectively, and are such that $s_{jj} = t_{jj}^*$. If such a pairing cannot be found due to roundoff errors, then an alternative is to sort s_{jj} and t_{jj}^* independently based on some desirable quantity such as their distance to a target, the size of the real part, etc. Substituting the above Krylov–Schur decompositions in (3.4) yields

$$\begin{aligned} AV_\ell &= V_\ell QSQ^* + \widetilde{\boldsymbol{v}}_{\ell+1} \boldsymbol{h}_\ell^*, \\ A^* W_\ell &= W_\ell ZTZ^* + \widetilde{\boldsymbol{w}}_{\ell+1} \boldsymbol{k}_\ell^*. \end{aligned}$$

Let $\widehat{V}_\ell = V_\ell Q$, $\widehat{W}_\ell = W_\ell Z$, $\widetilde{\boldsymbol{h}}_\ell = Q^* \boldsymbol{h}_\ell$, and $\widetilde{\boldsymbol{k}}_\ell = Z^* \boldsymbol{k}_\ell$, so that

$$\begin{aligned} A\widehat{V}_\ell &= \widehat{V}_\ell S + \widetilde{\boldsymbol{v}}_{\ell+1}\widetilde{\boldsymbol{h}}_\ell^*, \\ A^*\widehat{W}_\ell &= \widehat{W}_\ell T + \widetilde{\boldsymbol{w}}_{\ell+1}\widetilde{\boldsymbol{k}}_\ell^*, \end{aligned} \tag{3.6}$$

and in partitioned form

$$\begin{aligned} A\begin{bmatrix}\widehat{V}_1 & \widehat{V}_2\end{bmatrix} &= \begin{bmatrix}\widehat{V}_1 & \widehat{V}_2\end{bmatrix}\begin{bmatrix}S_{11} & S_{12}\\ 0 & S_{22}\end{bmatrix} + \widetilde{\boldsymbol{v}}_{\ell+1}\begin{bmatrix}\widetilde{\boldsymbol{h}}_1^* & \widetilde{\boldsymbol{h}}_2^*\end{bmatrix}, \\ A^*\begin{bmatrix}\widehat{W}_1 & \widehat{W}_2\end{bmatrix} &= \begin{bmatrix}\widehat{W}_1 & \widehat{W}_2\end{bmatrix}\begin{bmatrix}T_{11} & T_{12}\\ 0 & T_{22}\end{bmatrix} + \widetilde{\boldsymbol{w}}_{\ell+1}\begin{bmatrix}\widetilde{\boldsymbol{k}}_1^* & \widetilde{\boldsymbol{k}}_2^*\end{bmatrix}. \end{aligned} \tag{3.7}$$

We can now truncate the partitioned decompositions to

$$\begin{aligned} A\widehat{V}_1 &= \widehat{V}_1 S_{11} + \widetilde{\boldsymbol{v}}_{\ell+1}\widetilde{\boldsymbol{h}}_1^*, \\ A^*\widehat{W}_1 &= \widehat{W}_1 T_{11} + \widetilde{\boldsymbol{w}}_{\ell+1}\widetilde{\boldsymbol{k}}_1^*. \end{aligned} \tag{3.8}$$

The vector $\widetilde{\boldsymbol{v}}_{\ell+1}$ is in general not orthogonal to $\widehat{V}_1$, and $\widetilde{\boldsymbol{w}}_{\ell+1}$ is not orthogonal to $\widehat{W}_1$. This problem can be remedied by computing

$$\begin{aligned} A\widehat{V}_1 &= \widehat{V}_1 \widehat{H} + \widehat{\boldsymbol{v}}_{\ell+1}\widehat{\boldsymbol{h}}_1^*, \\ A^*\widehat{W}_1 &= \widehat{W}_1 \widehat{K} + \widehat{\boldsymbol{w}}_{\ell+1}\widehat{\boldsymbol{k}}_1^*, \end{aligned} \tag{3.9}$$

where $[\widehat{V}_1\ \widehat{\boldsymbol{v}}_{\ell+1}]$ and $[\widehat{W}_1\ \widehat{\boldsymbol{w}}_{\ell+1}]$ have orthonormal columns, and

$$\begin{aligned} \widehat{H} &= S_{11} + (\widehat{V}_1^*\widetilde{\boldsymbol{v}}_{\ell+1})\widetilde{\boldsymbol{h}}_1^*, \\ \widehat{K} &= T_{11} + (\widehat{W}_1^*\widetilde{\boldsymbol{w}}_{\ell+1})\widetilde{\boldsymbol{k}}_1^*, \\ \widehat{\boldsymbol{v}}_{\ell+1} &= \|(I - \widehat{V}_1\widehat{V}_1^*)\widetilde{\boldsymbol{v}}_{\ell+1}\|^{-1}(I - \widehat{V}_1\widehat{V}_1^*)\widetilde{\boldsymbol{v}}_{\ell+1}, \\ \widehat{\boldsymbol{w}}_{\ell+1} &= \|(I - \widehat{W}_1\widehat{W}_1^*)\widetilde{\boldsymbol{w}}_{\ell+1}\|^{-1}(I - \widehat{W}_1\widehat{W}_1^*)\widetilde{\boldsymbol{w}}_{\ell+1}, \\ \widehat{\boldsymbol{h}}_1 &= \|(I - \widehat{V}_1\widehat{V}_1^*)\widetilde{\boldsymbol{v}}_{\ell+1}\|\,\widetilde{\boldsymbol{h}}_1, \\ \widehat{\boldsymbol{k}}_1 &= \|(I - \widehat{W}_1\widehat{W}_1^*)\widetilde{\boldsymbol{w}}_{\ell+1}\|\,\widetilde{\boldsymbol{k}}_1. \end{aligned}$$

From here, the search spaces spanned by $\widehat{V}_1$ and $\widehat{W}_1$ can be expanded independently using the (one-sided) Arnoldi method. Below in Algorithm 3.2 we summarize the two-sided Krylov–Schur method for the computation of approximate right and left invariant subspaces.

Algorithm 3.2 (Two-sided Krylov–Schur)**.**
Input: Nonnormal $A \in \mathbb{C}^{n\times n}$, starting vectors $\boldsymbol{v}_1$ and $\boldsymbol{w}_1$, minimum and maximum dimensions m and ℓ.
Output: V_{m+1}, W_{m+1}, $\underline{H}_m$, and $\underline{K}_m$ such that $AV_m = V_{m+1}\underline{H}_m \approx V_m H_m$ and $A^*W_m = W_{m+1}\underline{K}_m \approx W_m K_m$.

1. **for** *number of restarts* **do**
2. Expand the Krylov decompositions to

$$AV_\ell = V_\ell H_\ell + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^*, \qquad A^*W_\ell = W_\ell K_\ell + \boldsymbol{w}_{\ell+1}\boldsymbol{k}_\ell^*,$$

   using the Arnoldi process, and update $M_\ell = W_\ell^* V_\ell$.
3. Compute $H_\ell = H_\ell + M_\ell^{-1}W_\ell^*\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^*$ and $\boldsymbol{v}_{\ell+1} = \boldsymbol{v}_{\ell+1} - V_\ell M_\ell^{-1}W_\ell^*\boldsymbol{v}_{\ell+1}$.
4. Compute $K_\ell = K_\ell + M_\ell^{-*}V_\ell^*\boldsymbol{w}_{\ell+1}\boldsymbol{k}_\ell^*$ and $\boldsymbol{w}_{\ell+1} = \boldsymbol{w}_{\ell+1} - W_\ell M_\ell^{-*}V_\ell^*\boldsymbol{w}_{\ell+1}$.
5. Compute the Schur decompositions $H_\ell = QSQ^*$ and $K_\ell = ZTZ^*$.
6. Partition Q, S, Z, and T as in (3.7).
7. Set $V_m = V_\ell Q_1$, $H_m = S_{11}$, $\boldsymbol{h}_m = Q_1^*\boldsymbol{b}_\ell$.
8. Set $W_m = W_\ell Z_1$, $K_m = T_{11}$, $\boldsymbol{k}_m = Z_1^*\boldsymbol{c}_\ell$.
9. Set $M_m = Z_1^* M_\ell Q_1$.
10. Set $H_m = H_m + (V_m^*\boldsymbol{v}_{\ell+1})\boldsymbol{h}_m^*$, $\boldsymbol{v}_{m+1} = (I - V_m V_m^*)\boldsymbol{v}_{\ell+1}$,
    $\boldsymbol{h}_m = \|\boldsymbol{v}_{m+1}\|\boldsymbol{h}_m$, $\boldsymbol{v}_{m+1} = \boldsymbol{v}_{m+1}/\|\boldsymbol{v}_{m+1}\|$.
11. Set $K_m = K_m + (W_m^*\boldsymbol{w}_{\ell+1})\boldsymbol{k}_m^*$, $\boldsymbol{w}_{m+1} = (I - W_m W_m^*)\boldsymbol{w}_{\ell+1}$,
    $\boldsymbol{k}_m = \|\boldsymbol{w}_{m+1}\|\boldsymbol{k}_m$, $\boldsymbol{w}_{m+1} = \boldsymbol{w}_{m+1}/\|\boldsymbol{w}_{m+1}\|$.
12. **if** *converged (e.g., cf.* (3.10)*)* **then break**
13. **end**

Usually, the oblique projections in steps 3 and 4 of Algorithm 3.2 must be repeated at least once in practice [85, Sec. 7], which can be seen as the oblique analogue of reorthogonalization. Step 12 requires extra attention as well, since properly measuring the convergence in two-sided Krylov–Schur is more complex than in one-sided Krylov–Schur. Luckily, we can rely on the work of Kahan, Parlett, and Jiang [46], who investigated the convergence of two-sided Lanczos and derived a set of termination criteria. We describe some of their results below.

For two unit vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ with $\boldsymbol{w}^*\boldsymbol{v} \neq 0$, define the two-sided Rayleigh quotient

$$\rho = \rho(\boldsymbol{v}, \boldsymbol{w}^*) = \frac{\boldsymbol{w}^* A \boldsymbol{v}}{\boldsymbol{w}^*\boldsymbol{v}}$$

and the right and left residuals

$$\boldsymbol{r} = (A - \rho I)\boldsymbol{v} \qquad \text{and} \qquad \boldsymbol{s} = (A - \rho I)^*\boldsymbol{w};$$

then the partial derivatives of ρ are

$$\partial_{\boldsymbol{v}}\rho(\boldsymbol{v}, \boldsymbol{w}^*) = \frac{\boldsymbol{s}^*}{\boldsymbol{w}^*\boldsymbol{v}} \qquad \text{and} \qquad \partial_{\boldsymbol{w}^*}\rho(\boldsymbol{v}, \boldsymbol{w}^*) = \frac{\boldsymbol{r}}{\boldsymbol{w}^*\boldsymbol{v}}.$$

Consequently, ρ should not be used as an approximate eigenvalue unless the value of $\max\{\|\boldsymbol{s}\|, \|\boldsymbol{r}\|\}/|\boldsymbol{w}^*\boldsymbol{v}|$ is sufficiently small relative to ρ. An additional result shows that for an eigenvalue λ near ρ the bound

$$|\lambda - \rho| \leq \kappa(\lambda)\|E\| + \mathcal{O}(\|E\|^2)$$

holds [46, Sec. 5], where $\|E\| \leq \max\{\|\boldsymbol{r}\|, \|\boldsymbol{s}\|\}$ and $\kappa(\lambda)$ is the condition number of λ. While $\kappa(\lambda)$ is unknown in practice, it can be approximated with $|\boldsymbol{w}^*\boldsymbol{v}|^{-1}$; see, e.g., Theorem 3.3 and [46, Sec. 8].

In the context of two-sided Krylov–Schur we compute

$$\begin{aligned} \widetilde{H}_\ell C &= C\Theta \qquad (\text{with } \Theta = \operatorname{diag}(\theta_1, \ldots, \theta_\ell) \text{ and } \|\boldsymbol{c}_j\| = 1), \\ \widetilde{K}_\ell D &= D\Gamma \qquad (\text{with } \Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell) \text{ and } \|\boldsymbol{d}_j\| = 1), \end{aligned}$$

where Θ would equal Γ^* in exact arithmetic, and take $\boldsymbol{v}_j = V_\ell \boldsymbol{c}_j$ and $\boldsymbol{w}_j = W_\ell \boldsymbol{d}_j$ as the right and left Ritz vectors. Then the Rayleigh quotients $\rho_j = \rho(\boldsymbol{v}_j, \boldsymbol{w}_j)$ can be shown to equal the Ritz values $\theta_j = \bar{\gamma}_j$, so that the residuals satisfy

$$\begin{aligned} \boldsymbol{r}_j &= \|(A - \rho_j I)\boldsymbol{v}_j\| = \|(A - \theta_j I)\boldsymbol{v}_j\| = \|\widetilde{\boldsymbol{v}}_{\ell+1}\| \; |\boldsymbol{h}_\ell^* \boldsymbol{c}_j|, \\ \boldsymbol{s}_j &= \|(A - \rho_j I)^* \boldsymbol{w}_j\| = \|(A - \bar{\gamma}_j I)^* \boldsymbol{w}_j\| = \|\widetilde{\boldsymbol{w}}_{\ell+1}\| \; |\boldsymbol{k}_\ell^* \boldsymbol{d}_j|. \end{aligned}$$

Using the sensitivities $\kappa_j = |\boldsymbol{w}_j^* \boldsymbol{v}_j|^{-1}$, we terminate, for example, if the relative error

$$\frac{\kappa_j}{|\rho_j|} \max\{\|\boldsymbol{r}_j\|, \|\boldsymbol{s}_j\|\} \tag{3.10}$$

is sufficiently small for the desired value(s) of ρ_j. In our tests using finite precision arithmetic, it was advantageous to use the right eigenvectors of both $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ instead of using the left and right eigenvectors of only one of the two. In some cases it may also be numerically preferable to use the Rayleigh quotients ρ_j instead of the Ritz values θ_j and γ_j [80, Sec. 4].

In this section we have discussed a two-sided version of the Krylov–Schur algorithm in addition to a suitable stopping criterion. In the subsequent section we focus on a two-sided Krylov–Schur restart with harmonic eigenvalue extraction.

3.4 Harmonic two-sided Krylov–Schur

The eigenvalue extraction from the previous section corresponds to imposing the Galerkin conditions

$$\begin{aligned} AV_\ell \boldsymbol{c} - \theta V_\ell \boldsymbol{c} &\perp \mathcal{W}_\ell, \\ A^* W_\ell \boldsymbol{d} - \eta W_\ell \boldsymbol{d} &\perp \mathcal{V}_\ell. \end{aligned} \tag{3.11}$$

Suppose one is interested in interior eigenvalues near a target τ not equal to an eigenvalue. These eigenvalues are exterior eigenvalues of the shifted and inverted matrix $(A - \tau I)^{-1}$; hence, it makes sense for the extraction to impose the Petrov–Galerkin conditions

$$\begin{aligned} (A - \tau I)^{-1} \boldsymbol{v} - (\theta - \tau)^{-1} \boldsymbol{v} &\perp \mathcal{U}_1, \\ (A - \tau I)^{-*} \boldsymbol{w} - (\eta - \tau)^{-*} \boldsymbol{w} &\perp \mathcal{U}_2 \end{aligned}$$

for certain test spaces $\mathcal{U}_1$ and $\mathcal{U}_2$; see also [31, Sec. 3.2]. It is straightforward to show that the choice $\boldsymbol{v} = V_\ell \boldsymbol{c}$, $\boldsymbol{w} = W_\ell \boldsymbol{d}$, $\mathcal{U}_1 = (A - \tau I)^* \mathcal{W}_\ell$, and $\mathcal{U}_2 = (A - \tau I)\mathcal{V}_\ell$ is equivalent with (3.11). For harmonic two-sided Rayleigh–Ritz one can take the test spaces

$$\begin{aligned} \mathcal{U}_1 &= (A - \tau I)^* (A - \tau I)^* \mathcal{W}_\ell, \\ \mathcal{U}_2 &= (A - \tau I)(A - \tau I)\mathcal{V}_\ell \end{aligned}$$

to obtain the equivalent conditions

$$\begin{aligned} (A - \theta I)\boldsymbol{v} &\perp (A - \tau I)^* \mathcal{W}_\ell, \\ (A - \eta I)^* \boldsymbol{w} &\perp (A - \tau I)\mathcal{V}_\ell \end{aligned}$$

after some manipulation. The former conditions lead to the generalized eigenvalue problems

$$\begin{aligned} W_\ell^* (A - \tau I) A V_\ell \boldsymbol{c} &= \theta W_\ell^* (A - \tau I) V_\ell \boldsymbol{c}, \\ V_\ell^* (A - \tau I)^* A^* W_\ell \boldsymbol{d} &= \eta V_\ell^* (A - \tau I)^* W_\ell \boldsymbol{d}. \end{aligned}$$

Since these are two conjugated generalized eigenvalue problems, it follows that they are satisfied by ℓ quadruples $(\theta, \eta, \boldsymbol{c}, \boldsymbol{d})$ with $\eta = \bar{\theta}$. If $W_\ell^*(A - \tau I)V_\ell$ is nonsingular, we receive the equivalent eigenvalue problems

$$\begin{aligned} (W_\ell^* (A - \tau I) V_\ell)^{-1} W_\ell^* (A - \tau I) A V_\ell \boldsymbol{c} &= \theta \boldsymbol{c}, \\ (V_\ell^* (A - \tau I)^* W_\ell)^{-1} V_\ell^* (A - \tau I)^* A^* W_\ell \boldsymbol{d} &= \bar{\theta} \boldsymbol{d}. \end{aligned}$$

Substituting the Arnoldi decompositions from (3.3) produces

$$\widetilde{H}_\ell \boldsymbol{c} = \theta \boldsymbol{c} \qquad \text{and} \qquad \widetilde{K}_\ell \boldsymbol{d} = \bar{\theta} \boldsymbol{d},$$

where $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ are rank-1 updates of H_ℓ and K_ℓ, defined by

$$\begin{aligned}
\widetilde{H}_\ell &= H_\ell + ((\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^* V_\ell)^{-1} (\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^* \boldsymbol{v}_{\ell+1} \boldsymbol{h}_\ell^*, \\
\widetilde{K}_\ell &= K_\ell + ((\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^* W_\ell)^{-1} (\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^* \boldsymbol{w}_{\ell+1} \boldsymbol{k}_\ell^*,
\end{aligned}$$

and $\underline{I}$ is the identity matrix with an additional zero bottom row. Next, define

$$\begin{aligned}
\widetilde{\boldsymbol{v}}_{\ell+1} &= (I - V_\ell((\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^* V_\ell)^{-1} (\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^*) \boldsymbol{v}_{\ell+1}, \\
\widetilde{\boldsymbol{w}}_{\ell+1} &= (I - W_\ell((\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^* W_\ell)^{-1} (\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^*) \boldsymbol{w}_{\ell+1},
\end{aligned}$$

so that

$$\begin{aligned}
AV_\ell &= V_\ell \widetilde{H}_\ell + \widetilde{\boldsymbol{v}}_{\ell+1} \boldsymbol{h}_\ell^*, \\
A^* W_\ell &= W_\ell \widetilde{K}_\ell + \widetilde{\boldsymbol{w}}_{\ell+1} \boldsymbol{k}_\ell^*,
\end{aligned}$$

and $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ are the Rayleigh quotients

$$\begin{aligned}
\widetilde{H}_\ell &= (W_\ell^*(A - \tau I)V_\ell)^{-1} W_\ell^*(A - \tau I)AV_\ell \\
&= ((\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^* V_\ell)^{-1} (\underline{K}_\ell - \bar{\tau}\underline{I})^* W_{\ell+1}^* AV_\ell, \\
\widetilde{K}_\ell &= (V_\ell^*(A - \tau I)^* W_\ell)^{-1} V_\ell^*(A - \tau I)^* A^* W_\ell \\
&= ((\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^* W_\ell)^{-1} (\underline{H}_\ell - \tau\underline{I})^* V_{\ell+1}^* A^* W_\ell.
\end{aligned} \tag{3.12}$$

As in Proposition 3.1, the eigenvalues of the $\widetilde{H}_\ell$ and $\widetilde{K}_\ell^*$ from this section coincide.

Proposition 3.2. *If $W_\ell^*(A - \tau I)V_\ell$ is nonsingular, then $\widetilde{H}_\ell$ and $\widetilde{K}_\ell^*$ in* (3.12) *are similar.*

Proof. The proof is comparable to the proof of Proposition 3.1, but with W_ℓ replaced by $(A - \tau I)^* W_\ell$. From (3.12) and $A(A - \tau I) = (A - \tau I)A$, it follows that

$$(W_\ell^*(A - \tau I)V_\ell)\widetilde{H}_\ell = W_\ell^*(A - \tau I)AV_\ell = \widetilde{K}_\ell^*(W_\ell^*(A - \tau I)^* V_\ell).$$

□

At this point we can compute Schur decompositions of $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ and continue analogously to the previous section. Algorithm 3.3 summarizes the harmonic two-sided Krylov–Schur method for the determination of approximate right and left invariant subspaces.

Algorithm 3.3 (Harmonic two-sided Krylov–Schur)**.**
Input: Nonnormal $A \in \mathbb{C}^{n\times n}$, starting vectors $\boldsymbol{v}_1$ and $\boldsymbol{w}_1$, minimum and maximum dimensions m and ℓ, target τ.
Output: V_{m+1}, W_{m+1}, $\underline{H}_m$, and $\underline{K}_m$ such that $AV_m = V_{m+1}\underline{H}_m \approx V_m H_m$ and $A^*W_m = W_{m+1}\underline{K}_m \approx W_m K_m$.

1. **for** *number of restarts* **do**
2. Expand the Krylov decompositions to

$$AV_\ell = V_\ell H_\ell + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^*, \qquad A^*W_\ell = W_\ell K_\ell + \boldsymbol{w}_{\ell+1}\boldsymbol{k}_\ell^*,$$

   using the Arnoldi process and update $M_{\ell+1} = W_{\ell+1}^* V_{\ell+1}$.
3. Compute $QR = \underline{K}_\ell - \bar{\tau}\underline{I}$, and set $\boldsymbol{p} = (Q^* M_{\ell+1,\ell})^{-1} Q^* M_{\ell+1}\boldsymbol{e}_{\ell+1}$.
4. Compute $QR = \underline{H}_\ell - \tau\underline{I}$, and set $\boldsymbol{q} = (Q^* M_{\ell,\ell+1}^*)^{-1} Q^* M_{\ell+1}^*\boldsymbol{e}_{\ell+1}$.
5. Let $H_\ell = H_\ell + \boldsymbol{p}\boldsymbol{h}_\ell^*$ and $\boldsymbol{v}_{\ell+1} = \boldsymbol{v}_{\ell+1} - V_\ell\boldsymbol{p}$.
6. Let $K_\ell = K_\ell + \boldsymbol{q}\boldsymbol{k}_\ell^*$ and $\boldsymbol{w}_{\ell+1} = \boldsymbol{w}_{\ell+1} - W_\ell\boldsymbol{q}$.
7. Compute the Schur decompositions $H_\ell = QSQ^*$ and $K_\ell = ZTZ^*$.
8. Partition Q, S, Z, and T as in (3.7)
9. Set $V_m = V_\ell Q_1$, $H_m = S_{11}$, $\boldsymbol{h}_m = Q_1^*\boldsymbol{b}_\ell$.
10. Set $W_m = W_\ell Z_1$, $K_m = T_{11}$, $\boldsymbol{k}_m = Z_1^*\boldsymbol{c}_\ell$.
11. Set $M_m = Z_1^* M_\ell Q_1$.
12. Set $H_m = H_m + (V_m^*\boldsymbol{v}_{\ell+1})\boldsymbol{h}_m^*$, $\boldsymbol{v}_{m+1} = (I - V_m V_m^*)\boldsymbol{v}_{\ell+1}$,
    $\boldsymbol{h}_m = \|\boldsymbol{v}_{m+1}\|\boldsymbol{h}_m$, $\boldsymbol{v}_{m+1} = \boldsymbol{v}_{m+1}/\|\boldsymbol{v}_{m+1}\|$.
13. Set $K_m = K_m + (W_m^*\boldsymbol{w}_{\ell+1})\boldsymbol{k}_m^*$, $\boldsymbol{w}_{m+1} = (I - W_m W_m^*)\boldsymbol{w}_{\ell+1}$,
    $\boldsymbol{k}_m = \|\boldsymbol{w}_{m+1}\|\boldsymbol{k}_m$, $\boldsymbol{w}_{m+1} = \boldsymbol{w}_{m+1}/\|\boldsymbol{w}_{m+1}\|$.
14. **if** *converged (see the discussion after Algorithm 3.2)* **then break**
15. **end**

Notice that in step 3 of the algorithm we attempt to improve the accuracy by using a QR factorization of $\underline{K}_\ell - \bar{\tau}\underline{I}$, so that we essentially work with the orthonormal basis $W_{\ell+1}Q$ instead of $W_{\ell+1}(\underline{K} - \bar{\tau}\underline{I})$. The approach of step 4 is comparable, and $M_{\ell,\ell+1}$ and $M_{\ell+1,\ell}$ denote the $\ell \times (\ell+1)$ and $(\ell+1)\times\ell$ leading principal submatrices of $M_{\ell+1}$, respectively. In step 14 the same stopping conditions from Section 3.3 can be used; however, in this case, using the Rayleigh quotients ρ_j in place of the Ritz values θ_j and γ_j is recommended (cf. [80, Sec. 4]).

We have now seen the regular and harmonic two-sided Krylov–Schur algorithms. In the following section we discuss the relation between these two algorithms and the two-sided Lanczos algorithm.

3.5 Relation with two-sided Lanczos

As discussed in Section 3.1, the two-sided Lanczos method and the two-sided Arnoldi method are closely related. Specifically, if (3.3) is in upper-Hessenberg form with $\boldsymbol{h}_\ell = \|\boldsymbol{h}_\ell\|\boldsymbol{e}_\ell$ and $\boldsymbol{k}_\ell = \|\boldsymbol{k}_\ell\|\boldsymbol{e}_\ell$, and $W_\ell^* V_\ell$ is nonsingular, then it can be verified that $\widetilde{H}$ and $\widetilde{K}$ in (3.4) are also upper-Hessenberg. Now let $W_\ell^* V_\ell = LU$ be a decomposition into lower and upper triangular factors, and define the biorthonormal bases $\widehat{V}_\ell = VU^{-1}$ and $\widehat{W}_\ell = WL^{-*}$. Furthermore, let $T = U\widetilde{H}_\ell U^{-1}$, then from the proof of Proposition 3.1 it follows that

$$T = U\widetilde{H}_\ell U^{-1} = L^{-1}LU\widetilde{H}_\ell U^{-1} = L^{-1}\widetilde{K}_\ell^* LUU^{-1} = (L^*\widetilde{K}L^{-*})^*.$$

Using this identity, (3.4) can be written as

$$\begin{aligned} A\widehat{V}_\ell &= \widehat{V}_\ell T + \widetilde{\boldsymbol{v}}_{\ell+1}\boldsymbol{h}_\ell^* U^{-1}, \\ A^*\widehat{W}_\ell &= \widehat{W}_\ell T^* + \widetilde{\boldsymbol{w}}_{\ell+1}\boldsymbol{k}_\ell^* L^{-*}, \end{aligned} \tag{3.13}$$

where T is tridiagonal since both $T = U\widetilde{H}U^{-1}$ and $T^* = L^*\widetilde{K}L^{-*}$ are upper-Hessenberg. The decompositions in (3.13) coincide with two-sided Lanczos. Assume for harmonic two-sided Arnoldi that $W_\ell^*(A - \tau I)V_\ell$ is nonsingular, let $W_\ell^*(A - \tau I)V_\ell = LU$ be a decomposition into lower and upper triangular factors, and define $\widehat{V}_\ell = V_\ell U^{-1}$ and $\widehat{W}_\ell = (A - \tau I)^* W_\ell L^{-*}$. Then Proposition 3.2 can be utilized to show that $T = U\widetilde{H}U^{-1} = (L^*\widetilde{K}L^{-*})^*$ is tridiagonal.

To summarize, two-sided Lanczos and two-sided Arnoldi generate bases for the same subspaces, although two-sided Lanczos uses biorthonormal bases and short recursions, while two-sided Arnoldi uses orthonormal bases and (full) reorthogonalization. The option to use short recursions with two-sided Lanczos makes it a computationally appealing method in situations where the computational cost or the memory requirements for (full) reorthogonalization would be prohibitive. On the other hand, using biorthonormal bases without (full) reorthogonalization may lead to numerical stability issues. Methods that have been developed to handle these issues include look-ahead techniques, selective reorthogonalization, and the detection of spurious Ritz values, see for instance Bai et al. [3, Section 4.4.4]. However, increases in memory capacity and computational power of computer hardware have diminished the necessity of such methods, and the modern approach is to favor (full) reorthogonalization when stability and accuracy are crucial. Even though two-sided Lanczos can be implemented with full re-biorthogonalization, Stewart [85] provides convincing reasons for preferring orthogonal bases. For example, orthogonal bases tend to be less sensitive to perturbations than biorthogonal bases. Furthermore, if X and Y are biorthonormal,

then the computation of

$$(I - XY^*)\boldsymbol{a}$$

for some vector $\boldsymbol{a}$ may incur a relative error up to $\gamma\|XY^*\|\epsilon$. Here, ϵ is the machine accuracy, and γ is a constant that depends on the accuracy of X and Y. The error bound implies that even if re-biorthogonalization is used, accuracy may be lost, especially if $\|XY^*\|$ is large and if errors accumulate.

Two-sided Krylov–Schur uses orthonormal bases and applies only orthonormal transformations to the bases. Some of the accuracy and stability issues are avoided as a result, especially if two-sided Krylov–Schur is implemented with full reorthogonalization. Unfortunately, we are not entirely clear of all stability issues associated with oblique projections, or more specifically, the terms $(W_\ell^* V_\ell)^{-1}$ for standard extraction and $(W_\ell^*(A - \tau I)V_\ell)^{-1}$ for harmonic extraction. The vectors $\widetilde{\boldsymbol{v}}_{\ell+1}$ and $\widetilde{\boldsymbol{w}}_{\ell+1}$ and the matrices $\widetilde{H}_\ell$ and $\widetilde{K}_\ell$ will depend on the previous matrix inverses, and therefore the computed Schur decompositions as well.

As it turns out, it is possible to avoid the explicit use of $(W_\ell^* V_\ell)^{-1}$ and improve the accuracy of the computations in Algorithms 3.2 and 3.3. For simplicity we consider only two-sided Rayleigh–Ritz extraction and note that the results can be adapted to two-sided harmonic Ritz. Suppose, for the moment, that we are given the orthonormal matrices Q and Z. The objective is to obtain the decompositions in (3.9) from (3.6) without using $(W_\ell^* V_\ell)^{-1}$. The update $\widehat{V}_1 = V_\ell Q_1$ can clearly be computed without using a matrix inverse; now

$$\begin{aligned}
A\widehat{V}_1 &= \widehat{V}_1 S_{11} + \widetilde{\boldsymbol{v}}_{\ell+1}\widetilde{\boldsymbol{h}}_1^* \\
&= \widehat{V}_1 Q_1^* \widetilde{H}_\ell Q_1 + (I - V_\ell (W_\ell^* V_\ell)^{-1} W_\ell^*)\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1 \\
&= \widehat{V}_1 Q_1^* H_\ell Q_1 + V_\ell Q_1 Q_1^* (W_\ell^* V_\ell)^{-1} W_\ell^* \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1 \\
&\quad + (I - V_\ell (W_\ell^* V_\ell)^{-1} W_\ell^*)\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1 \\
&= \widehat{V}_1 Q_1^* H_\ell Q_1 + (I - V_\ell Q_2 Q_2^* (W_\ell^* V_\ell)^{-1} W_\ell^*)\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1.
\end{aligned}$$

It is straightforward to verify that $\widehat{H} = Q_1^* H_\ell Q_1$ and

$$\widehat{\boldsymbol{v}}_{\ell+1}\widehat{\boldsymbol{h}}_1^* = (I - V_\ell Q_2 Q_2^* (W_\ell^* V_\ell)^{-1} W_\ell^*)\boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1.$$

On the other hand,

$$\begin{aligned}
AV_\ell Q_1 - V_\ell Q_1 Q_1^* H_\ell Q_1 &= V_\ell Q_2 Q_2^* H_\ell Q_1 + \boldsymbol{v}_{\ell+1}\boldsymbol{h}_\ell^* Q_1 \\
&= [V_\ell\ \boldsymbol{v}_{\ell+1}] \begin{bmatrix} Q_2 & \\ & 1 \end{bmatrix} \begin{bmatrix} Q_2^* & \\ & 1 \end{bmatrix} \begin{bmatrix} H_\ell \\ \boldsymbol{h}_\ell^* \end{bmatrix} Q_1.
\end{aligned}$$

It follows that it is possible to determine $\widehat{\boldsymbol{v}}_{\ell+1}\widehat{\boldsymbol{h}}_1^*$ by computing a rank-1 approximation $\boldsymbol{a}\boldsymbol{b}^*$ of

$$\begin{bmatrix} Q_2^* & \\ & 1 \end{bmatrix} \begin{bmatrix} H_\ell \\ \boldsymbol{h}_\ell^* \end{bmatrix} Q_1,$$

with $\|\boldsymbol{a}\| = 1$, and setting $\widehat{\boldsymbol{h}}_1 = \boldsymbol{b}$ and

$$\widehat{\boldsymbol{v}}_{\ell+1} = [V_\ell \, \boldsymbol{v}_{\ell+1}] \begin{bmatrix} Q_2 & \\ & 1 \end{bmatrix} \boldsymbol{a}.$$

An alternative is to use the relation

$$(I - V_\ell Q_2 Q_2^* (W_\ell^* V_\ell)^{-1} W_\ell^*) \boldsymbol{v}_{\ell+1} \|\boldsymbol{h}_\ell^* Q_1\|^2 = (V_\ell Q_2 Q_2^* H_\ell Q_1 + \boldsymbol{v}_{\ell+1} \boldsymbol{h}_\ell^* Q_1) Q_1^* \boldsymbol{h}_\ell$$

to determine $\widehat{\boldsymbol{v}}_{\ell+1}$, which is particularly appealing from a computational point of view. In our tests we found that the latter approach was faster and provided the best numerical performance. The vector $\widetilde{\boldsymbol{v}}_{\ell+1}$ is no longer needed with the above approaches, and its computation can be omitted. To summarize, the inverse of $W_\ell^* V_\ell$ can be bypassed once Q is known.

Computing Q without using $(W_\ell^* V_\ell)^{-1}$ is the remaining step. It is possible to avoid the explicit use of the inverse with the QZ decomposition

$$W_\ell^* A V_\ell = P S_\alpha Q^* \qquad \text{and} \qquad W_\ell^* V_\ell = P S_\beta Q^*$$

of the matrix pencil $(W_\ell^* A V_\ell, W_\ell^* V_\ell)$. Here P and Q are orthonormal, S_α and S_β are upper triangular, and $S = S_\beta^{-1} S_\alpha$. The QZ decomposition can be reordered if necessary. In our tests we found that the QZ approach did not improve the accuracy with sufficient significance and reliability to justify the increased computational cost.

In this section we have investigated the relation between two-sided Lanczos and two-sided Krylov–Schur and argued how most of the problems with the former are solved by a proper implementation of the latter.

3.6 Error bounds for Ritz values and Ritz vectors

In previous sections we have discussed the computation of Ritz values, Ritz vectors, and their harmonic counterparts. In this section we investigate the convergence of Ritz values and Ritz vectors with respect to the convergence of the search space to an invariant subspace. We will first focus on the convergence of Ritz values and address the convergence of the Ritz vectors later.

To investigate the convergence of a Ritz value θ to an eigenvalue of A, we could invoke, for instance, the Bauer–Fike theorem (cf., e.g., [78, Thm. 3.6]). The Bauer–Fike theorem is a key result in perturbation theory, and below we present a new two-sided version.

Theorem 3.3 (Two-sided Bauer–Fike)**.** *Suppose that A is diagonalizable such that*

$$A = X\Lambda X^{-1}.$$

Let $(\theta, \boldsymbol{v}, \boldsymbol{w})$ be an approximate eigentriplet of A with $\|\boldsymbol{v}\| = \|\boldsymbol{w}\| = 1$, and define the residuals

$$\boldsymbol{r} = A\boldsymbol{v} - \theta\boldsymbol{v} \qquad \text{and} \qquad \boldsymbol{s}^* = \boldsymbol{w}^*A - \theta\boldsymbol{w}^*.$$

Assume $\boldsymbol{w}^*\boldsymbol{v} \neq 0$ and define $\kappa_\theta = |\boldsymbol{w}^*\boldsymbol{v}|^{-1}$. If the condition number of X is denoted by $\kappa(X)$, then there exists an eigenvalue λ of A such that

$$|\lambda - \theta| \leq \sqrt{\kappa(X)\kappa_\theta \, \|\boldsymbol{r}\| \, \|\boldsymbol{s}\|}.$$

Proof. If θ is an eigenvalue of A the result is clear. Otherwise $A - \theta I$ is nonsingular and

$$|\boldsymbol{w}^*\boldsymbol{v}| = |\boldsymbol{s}^*(A - \theta I)^{-2}\boldsymbol{r}| = |\boldsymbol{s}^*X(\Lambda - \theta I)^{-2}X^{-1}\boldsymbol{r}| \leq \kappa(X) \, \|\boldsymbol{r}\| \, \|\boldsymbol{s}\| \, \|(\Lambda - \theta I)^{-2}\|.$$

Rearranging the terms gives

$$\min_{\mu \in \Lambda(A)} |\mu - \theta|^2 \leq \kappa(X)\kappa_\theta \, \|\boldsymbol{r}\| \, \|\boldsymbol{s}\|.$$

□

In particular, if $\max\{\|\boldsymbol{r}\|, \|\boldsymbol{s}\|\} \longrightarrow 0$, then θ converges to some eigenvalue λ of A, and κ_θ converges to the condition number $\kappa(\lambda)$ of λ. Theorem 3.3 can be used with Ritz vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ to match Ritz values to eigenvalues of A one at a time.

An alternative approach for studying the convergence of Ritz values is through Elsner's theorem [82, p. 38] .

Theorem 3.4 (Elsner [82])**.** *Let the eigenvalues of B be $\lambda_1, \ldots, \lambda_n$ and let the eigenvalues of $B + E$ be $\theta_1, \ldots, \theta_n$. Then there is a permutation $j_1, \ldots, j_n$ of the integers $1, \ldots, n$ such that*

$$|\lambda_i - \theta_{j_i}| \leq 4(\|B\| + \|B + E\|)^{1-1/n}\|E\|^{1/n} \qquad (i = 1, \ldots, n).$$

Hence, if the eigenvalues of B are in the spectrum of A and the eigenvalues of $B + E$ are the computed Ritz values, then $\theta_{j_1}, \ldots, \theta_{j_n}$ converge to $\lambda_1, \ldots, \lambda_n$, when $\|E\| \to 0$. An advantage of using Elsner's theorem is that we can match multiple θs to eigenvalues simultaneously.

At this point it is helpful to introduce notation that allows the uniform treatment of the remainder of this section. Let

$$V = \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \qquad \text{and} \qquad W = \begin{bmatrix} W_1 & W_2 & W_3 \end{bmatrix}$$

be full-rank orthonormal matrices, and introduce the shorthand notation $V_{1,2}$ and $W_{1,2}$ for the first two blocks of V and W respectively. In two-sided Krylov–Schur, the columns of V_1 and W_1 could, for instance, correspond to either the basis vectors retained after truncation or to a subset thereof. The next step is to make V and W biorthonormal, which is where the following proposition comes into play.

Proposition 3.5. *If $W_1^*V_1$ and $W_{1,2}^*V_{1,2}$ are nonsingular, then the 3×3 block LU decomposition of W^*V is given by*

$$L = \begin{bmatrix} W_1^*V_1 & & \\ W_2^*V_1 & W_2^*(I-P_1)V_2 & \\ W_3^*V_1 & W_3^*(I-P_1)V_2 & W_3^*(I-P_{1,2})V_3 \end{bmatrix},$$

$$U = \begin{bmatrix} I & (W_1^*V_1)^{-1}W_1^*V_2 & (W_1^*V_1)^{-1}W_1^*V_3 \\ & I & (W_2^*(I-P_1)V_2)^{-1}W_2^*(I-P_1)V_3 \\ & & I \end{bmatrix},$$

where $P_1 = V_1(W_1^*V_1)^{-1}W_1^*$ and $P_{1,2} = V_{1,2}(W_{1,2}^*V_{1,2})^{-1}W_{1,2}^*$.

Proof. Suppose for the moment that $W_2^*(I-P_1)V_2$ is nonsingular so that U is well defined. For most of the blocks it is straightforward to verify by direct computation that $LU = W^*V$. The only difficult block is

$$W_3^*V_3 = W_3^*P_1V_3 + W_3^*(I-P_1)V_2(W_2^*(I-P_1)V_2)^{-1}W_2^*(I-P_1)V_3 + W_3^*(I-P_{1,2})V_3.$$

To show that equality holds, it suffices to show that $P_{1,2} = Q$, where Q is the projector defined by

$$Q = P_1 + (I-P_1)V_2(W_2^*(I-P_1)V_2)^{-1}W_2^*(I-P_1).$$

From its definition we see that the range of Q must be a subset of the range of $V_{1,2}$, that is $\mathcal{R}(Q) \subset \mathcal{R}(V_{1,2})$, and likewise $\mathcal{R}(Q^*) \subset \mathcal{R}(W_{1,2})$. Furthermore,

notice that $QV_1 = V_1$, $QV_2 = V_2$, $Q^*W_1 = W_1$, and $Q^*W_2 = W_2$. Since a projector is uniquely defined by its column space and its row space, it follows from [85, Thm. 2.2] that $Q = P_{1,2}$.

To prove the Ansatz that $W_2^*(I - P_1)V_2$ is nonsingular, let $L_{1,2}$ and $U_{1,2}$ be the upper-left 2×2 blocks of L and U so that

$$\begin{aligned}\det(W_{1,2}^* V_{1,2}) = \det(L_{1,2}U_{1,2}) &= \det(L_{1,2})\det(U_{1,2}) \\ &= \det(W_1^* V_1)\det(W_2^*(I - P_1)V_2) \neq 0.\end{aligned}$$

□

Suppose that $W_1^*V_1$ and $W_{1,2}^*V_{1,2}$ are nonsingular and that L and U are given by Proposition 3.5; then the matrices defined by

$$\widetilde{V} = VU^{-1} = \begin{bmatrix}\widetilde{V}_1 & \widetilde{V}_2 & \widetilde{V}_3\end{bmatrix} \qquad \text{and} \qquad \widetilde{W} = WL^{-*} = \begin{bmatrix}\widetilde{W}_1 & \widetilde{W}_2 & \widetilde{W}_3\end{bmatrix}$$

are biorthonormal. Furthermore,

$$\widetilde{V}_1 = V_1, \qquad \widetilde{W}_1^* = (W_1^*V_1)^{-1}W_1^*, \qquad I - V_1(W_1^*V_1)^{-1}W_1^* = I - \widetilde{V}_1\widetilde{W}_1^*,$$

and

$$I - V_{1,2}(W_{1,2}^*V_{1,2})^{-1}W_{1,2}^* = I - \widetilde{V}_{1,2}\widetilde{W}_{1,2}^* = \widetilde{V}_3\widetilde{W}_3^*.$$

Assume that

$$S = (W_{1,2}^*V_{1,2})^{-1}W_{1,2}^*AV_{1,2} \qquad \text{and} \qquad T = (V_{1,2}^*W_{1,2})^{-1}V_{1,2}^*A^*W_{1,2}$$

are upper triangular; then an argument similar to the one at the beginning of Section 3.5 shows that

$$\widetilde{W}_{1,2}^*A\widetilde{V}_{1,2} = U_{1,2}SU_{1,2}^{-1} = (L_{1,2}^*TL_{1,2}^{-*})^*$$

is block diagonal, with

$$\begin{aligned}\widetilde{W}_{1,2}^*A\widetilde{V}_{1,2} &= \begin{bmatrix}S_{11} & \\ & S_{22}\end{bmatrix} \\ &= \begin{bmatrix}(W_1^*V_1)^{-1}T_{11}^*(W_1^*V_1) & \\ & (W_2^*(I - P_1)V_2)^{-1}T_{22}^*(W_2^*(I - P_1)V_2\end{bmatrix}.\end{aligned}$$

Finally, we assume for the remainder of this section that $\operatorname{rank}(X) \leq \operatorname{rank}(V_1)$, and we have the following definition.

Definition 3.6. Let $\mathcal{X}$ be an invariant subspace of A such that $A\mathcal{X} \subseteq \mathcal{X}$, and suppose that $[X\, X_\perp]$ is orthonormal, X is a basis of $\mathcal{X}$, and B is such that $AX = XB$. If the spectra of B and $X_\perp^* A X_\perp$ are disjoint, then (B, X) is called a simple orthonormal eigenpair of A.

Using the new notation, we are ready to state a generalization of Jia and Stewart [43, Thm. 4.1], which allows the application of Elsner's theorem to two-sided Arnoldi. The value δ can be interpreted as a measure of the angle between subspaces and will be analyzed in Theorem 3.11 and Proposition 3.13.

Theorem 3.7. *Let (B, X) be a simple orthonormal eigenpair of A. Define $Z = \widetilde{W}_1^* X$ and orthonormalize the columns of Z by setting*

$$\widetilde{Z} = ZQ, \qquad \text{where} \qquad Q = (Z^*Z)^{-1/2}.$$

Then there exists a matrix E satisfying

$$\|E\| = \|\widetilde{W}_1^* A(I - \widetilde{V}_1\widetilde{W}_1^*)XQ\|,$$

such that $(Q^{-1}BQ, \widetilde{Z})$ is an eigenpair of $S_{11} - E$. Furthermore, if $\delta = \|(I - \widetilde{V}_1\widetilde{W}_1^*)X\| < 1$, then

$$\|E\| \leq \|\widetilde{W}_1^* A\| \frac{\delta}{1-\delta}.$$

Proof. For the first part of the proof we multiply $AX = XB$ from the left by $\widetilde{W}_1^*$ and obtain

$$\widetilde{W}_1^* A(\widetilde{V}_1\widetilde{W}_1^* + (I - \widetilde{V}_1\widetilde{W}_1^*))X = \widetilde{W}_1^* XB.$$

Since $\widetilde{W}_1^* A\widetilde{V}_1 = S_{11}$, we can rearrange the terms to get

$$S_{11}\widetilde{W}_1^* X - \widetilde{W}_1^* XB = -\widetilde{W}_1^* A(I - \widetilde{V}_1\widetilde{W}_1^*)X,$$

which we use to define the residual

$$R = S_{11}\widetilde{Z} - \widetilde{Z}Q^{-1}BQ = -\widetilde{W}_1^* A(I - \widetilde{V}_1\widetilde{W}_1^*)XQ$$

and the perturbation matrix $E = R\widetilde{Z}^*$. Then $S_{11} - E$ satisfies

$$(S_{11} - E)\widetilde{Z} = \widetilde{Z}Q^{-1}BQ$$

and

$$\|E\| = \|R\| = \|\widetilde{W}_1^* A(I - \widetilde{V}_1\widetilde{W}_1^*)XQ\|,$$

which concludes the first part of the proof. For the second part of the proof we use the relation

$$\|Q\| = \sigma_{\min}^{-1}(Z) = \sigma_{\min}^{-1}(\widetilde{V}_1\widetilde{W}_1^*X).$$

To compute the smallest singular value of $\widetilde{V}_1\widetilde{W}_1^*X$, observe that

$$1 \leq \min_{\|\boldsymbol{z}\|=1} (\|\widetilde{V}_1\widetilde{W}_1^*X\boldsymbol{z}\| + \|(I - \widetilde{V}_1\widetilde{W}_1^*)X\boldsymbol{z}\|).$$

Therefore,

$$\sigma_{\min}(\widetilde{V}_1\widetilde{W}_1^*X) = \min_{\|\boldsymbol{z}\|=1} \|\widetilde{V}_1\widetilde{W}_1^*X\boldsymbol{z}\| \geq 1 - \max_{\|\boldsymbol{z}\|=1} \|(I - \widetilde{V}_1\widetilde{W}_1^*)X\boldsymbol{z}\| = 1 - \delta,$$

and

$$\|E\| \leq \|\widetilde{W}_1^*A\| \, \|(I - \widetilde{V}_1\widetilde{W}_1^*)\| \, \|Q\| \leq \|\widetilde{W}_1^*A\| \frac{\delta}{1-\delta}.$$

□

The key insight from Theorem 3.7 is that, under mild conditions, there exists a matrix E such that the eigenvalues of $\widetilde{Z}^*(S_{11} - E)\widetilde{Z} = Q^{-1}BQ$ are eigenvalues of A. By subsequently applying Theorem 3.4, the following corollary may be obtained.

Corollary 3.8. *Assume $r = \operatorname{rank}(X) = \operatorname{rank}(V_1)$, let the eigenvalues of B be $\lambda_1, \ldots, \lambda_r$, and let the eigenvalues of S_{11} be $\theta_1, \ldots, \theta_r$. Then there are integers $j_1, \ldots, j_r$ such that*

$$|\lambda_i - \theta_{j_i}| \leq 4(2\|(W_1^*V_1)^{-1}\| \, \|A\| + \|E\|)^{1-1/r}\|E\|^{1/r} \qquad (i = 1, \ldots, r).$$

Hence, if $\|(W_1^*V_1)^{-1}\|$ is asymptotically uniformly bounded and $\|E\| \to 0$, then there are Ritz values that converge to eigenvalues of A. In practice, the assumption on $(W_1^*V_1)^{-1}$ means that the corollary cannot be applied to defective eigenvalues.

The next proof relates the separation between Ritz values and eigenvalues to the convergence of the subspace V_1 to the invariant subspace X of A. The proof uses the following definition of the separation operator.

Definition 3.9. The separation between an $n \times n$ matrix N and an $m \times m$ matrix M is defined by

$$\operatorname{sep}(N, M) = \min_{\|Z\|=1} \|NZ - ZM\|.$$

For more information on the separation operator, see, for example, [82, p. 256].

Theorem 3.10. *Let (B, X) be a simple orthonormal eigenpair of A; then*

$$\operatorname{sep}(\widetilde{V}_1 S_{11} \widetilde{W}_1^*, B) \leq \frac{\|\widetilde{V}_1 \widetilde{W}_1^* A \widetilde{V}_3 \widetilde{W}_3^* X\|}{\|\widetilde{V}_1 \widetilde{W}_1^* X\|}.$$

Proof. Since $AX = XB$ we have that

$$\widetilde{W}_{1,2}^* A \widetilde{V} \widetilde{W}^* X = \widetilde{W}_{1,2}^* X B.$$

Rearranging the terms gives

$$\begin{bmatrix} S_{11} & \\ & S_{22} \end{bmatrix} \begin{bmatrix} \widetilde{W}_1^* X \\ \widetilde{W}_2^* X \end{bmatrix} - \begin{bmatrix} \widetilde{W}_1^* X \\ \widetilde{W}_2^* X \end{bmatrix} B = - \begin{bmatrix} \widetilde{W}_1^* A \widetilde{V}_3 \widetilde{W}_3^* X \\ \widetilde{W}_2^* A \widetilde{V}_3 \widetilde{W}_3^* X \end{bmatrix}. \tag{3.14}$$

From the first block row we see that

$$S_{11} \widetilde{W}_1^* X - \widetilde{W}_1^* X B = -\widetilde{W}_1^* A \widetilde{V}_3 \widetilde{W}_3^* X,$$

and hence

$$(\widetilde{V}_1 S_{11} \widetilde{W}_1^*) \widetilde{V}_1 \widetilde{W}_1^* X - \widetilde{V}_1 \widetilde{W}_1^* X B = -\widetilde{V}_1 \widetilde{W}_1^* A \widetilde{V}_3 \widetilde{W}_3^* X.$$

Using the definition of the separation operator we can now derive the bound

$$\operatorname{sep}(\widetilde{V}_1 S_{11} \widetilde{W}_1^*, B) \|\widetilde{V}_1 \widetilde{W}_1^* X\| \leq \|\widetilde{V}_1 \widetilde{W}_1^* A \widetilde{V}_3 \widetilde{W}_3^* X\|,$$

which concludes the proof. □

Theorem 3.10 tells us that the separation between $\widetilde{V}_1 S_{11} \widetilde{W}_1^*$ and B must go to zero as the span of X becomes contained in the span of $V_{1,2}$; this is true in particular if V_1 converges to X.

It is instructive to determine what can be said about $\|(I - P_1)X\|$ if it is known that $\|(I - P_{1,2})X\| \to 0$. Saad provides a bound in the case of Hermitian matrices; see, for example, [78, Thm 4.6]. Saad's theorem was generalized by Stewart for general matrices in [83]. In [31, Thm. 3] the theorem is further generalized to a two-sided result, but using $\|[\widetilde{W}_2 \ \widetilde{W}_3]^* X\|$ instead of $\|(I - P_1)X\|$, and restricted by the assumption that X is a vector. We therefore state a new two-sided Saad type theorem.

Theorem 3.11. *Let (B, X) be a simple orthonormal eigenpair of A. If $\operatorname{sep}(\widetilde{V}_2 S_{22} \widetilde{W}_2^*, B) > 0$, then*

$$\begin{aligned} \|(I - \widetilde{V}_1 \widetilde{W}_1^*) X\| &\leq \frac{\|\widetilde{V}_2 \widetilde{W}_2^* A \widetilde{V}_3 \widetilde{W}_3^* X\|}{\operatorname{sep}(\widetilde{V}_2 S_{22} \widetilde{W}_2^*, B)} + \|\widetilde{V}_3 \widetilde{W}_3^* X\| \\ &\leq \left(1 + \frac{\|\widetilde{V}_2 \widetilde{W}_2^* A\|}{\operatorname{sep}(\widetilde{V}_2 S_{22} \widetilde{W}_2^*, B)} \right) \|\widetilde{V}_3 \widetilde{W}_3^* X\|. \end{aligned}$$

Proof. From the second block row of (3.14) it follows that

$$S_{22}\widetilde{W}_2^*X - \widetilde{W}_2^*XB = -\widetilde{W}_2^*A\widetilde{V}_3\widetilde{W}_3^*X.$$

Using the fact that $\widetilde{V}_2^*\widetilde{W}_2 = I$, we can write

$$(\widetilde{V}_2S_{22}\widetilde{W}_2^*)\widetilde{V}_2\widetilde{W}_2^*X - \widetilde{V}_2\widetilde{W}_2^*XB = -\widetilde{V}_2\widetilde{W}_2^*A\widetilde{V}_3\widetilde{W}_3^*X,$$

so that

$$\operatorname{sep}(\widetilde{V}_2S_{22}\widetilde{W}_2^*, B)\|\widetilde{V}_2\widetilde{W}_2^*X\| \le \|\widetilde{V}_2\widetilde{W}_2^*A\widetilde{V}_3\widetilde{W}_3^*X\|.$$

Therefore, we acquire the bound

$$\|(I - \widetilde{V}_1\widetilde{W}_1^*)X\| = \|\widetilde{V}_2\widetilde{W}_2^*X + \widetilde{V}_3\widetilde{W}_3^*X\| \le \frac{\|\widetilde{V}_2\widetilde{W}_2^*A\widetilde{V}_3\widetilde{W}_3^*X\|}{\operatorname{sep}(\widetilde{V}_2S_{22}\widetilde{W}_2^*, B)} + \|\widetilde{V}_3\widetilde{W}_3^*X\|,$$

which concludes the proof. □

If there exists a positive constant α such that

$$\operatorname{sep}(\widetilde{V}_2S_{22}\widetilde{W}_{22}^*, B) \ge \alpha > 0$$

as $\|(I - P_{1,2})X\| \to 0$, then the bound

$$\|(I - P_1)X\| \lesssim \left(1 + \frac{\|\widetilde{V}_2\widetilde{W}_2^*A\|}{\alpha}\right)\|(I - P_{1,2})X\|$$

is asymptotically satisfied and $\|(I - P_1)X\| \to 0$ when $\|(I - P_{1,2})X\| \to 0$. The intuitive interpretation of the lower bound α is that there must be a gap between the spectra of B and S_{22} as V_1 and S_{11} converge.

Applying Theorems 3.7, 3.10, and 3.11 to two-sided Krylov–Schur yields the following bounds.

Corollary 3.12. *Suppose the relations in* (3.7) *are satisfied with* $V_1 = \widehat{V}_1$, $V_2 = \widehat{V}_2$, $W_1 = \widehat{W}_1$, *and* $W_2 = \widehat{W}_2$; *then the bound in Theorem 3.7 can be written as*

$$\|E\| \le \|P_1\|\,\|\widetilde{\boldsymbol{k}}_1\|\,\|(I - P_{1,2})XQ\|,$$

the bound in Theorem 3.10 as

$$\operatorname{sep}(\widetilde{V}_1S_{11}\widetilde{W}_1^*, B) \le \|P_1\|\,\|\widetilde{\boldsymbol{k}}_1\|\,\frac{\|(I - P_{1,2})X\|}{\|P_1X\|},$$

and the bound in Theorem 3.11 as

$$\|(I - P_1)X\| \le \left(1 + \frac{\|P_{1,2} - P_1\|\,\|\widetilde{\boldsymbol{k}}_2\|}{\operatorname{sep}(\widetilde{V}_2S_{22}\widetilde{W}_2^*, B)}\right)\|(I - P_{1,2})X\|.$$

The corollary shows that bounds of the form $\|\widetilde{V}_j\widetilde{W}_j^*\|\,\|\widetilde{\boldsymbol{k}}_j\|$ are obtained instead of $\|\widetilde{V}_j\widetilde{W}_j^*A\|$ when two–sided Krylov–Schur is used. This is an attractive result since $\|\widetilde{\boldsymbol{k}}_1\|$ can be expected to go to zero as V_1 converges to X.

It is possible to bound the norm of the oblique projections from the present section in terms of more common orthogonal projections; see, for example, the following proposition.

Proposition 3.13. *Suppose that X, V, and W have orthonormal columns and that W^*V is nonsingular; then for the 2-norm we have*

$$\|(I-VV^*)X\|_2 \le \|(I-V(W^*V)^{-1}W^*)X\|_2 \le \|(W^*V)^{-1}\|_2\,\|(I-VV^*)X\|_2,$$

and for the Frobenius norm

$$\|(I-VV^*)X\|_F \le \|(I-V(W^*V)^{-1}W^*)X\|_F \le \sqrt{1+\|(W^*V)^{-1}\|_F^2}\,\|(I-VV^*)X\|_F.$$

Proof. Define $Z=(I-VV^*)X$ and $P=V(W^*V)^{-1}W^*$; then

$$\begin{aligned}\|Z\|_F^2 &= \|(I-VV^*)(I-P)X\|_F^2\\ &= \operatorname{tr}(X^*(I-P)^*(I-VV^*)(I-P)X)\\ &= \|(I-P)X\|_F^2 - \|V^*(I-P)X\|_F^2 \le \|(I-P)X\|_F^2\end{aligned}$$

and

$$\|(I-P)X\|_F^2 = \|(I-P)Z\|_F^2 = \|Z-PZ\|_F^2 = \|Z\|_F^2+\|PZ\|_F^2 \le \|Z\|_F^2(1+\|(W^*V)^{-1}\|_F^2).$$

For the 2-norm we give a simplified and block version of the first part of the proof found in Chaturantabut and Sorensen [15, Lem. 3.2]. For a nontrivial projector P it holds that $\|I-P\|_2=\|P\|_2$; see, for example, Szyld [87]. Therefore

$$\|Z\|_2 = \|(I-VV^*)(I-P)X\|_2 \le \|(I-P)X\|_2$$

and

$$\|(I-P)X\|_2 = \|(I-P)Z\|_2 \le \|I-P\|_2\,\|Z\|_2 = \|P\|_2\,\|Z\|_2 \le \|(W^*V)^{-1}\|_2\,\|Z\|_2.$$

□

Consequently, if $\|(W^*V)^{-1}\|$ is sufficiently small, then the norms

$$\|(I-V(W^*V)^{-1}W^*)X\| \qquad \text{and} \qquad \|V(W^*V)^{-1}W^*X\|$$

can be seen as a generalization of $\sin(X,V)$ and $\cos(X,V)$, respectively; see Figure 3.1 for an illustration.

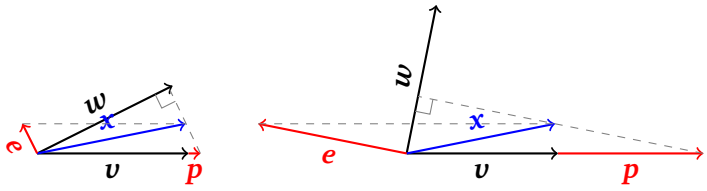


Figure 3.1: Consider the approximation $\boldsymbol{v}$ to $\boldsymbol{x}$, the projection $\boldsymbol{p} = \boldsymbol{v}(\boldsymbol{w}^*\boldsymbol{v})^{-1}\boldsymbol{w}^*\boldsymbol{x}$, and the complementary part $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{p}$. In the left diagram the angle between $\boldsymbol{v}$ and $\boldsymbol{w}$ is small and $\|\boldsymbol{p}\| \approx \|\boldsymbol{v}^*\boldsymbol{x}\| = \cos(\boldsymbol{x}, \boldsymbol{v})$ and $\|\boldsymbol{e}\| \approx \|(I - \boldsymbol{v}\boldsymbol{v}^*)\boldsymbol{x}\| = \sin(\boldsymbol{x}, \boldsymbol{v})$. In the right diagram the angle between $\boldsymbol{v}$ and $\boldsymbol{w}$ is large and $\|\boldsymbol{p}\|$ and $\|\boldsymbol{e}\|$ are no longer satisfactory approximations to the cosine and sine.

3.7 Two-sided distance properties

In the previous section we have considered the convergence of subspaces to invariant subspaces. The focus of this section is on the minimum distance between a given matrix A and a matrix with given invariant subspaces. Given a subspace $\mathcal{V}$, Noschese and Reichel [62] consider the problem of finding the matrix M closest to A satisfying

(3.15) $M\mathcal{V} \subseteq \mathcal{V}$.

In the two-sided case we impose the additional constraint

(3.16) $M^*\mathcal{W} \subseteq \mathcal{W}$

for a given subspace $\mathcal{W}$. Alternatively, this is the problem of finding the backward error $E = A - M$, where the norm of E can be seen as a measure for the quality of the subspaces as approximate invariant subspaces. Consider the following well-known theorem [46, Main Thm.].

Theorem 3.14 (Kahan, Parlett, and Jiang [46])**.** *Let A be an $n \times n$ matrix, and let two $n \times m$ matrices V and W having orthonormal columns be given. Suppose that W^*V is nonsingular. Let*

$$R = AV - VC, \qquad S^* = W^*A - DW^*,$$

where C and D are Rayleigh quotients

$$C = (W^*V)^{-1}W^*AV, \qquad D = W^*AV(W^*V)^{-1}.$$

Then the solution E of

$$(A - E)V = VC \quad \textit{and} \quad W^*(A - E) = DW^*$$

that simultaneously minimizes both

$$\|E\|_2 = \min_E \|E\|_2 = \max\{\|R\|_2, \|S\|_2\}$$

and

$$\|E\|_F = \min_E \|E\|_F = \sqrt{\|R\|_F^2 + \|S\|_F^2}$$

is given by

$$E = RV^* + WS^*.$$

Using the theorem we readily find the following result.

Corollary 3.15. *Suppose that V and W are orthonormal bases of the subspaces $\mathcal{V}$ and $\mathcal{W}$, respectively and that W^*V is nonsingular. Then, the matrix M closest to A that satisfies*

$$M\mathcal{V} \subseteq \mathcal{V} \qquad \text{and} \qquad M^*\mathcal{W} \subseteq \mathcal{W}$$

is given by

$$M = A - (I - V(W^*V)^{-1}W^*)AVV^* + WW^*A(I - V(W^*V)^{-1}W^*).$$

Furthermore, if two-sided Arnoldi is used to compute V, W, $\widetilde{H}$, and $\widetilde{K}$ so that

$$R = AV - V\widetilde{H} = \widetilde{\boldsymbol{v}}\widetilde{\boldsymbol{h}}^* \qquad \text{and} \qquad S = A^*W - W\widetilde{K} = \widetilde{\boldsymbol{w}}\widetilde{\boldsymbol{k}}^*,$$

where $\widetilde{\boldsymbol{v}}$, $\widetilde{\boldsymbol{w}}$, $\widetilde{\boldsymbol{h}}$, and $\widetilde{\boldsymbol{k}}$ are as in (3.8)*, then*

$$\|E\|_2 = \max\{\|\widetilde{\boldsymbol{v}}\|\,\|\widetilde{\boldsymbol{h}}\|, \|\widetilde{\boldsymbol{w}}\|\,\|\widetilde{\boldsymbol{k}}\|\} \qquad \text{and} \qquad \|E\|_F = \sqrt{\|\widetilde{\boldsymbol{v}}\|^2\|\widetilde{\boldsymbol{h}}\|^2 + \|\widetilde{\boldsymbol{w}}\|^2\|\widetilde{\boldsymbol{k}}\|^2}$$

for $E = A - M$.

The matrix M from Corollary 3.15 satisfies the additional constraint

$$W^*(A - zI)V = W^*(M - zI)V$$

for all scalars z. This kind of shift-invariance allows us to interpret $\|E\|_2$ and $\|E\|_F$ as a backward error for the approximation of pseudospectra in Section 3.8.2, where we compute

$$\sigma_{\min}(W^*(A - zI)V)$$

for a large number of complex shifts z near a region of interest. The matrix M is in general not of low rank, and instead we might be interested in the two-sided Arnoldi approximation

$$A_m = V(W^*V)^{-1}W^*AV(W^*V)^{-1}W^* = V\widetilde{H}(W^*V)^{-1}W = V(W^*V)^{-1}\widetilde{K}^*W^*,$$

which is the unique rank-m = rank(W^*AV) matrix satisfying

$$A_m\mathcal{V} \subseteq \mathcal{V}, \qquad A_m^*\mathcal{W} \subseteq \mathcal{W}, \qquad \text{and} \qquad W^*AV = W^*A_mV.$$

An alternative for the singular value problem is to consider the problem of finding the matrix N closest to A satisfying

$$N\mathcal{V} \subseteq \mathcal{W} \qquad \text{and} \qquad N^*\mathcal{W} \subseteq \mathcal{V},$$

as opposed to M satisfying (3.15) and (3.16). Noschese and Reichel [62, Sec. 3] show that

$$N = (I - WW^*)A(I - VV^*) + WW^*AVV^*$$

minimizes the distance to A with

$$\|A - N\|_F^2 = \|AV\|_F^2 + \|A^*W\|_F^2 - 2\|W^*AV\|_F^2.$$

As before, N satisfies the additional constraint

$$W^*(A - zI)V = W^*(N - zI)V$$

for any scalar z, making $\|A - N\|_F$ another backward error. The unique rank-m approximation B_m satisfying

$$B_m\mathcal{V} \subseteq \mathcal{W}, \qquad B_m^*\mathcal{W} \subseteq \mathcal{V}, \qquad \text{and} \qquad W^*AV = W^*B_mV$$

is given by the two-sided Arnoldi approximation

$$B_m = WW^*AVV^* = W(W^*V)\widetilde{H}V^* = W\widetilde{K}^*(W^*V)V^*.$$

The proposition below gives the distance between A and B_m; the proof closely follows the arguments in [62, Prop. 3.3] but uses general left-orthonormal V and W.

Proposition 3.16 (Generalization of [62, Prop. 3.3])**.** *Let V and W have orthonormal columns and define the matrix $B_m = WW^*AVV^*$; then*

$$\|A - B_m\|_F^2 = \|A\|_F^2 - \|B_m\|_F^2.$$

Proof. Using the cyclic property of the trace, we obtain

$$\begin{aligned}\|A - WW^*AVV^*\|_F^2 &= \operatorname{tr}(A^*A - A^*WW^*AVV^* - VV^*A^*WW^*A + VV^*A^*WW^*AVV^*)\\ &= \operatorname{tr}(A^*A) - \operatorname{tr}(V^*A^*WW^*AV) - \operatorname{tr}(V^*A^*WW^*AV)\\ &\quad + \operatorname{tr}(V^*A^*WW^*AV)\\ &= \|A\|_F^2 - \|B_m\|_F^2.\end{aligned}$$

□

We have given a two-sided analogue of Noschese and Reichel's result for (right) invariant subspaces. The bounds in Corollary 3.15 are particularly elegant and efficient to compute in the context of two-sided Krylov–Schur. Distances in the case of invariant singular subspaces are efficiently computable as well, assuming that the Frobenius norm is used and A is explicitly available. To summarize, the distance properties from this section can be used to gain insight into the quality of approximate invariant subspaces.

3.8 Applications and numerical experiments

3.8.1 Eigenvalue condition numbers

Suppose we wish to compute the best-conditioned eigenvalues of a nonnormal matrix A, which is effectively the opposite of the goal of the sensitive pole algorithm [74]. For instance, A might be constructed from uncertain data making the best-conditioned eigenvalues the most reliable ones. Alternatively, one may be focused on the least sensitive eigenvalues of some $A = A(\boldsymbol{p}_0)$ obtained from a parameterized problem for a specific set of parameters given by $\boldsymbol{p}_0$. Since the eigenvalue condition numbers are essential quantities, the choice of a two-sided method over a one-sided method may be appropriate.

In Table 3.1 we compare one-sided and two-sided Krylov–Schur for the computation of the best-conditioned eigenvalues. That is, we are looking for an approximation θ to an eigenvalue λ and an approximation κ_θ to $\kappa(\lambda)$, where $\kappa(\lambda)$ is as small as possible. We recognize that it may be more useful in practice to restrict the search to the best-conditioned eigenvalue near a target, but we make no such restriction here for the sake of simplicity. We measure the relative errors

$$\operatorname{err}_\lambda = \left|\frac{\lambda - \theta}{\lambda}\right| \qquad \text{and} \qquad \operatorname{err}_{\kappa(\lambda)} = \left|\frac{\kappa(\lambda) - \kappa_\theta}{\kappa(\lambda)}\right|,$$

as well as the number of matrix-vector products executed before the algorithms are terminated. We use (3.10) as a stopping criterion and terminate one-sided

and two-sided Krylov–Schur when

$$\frac{\|A\boldsymbol{v} - \theta\boldsymbol{v}\|}{|\theta|} \leq \epsilon 2^{10} \qquad \text{and} \qquad \frac{\max\{\|A\boldsymbol{v} - \theta\boldsymbol{v}\|,\, \|\boldsymbol{w}^*A - \theta\boldsymbol{w}^*\|\}}{|\theta\boldsymbol{w}^*\boldsymbol{v}|} \leq \epsilon 2^{10}$$

respectively, where the θ's are Ritz values, where $\boldsymbol{v}$ and $\boldsymbol{w}$ are right and left Ritz vectors with unit norm, and where ϵ is the machine accuracy. For example, $\epsilon \approx 2.22 \cdot 10^{-16}$ and $\epsilon 2^{10} \approx 2.27 \cdot 10^{-13}$ for IEEE double precision floating point numbers. We run the algorithms with minimum subspace dimension $m = 25$ and maximum subspace dimension $\ell = 50$ by default, and with $m = 50$ and $\ell = 100$ for problems marked by an asterisk (*). All the matrices, except randn, are from a test matrix collection of non-Hermitian eigenvalue problems [2] and are balanced first [16]. The matrix randn is generated using the identically named MATLAB function, and we use the same function to generate random starting vectors.

Table 3.1: Median results over 1000 runs with different random initial vectors for computing the best-conditioned eigenvalues of nonnormal matrices with one-sided and two-sided Krylov–Schur.

			One-sided			Two-sided		
Name	n	$\kappa(\lambda)$	err_λ	$\mathrm{err}_{\kappa(\lambda)}$	**MVs**	err_λ	$\mathrm{err}_{\kappa(\lambda)}$	**MVs**
randn*	1024	3.34	1.25e + 00	9.82e − 01	800	1.01e − 14	3.97e − 14	1100
bfw782a*	782	1.00	4.00e − 02	9.96e − 01	100	3.55e − 15	3.09e − 14	200
ck656	656	1.02	9.75e − 01	8.70e − 01	1275	4.84e − 16	3.35e − 03	50
pde900	900	4.04	2.35e − 01	1.00e + 00	1575	2.67e − 15	1.89e − 14	125
rdb1250l	1250	1.05	9.55e − 01	8.51e − 01	400	7.20e − 15	3.69e − 15	150
olm1000	1000	1.00	1.00e + 00	9.94e − 01	3300	2.99e − 14	2.94e − 14	7525
qh1484	1484	1.00	1.00e + 00	8.77e − 01	825	2.87e − 04	4.27e − 10	75
rdb1250	1250	1.01	9.88e − 01	8.41e − 01	350	5.32e − 15	2.65e − 15	150
qc2534	2534	1.01	1.41e + 00	1.00e + 00	10300	7.36e − 15	2.42e − 15	75
af23560*	23560	1.10	9.37e − 01	9.96e − 01	2700	4.98e − 14	6.80e − 07	350

The results in Table 3.1 show that two-sided Krylov–Schur computes more accurate approximations to both λ and $\kappa(\lambda)$ in every case, and does so using fewer matrix-vector products in seven out of 10 cases. In particular, the total number of matrix-vector products used by one-sided Krylov–Schur is 21625, versus 9800 used by two-sided Krylov–Schur. The high relative error of the one-sided approximations can be explained by the fact that one-sided Krylov–Schur converges to incorrect eigenvalues, a problem not shared by its two-sided counterpart. Evidently, two-sided Krylov–Schur benefits from the improved accuracy of the two-sided condition number estimates and the two-sided Rayleigh quotient.

3.8.2 Pseudospectra

When studying nonnormal matrices, computing pseudospectra rather than eigenvalues and condition numbers may be more insightful [89]. In particular, pseudospectra provide more detailed information regarding the behavior of the eigenvalues under matrix perturbations in the nonnormal case. Indeed, one possible definition of the ε-pseudospectrum of A that clearly shows its relation with matrix perturbations is

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \colon z \in \Lambda(A + E) \text{ for some } E \text{ with } \|E\| < \varepsilon\},$$

where $\Lambda(A + E)$ denotes the spectrum of $A + E$. An alternate definition that is more fitting for the computation of pseudospectra is

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \colon \sigma_{\min}(A - zI) < \varepsilon\}.$$

Ergo, one can simply compute $\sigma_{\min}(A - zI)$ for $z \in \mathbb{C}$ and plot ε-level curves; unfortunately, doing so for many grid points and large A is generally time- and memory-consuming. One method to improve performance is to use one-sided Krylov–Schur to obtain

$$AV_m = V_{m+1}\underline{H}_m,$$

with orthonormal V_{m+1}, and compute the approximation

$$\begin{aligned} \sigma_{\min}(A - zI) &\approx \sigma_{\min}((A - zI)V_m) \\ &= \sigma_{\min}(V_{m+1}^*(A - zI)V_m) = \sigma_{\min}(\underline{H}_m - z\underline{I}); \end{aligned} \tag{3.17}$$

see Wright and Trefethen [93]. Since the right and left singular subspaces differ for nonnormal matrices, it seems natural to project onto a subspace distinct from $\mathcal{V}_{m+1} = \operatorname{span}\{V_{m+1}\}$. At the same time, a shift-invariant subspace is ideal if the goal is to reduce computational effort. This suggests that the left Krylov subspace $\mathcal{W}_m$ belonging to A^* may be an excellent choice, especially if $\mathcal{W}_m$ approximates an invariant subspace belonging to the eigenvalues of interest. Hence, by using two-sided Krylov–Schur we can compute the approximation

$$\sigma_{\min}(A - zI) \approx \sigma_{\min}(W_m^*(A - zI)V_m)$$

or, if $\boldsymbol{w}_{m+1}$ and $\boldsymbol{v}_{m+1}$ have also been computed,

$$\sigma_{\min}(A - zI) \approx \min\{\sigma_{\min}(W_{m+1}^*(A - zI)V_m), \sigma_{\min}(W_m^*(A - zI)V_{m+1})\}. \tag{3.18}$$

The key idea is to use the shift-invariant subspaces $\mathcal{V}_m = \mathcal{K}_m(A, \boldsymbol{v}_1)$ and $\mathcal{W}_m = \mathcal{K}_m(A^*, \boldsymbol{w}_1)$ to compute the smallest singular values in a region surrounding the eigenvalues of interest by imposing the Galerkin conditions

$$\begin{aligned}(A - zI)V_m\boldsymbol{c} - \theta W_m\boldsymbol{d} &\perp \mathcal{W}_m,\\ (A - zI)^*W_m\boldsymbol{d} - \theta V_m\boldsymbol{c} &\perp \mathcal{V}_m\end{aligned}$$

for a large number of complex shifts z. Furthermore, the two-sided approach is symmetric in the sense that the same results are obtained if A is replaced by A^* and the starting vectors are swapped, which is not the case for the one-sided approximation.

We compute the pseudospectra of three disparate matrices in specific regions. The first matrix, randn, is generated using the identically named MATLAB function. The second and third matrices, rdb800l and pipe, are taken from Wright and Trefethen [93, Sec. 5]. For the Krylov–Schur algorithms we use minimum dimension $m = 25$ and maximum dimension $\ell = 50$. Table 3.2 lists additional details, including the number of restarts, which are hand-picked to achieve near optimal results. Because of the conditioning of the eigenvalues, we recompute $V_{m+1}^*AV_m$ and $W_{m+1}^*AV_m$ before computing the pseudospectra, as opposed to working with $(W_{m+1}^*V_{m+1})\underline{H}_m$ and $(V_{m+1}^*W_{m+1})\underline{K}_m$.

Table 3.2: The dimension size, region of interest, target, use of harmonic extraction, and number of restarts (#RS) for one-sided and two-sided Krylov–Schur for each matrix.

Name	n	**Region**	**Target**	**Harmonic**	**#RS-1**	**#RS-2**
randn	1024	$[-27, -17] \times [17, 25]$	$-22 + 21i$	Yes	100	25
rdb800l	800	$[-1.1, 1.1] \times [-0.25, 2.75]$	$+1.25i$	No	125	50
pipe	402	$[-0.15, 0.05] \times [-0.05, 0.05]$	$+0.05$	Yes	1500	1000

The pseudospectra of the test matrices can be seen in Figure 3.2, and their approximations with one-sided and two-sided Krylov–Schur in Figures 3.3 and 3.4, respectively. The latter two figures also include heat maps of the quantity

$$z \mapsto \log_{10}\left|\frac{\sigma_{\min}(A - zI) - \theta}{\sigma_{\min}(A - zI)}\right|, \tag{3.19}$$

where θ is the approximation from either (3.17) for one-sided Krylov–Schur or (3.18) for two-sided Krylov–Schur. The first two subplots in Figure 3.3 show that one-sided Krylov–Schur is capable of capturing the qualitative behavior of the pseudospectrum reasonably well, although the level curves appear to be

"shifted". For instance, the outermost level curve in the rdb800l approximation corresponds to $\varepsilon = 10^{-0.5}$, while the true pseudospectrum has the level curve for $\varepsilon = 10^{-0.8}$ at approximately the same position. The displacement of the level curves is presumably caused by the high relative errors in the approximate singular values. Indeed, for the same two examples, two-sided Krylov–Schur achieves lower relative errors and has better contour approximations. The approximation quality in the last example is comparable for both methods, with the two-sided Krylov–Schur approximation being more accurate near the eigenvalues, and the one-sided Krylov–Schur approximation being better further away. This contrast might be explained by the two-sided Rayleigh quotient having faster asymptotic convergence than the one-sided Rayleigh quotient. Finally, we remark that only the one-sided approximation of the singular value is monotonic in the sense that

$$\sigma_{\min}(A - zI) \leq \sigma_{\min}(V_{m+1}^*(A - zI)V_m),$$

and as a result, there are areas where

$$\sigma_{\min}(A-zI) \leq \sigma_{\min}(W_m^*(A-zI)V_m) \quad \text{or} \quad \sigma_{\min}(A-zI) \geq \sigma_{\min}(W_m^*(A-zI)V_m),$$

separated by curves where

$$\sigma_{\min}(A - zI) = \sigma_{\min}(W_m^*(A - zI)V_m).$$

These "zero-error" curves tend to connect accurate Ritz values and show up as dark(er) lines in the heat maps in Figure 3.4.

3.9 Conclusion

We have presented a two-sided Krylov–Schur method for nonnormal matrices as a natural generalization of the one-sided Krylov–Schur approach by Stewart. An advantage of two-sided Krylov–Schur over two-sided Lanczos is the use of orthonormal bases, and an advantage over one-sided Krylov–Schur is the simultaneous approximation of left and right eigenvectors or eigenspaces. The two-sided approximations may already give useful information concerning eigenvalue conditioning during the iterations. Furthermore, for some applications, the two-sided method may converge with fewer matrix-vector products than the standard Krylov–Schur method.

Primary disadvantages of the new method are the computational cost per iteration, which is roughly twice that of the one-sided Krylov–Schur method, and potential numerical stability and accuracy issues in the computation of the Ritz

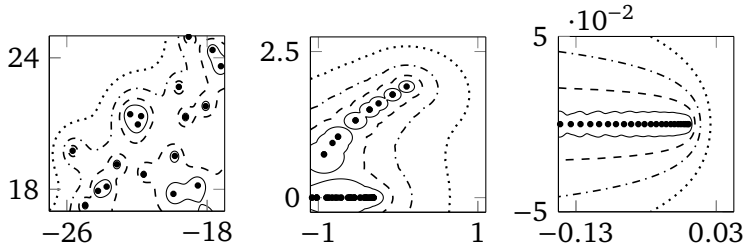


Figure 3.2: Pseudospectra of randn (left), rdb800l (middle), and pipe (right). The level curves range from $10^{-1.7}$ to $10^{-0.5}$, $10^{-1.4}$ to $10^{-0.5}$, and 10^{-5} to $10^{-3.5}$, respectively.

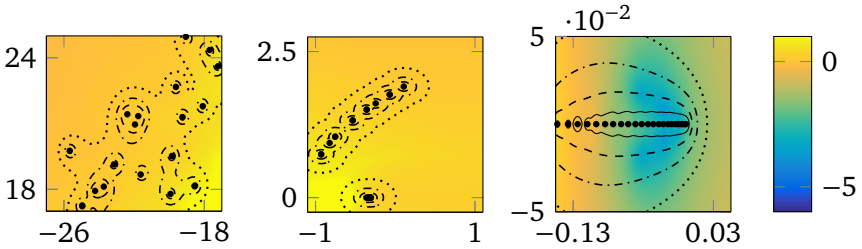


Figure 3.3: Level curves for the pseudospectra approximations obtained with one-sided Krylov–Schur, with randn (left), rdb800l (middle), and pipe (right). The heat maps show the value of the error measure defined in (3.19) and have the average values +0.249, +0.366, and −1.372, respectively.

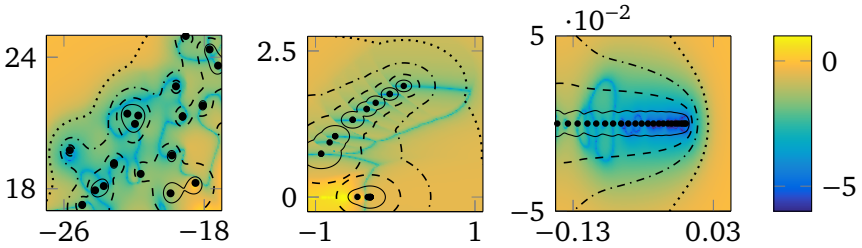


Figure 3.4: Level curves for the pseudospectra approximations obtained with one-sided Krylov–Schur, with randn (left), rdb800l (middle), and pipe (right). The heat maps show the value of the error measure defined in (3.19) and have the average values −1.464, −0.920, and −1.597, respectively.

values. The numerical issues caused by oblique projections can mainly be avoided with a proper implementation, as discussed in Section 3.5.

The two-sided Krylov–Schur method may be combined with either the standard two-sided Rayleigh–Ritz extraction or the harmonic two-sided Rayleigh–Ritz extraction. We have seen that the implementation of the standard two-sided extraction is relatively straightforward, while the implementation of the harmonic extraction is more complicated.

Theoretical convergence properties have been investigated and generalized and show when and how well we can expect two-sided methods to converge. Furthermore, numerical experiments demonstrate that two-sided Krylov–Schur may excel in finding the best-conditioned eigenvalues of nonnormal matrices. Additional numerical experiments show that the shift-invariant left and right Krylov spaces computed with two-sided Krylov–Schur may be useful for the approximation of pseudospectra.

Chapter 4

Multidirectional subspace expansion for one- and multiparameter Tikhonov regularization

Abstract. Tikhonov regularization is a popular method to approximate solutions of linear discrete ill-posed problems when the observed or measured data is contaminated by noise. Multiparameter Tikhonov regularization may improve the quality of the computed approximate solutions. We propose a new iterative method for large-scale multiparameter Tikhonov regularization with general regularization operators based on a multidirectional subspace expansion. The multidirectional subspace expansion may be combined with subspace truncation to avoid excessive growth of the search space. Furthermore, we introduce a simple and effective parameter selection strategy based on the discrepancy principle and related to perturbation results.

Key words. Tikhonov, multiparameter Tikhonov, generalized Krylov, multidirectional subspace expansion, subspace truncation, subspace method, linear discrete ill-posed problem, regularization, regularization parameter.

AMS subject classification. 15A29; 65F10; 65F22; 65F30; 65R30; 65R32

4.1 Introduction

We consider one-parameter and multiparameter Tikhonov regularization problems of the form

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \sum_{i=1}^{\ell} \mu^i \|L^i \boldsymbol{x}\|^2 \qquad (\ell \geq 1), \tag{4.1}$$

where $\|\cdot\|$ denotes the 2-norm and the superscript i is used as an index. We focus on large-scale discrete ill-posed problems such as the discretization of Fredholm integral equations of the first kind. More precisely, assume A is an ill-conditioned or even singular $m \times n$ matrix with $m \geq n$, L^i are $p^i \times n$ matrices such that the nullspaces of A and L^i intersect trivially, and μ^i are nonnegative regularization parameters. Furthermore, assume $\boldsymbol{b}$ is contaminated by an error $\boldsymbol{e}$ and satisfies

$\boldsymbol{b} = A\boldsymbol{x}_\star + \boldsymbol{e}$, where $\boldsymbol{x}_\star$ is the exact solution. Finally, we assume that a bound $\|\boldsymbol{e}\| \leq \epsilon$ is available, so that the discrepancy principle can be used.

In one-parameter Tikhonov regularization ($\ell = 1$), the choice of the regularization operator is typically significant, since frequencies in the nullspace of the operator remain unpenalized. Multiparameter Tikhonov can be used when a satisfactory choice of the regularization operator is unknown in advance, or can be seen as an attempt to combine the strengths of different regularization operators. In some applications, using more than one regularization operator and parameter allows for more accurate solutions [4, 13, 53, 57].

Solving (4.1) for large-scale problems may be challenging. In case the μ^i are fixed *a priori*, methods such as LSQR [67] or LSMR [20] may be used. However, the problem becomes more complicated when the regularization parameters are not fixed in advance [34, 47, 53]. In this chapter, we present a new subspace method consisting of three phases; a new expansion phase, a new extraction phase, and a new truncation phase. To be more specific, let $\mathcal{X}_k \subset \mathbb{R}^n$ be a subspace of dimension $k \ll n$, and let the columns of X_k form an orthonormal basis for $\mathcal{X}_k$. Then we can compute matrix decompositions

$$\begin{aligned} AX_k &= U_{k+1}\underline{H}_k, \\ L^iX_k &= V_k^iK_k^i \qquad (i = 1, 2, \ldots, \ell), \end{aligned} \tag{4.2}$$

where U_{k+1} and V_k^i have orthonormal columns, $\beta\boldsymbol{u}_1 = \boldsymbol{b}$, $\beta = \|\boldsymbol{b}\|$, $\underline{H}_k$ is a $(k+1)\times k$ Hessenberg matrix, and K_k^i is upper triangular. Denote $\boldsymbol{\mu} = (\mu^1, \ldots, \mu^\ell)$ for convenience. Now restrict the solution space to $\mathcal{X}_k$ so that $\boldsymbol{x}_k(\boldsymbol{\mu}) = X_k\boldsymbol{c}_k(\boldsymbol{\mu})$, where

$$\begin{aligned} \boldsymbol{c}_k(\boldsymbol{\mu}) &= \operatorname*{argmin}_{\boldsymbol{c}} \|AX_k\boldsymbol{c} - \boldsymbol{b}\|^2 + \sum_{i=1}^{\ell} \mu^i \|L^iX_k\boldsymbol{c}\|^2 \\ &= \operatorname*{argmin}_{\boldsymbol{c}} \|\underline{H}_k\boldsymbol{c} - \beta\boldsymbol{e}_1\|^2 + \sum_{i=1}^{\ell} \mu^i \|K_k^i\boldsymbol{c}\|^2. \end{aligned} \tag{4.3}$$

The vector $\boldsymbol{e}_1$ is the first standard basis vector of appropriate dimension. Our chapter has three contributions. First, a new expansion phase where we add multiple search directions to $\mathcal{X}_k$. Second, a new truncation phase which removes unwanted new search directions. Third, a new method for selecting the regularization parameters μ_k^i in the extraction phase. The three phases work alongside each other: the intermediate solution obtained in the extraction phase is preserved in the truncation phase, whereas the remaining perpendicular component(s) from the expansion phase are removed.

The chapter is organized as follows. In Section 4.2 an existing nonlinear subspace method is discussed, whereafter we propose the new multidirectional subspace expansion of the expansion phase. Discussion of the truncation phase follows immediately. Section 4.3 is focused on discrepancy principle based parameter selection for one-parameter regularization. New lower and upper bounds on the regularization parameter are provided. Sections 4.4 and 4.5 describe the extraction phase. In the former, a straightforward parameter selection strategy for multiparameter regularization is given, in the latter, a justification using perturbation analysis. Numerical experiments are performed in Section 4.6 and demonstrate the competitiveness of our new method. We end with concluding remarks in Section 4.7.

4.2 Subspace expansion for multiparameter Tikhonov

Let us first consider one-parameter Tikhonov regularization with a general regularization operator. Then $\ell = 1$ and we write $\mu = \mu^1$, $L = L^1$, and $K_k = K_k^1$, such that (4.1) simplifies to

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|L\boldsymbol{x}\|^2.$$

When $L = I$ we use the Golub–Kahan–Lanczos bidiagonalization procedure to generate the Krylov subspace

$$\mathcal{X}_k = \mathcal{K}_k(A^*A, A^*\boldsymbol{b}) = \operatorname{span}\{A^*\boldsymbol{b}, (A^*A)A^*\boldsymbol{b}, \ldots, (A^*A)^{k-1}A^*\boldsymbol{b}\}.$$

In this case $\underline{H}_k$ is lower bidiagonal and K_k is the identity and

$$\boldsymbol{x}_{k+1} = \frac{(I - X_kX_k^*)A^*\boldsymbol{u}_{k+1}}{\|(I - X_kX_k^*)A^*\boldsymbol{u}_{k+1}\|}.$$

If $L \neq I$ one can still try to use the above Krylov subspace [34], however, it may be more natural to consider a shift-independent generalized Krylov subspace of the form

$$\mathcal{X}_k = \mathcal{K}_k(A^*A, L^*L, A^*\boldsymbol{b}),$$

spanned by the first k vectors in

$$\begin{aligned}
&\text{Group 0} \quad A^*\boldsymbol{b} \\
&\text{Group 1} \quad (A^*A)A^*\boldsymbol{b},\ (L^*L)A^*\boldsymbol{b} \\
&\text{Group 2} \quad (A^*A)^2A^*\boldsymbol{b},\ (A^*A)(L^*L)A^*\boldsymbol{b},\ (L^*L)(A^*A)A^*\boldsymbol{b},\ (L^*L)^2A^*\boldsymbol{b} \\
&\qquad \ldots
\end{aligned}$$

This generalized Krylov subspace was first studied by Li and Ye [55] and later by Reichel, Sgallari, and Ye [71]. An orthonormal basis can be created with a generalization of Golub–Kahan–Lanczos bidiagonalization [35]. However, while the search space grows linearly as a function of the number of matrix-vector products, the dimension of the generalized Krylov subspace grows exponentially as a function of the total degree of a bivariate matrix polynomial. As a result, if we take any vector $\boldsymbol{x} \in \mathcal{K}_k(A^*A, L^*L, A^*\boldsymbol{b})$ and write it as $p(A^*A, L^*L)A^*\boldsymbol{b}$, where p is a bivariate polynomial, then p has at most degree $\lfloor \log_2 k \rfloor$. This low degree may be undesirable especially for small regularization parameters μ. Reichel and Yu [72, 73] solve this in part with algorithms that can prioritize one operator over the other. For instance, if $\boldsymbol{w}$ is a vector in a group j and B has priority over A, then group $j+1$ contains $(A^*A)\boldsymbol{w}$, $(B^*B)\boldsymbol{w}$, $(B^*B)^2\boldsymbol{w}$, …, $(B^*B)^\rho\boldsymbol{w}$. The downside is that ρ is a user defined constant, and that the expansion vectors are not necessarily optimal.

An alternative approach is a greedy nonlinear method described by Lampe, Reichel, and Voss [53]. We briefly review their method and state a straightforward extension to multiparameter Tikhonov regularization. First note that the low-dimensional minimization in (4.3) simplifies to

$$\begin{aligned}\boldsymbol{c}_k(\mu) &= \operatorname*{argmin}_{\boldsymbol{c}} \|AX_k\boldsymbol{c} - \boldsymbol{b}\|^2 + \mu\|LX_k\boldsymbol{c}\|^2 \\ &= \operatorname*{argmin}_{\boldsymbol{c}} \|\underline{H}_k\boldsymbol{c} - \beta\boldsymbol{e}_1\|^2 + \mu\|K_k\boldsymbol{c}\|^2\end{aligned}$$

in the one-parameter case. Next, compute a value $\mu = \mu_k$ using, e.g., the discrepancy principle. It is easy to verify that

$$\begin{aligned}&A^*\boldsymbol{b} - (A^*A + \mu_k L^*L)\boldsymbol{x}_k(\mu_k) \\ &\qquad = A^*U_{k+1}(\beta\boldsymbol{e}_1 - \underline{H}_k\boldsymbol{c}_k(\mu_k)) + \mu_k L^*V_k K_k\boldsymbol{c}_k(\mu_k)\end{aligned}$$

is perpendicular to $\mathcal{X}_k$, and is also the gradient of the cost function

$$\boldsymbol{x} \mapsto \frac{1}{2}(\|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|L\boldsymbol{x}\|^2)$$

in the point $\boldsymbol{x}_k(\mu_k)$. Therefore, this vector is used to expand the search space. As usual, expansion and extraction are repeated until suitable stopping criteria are met.

As indicated previously, Lampe, Reichel, and Voss [53] consider only one-parameter Tikhonov regularization, however, their method readily extends to multiparameter Tikhonov regularization. Again, the first step is to decide on

regularization parameters $\boldsymbol{\mu}_k$. Next, use the residual of the normal equations

$$A^*\boldsymbol{b} - \Big(A^*A + \sum_{i=1}^{\ell} \mu_k^i L^{i^*} L^i\Big)\boldsymbol{x}_k(\boldsymbol{\mu}_k)$$
$$= A^* U_{k+1}(\beta \boldsymbol{e}_1 - \underline{H}_k \boldsymbol{c}_k(\boldsymbol{\mu}_k)) - \sum_{i=1}^{\ell} \mu_k^i L^{i^*} V_k^i K_k^i \boldsymbol{c}_k(\boldsymbol{\mu}_k)$$

to expand the search space. Note that the residual is again orthogonal to $\mathcal{X}_k$ and also the gradient of the cost function

$$\boldsymbol{x} \mapsto \frac{1}{2}(\|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \sum_{i=1}^{\ell} \mu^i \|L^i \boldsymbol{x}\|^2).$$

We summarize this multiparameter method in Algorithm 4.1, but remark that in practice we initially use Golub–Kahan–Lanczos bidiagonalization until a $\boldsymbol{\mu}_k$ can be found that satisfies the discrepancy principle.

Algorithm 4.1 (Generalized Krylov subspace Tikhonov regularization; extension of [53])**.**
Input: Measurement matrix A, regularization operators $L^1, \ldots, L^\ell$, and data $\boldsymbol{b}$.
Output: Approximate solution $\boldsymbol{x}_k \approx \boldsymbol{x}_\star$.

Initialize $\beta = \|\boldsymbol{b}\|$, $U_1 = \boldsymbol{b}/\beta$, $X_0 = []$, $\boldsymbol{x}_0 = \boldsymbol{0}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$.
for $k = 1, 2, \ldots$ **do**
  Expand X_{k-1} with $A^*\boldsymbol{b} - (A^*A + \sum_{i=1}^{\ell} \mu_{k-1}^i L^{i^*} L^i)\boldsymbol{x}_{k-1}$.
  Update $AX_k = U_{k+1}\underline{H}_k$ and $L^i X_k = V_k^i K_k^i$.
  Select $\boldsymbol{\mu}_k$; see Section 4.4 and Algorithm 4.3.
  $\boldsymbol{c}_k = \operatorname{argmin}_{\boldsymbol{c}} \big\| \big[\underline{H}_k; \sqrt{\mu_k^1} K_k^1; \ldots; \sqrt{\mu_k^\ell} K_k^\ell\big]\boldsymbol{c} - \beta \boldsymbol{e}_1 \big\|$.
  $\boldsymbol{x}_k = X_k \boldsymbol{c}_k$.
  if $\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|/\|\boldsymbol{x}_k\|$ is sufficiently small **then break**

Suitable regularization operators often depend on the problem and its solution. Multiparameter regularization may be used when a priori information is lacking. In this case, it is not obvious that the residual vector above is a "good" expansion vector, in particular if the intermediate regularization parameters $\boldsymbol{\mu}_k$ are not necessarily accurate. Hence, we propose to remove the dependence on the parameters to some extent by expanding the search space with the vectors

$$A^*A\boldsymbol{x}_k(\boldsymbol{\mu}_k), \quad L^{1^*}L^1\boldsymbol{x}_k(\boldsymbol{\mu}_k), \quad \ldots, \quad L^{\ell^*}L^\ell\boldsymbol{x}_k(\boldsymbol{\mu}_k) \tag{4.4}$$

separately. Here, we omit $A^*\boldsymbol{b}$ as it is already contained in X_k. Since we expand the search space in multiple directions, we refer to this expansion as a "multidirectional" subspace expansion. Observe that the previous residual expansion vector is in the span of the multidirectional expansion vectors.

It is unappealing for the search space to grow with $\ell+1$ basis vectors per iteration, because the cost of orthogonalization and the cost of solving the projected problems depend on the dimension of the search space. Therefore, we wish to condense the best portions of the multiple directions in a single vector, and use the following approach. First we expand X_k with the vectors in (4.4) and obtain $\widetilde{X}_{k+\ell+1}$. Then we compute the decompositions

$$\begin{aligned} A\widetilde{X}_{k+\ell+1} &= \widetilde{U}_{k+\ell+2}\underline{\widetilde{H}}_{k+\ell+1}, \\ L^i\widetilde{X}_{k+\ell+1} &= \widetilde{V}^i_{k+\ell+1}\widetilde{K}^i_{k+\ell+1} \qquad (i = 1, 2, \ldots, \ell), \end{aligned}$$

analogously to (4.2), and determine parameters $\boldsymbol{\mu}_{k+1}$ and the approximate solution $\widetilde{\boldsymbol{c}}_{k+\ell+1}$. Next, we compute

$$\begin{aligned} A(\widetilde{X}_{k+\ell+1}Z^*) &= (\widetilde{U}_{k+\ell+2}P^*)(P\underline{\widetilde{H}}_{k+\ell+1}Z^*), \\ L^i(\widetilde{X}_{k+\ell+1}Z^*) &= (\widetilde{V}^i_{k+\ell+1}Q^{i*})(Q^i\widetilde{K}^i_{k+\ell+1}Z^*) \qquad (i = 1, 2, \ldots, \ell), \end{aligned} \tag{4.5}$$

where Z, P, and Q^i are orthonormal matrices of the form

$$Z = \begin{bmatrix} I_k & \\ & Z_{\ell+1} \end{bmatrix}, \quad P = \begin{bmatrix} I_{k+1} & \\ & P_{\ell+1} \end{bmatrix}, \quad Q^i = \begin{bmatrix} I_k & \\ & Q^i_{\ell+1} \end{bmatrix}. \tag{4.6}$$

Here I_k is the $k \times k$ identity matrix and $Z_{\ell+1}$ is an orthonormal matrix so that $Z_{\ell+1}\widetilde{\boldsymbol{c}}_{k+1:k+\ell+1} = \gamma\boldsymbol{e}_1$ for some scalar γ. The matrices $P_{\ell+1}$ and $Q^i_{\ell+1}$ are computed to make $\underline{\widetilde{H}}_{k+\ell+1}Z^*$ and $\widetilde{K}^i_{k+\ell+1}Z^*$ respectively upper-Hessenberg and upper-triangular again. At this point we can truncate (4.5) to obtain

$$\begin{aligned} AX_{k+1} &= U_{k+2}\underline{H}_{k+1}, \\ L^iX_{k+1} &= V^i_{k+1}K^i_{k+1} \qquad (i = 1, 2, \ldots, \ell), \end{aligned}$$

and truncate $Z\widetilde{\boldsymbol{c}}_{k+\ell+1}$ to obtain $\boldsymbol{c}_{k+1}$ so that $\widetilde{X}_{k+\ell+1}\widetilde{\boldsymbol{c}}_{k+\ell+1} = X_{k+1}\boldsymbol{c}_{k+1}$. The truncation is expected to keep important components, since the directions removed from $X_{k+\ell+1}$ are perpendicular to the current best approximation $\boldsymbol{x}_{k+1}$, and also to the previous best approximations $\boldsymbol{x}_k$, $\boldsymbol{x}_{k-1}$, …, $\boldsymbol{x}_1$. If the rotation and truncation are combined in one step, then the computational cost of the method is $\mathcal{O}((\ell+1)(n+m+p^1+\cdots+p^\ell))$, which quickly becomes smaller than the (re)orthogonalization cost as k grows.

To illustrate our approach, let us consider a one-parameter Tikhonov example where $\ell = 1$. First we expand $X_1 = \boldsymbol{x}_1$ with vectors $A^*A\boldsymbol{x}_1$ and $L^*L\boldsymbol{x}_1$. Let $A\widetilde{X}_{1+2} = \widetilde{U}_{2+2}\underline{\widetilde{H}}_{1+2}$ and $L\widetilde{X}_{1+2} = \widetilde{V}_{1+2}\widetilde{K}_{1+2}$, and use $\underline{\widetilde{H}}_{1+2}$ and $\widetilde{K}_{1+2}$ to compute $\widetilde{\boldsymbol{c}}_{1+2}$. We then compute a rotation matrix Z_2 so that $Z_2\widetilde{\boldsymbol{c}}_{2:3} = \pm\|\widetilde{\boldsymbol{c}}_{2:3}\|\boldsymbol{e}_1$, and let Z be defined as in (4.6). The matrices $\underline{\widetilde{H}}_{1+2}Z^*$ and $\widetilde{K}_{1+2}Z^*$ no longer have their original structure, hence, we need to compute orthonormal P and Q such that $P\underline{\widetilde{H}}_{1+2}Z^*$ is again upper-Hessenberg and $Q\widetilde{K}_{1+2}Z^*$ is upper-triangular. Schematically we have

$$\xrightarrow{\widetilde{\boldsymbol{c}}^*_{1+2}} \begin{bmatrix}\times & \times & \times\end{bmatrix} \xrightarrow{(Z\widetilde{\boldsymbol{c}}_{1+2})^*} \begin{bmatrix}\times & \times & 0\end{bmatrix},$$

$$\xrightarrow{\underline{\widetilde{H}}_{1+2}} \begin{bmatrix}\times & \times & \times\\ \times & \times & \times\\ 0 & \times & \times\\ 0 & 0 & \times\end{bmatrix} \xrightarrow{\widetilde{H}_{1+2}Z^*} \begin{bmatrix}\times & \times & \times\\ \times & \times & \times\\ 0 & \times & \times\\ 0 & \times & \times\end{bmatrix} \xrightarrow{P\underline{\widetilde{H}}_{1+2}Z^*} \begin{bmatrix}\times & \times & \times\\ \times & \times & \times\\ 0 & \times & \times\\ 0 & 0 & \times\end{bmatrix},$$

$$\xrightarrow{\widetilde{K}_{1+2}} \begin{bmatrix}\times & \times & \times\\ 0 & \times & \times\\ 0 & 0 & \times\end{bmatrix} \xrightarrow{\widetilde{K}_{1+2}Z^*} \begin{bmatrix}\times & \times & \times\\ 0 & \times & \times\\ 0 & \times & \times\end{bmatrix} \xrightarrow{Q\widetilde{K}_{1+2}Z^*} \begin{bmatrix}\times & \times & \times\\ 0 & \times & \times\\ 0 & 0 & \times\end{bmatrix},$$

accompanied by the decompositions

$$A(\widetilde{X}_{1+2}Z^*) = (\widetilde{U}_{2+2}P^*)(P\underline{\widetilde{H}}_{1+2}Z^*),$$
$$L(\widetilde{X}_{1+2}Z^*) = (\widetilde{V}_{1+2}Q^*)(Q\widetilde{K}_{1+2}Z^*).$$

At this point we truncate the subspaces by removing the last columns from $\widetilde{X}_{1+2}Z^*$, $\widetilde{U}_{2+2}P^*$, $P\underline{\widetilde{H}}_{1+2}Z^*$, $\widetilde{V}_{1+2}Q^*$, and $Q\widetilde{K}_{1+2}Z^*$, and the bottom rows of $P\underline{\widetilde{H}}_{1+2}Z^*$ and $Q\widetilde{K}_{1+2}Z^*$, to obtain

$$AX_2 = U_3\underline{H}_2,$$
$$LX_2 = V_2K_2.$$

Below we summarize the steps of the new algorithm for solving problem (4.1). In our implementation we take care to use full reorthogonalization and avoid extending X_k, U_{k+1}, and V_k^i with numerically linearly dependent vectors. We omit these steps from the pseudocode for brevity. In addition, we initially expand the search space solely with $A^*\boldsymbol{u}_{k+1}$ until the discrepancy principle can be satisfied conform Proposition 4.1 in Section 4.3.

Algorithm 4.2 (Multidirectional Tikhonov regularization)**.**
Input: Measurement matrix A, regularization operators. $L^1, \ldots, L^\ell$, and data $\boldsymbol{b}$.
Output: Approximate solution $\boldsymbol{x}_k \approx \boldsymbol{x}_\star$.

1. Initialize $\beta = \|\boldsymbol{b}\|$, $U_1 = \boldsymbol{b}/\beta$, $X_0 = []$, $\boldsymbol{x}_0 = \boldsymbol{0}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$.
   for $k = 0, 1, \ldots,$ **do**
2. Expand X_k with $A^*A\boldsymbol{x}_k, {L^1}^*L^1\boldsymbol{x}_k, \ldots, {L^\ell}^*L^\ell\boldsymbol{x}_k$.
3. Update $A\widetilde{X}_{k+\ell+1} = \widetilde{U}_{k+\ell+2}\underline{\widetilde{H}}_{k+\ell+1}$ and $L^i\widetilde{X}_{k+\ell+1} = \widetilde{V}^i_{k+\ell+1}\widetilde{K}^i_{k+\ell+1}$.
4. Select $\boldsymbol{\mu}_k$; see Section 4.4 and Algorithm 4.3.
5. $\widetilde{\boldsymbol{c}}_{k+\ell+1} = \operatorname{argmin}_{\boldsymbol{c}} \left\| \left[\underline{\widetilde{H}}_{k+\ell+1}; \sqrt{\mu^1_k}\widetilde{K}^1_{k+\ell+1}; \ldots; \sqrt{\mu^\ell_k}\widetilde{K}^\ell_{k+\ell+1} \right] \boldsymbol{c} - \beta\boldsymbol{e}_1 \right\|$.
6. Compute P, Q, and Z (see text).
7. Truncate $A(\widetilde{X}_{k+\ell+1}Z^*) = (\widetilde{U}_{k+\ell+2}P^*)(P\underline{\widetilde{H}}_{k+\ell+1}Z^*)$
   and $L^i(\widetilde{X}_{k+\ell+1}Z^*) = (\widetilde{V}^i_{k+\ell+1}Q^{i*})(Q^i\widetilde{K}^i_{k+\ell+1}Z^*)$
   to $AX_{k+1} = U_{k+2}\underline{H}_{k+1}$ and $L^iX_{k+1} = V^i_{k+1}K^i_{k+1}$.
8. Truncate $Z\widetilde{\boldsymbol{c}}_{k+\ell+1}$ to obtain $\boldsymbol{c}_{k+1}$ and set $\boldsymbol{x}_{k+1} = X_{k+1}\boldsymbol{c}_{k+1}$.
9. **if** $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|/\|\boldsymbol{x}_k\|$ is sufficiently small **then break**

We have completed our discussion of the expansion and truncation phase of our algorithm. In the following section we discuss the extraction phase for one-parameter Tikhonov regularization and discuss the multiparameter case in later sections.

4.3 Parameter selection in standard Tikhonov

In this section we investigate parameter selection for general form one-parameter Tikhonov, where $\ell = 1$, $\mu = \mu^1$, and $L = L^1$. Multiple methods exist in the one-parameter case to determine particular μ_k, including the discrepancy principle, the L-curve criterion and generalized cross validation; see, for example, Hansen [28, Ch. 7]. We focus on the discrepancy principle which states that μ_k must satisfy

$$\|A\boldsymbol{x}_k(\mu_k) - \boldsymbol{b}\| = \eta\epsilon, \tag{4.7}$$

where $\|\boldsymbol{e}\| \le \epsilon$ and $\eta > 1$ is a user supplied constant independent of ϵ.

Define the residual vector $\boldsymbol{r}_k(\mu) = A\boldsymbol{x}_k(\mu) - \boldsymbol{b}$ and the function $\varphi(\mu) = \|\boldsymbol{r}_k(\mu)\|^2$. A nonnegative μ_k satisfies the discrepancy principle if $\varphi(\mu_k) = \eta^2\epsilon^2$. It is known that root finding methods can find solutions, for example, Lampe, Reichel, and Voss [53] compare four of them. We prefer bisection for its reliability and straightforward analysis and implementation. The performance difference is not an issue because root finding requires a fraction of the total computation

time and is no bottleneck. A unique solution μ_k exists under mild conditions, see for instance [14]. Below we give a proof using our own notation.

Assume $\underline{H}_k$ and K_k are full rank and let $P_k\Sigma_k Q_k^*$ be the singular value decomposition of $\underline{H}_k K_k^{-1}$. Let the singular values be denoted by

$$\sigma_{\max} = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k = \sigma_{\min} > 0. \tag{4.8}$$

Now we can express $\boldsymbol{c}_k(\mu)$ and φ as

$$\begin{aligned}
\boldsymbol{c}_k(\mu) &= (\underline{H}_k^*\underline{H}_k + \mu K_k^* K_k)^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1 \\
&= K_k^{-1}(K_k^{-*}\underline{H}_k^*\underline{H}_k K_k^{-1} + \mu I)^{-1}K_k^{-*}\underline{H}_k^*\beta\boldsymbol{e}_1 \\
&= K_k^{-1}Q_k(\Sigma_k^2 + \mu I)^{-1}\Sigma_k P_k^*\beta\boldsymbol{e}_1
\end{aligned}$$

and

$$\begin{aligned}
\varphi(\mu) &= \|\beta\boldsymbol{e}_1 - \underline{H}_k\boldsymbol{c}_k(\mu)\|^2 \\
&= \beta^2\|\boldsymbol{e}_1 - \underline{H}_k K_k^{-1}Q_k(\Sigma_k^2 + \mu I)^{-1}\Sigma_k P_k^*\boldsymbol{e}_1\|^2 \\
&= \beta^2\|(I - P_kP_k^*)\boldsymbol{e}_1 + P_kP_k^*\boldsymbol{e}_1 - P_k\Sigma_k(\Sigma_k^2 + \mu I)^{-1}\Sigma_k P_k^*\boldsymbol{e}_1\|^2 \\
&= \beta^2\|(I - P_kP_k^*)\boldsymbol{e}_1\|^2 + \beta^2\|\mu(\Sigma_k^2 + \mu I)^{-1}P_k^*\boldsymbol{e}_1\|^2.
\end{aligned}$$

Or alternatively,

$$\varphi(\mu) = \beta^2\|(I - P_kP_k^*)\boldsymbol{e}_1\|^2 + \beta^2\sum_{j=1}^{k}\left(\frac{\mu}{\sigma_j^2 + \mu}\right)^2|P_k|_{1j}^2. \tag{4.9}$$

Observe that P_k is a basis for the range of $\underline{H}_k$ and $I - P_kP_k^*$ is the orthogonal projection onto the nullspace $\mathcal{N}(\underline{H}_k^*)$ and is sometimes denoted by $\mathcal{P}_{\mathcal{N}(\underline{H}_k^*)}$. Furthermore, it can be verified that $\underline{H}_k\beta\boldsymbol{e}_1 \neq \boldsymbol{0}$ if $A^*\boldsymbol{b} \neq \boldsymbol{0}$, that is, $\boldsymbol{b} \notin \mathcal{N}(A^*)$.

Proposition 4.1. *If $\beta^2\|(I - P_kP_k^*)\boldsymbol{e}_1\|^2 \leq \eta^2\epsilon^2 < \|\boldsymbol{b}\|^2$, then there exists a unique $\mu_k \geq 0$ such that $\varphi(\mu_k) = \eta^2\epsilon^2$.*

Proof. (See also [14] and references therein). From (4.9) it follows that φ is a rational function with poles $\mu = -\sigma_j^2$ for all $\sigma_j > 0$, therefore, φ is C^∞ on the interval $[0, \infty)$. Additionally, φ is a strictly increasing and bounded function on the same interval, since

$$\frac{d}{d\mu}\left(\frac{\mu}{\sigma_j^2 + \mu}\right)^2 = 2\frac{\mu\sigma_j^2}{(\sigma_j^2 + \mu)^3} > 0 \quad \text{for all} \quad \mu > 0$$

implies $\varphi'(\mu) > 0$ and

$$\varphi(0) = \beta^2 \|(I - P_k P_k^*)\boldsymbol{e}_1\|^2 \quad \text{and} \quad \lim_{\mu\to\infty} \varphi(\mu) = \beta^2 = \|\boldsymbol{b}\|^2.$$

Consequently, there exists a unique $\mu_k \in [0, \infty)$ such that $\varphi(\mu_k) = \eta^2\epsilon^2$. □

Beyond nonnegativity, the proposition above provides little insight on the location of μ_k on the real axis, and we would like to have lower and upper bounds. We determine bounds in Proposition 4.2 and believe the results to be new. Both in practice and for the proof of the subsequent proposition, it is useful to remove nonessential parts of $\varphi(\mu)$ and instead work with the function

$$\widetilde{\varphi}(\mu) = \frac{\varphi(\mu) - \varphi(0)}{\beta^2} = \sum_{j=1}^{k} \left(\frac{\mu}{\sigma_j^2 + \mu}\right)^2 |P_k|_{1j}^2$$

and the quantity

$$\widetilde{\epsilon}^2 = \frac{\eta^2\epsilon^2 - \varphi(0)}{\beta^2}. \tag{4.10}$$

Then $0 \le \widetilde{\varphi}(\mu) \le \rho$, where $\rho = \|P_k^*\boldsymbol{e}_1\| \le 1$, and $\eta^2\epsilon^2$ satisfies the bounds in Proposition 4.1 if and only if $0 \le \widetilde{\epsilon} < \rho$, and $\varphi(\mu_k) = \eta^2\epsilon^2$ if and only if $\widetilde{\varphi}(\mu_k) = \widetilde{\epsilon}^2$.

Proposition 4.2. *If $0 \le \widetilde{\epsilon} < \rho$, and μ_k is such that $\widetilde{\varphi}(\mu_k) = \widetilde{\epsilon}^2$, then*

$$\frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}}\sigma_{\min}^2 \le \mu_k \le \frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}}\sigma_{\max}^2, \tag{4.11}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are as in (4.8).

Proof. Observe that

$$\frac{\mu}{\sigma_{\max}^2 + \mu} \le \frac{\mu}{\sigma_j^2 + \mu} \le \frac{\mu}{\sigma_{\min}^2 + \mu}$$

for all $j = 1, \ldots, k$. Combining this observation with the definition of $\widetilde{\varphi}$ yields

$$\left(\frac{\mu_k}{\sigma_{\max}^2 + \mu_k}\right)^2 \sum_{j=1}^{k} |P_k|_{1j}^2 \le \sum_{j=1}^{k} \left(\frac{\mu_k}{\sigma_j^2 + \mu_k}\right)^2 |P_k|_{1j}^2 \le \left(\frac{\mu_k}{\sigma_{\min}^2 + \mu_k}\right)^2 \sum_{j=1}^{k} |P_k|_{1j}^2.$$

Since $\sum_{j=1}^{k} |P_k|_{1j}^2 = \|P_k^*\boldsymbol{e}_1\|^2 = \rho^2$ and $\widetilde{\varphi}(\mu_k) = \widetilde{\epsilon}^2$, it follows that

$$\frac{\mu_k}{\sigma_{\max}^2 + \mu_k}\rho \le \widetilde{\epsilon} \le \frac{\mu_k}{\sigma_{\min}^2 + \mu_k}\rho.$$

Hence, if $\widetilde{\epsilon} = 0$, then $\mu_k = 0$ and we are done. Otherwise $\mu_k \neq 0$ and we can divide by ρ, take the reciprocals, and subtract 1 to arrive at

$$\frac{\sigma_{\max}^2}{\mu_k} \geq \frac{\rho}{\widetilde{\epsilon}} - 1 \geq \frac{\sigma_{\min}^2}{\mu_k},$$

so that

$$\frac{\mu_k}{\sigma_{\max}^2} \leq \frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}} \leq \frac{\mu_k}{\sigma_{\min}^2},$$

and the proposition follows. □

It is undesirable to work with the inverse of K_k when it becomes ill-conditioned. Instead it may be preferred to use the generalized singular value decomposition (GSVD)

$$\begin{aligned} \underline{H}_k &= P_k C_k Z_k^{-1}, \\ K_k &= Q_k S_k Z_k^{-1}, \end{aligned}$$

where P_k and Q_k have orthogonal columns and Z_k is nonsingular. The matrices C_k and S_k are diagonal with entries $0 \leq c_1 \leq c_2 \leq \cdots \leq c_k$ and respectively $s_1 \geq \cdots \geq s_k \geq 0$, such that $c_i^2 + s_i^2 = 1$. The generalized singular values are given by c_i/s_i and are understood to be infinite when $s_i = 0$. If K_k is nonsingular, then the generalized singular values coincide with the singular values of $\underline{H}_k K_k^{-1}$. See Golub and Van Loan [24, Section 8.7.4] for more information.

Using a similar derivation as before, we can show that

$$\varphi(\mu) = \beta^2 \|(I - P_k P_k^*)\boldsymbol{e}_1\|^2 + \beta^2 \sum_{j=1}^{k} \left(\frac{\mu s_j^2}{c_j^2 + \mu s_j^2} \right)^2 |P_k|_{1j}^2$$

and that the new bounds are given by

$$\frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}} \left(\frac{c_1}{s_1} \right)^2 \leq \mu_k \leq \frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}} \left(\frac{c_k}{s_k} \right)^2.$$

Here μ_k is unbounded from above if $s_k = 0$, that is, if K_k becomes singular.

The bounds in this section can be readily computed and used to implement bisection and the secant method. We consider parameter selection for multiparameter regularization in the following section.

4.4 A multiparameter selection strategy

Choosing satisfactory μ_k^i in multiparameter regularization is more difficult than the corresponding one-parameter problem. See for example [4, 13, 22, 41, 52, 57]. In particular, there is no obvious multiparameter extension of the discrepancy principle. Nevertheless, methods based on the discrepancy principle exist and we will discuss three of them.

Brezinski et al. [13] have had some success with operator splitting. Substituting $\mu_k^i = \nu_k^i \omega_k^i$ in (4.3) with nonnegative weights ω_k^i and $\sum_{i=1}^{\ell} \omega_k^i = 1$ leads to

$$\operatorname*{argmin}_{\boldsymbol{c}} \sum_{i=1}^{\ell} \omega_k^i (\|\underline{H}_k \boldsymbol{c} - \beta \boldsymbol{e}_1\|^2 + \nu_k^i \|K_k^i \boldsymbol{c}\|^2).$$

This form of the minimization problem suggests the approximation of $X_k^* \boldsymbol{x}_\star$ by a linear combination [13, Sec. 3] of $\boldsymbol{c}_k^i(\nu_k^i)$, where

$$\boldsymbol{c}_k^i(\nu) = \operatorname*{argmin}_{\boldsymbol{c}} \|\underline{H}_k \boldsymbol{c} - \beta \boldsymbol{e}_1\|^2 + \nu \|K_k^i \boldsymbol{c}\|^2 \qquad (i = 1, 2, \ldots, \ell) \tag{4.12}$$

and where ν_k^i is such that $\|\underline{H}_k \boldsymbol{c}_k^i(\nu_k^i) - \beta \boldsymbol{e}_1\| = \eta\epsilon$. Alternatively, Brezinski et al. consider solving

$$\boldsymbol{c}_k = \operatorname*{argmin}_{\boldsymbol{c}} \left\| \left[\underline{H}_k; \sqrt{\nu_k^1} K_k^1; \ldots; \sqrt{\nu_k^\ell} K_k^\ell \right] \boldsymbol{c} - \beta \boldsymbol{e}_1 \right\|,$$

where ν^i are fixed and obtained from (4.12). The latter approach provides better results in exchange for an additional QR decomposition. In either case, operator splitting is a straightforward approach, but does not necessarily satisfy the discrepancy principle exactly.

Lu and Pereverzyev [56] and later Fornasier, Naumova, and Pereverzyev [21] rewrite the constrained minimization problem as a differential equation and approximate

$$F(\boldsymbol{\mu}) = \|\underline{H}_k \boldsymbol{c}_k(\boldsymbol{\mu}) - \beta \boldsymbol{e}_1\|^2 + \sum_{i=1}^{\ell} \mu^i \|K_k^i \boldsymbol{c}_k(\boldsymbol{\mu})\|^2$$

by a model function $m(\boldsymbol{\mu})$ which admits a straightforward solution to the constructed differential equation. However, it is unclear which $\boldsymbol{\mu}$ the method finds and its solution may depend on the initial guess. On the other hand, it is possible to keep all but one parameter fixed and compute a value for the free parameter

such that the discrepancy principle is satisfied. This allows one to trace discrepancy hypersurfaces to some extent.

Gazzola and Novati [22] describe another interesting method. They start with a one-parameter problem and successively add parameters in a novel way, until each parameter of the full multiparameter problem has a value assigned. Especially in early iterations the discrepancy principle is not satisfied, but the parameters are updated in each iteration so that the norm of the residual is expected to approach $\eta\epsilon$. Unfortunately, we observed some issues in our implementation. For example, the quality of the result depends on initial values, as well as on the order in which the operators are added (that is, the indexing of the operators). The latter problem has been solved by a recently published and improved version of the method [23].

We propose a new method that satisfies the discrepancy principle exactly, does not depend on an initial guess, and is independent of the scaling or indexing of the operators. The method uses the operator splitting approach in combination with new weights. Let us omit all k subscripts for the remainder of this section, and suppose $\mu^i = \mu\omega^i$, where ω^i are nonnegative, but do not necessarily sum to one, and μ is such that the discrepancy principle is satisfied. Then (4.3) can be written as

$$\operatorname*{argmin}_{\boldsymbol{c}} \|\underline{H}\boldsymbol{c} - \beta\boldsymbol{e}_1\|^2 + \mu \sum_{i=1}^{\ell} \omega^i \|K^i\boldsymbol{c}\|^2. \tag{4.13}$$

Since the goal of regularization is to reduce sensitivity of the solution to noise, we use the weights

$$\omega^i = \frac{\|\boldsymbol{c}^i(\nu^i)\|}{\|D\boldsymbol{c}^i(\nu^i)\|}, \tag{4.14}$$

which bias the regularization parameters in the direction of lower sensitivity with respect to changes in ν^i. Here D denotes the (total) derivative with respect to regularization parameter(s), and $\boldsymbol{c}^i$ and ν^i are defined as before, consequently

$$D\boldsymbol{c}^i(\nu^i) = -(\underline{H}^*\underline{H} + \nu^i K^{i*}K^i)^{-1}K^{i*}K^i\boldsymbol{c}^i(\nu^i).$$

If for some indices $D\boldsymbol{c}^i(\nu^i) = \boldsymbol{0}$, then we take a $\boldsymbol{c}^i(\nu^i)$ as the solution, or replace $\|D\boldsymbol{c}^i(\nu^i)\|$ by a small positive constant. With this parameter choice, the solution does not depend on the indexing of the operators, nor, up to a constant, on the scaling of A, $\boldsymbol{b}$, or any of the L^i. The former is easy to see; for the latter, let α, γ, and λ^i be positive constants, and consider the scaled problem

$$\operatorname*{argmin}_{\widehat{\boldsymbol{x}}} \|\gamma\boldsymbol{b} - \alpha A\widehat{\boldsymbol{x}}\|^2 + \mu \sum_{i=1}^{\ell} \widehat{\omega}^i \|\lambda^i L^i\widehat{\boldsymbol{x}}\|^2.$$

The noisy component of $\gamma \boldsymbol{b}$ is $\gamma \boldsymbol{e}$ and $\|\gamma \boldsymbol{e}\| \leq \gamma\epsilon$, hence the new discrepancy bound becomes

$$\|\alpha A\widehat{\boldsymbol{x}} - \gamma \boldsymbol{b}\| = \gamma\eta\epsilon.$$

The bound is satisfied when $\widehat{\omega}^i = \alpha^2/(\lambda^i)^2\ \omega^i$, since in this case

$$\widehat{\boldsymbol{x}} = \left(\alpha^2 A^*A + \mu \sum_{i=1}^{\ell} \omega^i \frac{\alpha^2}{(\lambda^i)^2} (\lambda^i)^2 L^{i^*} L^i\right)^{-1} \alpha A^* \gamma \boldsymbol{b} = \frac{\gamma}{\alpha} \boldsymbol{x}$$

and

$$\min_{\widehat{\boldsymbol{x}}} \|\gamma \boldsymbol{b} - \alpha A\widehat{\boldsymbol{x}}\|^2 + \mu \sum_{i=1}^{\ell} \widehat{\omega}^i \|\lambda^i L^i \widehat{\boldsymbol{x}}\|^2 = \gamma^2 \left(\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu \sum_{i=1}^{\ell} \omega^i \|L^i \boldsymbol{x}\|^2 \right).$$

It may be checked that the weights in (4.14) are indeed proportional to $\alpha^2/(\lambda^i)^2$, that is

$$\omega^i = \frac{\|\boldsymbol{c}^i(\nu^i)\|}{\|D\boldsymbol{c}^i(\nu^i)\|} \sim \frac{\alpha^2}{(\lambda^i)^2}.$$

There are additional viable choices for ω^i, including two smoothed versions of the above:

$$\omega^i = \frac{\|\underline{H}\boldsymbol{c}^i(\nu^i)\|}{\|\underline{H}D\boldsymbol{c}^i(\nu^i)\|} \quad \text{and} \quad \omega^i = \frac{\|K^i\boldsymbol{c}^i(\nu^i)\|}{\|K^iD\boldsymbol{c}^i(\nu^i)\|},$$

which consider the sensitivity of $\boldsymbol{c}^i(\nu^i)$ in the range of $\underline{H}$ and K^i respectively. We summarize the new parameter selection in Algorithm 4.3 below.

Algorithm 4.3 (Multiparameter selection)**.**
Input: Projected matrices $\underline{H}$, $K^1, \ldots, K^\ell$, $\beta = \|\boldsymbol{b}\|$, noise estimate ϵ, uncertainty parameter η, and threshold τ.
Output: Regularization parameters $\mu^1, \ldots, \mu^\ell$.

1. Use (4.12) to compute $\boldsymbol{c}^i$ and ν^i.
   if $\|D\boldsymbol{c}^i(\nu^i)\| \leq \tau\|\boldsymbol{c}^i(\nu^i)\|$ for some i **then**
2. Set $\omega^i = \tau^{-1}$; or set $\mu^i = \nu^i$ and $\mu^j = 0$ for $j \neq i$.
   else
3. Let $\omega^i = \|\boldsymbol{c}^i(\nu^i)\|/\|D\boldsymbol{c}^i(\nu^i)\|$.
4. Compute μ in (4.13) such that the discrepancy principle is satisfied.
5. Set $\mu^i = \mu\omega^i$.

An interesting property of Algorithm 4.3 is that, under certain conditions, $\boldsymbol{c}(\boldsymbol{\mu}(\widetilde{\epsilon}))$ converges to the unregularized least squares solution

$$\boldsymbol{c}(\mathbf{0}) = (\underline{H}^*\underline{H})^{-1}\underline{H}^*\beta\boldsymbol{e}_1 = \underline{H}^+\beta\boldsymbol{e}_1$$

as $\widetilde{\epsilon}$ goes to zero. Here $\underline{H}^+$ denotes the Moore–Penrose pseudoinverse and $\boldsymbol{c}(\mathbf{0})$ is the minimum norm solution of the unregularized problem. The following proposition formalizes this observation.

Proposition 4.3. *Assume that $\underline{H}$ is full rank, $\underline{H}^*\beta\boldsymbol{e}_1 \neq \mathbf{0}$, and that K^i is nonsingular for $i = 1, \dots \ell$. Let $\widetilde{\epsilon}$ and ρ be defined as in Section 4.3, let $\eta > 1$ be fixed, and suppose that $\nu^i(\widetilde{\epsilon})$ and*

$$\boldsymbol{\mu}(\widetilde{\epsilon}) = (\mu^1(\widetilde{\epsilon}), \dots, \mu^\ell(\widetilde{\epsilon})) = \mu(\widetilde{\epsilon})(\omega^1(\nu^1(\widetilde{\epsilon})), \dots, \omega^\ell(\nu^\ell(\widetilde{\epsilon})))$$

are computed according to Algorithm 4.3 for all $0 \leq \widetilde{\epsilon} < \rho$. Then

$$\lim_{\widetilde{\epsilon}\downarrow 0} \omega^i(\nu^i(\widetilde{\epsilon})) = \omega^i(0) \quad \textit{and} \quad \lim_{\widetilde{\epsilon}\downarrow 0} \boldsymbol{c}(\boldsymbol{\mu}(\widetilde{\epsilon})) = \boldsymbol{c}(\mathbf{0}).$$

Proof. First note that $\underline{H}^*\beta\boldsymbol{e}_1 \neq \mathbf{0}$ implies that $\beta > 0$ and $\rho > 0$. Since $\underline{H}$ is full rank, the maps

$$\nu \mapsto \boldsymbol{c}^i(\nu), \quad \nu \mapsto D\boldsymbol{c}^i(\nu), \quad \text{and} \quad \boldsymbol{\mu} \mapsto \boldsymbol{c}(\boldsymbol{\mu})$$

are continuous for all $\nu \geq 0$ and $\boldsymbol{\mu} \geq \mathbf{0}$, where the latter bound should be interpreted element-wise. Hence

$$\lim_{\nu\downarrow 0} \boldsymbol{c}^i(\nu) = \boldsymbol{c}^i(0), \quad \lim_{\nu\downarrow 0} D\boldsymbol{c}^i(\nu) = D\boldsymbol{c}^i(0), \quad \text{and} \quad \lim_{\boldsymbol{\mu}\downarrow \mathbf{0}} \boldsymbol{c}(\boldsymbol{\mu}) = \boldsymbol{c}(\mathbf{0}).$$

It remains to be shown that

$$\lim_{\widetilde{\epsilon}\downarrow 0} \nu^i(\widetilde{\epsilon}) = 0, \quad \|D\boldsymbol{c}^i(0)\| \neq 0, \quad \text{and} \quad \lim_{\widetilde{\epsilon}\downarrow 0} \boldsymbol{\mu}(\widetilde{\epsilon}) = \mathbf{0}. \tag{4.15}$$

Let $\widetilde{\epsilon}$ be restricted to the interval $[0, \rho/2]$ and define $\nu^i_{\max} = \sigma^2_{\max}(\underline{H}(K^i)^{-1})$. By Proposition 4.2,

$$0 \leq \nu^i(\widetilde{\epsilon}) \leq \frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}}\nu^i_{\max} \leq \nu^i_{\max},$$

which proves the first limit in (4.15). Furthermore, using the definitions of $\boldsymbol{c}^i(\nu^i(\widetilde{\epsilon}))$ and $D\boldsymbol{c}^i(\nu^i(\widetilde{\epsilon}))$ we find the bounds

$$0 < \rho\beta\frac{\sigma_{\min}(\underline{H})}{\|\underline{H}\|^2 + \nu^i_{\max}\|K^i\|^2} \leq \|\boldsymbol{c}^i(\nu^i(\widetilde{\epsilon}))\| \leq \rho\beta\|\underline{H}^+\boldsymbol{e}_1\|,$$

$$0 < \rho\beta\frac{\sigma_{\min}(\underline{H})\sigma^2_{\min}(K^i)}{(\|\underline{H}\|^2 + \nu^i_{\max}\|K^i\|^2)^2} \leq \|D\boldsymbol{c}^i(\nu^i(\widetilde{\epsilon}))\| \leq \rho\beta\frac{\|K^i\|^2\,\|\underline{H}^+\boldsymbol{e}_1\|}{\sigma^2_{\min}(\underline{H})},$$

which show that the inequality in (4.15) is satisfied. Moreover, the bounds show there exist $\omega_{\min}$ and $\omega_{\max}$ such that

$$0 < \omega_{\min} \leq \omega^i(\widetilde{\epsilon}) \leq \omega_{\max} < \infty.$$

Now, let $\mathbf{K}(\widetilde{\epsilon})$ be the nonsingular matrix satisfying

$$\mathbf{K}(\widetilde{\epsilon})^* \mathbf{K}(\widetilde{\epsilon}) = \sum_{i=1}^{\ell} \omega^i(\widetilde{\epsilon}) K^{i^*} K^i;$$

then it can be checked that

$$\|\underline{H}\mathbf{K}(\widetilde{\epsilon})^{-1}\|^2 \leq \frac{\|\underline{H}\|^2}{\min_i \omega_{\min} \sigma^2_{\min}(K^i)} < \infty.$$

Define the right-hand side of the equation above as M, then by Proposition 4.2, each entry of $\boldsymbol{\mu}(\widetilde{\epsilon})$ is bounded from below by 0 and from above by

$$\frac{\widetilde{\epsilon}}{\rho - \widetilde{\epsilon}} M \omega_{\max},$$

which goes to 0 as $\widetilde{\epsilon} \downarrow 0$. This proves the second limit in (4.15). □

Proposition 4.3 is related to [25, Thm 3.3.3], where it is shown that the solution of a standard form Tikhonov regularization problem converges to a minimum norm least squares solution when the discrepancy principle is used and the noise converges to zero.

In this section we have discussed a new parameter selection method. In the next section we will look at the effect of perturbations in the parameters on the obtained solutions.

4.5 Perturbation analysis

The goal of regularization is to make reconstruction robust with respect to noise. By extension, a high sensitivity to the regularization parameters is undesirable. Consider a set of perturbed parameters $\boldsymbol{\mu}_k + \Delta\boldsymbol{\mu}$; if $\|\Delta\boldsymbol{\mu}\|$ is sufficiently small

$$\begin{aligned} \boldsymbol{c}_k(\boldsymbol{\mu}_k + \Delta\boldsymbol{\mu}) &= \boldsymbol{c}_k(\boldsymbol{\mu}_k) + D\boldsymbol{c}_k(\boldsymbol{\mu}_k)\Delta\boldsymbol{\mu} + \mathcal{O}(\|\Delta\boldsymbol{\mu}\|^2) \\ &= \boldsymbol{c}_k(\boldsymbol{\mu}_k) - M^{-1}\Delta M \boldsymbol{c}_k(\boldsymbol{\mu}_k) + \mathcal{O}(\|\Delta\boldsymbol{\mu}\|^2), \end{aligned}$$

where M and ΔM are defined as

$$M = \underline{H}_k^* \underline{H}_k + \sum_{i=1}^{\ell} \mu_k^i K_k^{i^*} K_k^i \quad \text{and} \quad \Delta M = \sum_{i=1}^{\ell} \Delta\mu_k^i K_k^{i^*} K_k^i. \tag{4.16}$$

Therefore, one might choose $\boldsymbol{\mu}_k$ to minimize the sensitivity measure

$$\|D\boldsymbol{c}(\boldsymbol{\mu}_k)\Delta\boldsymbol{\mu}\| = \|M^{-1}\Delta M\boldsymbol{c}(\boldsymbol{\mu}_k)\|.$$

To see the connection with the previous section, suppose that $\boldsymbol{\mu}_k = \nu_k^i \boldsymbol{e}_i$ and $\Delta\boldsymbol{\mu} = \pm\|\Delta\boldsymbol{\mu}\|\boldsymbol{e}_i$, then

$$\|M^{-1}\Delta M\| \geq \frac{\|M^{-1}\Delta M\boldsymbol{c}_k(\boldsymbol{\mu}_k)\|}{\|\boldsymbol{c}_k(\boldsymbol{\mu}_k)\|} = \frac{\|D\boldsymbol{c}_k(\boldsymbol{\mu}_k)\Delta\boldsymbol{\mu}\|}{\|\boldsymbol{c}_k(\boldsymbol{\mu}_k)\|} = \frac{\|D\boldsymbol{c}_k^i(\nu_k^i)\|\,\|\Delta\boldsymbol{\mu}\|}{\|\boldsymbol{c}_k^i(\nu_k^i)\|} = \frac{\|\Delta\boldsymbol{\mu}\|}{\omega_k^i}.$$

Thus, larger weights ω_k^i correspond to smaller lower bounds on $\|M^{-1}\Delta M\|$. Having small lower bounds is desirable, since we show in Proposition 4.4 and 4.5 that minimizing $\|M^{-1}\Delta M\|$ is equivalent to minimizing upper bounds on the forward and backward errors respectively.

Proposition 4.4. *Given regularization parameters μ_k^i and perturbations $\mu_\star^i = \mu_k^i + \Delta\mu_k^i$, let $\boldsymbol{c}_k = \boldsymbol{c}_k(\boldsymbol{\mu}_k)$, $\boldsymbol{c}_\star = \boldsymbol{c}_k(\boldsymbol{\mu}_\star)$, $\boldsymbol{x}_k = X_k\boldsymbol{c}_k$, and $\boldsymbol{x}_\star = X_k\boldsymbol{c}_\star$. Assume $\underline{H}_k$ and all K_k^i are of full rank and define matrices M and ΔM as in* (4.16)*. If M and $M + \Delta M$ are nonsingular and the $\Delta\mu_k^i$ are sufficiently small so that $\|M^{-1}\Delta M\| < 1$, then*

$$\frac{\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|}{\|\boldsymbol{x}_k\|} \leq \frac{\|M^{-1}\Delta M\|}{1 - \|M^{-1}\Delta M\|}.$$

Proof. Observe that $\boldsymbol{c}_k = M^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1$ and $\boldsymbol{c}_\star = (M + \Delta M)^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1$. With a little manipulation we obtain

$$\boldsymbol{c}_\star = (M + \Delta M)^{-1}M\boldsymbol{c}_k = (I + M^{-1}\Delta M)^{-1}\boldsymbol{c}_k = \sum_{j=0}^{\infty}(-M^{-1}\Delta M)^j\boldsymbol{c}_k.$$

It follows that

$$\frac{\|\boldsymbol{c}_k - \boldsymbol{c}_\star\|}{\|\boldsymbol{c}_k\|} = \frac{1}{\|\boldsymbol{c}_k\|}\left\|\sum_{j=1}^{\infty}(-M^{-1}\Delta M)^j\boldsymbol{c}_k\right\| \leq \sum_{j=1}^{\infty}\|M^{-1}\Delta M\|^j \leq \frac{\|M^{-1}\Delta M\|}{1 - \|M^{-1}\Delta M\|}.$$

Since X_k has orthonormal columns, the result of the proposition follows. □

One may wonder if it is possible to pick a vector $\boldsymbol{f}$ close to $\beta\boldsymbol{e}_1$ such that

$$\boldsymbol{c}_k = (M + \Delta M)^{-1}\underline{H}_k^*\boldsymbol{f}.$$

Or in other words, given perturbed regularization parameters, is there a perturbation of $\beta\boldsymbol{e}_1$ such that the optimal approximation to the exact solution is obtained? The following proposition provides a positive answer.

Proposition 4.5. *Under the assumptions of Proposition 4.4, there exist vectors* $\boldsymbol{f}$ *and* $\boldsymbol{g}$ *such that* $\boldsymbol{c}_k = (M + \Delta M)^{-1}\underline{H}_k^*\boldsymbol{f}$ *and* $\boldsymbol{c}_\star = M^{-1}\underline{H}_k^*\boldsymbol{g}$. *Furthermore,* $\boldsymbol{f}$ *and* $\boldsymbol{g}$ *satisfy*

$$\frac{\|\beta\boldsymbol{e}_1 - \boldsymbol{f}\|}{\|\beta\boldsymbol{e}_1\|} \leq \kappa(\underline{H}_k)\frac{\|M^{-1}\Delta M\|}{1 - \|M^{-1}\Delta M\|},$$
$$\frac{\|\beta\boldsymbol{e}_1 - \boldsymbol{g}\|}{\|\beta\boldsymbol{e}_1\|} \leq \kappa(\underline{H}_k)\|M^{-1}\Delta M\|,$$

where $\kappa(\underline{H}_k)$ *is the condition number of* $\underline{H}_k$.

Proof. The vector $\boldsymbol{f}$ is easy to derive using the *Ansatz*

$$(M + \Delta M)^{-1}\underline{H}_k^*\boldsymbol{f} = M^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1.$$

Let $\underline{H}_k = QR$ denote the reduced QR-decomposition of $\underline{H}_k$, then

$$R^*Q^*\boldsymbol{f} = (M + \Delta M)M^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1$$

and

$$\boldsymbol{f} = QR^{-*}(M + \Delta M)M^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1 + (I - QQ^*)\boldsymbol{v}$$

for arbitrary $\boldsymbol{v}$. Indeed, it is easy to verify that the above vector satisfies

$$\boldsymbol{c}_k = (M + \Delta M)^{-1}\underline{H}_k^*\boldsymbol{f}.$$

If we choose $\boldsymbol{v} = \beta\boldsymbol{e}_1$, then

$$\boldsymbol{f} = QR^{-*}\Delta MM^{-1}R^*Q^*\beta\boldsymbol{e}_1 + \beta\boldsymbol{e}_1,$$

so that

$$\frac{\|\beta\boldsymbol{e}_1 - \boldsymbol{f}\|}{\|\beta\boldsymbol{e}_1\|} = \|QR^{-*}\Delta MM^{-1}R^*Q^*\boldsymbol{e}_1\| \leq \|R^{-*}\|\,\|R^*\|\,\|\Delta MM^{-1}\|.$$

Here $\|R^{-*}\|\,\|R^*\|$ is the condition number $\kappa(\underline{H}_k)$ and $\|\Delta MM^{-1}\| = \|M^{-1}\Delta M\|$, since both M and ΔM are symmetric. This proves the first part of the proposition.

The second part is analogous. In particular, we use the *Ansatz*

$$M^{-1}\underline{H}_k^*\boldsymbol{g} = (M + \Delta M)^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1$$

and derive

$$\boldsymbol{g} = R^{-*}QM(M + \Delta M)^{-1}\underline{H}_k^*\beta\boldsymbol{e}_1 + (I - QQ^*)\beta\boldsymbol{e}_1.$$

Again it is easy to verify that $\boldsymbol{c}_\star = M^{-1}\underline{H}_k^*\boldsymbol{g}$. Observe that $\boldsymbol{g}$ can be rewritten as

$$\boldsymbol{g} = R^{-*}Q((I + \Delta MM^{-1})^{-1} - I)R^*Q^*\beta\boldsymbol{e}_1 + \beta\boldsymbol{e}_1,$$

such that

$$\begin{aligned}\frac{\|\beta\boldsymbol{e}_1 - \boldsymbol{f}\|}{\|\beta\boldsymbol{e}_1\|} &= \|R^{-*}((I + \Delta MM^{-1})^{-1} - I)R^*Q^*\boldsymbol{e}_1\| \\ &\leq \|R^{-*}\|\ \|R^*\|\ \|(I + \Delta MM^{-1})^{-1} - I\|.\end{aligned}$$

Since $\|\Delta MM^{-1}\| = \|M^{-1}\Delta M\| < 1$, it follows that

$$\|(I + \Delta MM^{-1})^{-1} - I\| \leq \sum_{j=1}^{\infty} \|-\Delta MM^{-1}\|^j = \frac{\|M^{-1}\Delta M\|}{1 - \|M^{-1}\Delta M\|},$$

which concludes the proof. □

We have discussed forward and backward error bounds which help to choose our parameters. Now that we have investigated each of the three phases of our method, we are ready to show numerical results.

4.6 Numerical experiments

We benchmark our algorithm with problems from Regularization Tools by Hansen [27]. Each problem provides an ill-conditioned $n \times n$ matrix A, a solution vector $\boldsymbol{x}_\star$ of length n and a corresponding measured vector $\boldsymbol{b}$. We take $n = 1024$ and add a noise vector $\boldsymbol{e}$ to $\boldsymbol{b}$. The entries of $\boldsymbol{e}$ are drawn independently from the standard normal distribution. The noise vector is then scaled such that $\epsilon = \|\boldsymbol{e}\|$ equals $0.01\|\boldsymbol{b}\|$ or $0.05\|\boldsymbol{b}\|$ for 1% and 5% noise respectively. We use $\eta = 1.01$ for the discrepancy bound in (4.7). We test the algorithms with 1000 different noise vectors for every triplet A, $\boldsymbol{x}_\star$, and $\boldsymbol{b}$ and report the median results.

The algorithms terminate when the relative difference between two subsequent approximations is less than 0.01, when $\boldsymbol{x}_{k+1}$ is (numerically) linear dependent in X_k, when both U_{k+1} and none of the V_k^i can be expanded, or when a maximum number of iterations is reached. For Algorithm 4.2 we use a maximum of 20 iterations and for Algorithm 4.1 a maximum of $(\ell + 1) \times 20$ iterations. For the sake of a fair comparison, the algorithms return the best obtained approximations and their iteration numbers.

For each test problem, the tables below list the relative error obtained with Algorithm 4.1, abbreviated by E_{od}, and Algorithm 4.2, abbreviated by E_{md}. OD and MD stand for one direction and multidirectional respectively. Also listed

are the ratio ρ_E of E_{md} to E_{od} and the ratio ρ_{mv} of the number of matrix-vector products. That is,

$$\rho_E = \frac{E_{\text{md}}}{E_{\text{od}}} \quad \text{and} \quad \rho_{\text{mv}} = \frac{\#\text{ MVs Algorithm 4.2}}{\#\text{ MVs Algorithm 4.1}}.$$

Only matrix-vector multiplications with A, A^*, L^i, and L^{i^*} count towards the total number of MVs used by each algorithm. We note, however, that multiplications with L^i and L^{i^*} are often less costly than multiplications with A and A^*.

Table 4.1: One-parameter Tikhonov regularization results.

Noise	**1%**				**5%**			
Problem	E_{od}	E_{md}	ρ_E	ρ_{mv}	E_{od}	E_{md}	ρ_E	ρ_{mv}
Baart	$1.73 \cdot 10^{-1}$	$1.11 \cdot 10^{-1}$	0.64	1.93	$2.91 \cdot 10^{-1}$	$2.71 \cdot 10^{-1}$	0.93	1.53
Deriv2-1	$2.44 \cdot 10^{-1}$	$2.44 \cdot 10^{-1}$	1.00	1.00	$3.32 \cdot 10^{-1}$	$3.32 \cdot 10^{-1}$	1.00	0.78
Deriv2-2	$2.35 \cdot 10^{-1}$	$2.35 \cdot 10^{-1}$	1.00	0.83	$3.22 \cdot 10^{-1}$	$3.22 \cdot 10^{-1}$	1.00	0.78
Deriv2-3	$4.35 \cdot 10^{-2}$	$4.35 \cdot 10^{-2}$	1.00	0.92	$7.97 \cdot 10^{-2}$	$7.64 \cdot 10^{-2}$	0.96	1.17
Foxgood	$3.31 \cdot 10^{-2}$	$3.30 \cdot 10^{-2}$	1.00	0.67	$6.64 \cdot 10^{-2}$	$6.63 \cdot 10^{-2}$	1.00	0.67
Gravity-1	$3.85 \cdot 10^{-2}$	$3.41 \cdot 10^{-2}$	0.88	1.08	$7.39 \cdot 10^{-2}$	$6.86 \cdot 10^{-2}$	0.93	1.11
Gravity-2	$5.53 \cdot 10^{-2}$	$5.26 \cdot 10^{-2}$	0.95	1.10	$8.66 \cdot 10^{-2}$	$8.39 \cdot 10^{-2}$	0.97	1.11
Gravity-3	$1.03 \cdot 10^{-1}$	$9.21 \cdot 10^{-2}$	0.90	1.08	$1.14 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	0.97	1.11
Heat	$9.26 \cdot 10^{-2}$	$9.12 \cdot 10^{-2}$	0.99	1.05	$2.02 \cdot 10^{-1}$	$1.91 \cdot 10^{-1}$	0.95	1.37
Phillips	$2.50 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$	1.00	1.00	$4.52 \cdot 10^{-2}$	$4.52 \cdot 10^{-2}$	1.00	1.00

Table 4.1 lists the results for one-parameter Tikhonov regularization, where we used the following regularization operators. The first derivative operator L_1 with stencil $[1, -1]$ for Gravity-3, Heat-5, Heat, and Phillips. The second derivative operator L_2 with stencil $[1, -2, 1]$ for Deriv2-1, Deriv2-2, Foxgood, Gravity-1, and Gravity-2. The third derivative operator L_3 with stencil $[-1, 3, -3, 1]$ for Baart. The fifth derivative operator L_5 with stencil $[-1, 5, -10, 10, -5, 1]$ and Deriv2-3. The derivative operators L_d are of size $(n - d) \times n$.

The table shows that multidirectional subspace expansion can obtain small improvements in the relative error at the cost of a small number of extra matrix-vector products, especially for 1% noise. We stress that in these cases, Algorithm 4.1 is allowed to perform additional MVs, but converges with a higher relative error. If there is no improvement in the relative error, we see that multidirectional subspace expansion can improve convergence, for example, for the Deriv2 problems as well as Foxgood.

Table 4.2: Multiparameter Tikhonov regularization results.

Noise	1%				5%			
Problem	E_{od}	E_{md}	ρ_E	ρ_{mv}	E_{od}	E_{md}	ρ_E	ρ_{mv}
Baart	$1.72 \cdot 10^{-1}$	$5.39 \cdot 10^{-2}$	0.31	2.60	$2.84 \cdot 10^{-1}$	$2.59 \cdot 10^{-1}$	0.91	2.60
Deriv2-1	$2.27 \cdot 10^{-1}$	$5.82 \cdot 10^{-3}$	0.03	1.81	$3.21 \cdot 10^{-1}$	$2.91 \cdot 10^{-2}$	0.09	2.20
Deriv2-2	$2.29 \cdot 10^{-1}$	$2.03 \cdot 10^{-2}$	0.09	1.55	$2.95 \cdot 10^{-1}$	$4.91 \cdot 10^{-2}$	0.17	1.72
Deriv2-3	$4.35 \cdot 10^{-2}$	$4.32 \cdot 10^{-2}$	0.99	1.00	$7.71 \cdot 10^{-2}$	$7.71 \cdot 10^{-2}$	1.00	1.00
Foxgood	$3.29 \cdot 10^{-2}$	$1.10 \cdot 10^{-2}$	0.34	1.35	$6.26 \cdot 10^{-2}$	$5.44 \cdot 10^{-2}$	0.87	1.35
Gravity-1	$3.69 \cdot 10^{-2}$	$1.83 \cdot 10^{-2}$	0.50	1.18	$7.24 \cdot 10^{-2}$	$4.52 \cdot 10^{-2}$	0.63	1.63
Gravity-2	$5.52 \cdot 10^{-2}$	$3.97 \cdot 10^{-2}$	0.72	2.04	$8.52 \cdot 10^{-2}$	$6.96 \cdot 10^{-2}$	0.82	2.26
Gravity-3	$1.02 \cdot 10^{-1}$	$9.24 \cdot 10^{-2}$	0.91	1.89	$1.14 \cdot 10^{-1}$	$1.08 \cdot 10^{-1}$	0.95	1.72
Heat	$8.79 \cdot 10^{-2}$	$8.77 \cdot 10^{-2}$	1.00	1.19	$1.97 \cdot 10^{-1}$	$1.83 \cdot 10^{-1}$	0.93	1.40
Phillips	$2.49 \cdot 10^{-2}$	$2.47 \cdot 10^{-2}$	0.99	1.21	$4.08 \cdot 10^{-2}$	$4.01 \cdot 10^{-2}$	0.98	1.40

Table 4.2 lists the results for multiparameter Tikhonov regularization. We have used the following regularization operators for each problem: the derivative operator L_d as listed above, the identity operator I, and the orthogonal projection $I - N_d N_d^*$, where the columns of N_d are an orthonormal basis for the nullspace $\mathcal{N}(L_d)$.

Overall, we observe larger improvements in the relative error for multidirectional subspace expansion, but also a larger number of MVs. We no longer see cases where multidirectional subspace expansion terminates with fewer MVs. In fact, the relative error is the same for Heat, although more MVs are required. Finally, Figure 4.1 illustrates an example of the improved results which can be obtained by using multidirectional subspace expansion.

In the next tests we attempt to reconstruct the original image from a blurred and noisy observation. Consider an $n \times n$ grayscale image with pixel values in the interval $[0, 1]$. Then $\boldsymbol{x}$ is a vector of length n^2 obtained by stacking the columns of the image below each other. The matrix A represents a Gaussian blurring operator, generated with `blur` from Regularization Tools. The matrix A is block-Toeplitz with half-bandwidth `band=11` and the amount of blurring is given by the variance `sigma=5`. The entries of the noise vector $\boldsymbol{e}$ are independently drawn from the standard normal distribution after which the vector is scaled such that $\epsilon = \mathbb{E}[\|\boldsymbol{e}\|] = 0.05\|\boldsymbol{b}\|$. We take η such that $\|\boldsymbol{e}\| \leq \eta\epsilon$ in 99.9% of the cases, that is,

$$\eta = 1 + \frac{3.090232}{\sqrt{2n^2}}. \tag{4.17}$$

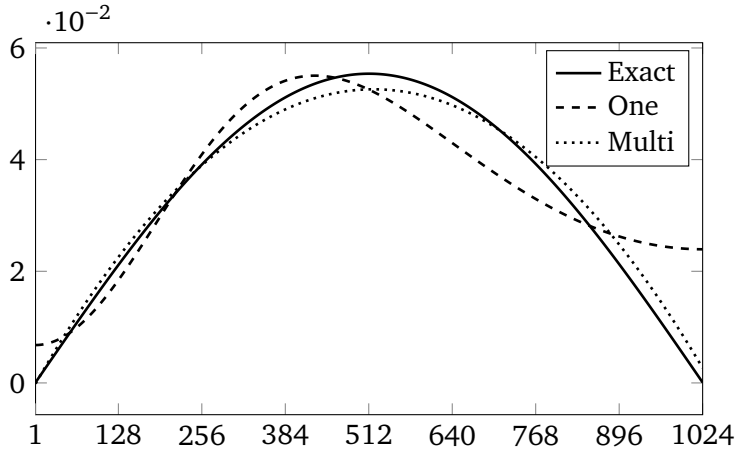


Figure 4.1: baart test matrix with $n = 1024$ and 1% noise. The solid line is the exact solution. The dashed line is the solution obtained with multiparameter regularization and the residual subspace expansion (Algorithm 4.1). The dotted line is the solution obtained with multiparameter regularization and multidirectional subspace expansion (Algorithm 4.2).

For regularization we choose an approximation to the Perona–Malik [70] operator

$$\mathcal{L}(\boldsymbol{x}) = \operatorname{div}(g(|\nabla \boldsymbol{x}|^2)\nabla \boldsymbol{x}),$$
$$g(s) = e^{-s/\rho^2} \qquad (\rho > 0),$$

where ρ is a small positive constant. Because $\mathcal{L}$ is a nonlinear operator, we first perform a small number of iterations with a finite difference approximation $L_{\boldsymbol{b}}$ of $\mathcal{L}(\boldsymbol{b})$. The resulting intermediate solution $\widetilde{\boldsymbol{x}}$ is used for a new approximation $L_{\tilde{\boldsymbol{x}}}$ of $\mathcal{L}(\widetilde{\boldsymbol{x}})$. Finally, we run the algorithms a second time with $L_{\widetilde{\boldsymbol{x}}}$ and more iterations; see Reichel, Sgallari, and Ye [71] for more information regarding the implementation of the Perona–Malik operator. As a measure of the reconstruction quality we use the peak signal-to-noise ratio (PSNR) given by

$$-20 \log_{10}\left(\frac{\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|}{n}\right),$$

where a higher value corresponds to a higher quality reconstruction.

We use the Chinese Lu symbol as a test image, see Figure 4.2; $\rho = 0.1$ and $\rho = 0.075$ for the Perona–Malik operator, 25 iterations for the first run, and 500 iterations for the second run. The convergence history in Figure 4.3 shows that $\rho = 0.1$ leads to faster convergence, while $\rho = 0.075$ leads to a higher PSNR. We also observe that, depending on the regularization operator, multidirectional

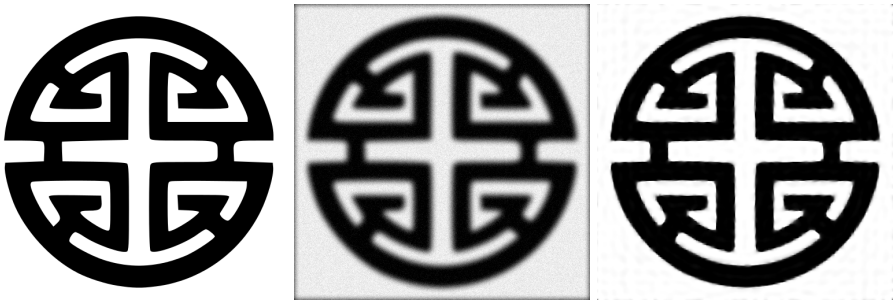


Figure 4.2: Deblurring results for lu. The original (left), observed (middle), and reconstructed images (right).

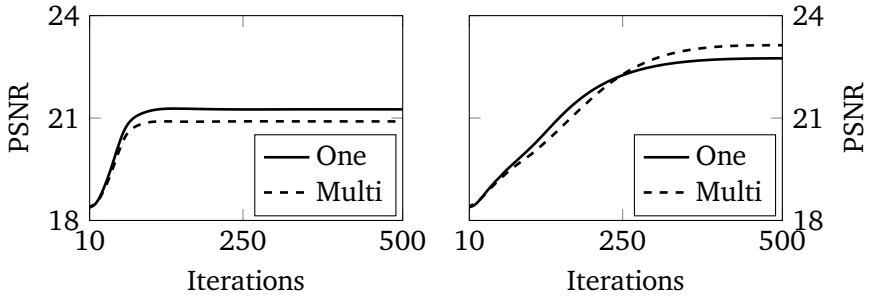


Figure 4.3: Convergence history for lu with $\rho = 0.1$ (left) and $\rho = 0.075$ (right).

Table 4.3: The number of iterations, matrix-vector products, and wall-clock time required by Algorithms 4.1 and 4.2 for deblurring lu and reaching the highest PSNR achieved by the less accurate algorithm. Results for $\rho = 0.1$ and the PSNR 20.90 in the upper rows, and $\rho = 0.075$ and the PSNR 22.75 in the lower rows.

Method	**#Itn**	**A**	**A***	**L**	**L***	**Total**	**Time (s)**
Alg 4.1	74	74	74	74	73	295	27
Alg 4.2	286	286	286	563	277	1689	312
Alg 4.1	500	500	500	500	499	1999	662
Alg 4.2	303	597	303	597	294	1791	360

subspace expansion may allow the convergence to a more accurate solution. In both cases, the algorithm with the most accurate solution obtains the best solution of the other algorithm with fewer iterations. However, because multidirectional subspace expansion requires extra matrix-vector products we need to be careful when investigating the performance difference. A detailed breakdown of the number of matrix-vector products used, as well as the wall-clock time, is shown in Table 4.3. It can be seen that although there is only a small difference in the number of total matrix-vector products used for $\rho = 0.075$, there is a large improvement in wall-clock time. This improvement can be explained by the lower orthogonalization cost for smaller subspaces, and the use of block operations which can only be used in Algorithm 4.2. For reference, the runtimes were obtained on an Intel Core i7-3770 and with MATLAB R2015b on 64-bit Linux 4.2.5.

4.7 Conclusion

We have presented a new method for large-scale Tikhonov regularization problems. In accordance with Algorithm 4.2, the method combines a new multidirectional subspace expansion with optional truncation to produce a higher quality search space. The multidirectional expansion generates a richer search space, whereas the truncation ensures moderate growth. Numerical results illustrate that our method can yield more accurate results or faster convergence. Furthermore, using block methods can partially offset the increased amount of work per iteration of the multidirectional method. In addition, the total orthogonalization cost may be lower when higher quality approximations are available in smaller subspaces. We have also presented lower and upper bounds on the regularization parameter when the discrepancy principle is applied to one-parameter regularization. These lower and upper bounds can be used in particular to initiate the bisection or the secant method. In addition, we have introduced a straightforward parameter choice for multiparameter regularization, as summarized by Algorithm 4.3. The parameter selection satisfies the discrepancy principle, and is based on easy to compute derivatives that are related to the perturbation results of Section 4.5.

Chapter 5

Generalized Davidson and multidirectional-type methods for the generalized singular value decomposition

Abstract. We propose new iterative methods for computing nontrivial extremal generalized singular values and vectors. The first method is a generalized Davidson-type algorithm and the second method employs a multidirectional subspace expansion technique. Essential to the latter is a fast truncation step designed to remove a low quality search direction and to ensure moderate growth of the search space. Both methods rely on thick restarts and may be combined with two different deflation approaches. We argue that the methods have monotonic and (asymptotic) linear convergence, derive and discuss locally optimal expansion vectors, and explain why the fast truncation step ideally removes search directions orthogonal to the desired generalized singular vector. Furthermore, we identify the relation between our generalized Davidson-type algorithm and the Jacobi–Davidson algorithm for the generalized singular value decomposition. Finally, we generalize several known convergence results for the Hermitian eigenvalue problem to the Hermitian positive definite generalized eigenvalue problem. Numerical experiments indicate that both methods are competitive.

Key words. Generalized singular value decomposition, GSVD, generalized singular value, generalized singular vector, generalized Davidson, multidirectional subspace expansion, subspace truncation, thick restart.

AMS subject classification. 15A18, 15A23, 15A29, 65F15, 65F22, 65F30, 65F50.

5.1 Introduction

The generalized singular value decomposition (GSVD) [66] is a generalization of the standard singular value decomposition (SVD), and is used in, for example, linear discriminant analysis [40], the method of particular solutions [8], general form Tikhonov regularization [28, Sec. 5.1], and more [1]. Computing the full GSVD with direct methods can be prohibitively time-consuming for large problem sizes; however, for many applications is suffices to compute only a few of the

largest or smallest generalized singular values and vectors. As a result, iterative methods may become attractive when the matrices involved are large and sparse.

An early iterative approach based on a modified Lanczos method was introduced by Zha [95], and later a variation by Kilmer, Hansen, and Español [47]. Both methods are inner-outer methods that require the solution to a least squares problem in each iteration, which may be computationally expensive. An approach that naturally allows for inexact solutions is the Jacobi–Davidson-type method (JDGSVD) introduced in [32]; however, this is still an inner-outer method. Alternatives to the previously mentioned methods include iterative methods designed for (symmetric positive definite) generalized eigenvalue problems, in particular generalized Davidson [48, 61] and LOBPCG [49]. These methods compute only the right generalized singular vectors and require additional steps to determine the left generalized singular vectors. More importantly, applying these methods involves squaring potentially ill-conditioned matrices.

In this chapter we discuss two new and competitive iterative methods for the computation of extremal generalized singular values and corresponding generalized singular vectors. The first can be seen as a generalized Davidson-type algorithm for the GSVD, while the second method builds upon the first, but uses multidirectional subspace expansion alongside a fast subspace truncation. The multidirectional subspace expansion is intended to produce improved search directions, whereas the subspace truncation is designed to remove low-quality search directions that are ideally orthogonal to the desired generalized singular vector. Both methods can be used to compute either the smallest or the largest generalized singular values of a matrix pair, or to approximate the truncated GSVD (TGSVD). A crucial part of both methods is a thick restart that allows for the removal of unwanted elements.

The remainder of this chapter is organized as follows. We derive a generalized Davidson-type algorithm for the GSVD in the next section, and prove multiple related theoretical properties. We subsequently discuss a B^*B-orthonormal version of the algorithm and its connection to JDGSVD in Section 5.3. In Section 5.4, we examine locally optimal search directions and argue for a multidirectional subspace expansion followed by a fast subspace truncation; then we present our second algorithm. In Section 5.5, we explore the deflation of generalized singular values and generalized singular vectors. We generalize several known error bounds for the Hermitian eigenvalue problem to results for the generalized singular value decomposition in Section 5.6. Finally, we consider numerical examples and experiments in Section 5.7, and end with conclusions in Section 5.8.

5.2 Generalized Davidson for the GSVD

Triangular and diagonal are two closely related forms of the GSVD. The triangular form is practical for the derivation and implementation of our methods, while the diagonal form is particularly relevant for the analysis. We adopt the definitions from Bai [1], but with a slightly more compact presentation. Let A be an $m \times n$ matrix, B a $p \times n$ matrix, and assume for the sake of simplicity that $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$; then $\operatorname{rank}([A^T\ B^T]^T) = n$ and there exist unitary matrices U, V, W, an $m \times n$ matrix Σ_A, a $p \times n$ matrix Σ_B, and a nonsingular upper-triangular $n \times n$ matrix R such that

$$AW = U\Sigma_A R \quad \text{and} \quad BW = V\Sigma_B R. \tag{5.1}$$

The matrices Σ_A and Σ_B satisfy

$$\Sigma_A^T\Sigma_A = \operatorname{diag}(c_1^2, \ldots, c_n^2), \qquad \Sigma_B^T\Sigma_B = \operatorname{diag}(s_1^2, \ldots, s_n^2), \qquad \Sigma_A^T\Sigma_A + \Sigma_B^T\Sigma_B = I,$$

and can be partitioned as

$$\begin{array}{c} \\ l \\ (n-p)_+ \\ (m-n)_+ \end{array} \begin{array}{c} \begin{array}{ccc} l & (n-p)_+ & (n-m)_+ \end{array} \\ \begin{bmatrix} D_A & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \begin{array}{c} \\ l \\ (n-m)_+ \\ (p-n)_+ \end{array} \begin{array}{c} \begin{array}{ccc} l & (n-p)_+ & (n-m)_+ \end{array} \\ \begin{bmatrix} D_B & 0 & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{bmatrix} \end{array},$$

where $l = \min\{m, p, n, m+p-n\}$, $(\cdot)_+ = \max\{\cdot, 0\}$, and D_A and D_B are diagonal matrices with nonnegative entries. The generalized singular pairs (c_j, s_j) are nonnegative and define the regular generalized singular values $\sigma_j = \infty$ if $s_j = 0$ and $\sigma_j = c_j/s_j$ otherwise. Hence, we call a generalized singular pair (c_j, s_j) large if σ_j is large and small if σ_j is small, and additionally refer to the largest and smallest σ_j as $\sigma_{\max}$ and $\sigma_{\min}$, respectively. The diagonal counterpart of (5.1) is

$$AX = U\Sigma_A \quad \text{and} \quad BX = V\Sigma_B \quad \text{with} \quad X = WR^{-1}, \tag{5.2}$$

and is useful because the columns of X are the (right) singular vectors $\boldsymbol{x}_j$ and satisfy, for instance,

$$s_j^2 A^*A\boldsymbol{x}_j = c_j^2 B^*B\boldsymbol{x}_j. \tag{5.3}$$

The assumption $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$ is not necessary for the implementation of our algorithm; nevertheless, we will make this assumption for the remainder of the chapter to simplify our discussion and analysis. We may also assume without loss of generality that the desired generalized singular values are contained in the leading principal submatrices of the factors. Consequently, if $k < l$ and C_k, S_k, and R_k denote the leading $k \times k$ principal submatrices of Σ_A, Σ_B, and R; and U_k, V_k, W_k, and X_k denote the first k columns of U, V, W, and X; then $X_k = W_k R_k^{-1}$ and we can define the partial (or truncated) GSVD of (A, B) as

$$AW_k = U_k C_k R_k \quad \text{and} \quad BW_k = V_k S_k R_k.$$

We aim is to approximate this partial GSVD for a $k \ll n$.

Since (5.3) can be interpreted as a generalized eigenvalue problem, it appears reasonable to consider the search space

$$\begin{aligned}\mathcal{W}_k = \operatorname{span}\{&\widetilde{\boldsymbol{x}}_{(0)}, (\widetilde{s}_{(0)}^2 A^*A - \widetilde{c}_{(0)}^2 B^*B)\widetilde{\boldsymbol{x}}_{(0)},\\ &(\widetilde{s}_{(1)}^2 A^*A - \widetilde{c}_{(1)}^2 B^*B)\widetilde{\boldsymbol{x}}_{(1)}, \ldots, (\widetilde{s}_{(k-1)}^2 A^*A - \widetilde{c}_{(k-1)}^2 B^*B)\widetilde{\boldsymbol{x}}_{(k-1)}\},\end{aligned}$$

consisting of homogeneous residuals generated by the generalized Davidson method (c.f., e.g., [61, Sec. 11.2.4] and [48, Sec. 11.3.6]) applied to the matrix pencil (A^*A, B^*B). The quantities $\widetilde{\boldsymbol{x}}_{(j)}$, $\widetilde{c}_{(j)}$, and $\widetilde{s}_{(j)}$ are approximations to $\boldsymbol{x}_1$, c_1, and s_1 with respect to the search space $\mathcal{W}_j$. The challenge is to compute a basis W_k with orthonormal columns for $\mathcal{W}_k$ without using the products A^*A and B^*B; however, let us focus on the extraction phase first. We will later see that a natural subspace expansion follows as a consequence.

Given W_k, we can compute the reduced QR decompositions

$$AW_k = U_k H_k, \qquad BW_k = V_k K_k, \tag{5.4}$$

where U_k and V_k have k orthonormal columns and H_k and K_k are $k \times k$ and upper-triangular. To compute the approximate generalized singular values, let the triangular form GSVD of (H_k, K_k) be given by

$$H_k\widetilde{W} = \widetilde{U}\widetilde{C}\widetilde{R}, \qquad K_k\widetilde{W} = \widetilde{V}\widetilde{S}\widetilde{R},$$

where $\widetilde{U}$, $\widetilde{V}$, and $\widetilde{W}$ are orthonormal, C and S are diagonal, and R is upper triangular. At this point, we can readily form the approximate partial GSVD

$$A(W_k\widetilde{W}) = (U_k\widetilde{U})\widetilde{C}\widetilde{R}, \qquad B(W_k\widetilde{W}) = (V_k\widetilde{V})\widetilde{S}\widetilde{R}, \tag{5.5}$$

and determine the leading approximate generalized singular values and vectors. When the dimension of the search space $\mathcal{W}_k$ grows large, a thick restart can be

performed by partitioning the decompositions in (5.5) as

$$\begin{aligned} A\begin{bmatrix} W_k\widetilde{W}_1 & W_k\widetilde{W}_2 \end{bmatrix} &= \begin{bmatrix} U_k\widetilde{U}_1 & U_k\widetilde{U}_2 \end{bmatrix} \begin{bmatrix} \widetilde{C}_1 & \\ & \widetilde{C}_2 \end{bmatrix} \begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ & \widetilde{R}_{22} \end{bmatrix}, \\ B\begin{bmatrix} W_k\widetilde{W}_1 & W_k\widetilde{W}_2 \end{bmatrix} &= \begin{bmatrix} V_k\widetilde{V}_1 & V_k\widetilde{V}_2 \end{bmatrix} \begin{bmatrix} \widetilde{S}_1 & \\ & \widetilde{S}_2 \end{bmatrix} \begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ & \widetilde{R}_{22} \end{bmatrix}, \end{aligned} \tag{5.6}$$

and truncating to

$$A(W_k\widetilde{W}_1) = (U_k\widetilde{U}_1)\widetilde{C}_1\widetilde{R}_{11}, \qquad B(W_k\widetilde{W}_1) = (V_k\widetilde{V}_1)\widetilde{S}_1\widetilde{R}_{11}.$$

If there is need to reorder the c_j and s_j, then we can simply use the appropriate permutation matrix P and compute

$$\begin{aligned} A(W_k\widetilde{W}Q) &= (U_k\widetilde{U}P)(P^*CP)(P^*RQ), \\ B(W_k\widetilde{W}Q) &= (V_k\widetilde{V}P)(P^*SP)(P^*RQ), \end{aligned}$$

where Q is unitary and such that P^*RQ is upper triangular.

For a subsequent generalized Davidson-type expansion of the search space, let

$$\widetilde{\boldsymbol{u}}_1 = U_k\widetilde{U}_1\boldsymbol{e}_1, \qquad \widetilde{\boldsymbol{v}}_1 = V_k\widetilde{V}_1\boldsymbol{e}_1, \qquad \widetilde{\boldsymbol{w}}_1 = W_k\widetilde{W}_1\boldsymbol{e}_1, \quad \text{and} \quad \widetilde{\boldsymbol{x}}_1 = \widetilde{\boldsymbol{w}}_1/\widetilde{r}_{11}$$

be the approximate generalized singular vectors satisfying

$$A\widetilde{\boldsymbol{x}}_1 = \widetilde{c}_1\widetilde{\boldsymbol{u}}_1 \quad \text{and} \quad B\widetilde{\boldsymbol{x}}_1 = \widetilde{s}_1\widetilde{\boldsymbol{v}}_1.$$

Then the homogeneous residual given by

$$\boldsymbol{r} = (\widetilde{s}_1^2A^*A - \widetilde{c}_1^2B^*B)\widetilde{\boldsymbol{x}}_1 = \widetilde{c}_1\widetilde{s}_1(\widetilde{s}_1A^*\widetilde{\boldsymbol{u}}_1 - \widetilde{c}_1B^*\widetilde{\boldsymbol{v}}_1) \tag{5.7}$$

suggests the expansion vector $\widetilde{\boldsymbol{r}} = \widetilde{s}_1A^*\widetilde{\boldsymbol{u}}_1 - \widetilde{c}_1B^*\widetilde{\boldsymbol{v}}_1$, which is orthogonal to W_k. The residual norm $\|\boldsymbol{r}\|$ goes to zero as the generalized singular value and vector approximations converge, and we recommend terminating the iterations when the right-hand side of

$$\frac{\|\boldsymbol{r}\|}{(\widetilde{s}_1^2\|A^*A\| + \widetilde{c}_1^2\|B^*B\|)\|\widetilde{\boldsymbol{x}}_1\|} \le \frac{\sqrt{n}\,|\widetilde{r}_{11}|\,\|\boldsymbol{r}\|}{\widetilde{s}_1^2\|A^*A\|_1 + \widetilde{c}_1^2\|B^*B\|_1} \tag{5.8}$$

is sufficiently small. The left-hand side is the normwise backward error by Tisseur [88], and the right-hand side is an alternative that can be approximated efficiently; for example, using the `normest1` function in MATLAB, which does not require computing the matrix products A^*A and B^*B explicitly. The GDGSVD algorithm is summarized in Algorithm 5.1.

Algorithm 5.1 (Generalized Davidson for the GSVD (GDGSVD))**.**
Input: Matrix pair (A, B), starting vector $\boldsymbol{w}_0$, minimum and maximum dimensions $j < \ell$.
Output: $AW_j = U_j C_j R_j$ and $BW_j = V_j S_j R_j$ approximating a partial GSVD.

```
1.  Let r̃ = w_0.
2.  for number of restarts and not converged (cf., e.g., (5.8)) do
3.      for k = 1, 2, ..., ℓ do
4.          w_k = r̃/‖r̃‖.
5.          Update AW_k = U_k H_k and BW_k = V_k K_k.
6.          Compute H_k = Ũ C̃ R̃ W̃* and K_k = Ṽ S̃ R̃ W̃*.
7.          Let r̃ = s_1 A* ũ_1 − c_1 B* ṽ_1.
8.          if j ≤ k and converged (cf., e.g., (5.8)) then break
9.      end
10.     Partition Ũ, Ṽ, W̃, C̃, S̃, and R̃ according to (5.6).
11.     Let U_j = U_k Ũ_1, V_j = V_k Ṽ_1, and W_j = W_k W̃_1.
12.     Let H_j = C̃_1 R̃_11 and K_j = S̃_1 R̃_11.
13. end
```

By design, the largest (or smallest) Ritz values are preserved after the restart; moreover, the generalized singular values increase (or decrease) monotonically per iteration as indicated by the proposition below. We wish to emphasize that the proof of the proposition does not require B^*B to be nonsingular, as opposed to the Courant–Fischer minimax principles for the generalized eigenvalue problem.

Proposition 5.1. *Let $\mathcal{W}_k$ and $\mathcal{W}_{k+1}$ be subspaces of dimensions k and $k+1$, respectively, and such that $\mathcal{W}_k \subset \mathcal{W}_{k+1}$. If $\sigma_{\max}(\mathcal{W})$ and $\sigma_{\min}(\mathcal{W})$ denote the maximum and minimum generalized singular values of A and B with respect to the subspace $\mathcal{W}$, then*

$$\sigma_{\max} \geq \sigma_{\max}(\mathcal{W}_{k+1}) \geq \sigma_{\max}(\mathcal{W}_k) \geq \sigma_{\min}(\mathcal{W}_k) \geq \sigma_{\min}(\mathcal{W}_{k+1}) \geq \sigma_{\min}.$$

Proof. Both A^*A and B^*B may be singular; therefore, we consider the pencil

$$(A^*A,\, A^*A + B^*B) = (A^*A,\, X^{-*}X^{-1})$$

with generalized eigenvalues c_i^2 and note that $\sigma_i^2 = c_i^2/(1-c_i^2)$ with the convention that $1/0 = \infty$. Applying the Courant–Fischer minimax principles yields

$$\begin{aligned} c_1 \geq \max_{\mathbf{0}\neq\boldsymbol{w}\in\mathcal{W}_{k+1}} \frac{\|A\boldsymbol{w}\|}{\|X^{-1}\boldsymbol{w}\|} &\geq \max_{\mathbf{0}\neq\boldsymbol{w}\in\mathcal{W}_k} \frac{\|A\boldsymbol{w}\|}{\|X^{-1}\boldsymbol{w}\|} \\ &\geq \min_{\mathbf{0}\neq\boldsymbol{w}\in\mathcal{W}_k} \frac{\|A\boldsymbol{w}\|}{\|X^{-1}\boldsymbol{w}\|} \geq \min_{\mathbf{0}\neq\boldsymbol{w}\in\mathcal{W}_{k+1}} \frac{\|A\boldsymbol{w}\|}{\|X^{-1}\boldsymbol{w}\|} \geq c_n. \end{aligned}$$

□

Proposition 5.1 implies that if a basis W_k for a subspace $\mathcal{W}_k$ is computed by Algorithm 5.1, then

$$\sigma_{\max}(\mathcal{W}_k) = \max_{\mathbf{0}\neq \boldsymbol{w}\in\mathcal{W}_k} \frac{\|A\boldsymbol{w}\|}{\|X^{-1}\boldsymbol{w}\|} = \max_{\boldsymbol{c}\neq\mathbf{0}} \frac{\|AW_k\boldsymbol{c}\|}{\|[A^T\ B^T]^T W_k\boldsymbol{c}\|} = \max_{\boldsymbol{c}\neq\mathbf{0}} \frac{\|H_k\boldsymbol{c}\|}{\|[H_k^T\ K_k^T]^T\boldsymbol{c}\|};$$

that is, the largest generalized singular value of the matrix pair (A, B) with respect to the subspace $\mathcal{W}_k$ is the largest generalized singular value of (H_k, K_k). A similar statement holds for the smallest generalized singular value. Furthermore, the matrix pair (H_k, K_k) is optimal in the sense of the following proposition.

Proposition 5.2. *Let the M-Frobenius norm for a Hermitian positive definite matrix M be defined as $\|Y\|_{F,M}^2 = \operatorname{trace}(Y^*MY)$. Now consider the decompositions from* (5.4) *and define the residuals*

$$\begin{aligned} R_1(G) &= AW_k - U_kG, \\ R_2(G) &= BW_k - V_kG, \\ R_3(G) &= A^*U_k - B^*V_kG^*, \\ R_4(G) &= B^*V_k - A^*U_kG^*; \end{aligned}$$

then the following results hold.

1. *$G = H_k = U_k^*AW_k$ minimizes $\|R_1(G)\|_2$ and is the unique minimizer of $\|R_1(H_k)\|_F$.*
2. *$G = K_k = V_k^*BW_k$ minimizes $\|R_2(G)\|_2$ and is the unique minimizer of $\|R_2(K_k)\|_F$.*
3. *If B^*B is nonsingular, then $G = H_kK_k^{-1}$ minimizes $\|R_3(G)\|_{(B^*B)^{-1}}$ and is the unique minimizer of R_3 with respect to the $(B^*B)^{-1}$-Frobenius norm.*
4. *If A^*A is nonsingular, then $G = K_kH_k^{-1}$ minimizes $\|R_4(G)\|_{(A^*A)^{-1}}$ and is the unique minimizer of R_4 with respect to the $(A^*A)^{-1}$-Frobenius norm.*

Proof. With the observation that $A^*U_k = A^*AW_kH_k^{-1}$ and $B^*V_k = B^*BW_kK_k^{-1}$, the proof becomes a straightforward adaptation of [32, Thm 2.1]. □

Propositions 5.1 and 5.2 demonstrate that the convergence behavior of Algorithm 5.1 is monotonic, and that the computed H_k and K_k are in some sense optimal for the search space $\mathcal{W}_k = \operatorname{span}(W_k)$; however, the propositions make no statement regarding the quality of the subspace expansion. A locally optimal residual-type subspace expansion can be derived with inspiration from Ye [94].

Proposition 5.3. *Define*

$$R_k = A^*AW_k(H_k^*H_k + K_k^*K_k)^{-1}K_k^*K_k - B^*BW_k(H_k^*H_k + K_k^*K_k)^{-1}H_k^*H_k$$

and let $\boldsymbol{r} = R_k\boldsymbol{c}$*; then*

$$\cos^2(\boldsymbol{x}_1, [W_k\ \boldsymbol{r}]) = \cos^2(\boldsymbol{x}_1, W_k) + \cos^2(\boldsymbol{x}_1, \boldsymbol{r})$$

is maximized for $\boldsymbol{c} = R_k^+\boldsymbol{x}_1$.

Proof. Since $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\boldsymbol{0}\}$ we also have $\mathcal{N}(H_k) \cap \mathcal{N}(K_k) = \{\boldsymbol{0}\}$, which implies that $H_k^*H_k + K_k^*K_k$ is invertible and R_k is well-defined. Furthermore, it is now straightforward to verify that

$$W_k^*R_k = H_k^*H_k(H_k^*H_k + K_k^*K_k)^{-1}K_k^*K_k - K_k^*K_k(H_k^*H_k + K_k^*K_k)^{-1}H_k^*H_k = 0$$

using the GSVD of H_k and K_k. It follows that

$$\|[W_k\ \boldsymbol{r}]^*\boldsymbol{x}_1\|^2 = \|W_k^*\boldsymbol{x}_1\|^2 + |\boldsymbol{r}^*\boldsymbol{x}_1|^2,$$

which realizes its maximum for $\boldsymbol{c} = R_k^+\boldsymbol{x}_1$. □

Different choices for R_k in Proposition 5.3 are possible; however, the current choice does not require additional assumptions on, for instance, H_k and K_k. Regardless of the choice of R_k, computing the optimal expansion vector is generally impossible without a priori knowledge of the desired generalized singular vector $\boldsymbol{x}_1$. Therefore, we expand the search space with a residual-type vector similar to generalized Davidson. The convergence of generalized Davidson is closely connected to steepest descent and has been studied extensively; see, for example, Ovtchinnikov [64, 65] and references therein. For completeness, we add the following asymptotic bound for the GSVD.

Proposition 5.4. *Let* (c_1, s_1) *be the smallest generalized singular pair of* (A, B) *with corresponding generalized singular vector* $\boldsymbol{x}_1$*, and assume the pair is simple. Define the Hermitian positive definite operator* $M = s_1^2A^*A - c_1^2B^*B$ *restricted to the domain perpendicular to* $(A^*A + B^*B)\boldsymbol{x}_1 = X^{-*}\boldsymbol{e}_1$*, and let the eigenvalues of* M *be given by*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} > 0.$$

Furthermore, let $\widetilde{\boldsymbol{x}}_1$*,* $\widetilde{c}_1 = \|A\widetilde{\boldsymbol{x}}_1\|$*, and* $\widetilde{s}_1 = \|B\widetilde{\boldsymbol{x}}_1\|$ *approximate* $\boldsymbol{x}_1$*,* c_1*, and* s_1*, respectively, and be such that* $\widetilde{c}_1^2 + \widetilde{s}_1^2 = 1$*. If* $\widetilde{\boldsymbol{x}}_1 = \xi\boldsymbol{x}_1 + \boldsymbol{f}$ *for some scalar* ξ *and vector* $\boldsymbol{f} \perp X^{-*}\boldsymbol{e}_1$*; then*

$$\sin^2([\widetilde{\boldsymbol{x}}_1\ \boldsymbol{r}], \boldsymbol{x}_1) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1) + \mathcal{O}(\|\boldsymbol{f}\|^3),$$

where $\kappa = \lambda_1/\lambda_{n-1}$ is the condition number of M, and $\boldsymbol{r} = (\widetilde{s}_1^2 A^*A - \widetilde{c}_1^2 B^*B)\widetilde{\boldsymbol{x}}_1$ is the homogeneous residual.

Proof. We have

$$\widetilde{c}_1^2 = \widetilde{\boldsymbol{x}}_1^* A^*A \widetilde{\boldsymbol{x}}_1 = \xi^2 c_1^2 + \|A\boldsymbol{f}\|^2 \quad \text{and} \quad \widetilde{s}_1^2 = \widetilde{\boldsymbol{x}}_1^* B^*B \widetilde{\boldsymbol{x}}_1 = \xi^2 s_1^2 + \|B\boldsymbol{f}\|^2,$$

and it follows that

$$\boldsymbol{r} = \xi^2(s_1^2 A^*A - c_1^2 B^*B)\boldsymbol{f} + (\|B\boldsymbol{f}\|^2 A^*A - \|A\boldsymbol{f}\|^2 B^*B)(\xi\boldsymbol{x}_1 + \boldsymbol{f}) = \xi^2 M\boldsymbol{f} + \mathcal{O}(\|\boldsymbol{f}\|^2)$$

and

$$\boldsymbol{r}^*\boldsymbol{x}_1 = \xi \boldsymbol{f}^*(s_1^2 A^*A - c_1^2 B^*B)\boldsymbol{f} = \xi \boldsymbol{f}^* M \boldsymbol{f}.$$

Hence, $\widetilde{\boldsymbol{x}}_1 = \boldsymbol{x}_1$ if $\|\boldsymbol{r}\| = 0$ for $\widetilde{\boldsymbol{x}}_1$ sufficiently close to $\boldsymbol{x}_1$ and we are done. Otherwise, $\boldsymbol{r}$ is nonzero and perpendicular to $\widetilde{\boldsymbol{x}}_1$, so that

$$\sin^2([\widetilde{\boldsymbol{x}}_1\ \boldsymbol{r}], \boldsymbol{x}_1) = 1 - \cos^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1) - \cos^2(\boldsymbol{r}, \boldsymbol{x}_1) = \left(1 - \frac{\cos^2(\boldsymbol{r}, \boldsymbol{x}_1)}{\sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1)}\right) \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1).$$

Combining the above expressions, and using the fact that nontrivial orthogonal projectors have unit norm, yields

$$\frac{\cos^2(\boldsymbol{r}, \boldsymbol{x}_1)}{\sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1)} = \frac{|\boldsymbol{r}^*\boldsymbol{x}_1|^2}{\|\boldsymbol{r}\|^2 \|(I - \widetilde{\boldsymbol{x}}_1\widetilde{\boldsymbol{x}}_1^*)\boldsymbol{x}_1\|^2} \geq \frac{|\boldsymbol{f}^* M \boldsymbol{f}|^2}{\|M\boldsymbol{f}\|^2 \|\boldsymbol{f}\|^2} + \mathcal{O}(\|\boldsymbol{f}\|).$$

Using the Kantorovich inequality (cf., e.g., [26, p. 68]) we obtain

$$\begin{aligned}\sin^2([\widetilde{\boldsymbol{x}}_1\ \boldsymbol{r}], \boldsymbol{x}_1) &\leq \left(1 - \frac{4\lambda_1\lambda_{n-1}}{(\lambda_1 + \lambda_{n-1})^2}\right) \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1) + \mathcal{O}(\|\boldsymbol{f}\| \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1)) \\ &= \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1) + \mathcal{O}(\|\boldsymbol{f}\| \sin^2(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1)).\end{aligned}$$

Finally, $\widetilde{c}_1^2 + \widetilde{s}_1^2 = 1$ implies $\|\widetilde{\boldsymbol{x}}_1\| \geq \sigma_{\min}(X)$, so that

$$\sin(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}) \leq \sigma_{\min}^{-1}(X)\, \|\boldsymbol{f}\| = \mathcal{O}(\|\boldsymbol{f}\|).$$

□

The condition number κ from Proposition 5.4 may be large in practice, in which case the quantity $(\kappa - 1)/(\kappa + 1)$ is close to 1. However, this upper bound may be rather pessimistic and we will see considerably faster convergence during the numerical tests in Section 5.7.

5.3 B^*B-orthonormal GDGSVD

In the previous section we have derived the GDGSVD algorithm for an orthonormal basis of $\mathcal{W}_k$. An alternative is to construct a B^*B-orthonormal basis of $\mathcal{W}_k$, which allows us to use the SVD instead of the slower GSVD for the projected problem, as well as reduce the amount of work necessary for a restart. Another benefit is that the B^*B-orthonormality reveals the connection between GDGSVD and JDGSVD, a Jacobi–Davidson-type algorithm for the GSVD [32].

The derivation of B^*B-orthonormal GDGSVD is similar to the derivation of Algorithm 5.1. Suppose that B^*B is nonsingular, let $\widehat{W}_k$ be a basis of $\mathcal{W}_k$ satisfying $\widehat{W}_k^* B^*B \widehat{W}_k = I$, and compute the QR-decomposition

$$A\widehat{W}_k = \widehat{U}_k \widehat{H}_k, \tag{5.9}$$

where $\widehat{U}$ has orthonormal columns and $\widehat{H}_k$ is upper-triangular. Note that (5.9) can be obtained from the QR-decompositions in (5.4) by setting $\widehat{W}_k = W_k K_k^{-1}$, $\widehat{U}_k = U_k$, and $\widehat{H}_k = H_k K_k^{-1}$. If $\widehat{H}_k = \widetilde{U}\Sigma\widetilde{W}^*$ is the SVD of $\widehat{H}_k$; then

$$A(\widehat{W}_k\widetilde{W}) = (\widehat{U}_k\widetilde{U})\Sigma,$$

which can be partitioned as

$$A\begin{bmatrix}\widehat{W}_k\widetilde{W}_1 & \widehat{W}_k\widetilde{W}_2\end{bmatrix} = \begin{bmatrix}\widehat{U}_k\widetilde{U}_1 & \widehat{U}_k\widetilde{U}_2\end{bmatrix}\begin{bmatrix}\Sigma_1 & \\ & \Sigma_2\end{bmatrix} \tag{5.10}$$

and truncated to $A\widehat{W}_k\widetilde{W}_1 = \widehat{U}_k\widetilde{U}_1\Sigma_1$. With $\widehat{\boldsymbol{u}}_1 = \widehat{U}_k\widetilde{U}\boldsymbol{e}_1$ and $\widehat{\boldsymbol{w}}_1 = \widehat{W}_k\widetilde{W}\boldsymbol{e}_1$ we get the residual

$$\boldsymbol{r} = (A^*A - \sigma_1^2 B^*B)\widehat{\boldsymbol{w}}_1 = \sigma_1(A\widehat{\boldsymbol{u}}_1 - \sigma_1 B^*B\widehat{\boldsymbol{w}}_1)$$

and the expansion vector $\widehat{\boldsymbol{r}} = A\widehat{\boldsymbol{u}}_1 - \sigma_1 B^*B\widehat{\boldsymbol{w}}_1$. The expansion vector $\widehat{\boldsymbol{r}}$ is orthogonal to $\widehat{W}_k$ in exact arithmetic, but should in practice still be orthogonalized with respect to $\widehat{W}_k$ prior to B^*B-orthogonalization in order to improve numerical stability and accuracy [30, Sec. 3.5]. Finally, in the B^*B-orthonormal case the suggested stopping condition (5.8) becomes

$$\frac{\|\boldsymbol{r}\|}{(\|A^*A\| + \sigma_1^2\|B^*B\|)\|\widehat{\boldsymbol{w}}_1\|} \leq \frac{\sqrt{n}\,\|\boldsymbol{r}\|}{(\|A^*A\|_1 + \sigma_1^2\|B^*B\|_1)\|\widehat{\boldsymbol{w}}_1\|} \leq \tau \tag{5.11}$$

for some tolerance τ. The algorithm is summarized below in Algorithm 5.2, where $\widehat{V}_k = B\widehat{W}_k$ has orthonormal columns.

Algorithm 5.2 (B^*B-orthonormal GDGSVD)**.**
Input: Matrix pair (A, B), starting vector $\boldsymbol{w}_0$, minimum and maximum dimensions $j < \ell$.
Output: Orthonormal $\widehat{U}_j$, B^*B-orthonormal $\widehat{W}_j$, and diagonal Σ_j satisfying $A\widehat{W}_j = \widehat{U}_j\Sigma_j$.

1. Let $\widehat{W}_0 = \widehat{V}_0 = []$ and $\widehat{\boldsymbol{r}} = \boldsymbol{w}_0$.
2. **for** *number of restarts* **and** *not converged (cf., e.g.,* (5.11)*)* **do**
3. **for** $k = 1, 2, \ldots, \ell$ **do**
4. $\widehat{\boldsymbol{w}}_k = (I - \widehat{W}_{k-1}(\widehat{W}_{k-1}^*\widehat{W}_{k-1})^{-1}\widehat{W}_{k-1}^*)\widehat{\boldsymbol{r}}$.
5. Compute $\widehat{\boldsymbol{v}}_k = B\widehat{\boldsymbol{w}}_k$.
6. B^*B-orthogonalize: $\widehat{\boldsymbol{w}}_k = \widehat{\boldsymbol{w}}_k - \widehat{W}_{k-1}\widehat{V}_{k-1}^*\widehat{\boldsymbol{v}}_k$.
7. $\widehat{\boldsymbol{v}}_k = (I - \widehat{V}_{k-1}\widehat{V}_{k-1}^*)\widehat{\boldsymbol{v}}_k$.
8. $\widehat{\boldsymbol{w}}_k = \widehat{\boldsymbol{w}}_k/\|\widehat{\boldsymbol{v}}_k\|$ and $\widehat{\boldsymbol{v}}_k = \widehat{\boldsymbol{v}}_k/\|\widehat{\boldsymbol{v}}_k\|$.
9. Update the QR-decomposition $A\widehat{W}_k = \widehat{U}_k\widehat{H}_k$.
10. Compute the SVD $\widehat{H}_k = \widetilde{U}\Sigma\widetilde{W}^*$.
11. $\widehat{\boldsymbol{r}} = A^*\widehat{U}_k\widetilde{\boldsymbol{u}}_1 - \sigma_1 B^*\widehat{V}_k\widetilde{\boldsymbol{w}}_1$.
12. **if** $j \leq k$ **and** *converged (cf., e.g.,* (5.11)*)* **then break**
13. **end**
14. Partition $\widetilde{U}$, Σ, and $\widetilde{W}$ according to (5.10).
15. Let $\widehat{U}_j = \widehat{U}_k\widetilde{U}_1$, $\widehat{V}_j = \widehat{V}_k\widetilde{W}_1$, and $\widehat{W}_j = \widehat{W}_k\widetilde{W}_1$.
16. Let $H_j = \Sigma_1$.
17. **end**

The product B^*B may be arbitrarily close to singularity, and a severely ill-conditioned B^*B may prove to be problematic despite the additional orthogonalization step in Algorithm 5.2. Therefore, we would generally advise against using Algorithm 5.2, and recommend using Algorithm 5.1 and orthonormal bases instead. However, B^*B-orthonormal GDGSVD relates nicely to JDGSVD on a theoretical level, regardless of the potential practical issues. In JDGSVD the search spaces $\widehat{U}_k$ and $\widehat{W}_k$ are repeatedly updated with the vectors $\boldsymbol{s} \perp \widehat{\boldsymbol{u}}_1$ and $\boldsymbol{t} \perp \widehat{\boldsymbol{w}}_1$, which are obtained by solving correction equations. Picking the updates

$$\boldsymbol{s} = (I - \widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^*)A\boldsymbol{r} \quad \text{and} \quad \boldsymbol{t} = \boldsymbol{r},$$

instead of solving the correction equations gives JDGSVD the same subspace expansions as B^*B-orthogonal GDGSVD. Furthermore, standard extraction in JDGSVD is performed by computing the SVD of $\widehat{U}_k^*A\widehat{W}_k$, which is identical to the extraction in B^*B-orthonormal GDGSVD. For harmonic Ritz extraction, JDGSVD uses the harmonic Ritz vectors $\widehat{U}_k\boldsymbol{c}$ and $\widehat{W}_k\boldsymbol{d}$, where $\boldsymbol{c}$ and $\boldsymbol{d}$ solve

$$\widehat{W}_k^*A^*A\widehat{W}_k\boldsymbol{d} = \sigma^2\widehat{W}_k^*B^*B\widehat{W}_k\boldsymbol{d} \quad \text{and} \quad \boldsymbol{c} = \sigma(\widehat{W}_k^*A^*\widehat{U}_k)^{-1}\widehat{W}_k^*B^*B\widehat{W}_k\boldsymbol{d}.$$

The above simplifies to

$$\widetilde{W}\Sigma^2\widetilde{W}^*\boldsymbol{d} = \sigma^2\boldsymbol{d} \quad \text{and} \quad \boldsymbol{c} = \sigma\widetilde{U}\Sigma^{-1}\widetilde{W}^*\boldsymbol{d},$$

for B^*B-orthonormal GDGSVD and produces the same primitive Ritz vectors as the standard extraction. To summarize, JDGSVD coincides with B^*B-orthonormal GDGSVD for specific expansion vectors, and there is no difference between standard and harmonic extraction in B^*B-orthonormal GDGSVD. The difference in practice between the two methods is primarily caused by the different expansion phases, where GDGSVD uses residual-type vectors and JDGSVD normally solves correction equations. In the next section we will discuss how the subspace expansion for GDGSVD may be further improved.

5.4 Multidirectional subspace expansion

While the residual vector $\boldsymbol{r}$ from (5.7) is a practical choice for the subspace expansion, it is not necessarily optimal. In fact, neither is the vector given by Proposition 5.3, which is only the optimal "residual-type" expansion vector. In their most general form, the desired expansion vectors are

(5.12)
$$\boldsymbol{a} - \boldsymbol{b}, \quad \text{where} \quad \boldsymbol{a} = (I - W_kW_k^*)A^*AW_k\boldsymbol{c}_\star \quad \text{and} \quad \boldsymbol{b} = (I - W_kW_k^*)B^*BW_k\boldsymbol{d}_\star,$$

for some "optimal" choice of $\boldsymbol{c}_\star$ and $\boldsymbol{d}_\star$. The following proposition characterizes $\boldsymbol{c}_\star$ and $\boldsymbol{d}_\star$.

Proposition 5.5. *Let R_k and $\boldsymbol{r}$ be defined as in Proposition 5.3, and assume that R_k has full column rank. If $R_k^*A^*AW_k$ and $R_k^*B^*BW_k$ are nonsingular and if $\boldsymbol{s} = S_k\boldsymbol{d}$ with*

$$S_k = (A^*AW_k - W_kH_k^*H_k)(R_k^*A^*AW_k)^{-1} - (B^*BW_k - W_kK_k^*K_k)(R_k^*B^*BW_k)^{-1};$$

then

$$\cos^2(\boldsymbol{x}_1, [W_k\ \boldsymbol{r}\ \boldsymbol{s}]) = \cos^2(\boldsymbol{x}_1, W_k) + \cos^2(\boldsymbol{x}_1, \boldsymbol{r}) + \cos^2(\boldsymbol{x}_1, \boldsymbol{s})$$

is maximized for $\boldsymbol{c} = R_k^+\boldsymbol{x}_1$ and $\boldsymbol{d} = S_k^+\boldsymbol{x}_1$. Moreover, for any $\boldsymbol{c}$, $\boldsymbol{d}$, and scalar t, the linear combination $R_k\boldsymbol{c} + tS_k\boldsymbol{d}$ can be written in the form of (5.12)*. The mapping from $\boldsymbol{c}$ and $\boldsymbol{d}$ to $\boldsymbol{c}_\star$ and $\boldsymbol{d}_\star$ is one-to-one if $t \neq 0$.*

Proof. For the first part of the proof, use that $W_k^*R_k = W_k^*S_k = R_k^*S_k = 0$. For the second part, define the shorthand $M = H_k^*H_k + K_k^*K_k$ and recall that

$$R_k = A^*AW_kM^{-1}K_k^*K_k - B^*BW_kM^{-1}H_k^*H_k.$$

Hence, for any $\boldsymbol{c}$, $\boldsymbol{d}$ and scalar t we have

$$\begin{aligned} R_k\boldsymbol{c} + tS_k\boldsymbol{d} &= (I - W_kW_k^*)R_k\boldsymbol{c} + t(I - W_kW_k^*)S_k\boldsymbol{d} \\ &= (I - W_kW_k^*)A^*AW_k\left(M^{-1}K_k^*K_k\boldsymbol{c} + t(R_k^*A^*AW_k)^{-1}\boldsymbol{d}\right) \\ &- (I - W_kW_k^*)B^*BW_k\left(M^{-1}H_k^*H_k\boldsymbol{c} + t(R_k^*B^*BW_k)^{-1}\boldsymbol{d}\right) \\ &= \boldsymbol{a} - \boldsymbol{b}, \end{aligned}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are defined as in (5.12) for the $\boldsymbol{c}_\star$ and $\boldsymbol{d}_\star$ satisfying

$$\begin{bmatrix} \boldsymbol{c}_\star \\ \boldsymbol{d}_\star \end{bmatrix} = \begin{bmatrix} M^{-1}H_k^*H_k & t(R_k^*A^*AW_k)^{-1} \\ M^{-1}K_k^*K_k & t(R_k^*B^*BW_k)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix}.$$

Finally, the matrix above is invertible if

$$t\det\begin{bmatrix} (R_k^*A^*AW_k)^{-1} & \\ & (R_k^*B^*BW_k)^{-1} \end{bmatrix} \cdot \det\begin{bmatrix} R_k^*A^*AW_kM^{-1}H_k^*H_k & I \\ R_k^*B^*BW_kM^{-1}K_k^*K_k & I \end{bmatrix} \neq 0,$$

where the first determinant is nonzero because its subblocks are invertible, and the second determinant equals

$$\det(R_k^*A^*AW_kM^{-1}H_k^*H_k - R_k^*B^*BW_kM^{-1}K_k^*K_k) = \det(R_k^*R_k) \neq 0$$

since R_k has full column rank. □

Let $\boldsymbol{r}$ and $\boldsymbol{s}$ be two nonzero orthogonal vectors; then the locally optimal search direction in $\mathcal{S} = \operatorname{span}\{\boldsymbol{r}, \boldsymbol{s}\}$ is the projection of the desired generalized singular vector $\boldsymbol{x}_1$ onto $\mathcal{S}$, and is given by

$$\frac{\boldsymbol{r}^*\boldsymbol{x}_1}{\boldsymbol{r}^*\boldsymbol{r}}\boldsymbol{r} + \frac{\boldsymbol{s}^*\boldsymbol{x}_1}{\boldsymbol{s}^*\boldsymbol{s}}\boldsymbol{s}. \tag{5.13}$$

The remaining orthogonal direction in $\mathcal{S}$ is

$$(\boldsymbol{x}_1^*\boldsymbol{s})\boldsymbol{r} - (\boldsymbol{x}_1^*\boldsymbol{r})\boldsymbol{s}, \tag{5.14}$$

which is perpendicular to $\boldsymbol{x}_1$. It is usually impossible to compute the vectors from Proposition 5.5 and the linear combination in (5.13) without a priori knowledge of $\boldsymbol{x}_1$. Therefore, the idea is to pick $\boldsymbol{r}$ and $\boldsymbol{s}$ or $\boldsymbol{a}$ and $\boldsymbol{b}$ based on a different criterion, expand the search space with both vectors, and to rely on the extraction process to determine a good new search direction. If successful, then (5.14) suggests that there is at least one direction in the enlarged search space that is (nearly)

perpendicular to $\boldsymbol{x}_1$. This direction may be removed to avoid excessive growth of the search space.

For example, we could use the approximate generalized singular pair and corresponding vectors from Section 5.2 and choose the vectors

$$\boldsymbol{a} = \widetilde{s}_1^2(I - W_k W_k^*)A^*A\widetilde{\boldsymbol{x}}_1 \quad \text{and} \quad \boldsymbol{b} = \widetilde{c}_1^2(I - W_k W_k^*)B^*B\widetilde{\boldsymbol{x}}_1$$

for expansion, and set

$$\boldsymbol{r} = \boldsymbol{a} - \boldsymbol{b} \quad \text{and} \quad \boldsymbol{s} = (\boldsymbol{r}^*\boldsymbol{b})\boldsymbol{a} - (\boldsymbol{r}^*\boldsymbol{a})\boldsymbol{b},$$

since the residual norm $\|\boldsymbol{r}\|$ is required anyway. Moreover, this choice ensures at least the same improvement per iteration as the residual expansion from generalized Davidson. After the expansion and extraction, a low-quality search direction may be removed. Below we describe the process in more detail.

In Section 5.2 we have seen that $A^*A\widetilde{\boldsymbol{x}}_1 = \widetilde{c}_1 A^*\widetilde{\boldsymbol{u}}_1$ and $B^*B\widetilde{\boldsymbol{x}}_1 = \widetilde{s}_1 B^*\widetilde{\boldsymbol{v}}_1$; hence, suppose that W_{k+2} is obtained by extending W_k with the $A^*\widetilde{\boldsymbol{u}}_1$ and $B^*\widetilde{\boldsymbol{v}}_1$ after orthonormalization. Then we can compute the reduced QR-decompositions

$$AW_{k+2} = U_{k+2}H_{k+2} \quad \text{and} \quad BW_{k+2} = V_{k+2}K_{k+2}, \tag{5.15}$$

and the triangular-form GSVD

$$H_{k+2}\begin{bmatrix}\widetilde{W}_{k+1} & \widetilde{\boldsymbol{w}}_{k+2}\end{bmatrix} = \begin{bmatrix}\widetilde{U}_{k+1} & \widetilde{\boldsymbol{u}}_{k+2}\end{bmatrix}\begin{bmatrix}\widetilde{C}_{k+1} & \\ & \widetilde{c}_{k+2}\end{bmatrix}\begin{bmatrix}\widetilde{R}_{k+1} & \widetilde{\boldsymbol{r}}_{k+1,k+2} \\ & \widetilde{r}_{k+2,k+2}\end{bmatrix},$$

$$K_{k+2}\begin{bmatrix}\widetilde{W}_{k+1} & \widetilde{\boldsymbol{w}}_{k+2}\end{bmatrix} = \begin{bmatrix}\widetilde{V}_{k+1} & \widetilde{\boldsymbol{v}}_{k+2}\end{bmatrix}\begin{bmatrix}\widetilde{S}_{k+1} & \\ & \widetilde{s}_{k+2}\end{bmatrix}\begin{bmatrix}\widetilde{R}_{k+1} & \widetilde{\boldsymbol{r}}_{k+1,k+2} \\ & \widetilde{r}_{k+2,k+2}\end{bmatrix},$$

where we may assume without loss of generality that $(\widetilde{c}_{k+2}, \widetilde{s}_{k+2})$ is the generalized singular pair furthest from the desired pair. By combining the partitioned decompositions above with (5.15), we see that the objective becomes the removal of $\operatorname{span}\{W_{k+2}\widetilde{\boldsymbol{w}}_{k+2}\}$ from the search space. One way to truncate this unwanted direction from the search space, is to perform a restart conform Section 5.2 and compute

$$U_{k+2}\widetilde{U}_{k+1}, \qquad V_{k+2}\widetilde{V}_{k+1}, \qquad W_{k+2}\widetilde{W}_{k+1}, \qquad \widetilde{C}_{k+1}\widetilde{R}_{k+1}, \quad \text{and} \quad \widetilde{S}_{k+1}\widetilde{R}_{k+1} \tag{5.16}$$

explicitly. However, with $\mathcal{O}(nk^2)$ floating-point operations per iteration, the computational cost of this approach is too high. The key to a faster method is to realize that we only need to be able to truncate

$$U_{k+2}\widetilde{\boldsymbol{u}}_{k+2}, \qquad V_{k+2}\widetilde{\boldsymbol{v}}_{k+2}, \qquad W_{k+2}\widetilde{\boldsymbol{w}}_{k+2}, \qquad \widetilde{c}_{k+2}, \quad \text{and} \quad \widetilde{s}_{k+2},$$

but do not require the matrices in (5.16). To this end, let P, Q, and Z be Householder reflections of the form

$$P = I - 2\frac{\boldsymbol{p}\boldsymbol{p}^*}{\boldsymbol{p}^*\boldsymbol{p}}, \qquad Q = I - 2\frac{\boldsymbol{q}\boldsymbol{q}^*}{\boldsymbol{q}^*\boldsymbol{q}}, \quad \text{and} \quad Z = I - 2\frac{\boldsymbol{z}\boldsymbol{z}^*}{\boldsymbol{z}^*\boldsymbol{z}},$$

with $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{z}$ such that

$$P\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{u}}_{k+2}, \qquad Q\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{v}}_{k+2}, \qquad \text{and} \qquad Z\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{w}}_{k+2}.$$

Applying the Householder matrices yields

$$A(W_{k+2}Z) = (U_{k+2}P)(P^*H_{k+2}Z) \quad \text{and} \quad B(W_{k+2}Z) = (V_{k+2}Q)(Q^*K_{k+2}Z), \tag{5.17}$$

which can be computed in $\mathcal{O}(nk)$ through rank-1 updates. It is straightforward to verify that the bottom rows of $P^*H_{k+2}Z$ and $Q^*K_{k+2}Z$ are multiples of $\boldsymbol{e}_{k+2}^*$, e.g.,

$$\boldsymbol{e}_{k+2}^*P^*H_{k+2}Z = \widetilde{\boldsymbol{u}}_{k+2}^*(\widetilde{U}\widetilde{C}\widetilde{R}\widetilde{W}^*)Z = \widetilde{c}_{k+2}\widetilde{r}_{k+2,k+2}\widetilde{\boldsymbol{w}}_{k+2}^*Z = \widetilde{c}_{k+2}\widetilde{r}_{k+2,k+2}\boldsymbol{e}_{k+2}^*.$$

As a result, (5.17) can be partitioned as

$$A\begin{bmatrix} W_{k+1} & W_{k+2}\widetilde{\boldsymbol{w}}_{k+2} \end{bmatrix} = \begin{bmatrix} U_{k+1} & U_{k+2}\widetilde{\boldsymbol{u}}_{k+2} \end{bmatrix} \begin{bmatrix} H_{k+1} & \times \\ & \widetilde{c}_{k+2}\widetilde{r}_{k+2,k+2} \end{bmatrix},$$

$$B\begin{bmatrix} W_{k+1} & W_{k+2}\widetilde{\boldsymbol{w}}_{k+2} \end{bmatrix} = \begin{bmatrix} V_{k+1} & V_{k+2}\widetilde{\boldsymbol{v}}_{k+2} \end{bmatrix} \begin{bmatrix} K_{k+1} & \times \\ & \widetilde{c}_{k+2}\widetilde{r}_{k+2,k+2} \end{bmatrix},$$

defining U_{k+1}, V_{k+1}, W_{k+1}, H_{k+1}, and K_{k+1}. This partitioning can be truncated to obtain

$$AW_{k+1} = U_{k+1}H_{k+1} \quad \text{and} \quad BW_{k+1} = V_{k+1}K_{k+1}, \tag{5.18}$$

where U_{k+1}, V_{k+1}, and W_{k+1} have orthonormal columns, but H_{k+1} and K_{k+1} are not necessarily upper-triangular. The algorithm is summarized below in Algorithm 5.3.

Algorithm 5.3 (Multidirectional GSVD (MDGSVD))**.**
Input: Matrix pair (A, B), starting vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, minimum and maximum dimensions $j < \ell$.
Output: $AW_j = U_jC_jR_j$ and $BW_j = V_jS_jR_j$ approximating a partial GSVD.

1. Set $W_0 = []$.
2. **for** *number of restarts* **and** *not converged (cf., e.g.,* (5.8)*)* **do**
3. **for** $k = 0, 1, \ldots, \ell - 2$ **do**

Let $\boldsymbol{w}_{k+1} = (I - W_k W_k^*)\boldsymbol{w}_{k+1}$, and $\boldsymbol{w}_{k+1} = \boldsymbol{w}_{k+1}/\|\boldsymbol{w}_{k+1}\|$.
Let $\boldsymbol{w}_{k+2} = (I - W_{k+1} W_{k+1}^*)\boldsymbol{w}_{k+2}$ and $\boldsymbol{w}_{k+2} = \boldsymbol{w}_{k+2}/\|\boldsymbol{w}_{k+2}\|$.
Update the QR-decompositions
$$AW_{k+2} = U_{k+2}H_{k+2} \text{ and } BW_{k+2} = V_{k+2}K_{k+2}.$$
Compute the GSVD $H_{k+2} = \widetilde{U}\widetilde{C}\widetilde{R}\widetilde{W}^*$ and $K_{k+2} = \widetilde{V}\widetilde{S}\widetilde{R}\widetilde{W}^*$.
Let P, Q, and Z be Householder reflections such that
$$P\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{u}}_{k+2},\ Q\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{v}}_{k+2}, \text{ and } Z\boldsymbol{e}_{k+2} = \widetilde{\boldsymbol{z}}_{k+2}.$$
Let $U_{k+2} = U_{k+2}P$, $V_{k+2} = V_{k+2}Q$, $W_{k+2} = W_{k+2}Z$,
$$H_{k+2} = P^*H_{k+2}Z, \text{ and } K_{k+2} = Q^*K_{k+2}Z.$$
$\boldsymbol{w}_{k+2} = A^*\widetilde{\boldsymbol{u}}_1$ and $\boldsymbol{w}_{k+3} = B^*\widetilde{\boldsymbol{v}}_1$.
if $j \leq k$ **and** *converged (cf., e.g.,* (5.8)*)* **then break**
end
Partition $\widetilde{U}$, $\widetilde{V}$, $\widetilde{W}$, $\widetilde{C}$, $\widetilde{S}$, and $\widetilde{R}$ according to (5.6).
Let $U_j = U_k\widetilde{U}_1$, $V_j = V_k\widetilde{V}_1$, and $W_j = W_k\widetilde{W}_1$.
Let $H_j = \widetilde{C}_1\widetilde{R}_{11}$ and $K_j = \widetilde{S}_1\widetilde{R}_{11}$.
end

Algorithm 5.3 is a simplified description for the sake of clarity. For instance, the expansion vectors may be linearly dependent in practice, and it may be desirable to expand a search space of dimension $\ell - 1$ with only the residual instead of two vectors. Another missing feature that might be required in practice is deflation, which is the topic of the next section.

5.5 Deflation and the truncated GSVD

Deflation is used in eigenvalue computations to prevent iterative methods from recomputing known eigenpairs. Since Algorithm 5.1 and Algorithm 5.3 compute generalized singular values and vectors one at a time, deflation may be necessary for applications where more than one generalized singular pair is required. The truncated GSVD is an example of such an application. There are at least two ways in which generalized singular values and vectors can be deflated, namely by transformation and by restriction. These two approaches have been inspired by their counterparts for the symmetric eigenvalue problem (cf., e.g., Parlett [68, Ch. 5]). We only describe the two approaches for m, $p \geq n$ to avoid clutter, but note that they can be adapted to the general case.

The restriction approach is related to the truncation described in the previous section and may be used to deflate a single generalized singular pair at a time. Suppose we wish to deflate the simple pair (c_1, s_1) and let the GSVD of (A, B) be

partitioned as

$$A = \begin{bmatrix} \boldsymbol{u}_1 & U_2 \end{bmatrix} \begin{bmatrix} c_1 & \\ & C_2 \end{bmatrix} \begin{bmatrix} r_{11} & \boldsymbol{r}_{12}^* \\ & R_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1^* \\ W_2^* \end{bmatrix},$$
$$B = \begin{bmatrix} \boldsymbol{v}_1 & V_2 \end{bmatrix} \begin{bmatrix} s_1 & \\ & S_2 \end{bmatrix} \begin{bmatrix} r_{11} & \boldsymbol{r}_{12}^* \\ & R_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1^* \\ W_2^* \end{bmatrix},$$

where C_2 and S_2 may be rectangular. Then, with Householder reflections P, Q, and Z, satisfying

$$P\boldsymbol{u}_1 = \boldsymbol{e}_1, \qquad Q\boldsymbol{v}_1 = \boldsymbol{e}_1, \quad \text{and} \quad Z\boldsymbol{w}_1 = \boldsymbol{e}_1,$$

it holds that

$$PAZ = \begin{bmatrix} c_1 r_{11} & \times \\ & \widehat{A} \end{bmatrix} \quad \text{and} \quad QBZ = \begin{bmatrix} s_1 r_{11} & \times \\ & \widehat{B} \end{bmatrix},$$

defining $\widehat{A}$ and $\widehat{B}$. At this point, the generalized singular pairs of $(\widehat{A}, \widehat{B})$ are the generalized singular pairs of (A, B) other than (c_1, s_1). Additional generalized singular pairs can be deflated inductively.

An alternative that allows for the deflation of multiple generalized singular pairs simultaneously is the restriction approach. To derive this approach, let the GSVD of (A, B) be partitioned as

$$\begin{aligned} A &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} C_1 & \\ & C_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix} \begin{bmatrix} W_1^* \\ W_2^* \end{bmatrix}, \\ B &= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} S_1 & \\ & S_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix} \begin{bmatrix} W_1^* \\ W_2^* \end{bmatrix}, \end{aligned} \tag{5.19}$$

where C_1 and S_1 are square and must be deflated, while C_2 and S_2 may be rectangular and must be retained. Therefore, the desired generalized singular pairs are deflated by working with the operators

$$\begin{aligned} \widehat{A} &= U_2 C_2 R_{22} W_2^* = U_2 U_2^* A W_2 W_2^* = (I - U_1 U_1^*) A (I - W_1 W_1^*), \\ \widehat{B} &= V_2 S_2 R_{22} W_2^* = V_2 V_2^* B W_2 W_2^* = (I - V_1 V_1^*) B (I - W_1 W_1^*), \end{aligned} \tag{5.20}$$

restricted to $\mathcal{W}_2 = \operatorname{span}\{W_2\}$. An important benefit of this approach is that the restriction may be performed implicitly during the iterations. For example, if (5.6) is such that

$$U_k \widetilde{U}_1 = U_1, \quad V_k \widetilde{V}_1 = V_1, \quad W_k \widetilde{W}_1 = W_1, \quad \widetilde{C}_1 = C_1, \quad \widetilde{S}_1 = S_1, \quad \text{and} \quad \widetilde{R}_{11} = R_{11};$$

then

$$\widehat{A}W_k\widetilde{W}_2 = U_k\widetilde{U}_2\widetilde{C}_2\widetilde{R}_{22} \quad \text{and} \quad \widehat{B}W_k\widetilde{W}_2 = V_k\widetilde{V}_2\widetilde{S}_2\widetilde{R}_{22},$$

where the right-hand sides are available without explicitly working with $\widehat{A}$ and $\widehat{B}$. In addition, if we define the approximations for the next generalized singular pair and corresponding vectors as

$$\begin{aligned} \alpha &= \boldsymbol{e}_1^*\widetilde{C}_2\boldsymbol{e}_1, \qquad \beta = \boldsymbol{e}_1^*\widetilde{S}_2\boldsymbol{e}_1, \qquad \rho = \boldsymbol{e}_1^*\widetilde{R}_{22}\boldsymbol{e}_1, \\ \widetilde{\boldsymbol{u}} &= U_k\widetilde{U}_2\boldsymbol{e}_1, \qquad \widetilde{\boldsymbol{v}} = V_k\widetilde{V}_2\boldsymbol{e}_1, \qquad \widetilde{\boldsymbol{w}} = W_k\widetilde{W}_2\boldsymbol{e}_1, \end{aligned}$$

cf. Section 5.2, and

$$\widetilde{\boldsymbol{x}} = \rho^{-1}W_k\widetilde{W}\begin{bmatrix}\widetilde{R}_{11}^{-1}\widetilde{R}_{12}\boldsymbol{e}_1 \\ \boldsymbol{e}_1\end{bmatrix} = \rho^{-1}(W_k\widetilde{W}_1\widetilde{R}_{11}^{-1}\widetilde{R}_{12}\boldsymbol{e}_1 + \widetilde{\boldsymbol{w}});$$

then the residual

$$\begin{aligned} \boldsymbol{r} = \rho^{-1}(\beta^2\widehat{A}^*\widehat{A} - \alpha^2\widehat{B}^*\widehat{B})\widetilde{\boldsymbol{w}} &= \alpha\beta(\beta\widehat{A}^*\widetilde{\boldsymbol{u}} - \alpha\widehat{B}^*\widetilde{\boldsymbol{v}}) \\ &= \alpha\beta(\beta A^*\widetilde{\boldsymbol{u}} - \alpha B^*\widetilde{\boldsymbol{v}}) = (\beta^2A^*A - \alpha^2B^*B)\widetilde{\boldsymbol{x}} \end{aligned}$$

and expansion vector(s) can also be computed without $\widehat{A}$ and $\widehat{B}$.

It may be instructive to point out that the restriction approach for deflation corresponds to a splitting method for general form Tikhonov regularization described in [34] and references. This method separates the penalized part of the solution from the unpenalized part associated with the nullspace of the regularization operator, essentially deflating specific generalized singular values and vectors. Consider, for instance, the minimization problem

$$\operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|B\boldsymbol{x}\|^2$$

for some $\mu > 0$. Assume for the sake of simplicity that $p \geq n$, adding zero rows to B if necessary, and suppose that W_1 is a basis for the nullspace of B; then we obtain

(5.21)

$$\begin{aligned} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|B\boldsymbol{x}\|^2 = {} & \|U_1U_1^*AW_1W_1^*\boldsymbol{x} - (U_1U_1^*\boldsymbol{b} - U_1U_1^*AW_2W_2^*\boldsymbol{x})\|^2 \\ & + \|U_2U_2^*AW_2W_2^*\boldsymbol{x} - U_2U_2^*\boldsymbol{b}\|^2 + \mu\|V_2V_2^*BW_2W_2^*\boldsymbol{x}\|^2 \end{aligned}$$

by following the splitting approach and using that $U_2U_2^*AW_1W_1^*\boldsymbol{x} = \boldsymbol{0}$. Furthermore, with $\boldsymbol{y}_1 = W_1^*\boldsymbol{x}$ and $\boldsymbol{y}_2 = W_2^*\boldsymbol{x}$, the first part of the right-hand side of (5.21) reduces to

$$\|(U_1^*AW_1)\boldsymbol{y}_1 - (U_1^*\boldsymbol{b} - U_1^*AW_2\boldsymbol{y}_2)\|^2 = \|R_{11}\boldsymbol{y}_1 - U_1^*(\boldsymbol{b} - AW_2\boldsymbol{y}_2)\|^2,$$

which vanishes for $\boldsymbol{y}_1 = R_{11}^{-1} U_1^* (\boldsymbol{b} - AW_2\boldsymbol{y}_2)$. The remaining part may be written as

$$\|\widehat{A}W_2\boldsymbol{y}_2 - U_2U_2^*\boldsymbol{b}\|^2 + \mu\|\widehat{B}W_2\boldsymbol{y}_2\|^2,$$

where we recognize the deflated matrices from (5.20). A similar expression can be derived for deflation through restriction, but does not provide additional insight.

5.6 Error analysis

In this section we are concerned with the quality of the computed approximations, and develop Rayleigh–Ritz theory that is useful for the GSVD. In particular, we will generalize several known results for the $n \times n$ standard Hermitian eigenvalue problem to the Hermitian positive definite generalized eigenvalue problem

$$N\boldsymbol{x} = \lambda M\boldsymbol{x}, \qquad M > 0, \qquad M = L^2, \tag{5.22}$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. This generalized problem is applicable in our context with $N = A^*A$ and $M = X^{-*}X^{-1}$ if we are interested in the largest generalized singular values, or with $N = B^*B$ and $M = X^{-*}X^{-1}$ if we are interested in the smallest generalized singular values; and corresponds to the standard problem

$$L^{-1}NL^{-1}\boldsymbol{y} = \lambda\boldsymbol{y}, \qquad \boldsymbol{y} = L\boldsymbol{x}, \tag{5.23}$$

with the same eigenvalues. Hence, if the subspace $\mathcal{W}$ is a search space for (5.22), then it is natural to consider $\mathcal{Z} = L\mathcal{W}$ as a search space for (5.23) and to associate every approximate generalized eigenvector $\boldsymbol{w} \in \mathcal{W}$ with an approximate eigenvector $\boldsymbol{z} = L\boldsymbol{w} \in \mathcal{Z}$. The corresponding Rayleigh quotients satisfy

$$\theta = \frac{\boldsymbol{w}^*N\boldsymbol{w}}{\boldsymbol{w}^*M\boldsymbol{w}} = \frac{\boldsymbol{z}^*L^{-1}NL^{-1}\boldsymbol{z}}{\boldsymbol{z}^*\boldsymbol{z}} \tag{5.24}$$

and define the approximate eigenvalue θ.

Key to extending results for the generalized problem (5.23) to results for the standard problem (5.22), is to introduce generalized sines, cosines, and tangents, with respect to the M-norm defined by $\|\boldsymbol{x}\|_M^2 = \boldsymbol{x}^*M\boldsymbol{x} = \|L\boldsymbol{x}\|^2$. Generalizations of these trigonometric functions have previously been considered by Berns–Müller and Spence [6], and the generalized tangent can also be found in [68, Thm. 15.9.3]; however, we believe the treatment and results presented here to be new. The regular sine for two nonzero vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ can be defined as

$$\sin(\boldsymbol{z}, \boldsymbol{y}) = \frac{\left\|\left(I - \frac{\boldsymbol{z}\boldsymbol{z}^*}{\boldsymbol{z}^*\boldsymbol{z}}\right)\boldsymbol{y}\right\|}{\|\boldsymbol{y}\|} = \frac{\left\|\left(I - \frac{\boldsymbol{y}\boldsymbol{y}^*}{\boldsymbol{y}^*\boldsymbol{y}}\right)\boldsymbol{z}\right\|}{\|\boldsymbol{z}\|},$$

where it is easily verified that the above two expressions are equal indeed. Substituting $L\boldsymbol{x}$ for $\boldsymbol{y}$ and $L\boldsymbol{w}$ for $\boldsymbol{z}$ yields the M-sine defined by

$$\sin_M(\boldsymbol{w}, \boldsymbol{x}) = \sin(L\boldsymbol{w}, L\boldsymbol{x}) = \frac{\left\|\left(I - \frac{\boldsymbol{w}\boldsymbol{w}^*M}{\boldsymbol{w}^*M\boldsymbol{w}}\right)\boldsymbol{x}\right\|_M}{\|\boldsymbol{x}\|_M} = \frac{\left\|\left(I - \frac{\boldsymbol{x}\boldsymbol{x}^*M}{\boldsymbol{x}^*M\boldsymbol{x}}\right)\boldsymbol{w}\right\|_M}{\|\boldsymbol{w}\|_M}.$$

Again, it may be checked that the above two expressions are equal. The regular cosine is given by

$$\cos(\boldsymbol{z}, \boldsymbol{y}) = \frac{\left\|\frac{\boldsymbol{z}\boldsymbol{z}^*}{\boldsymbol{z}^*\boldsymbol{z}}\boldsymbol{y}\right\|}{\|\boldsymbol{y}\|} = \frac{\left\|\frac{\boldsymbol{y}\boldsymbol{y}^*}{\boldsymbol{y}^*\boldsymbol{y}}\boldsymbol{z}\right\|}{\|\boldsymbol{z}\|} = \frac{|\boldsymbol{z}^*\boldsymbol{y}|}{\|\boldsymbol{z}\|\,\|\boldsymbol{y}\|},$$

and with the same substitution we find the M-cosine

$$\cos_M(\boldsymbol{w}, \boldsymbol{x}) = \cos(L\boldsymbol{w}, L\boldsymbol{x}) = \frac{\left\|\frac{\boldsymbol{w}\boldsymbol{w}^*M}{\boldsymbol{w}^*M\boldsymbol{w}}\boldsymbol{x}\right\|_M}{\|\boldsymbol{x}\|_M} = \frac{\left\|\frac{\boldsymbol{x}\boldsymbol{x}^*M}{\boldsymbol{x}^*M\boldsymbol{x}}\boldsymbol{w}\right\|_M}{\|\boldsymbol{w}\|_M} = \frac{|\boldsymbol{w}^*M\boldsymbol{x}|}{\|\boldsymbol{w}\|_M\|\boldsymbol{x}\|_M}.$$

The M-tangent is now naturally defined as $\tan_M(\boldsymbol{w}, \boldsymbol{x}) = \sin_M(\boldsymbol{w}, \boldsymbol{x})/\cos_M(\boldsymbol{w}, \boldsymbol{x})$. We can derive the M-sines, M-cosines, and M-tangents between subspaces and vectors with a similar approach. For instance, let W and LW denote bases for $\mathcal{W}$ and $\mathcal{Z}$, respectively; then

$$\sin(\mathcal{Z}, \boldsymbol{y}) = \frac{\left\|\left(I - \frac{\boldsymbol{y}\boldsymbol{y}^*}{\boldsymbol{y}^*\boldsymbol{y}}\right)Z(Z^*Z)^{-1}Z^*\boldsymbol{y}\right\|}{\|Z(Z^*Z)^{-1}Z^*\boldsymbol{y}\|} \quad \text{and} \quad \cos(\mathcal{Z}, \boldsymbol{y}) = \frac{\|Z(Z^*Z)^{-1}Z^*\boldsymbol{y}\|}{\|\boldsymbol{y}\|},$$

so that

$$\sin_M(\mathcal{W}, \boldsymbol{x}) = \sin(L\mathcal{W}, L\boldsymbol{x}) = \frac{\left\|\left(I - \frac{\boldsymbol{x}\boldsymbol{x}^*M}{\boldsymbol{x}^*M\boldsymbol{x}}\right)W(W^*MW)^{-1}W^*M\boldsymbol{x}\right\|_M}{\|W(W^*MW)^{-1}W^*M\boldsymbol{x}\|_M},$$

$$\cos_M(\mathcal{W}, \boldsymbol{x}) = \cos(L\mathcal{W}, L\boldsymbol{x}) = \frac{\|W(W^*MW)^{-1}W^*M\boldsymbol{x}\|_M}{\|\boldsymbol{x}\|_M},$$

and $\tan_M(\mathcal{W}, \boldsymbol{x}) = \sin_M(\mathcal{W}, \boldsymbol{x})/\cos_M(\mathcal{W}, \boldsymbol{x})$. It is important to note that $\sin_M$, $\cos_M$, and $\tan_M$ can all be computed without the matrix square root L of M.

Since our M-sines, M-cosines, and M-tangents equal their regular counterparts, the extension of several known results for the standard problem (5.23) to results for the generalized problem (5.22) is immediate. Below is a selection of error bounds, where we assume that the largest generalized eigenpair $(\lambda_1, \boldsymbol{x}_1)$ is simple and is approximated by the Ritz pair $(\theta_1, \boldsymbol{w}_1)$ of (5.22) with respect to the search space $\mathcal{W}$.

Proposition 5.6 (Generalization of, e.g., [68, Lemma 11.9.2])**.**

$$\sin_M^2(\boldsymbol{w}_1, \boldsymbol{x}_1) \leq \frac{\lambda_1 - \theta_1}{\lambda_1 - \lambda_2}.$$

Proposition 5.7 (Generalization of [79, Thm. 2.1])**.**

$$\lambda_1 - \theta_1 \leq (\lambda_1 - \lambda_n) \sin_M^2(\mathcal{W}, \boldsymbol{x}_1).$$

The two propositions imply that $\boldsymbol{w}_1 \to \boldsymbol{x}_1$ when $\theta_1 \to \lambda_1$, with θ_1 tending to λ_1 when $\sin(\mathcal{W}, \boldsymbol{x}_1) \to 0$. The next corollary is a straightforward consequence.

Corollary 5.8 (Generalization of [79, Thm. 2.1])**.**

$$\sin_M^2(\boldsymbol{w}_1, \boldsymbol{x}_1) \leq \frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_2} \sin_M^2(\mathcal{W}, \boldsymbol{x}_1) = \left(1 + \frac{\lambda_2 - \lambda_n}{\lambda_1 - \lambda_2}\right) \sin_M^2(\mathcal{W}, \boldsymbol{x}_1).$$

As a result of Corollary 5.8, we can expect $\sin_M(\boldsymbol{w}_1, \boldsymbol{x}_1)$ to be close to $\sin_M(\mathcal{W}, \boldsymbol{x}_1)$ if the eigenvalue λ_1 is well separated from the rest of the spectrum. A sharper bound can be obtained by generalizing the optimal bound from Sleijpen, Eshof, and Smit [79].

Proposition 5.9 (Generalization of [79, Thm. 3.2])**.** *Let $(\theta_j, \boldsymbol{w}_j)$ denote the Ritz pairs of the generalized problem* (5.22) *with respect to $\mathcal{W}$, and define*

$$\delta_{\mathcal{W}} = \min \sin_M(\boldsymbol{w}_j, \boldsymbol{x}_1)$$

as the smallest of all M-sines between the Ritz vectors $\boldsymbol{w}_j$ and the generalized eigenvector $\boldsymbol{x}_1$. Furthermore, define for any $\epsilon > 0$ the maximum

$$\delta_k(\epsilon) = \max_{\mathcal{W}} \{\delta_{\mathcal{W}} \mid \dim(\mathcal{W}) = k, \sin_M(\mathcal{W}, \boldsymbol{x}_1) \leq \epsilon\}.$$

If $(\theta_{\mathcal{W}}, \boldsymbol{w}_{\mathcal{W}})$ is the Ritz pair for which $\delta_{\mathcal{W}}$ is realized and

$$0 \leq \epsilon < (\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_n);$$

then $\theta_{\mathcal{W}} = \theta_1 > \lambda_2$ and

$$\delta_k^2(\epsilon) = \frac{1}{2}(1 + \epsilon^2) - \frac{1}{2}\sqrt{(1 - \epsilon^2)^2 - \kappa\epsilon^2} \quad \textit{with} \quad \kappa = \frac{(\lambda_2 - \lambda_n)^2}{(\lambda_1 - \lambda_n)(\lambda_1 - \lambda_2)},$$

for all $k \in \{2, \ldots, n-1\}$.

The quantity $\delta_k^2(\epsilon)$ is not particularly elegant, but is sharp and can be used to obtain the following upper bound, which is sharper than the bound in Corollary 5.8.

Corollary 5.10 (Generalization of [79, Cor. 3.3])**.** *If the conditions in Proposition 5.9 are satisfied, then*

$$\sin^2_M(\boldsymbol{w}_1, \boldsymbol{x}_1) \leq \sin^2_M(\mathcal{W}, \boldsymbol{x}_1) + \frac{\kappa}{2} \tan^2_M(\mathcal{W}, \boldsymbol{x}_1).$$

Now that we have extended a number of results for the standard problem (5.23) to the generalized problem (5.22), it may be worthwhile to bound the generalized sine $\sin_M$ in terms of the standard sine.

Proposition 5.11. *Let $\kappa = \kappa(M)$ be the condition number of M, then*

$$\frac{1}{\kappa} \sin^2(\boldsymbol{w}, \boldsymbol{x}) \leq \sin^2_M(\boldsymbol{w}, \boldsymbol{x}) \leq \frac{1}{4}(\kappa + 1)^2 \sin^2(\boldsymbol{w}, \boldsymbol{x}).$$

Proof. Without loss of generality we assume $\|\boldsymbol{w}\| = \|\boldsymbol{x}\| = 1$, so that

$$\lambda_{\min}(M) \leq \|\boldsymbol{x}\|^2_M \leq \lambda_{\max}(M).$$

The first inequality follows from

$$\begin{aligned}\sin^2(\boldsymbol{w}, \boldsymbol{x}) &= \left\|(I - \boldsymbol{w}\boldsymbol{w}^*)\left(I - \tfrac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}}\right)\boldsymbol{x}\right\|^2 \\ &\leq \|\boldsymbol{x}\|^2_M \frac{\left\|\left(I - \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}}\right)\boldsymbol{x}\right\|^2}{\left\|\left(I - \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}}\right)\boldsymbol{x}\right\|^2_M} \sin^2_M(\boldsymbol{w}, \boldsymbol{x}) \leq \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \sin^2_M(\boldsymbol{w}, \boldsymbol{x}).\end{aligned}$$

For the second inequality, it follows from, e.g., [87] that

$$\left\| I - \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}} \right\| = \left\| \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}} \right\| = \frac{\|M\boldsymbol{w}\|}{\boldsymbol{w}^* M \boldsymbol{w}} = \cos^{-1}(\boldsymbol{w}, M\boldsymbol{w}) \leq \mu^{-1},$$

where μ^{-1} is the inverse of the first anti-eigenvalue [26, Ch. 3.6]

$$\mu = \min_{\|\boldsymbol{w}\|=1} \frac{\boldsymbol{w}^* M \boldsymbol{w}}{\|M\boldsymbol{w}\|}.$$

By applying Kantorovich' inequality we find [26, p. 68]

$$\mu^{-1} = \frac{1}{2} \frac{\lambda_{\min}(M) + \lambda_{\max}(M)}{\sqrt{\lambda_{\min}(M)\, \lambda_{\max}(M)}} = \frac{1}{2} \frac{\kappa + 1}{\sqrt{\kappa}}.$$

Finally, by combining the above and using

$$\left(I - \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}}\right) = \left(I - \frac{\boldsymbol{w}\boldsymbol{w}^* M}{\boldsymbol{w}^* M \boldsymbol{w}}\right)(I - \boldsymbol{w}\boldsymbol{w}^*)\boldsymbol{x},$$

we see that

$$\sin^2_M(\boldsymbol{w}, \boldsymbol{x}) = \frac{\left\|\left(I - \frac{\boldsymbol{x}\boldsymbol{x}^*M}{\boldsymbol{x}^*M\boldsymbol{x}}\right)\boldsymbol{x}\right\|^2_M}{\|\boldsymbol{x}\|^2_M} \leq \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \left\|\left(I - \frac{\boldsymbol{w}\boldsymbol{w}^*M}{\boldsymbol{w}^*M\boldsymbol{w}}\right)\boldsymbol{x}\right\|^2$$
$$\leq \kappa \left\|I - \frac{\boldsymbol{w}\boldsymbol{w}^*M}{\boldsymbol{w}^*M\boldsymbol{w}}\right\|^2 \|(I - \boldsymbol{w}\boldsymbol{w}^*)\boldsymbol{x}\|^2 \leq \frac{1}{4}(\kappa + 1)^2 \sin^2(\boldsymbol{w}, \boldsymbol{x}),$$

which concludes the proof. □

An interesting observation about $\sin_M$ in the context of the GSVD is that $\|\boldsymbol{f}\|$ from Proposition 5.4 equals $\sin_M(\widetilde{\boldsymbol{x}}_1, \boldsymbol{x}_1)$ if $M = A^*A + B^*B = X^{-*}X^{-1}$. Furthermore, it has been shown in the proof of Proposition 5.4 that the error in $\widetilde{c}_1^2 = \|A\widetilde{\boldsymbol{x}}_1\|^2$ and $\widetilde{s}_1^2 = \|B\widetilde{\boldsymbol{x}}_1\|^2$ is quadratic in $\|\boldsymbol{f}\|$. An alternative is to express the approximation error in terms of the residual. We have, for example, the following straightforward Bauer–Fike-type result.

Proposition 5.12 (Bauer–Fike for the GSVD)**.** *Let $(\widetilde{c}, \widetilde{s})$ be an approximate generalized singular pair with corresponding generalized singular vector $\widetilde{\boldsymbol{x}}$ and residual*

$$\boldsymbol{r} = (\widetilde{s}^2 A^*A - \widetilde{c}^2 B^*B)\widetilde{\boldsymbol{x}};$$

then there exists a generalized singular pair $(c_\star, s_\star)$ of (A, B) such that

$$|\widetilde{s}^2 c_\star^2 - \widetilde{c}^2 s_\star^2| \leq \|X\|^2 \frac{\|\boldsymbol{r}\|}{\|\widetilde{\boldsymbol{x}}\|}.$$

Proof. The result follows from

$$\frac{\|\boldsymbol{r}\|}{\|\widetilde{\boldsymbol{x}}\|} \geq \sigma_{\min}(\widetilde{s}^2 A^*A - \widetilde{c}^2 B^*B)$$
$$= \sigma_{\min}(X^{-*}(\widetilde{s}^2 \Sigma_A^T \Sigma_A - \widetilde{c}^2 \Sigma_B^T \Sigma_B)X^{-1}) \geq \sigma^2_{\min}(X^{-1}) \min_j |\widetilde{s}^2 c_j^2 - \widetilde{c}^2 s_j^2|.$$

□

An additional interesting observation is that if $\widetilde{c}$ and $\widetilde{s}$ are scaled such that $\widetilde{c}^2 + \widetilde{s}^2 = c_\star^2 + s_\star^2 = 1$, and the generalized singular values are given by $\widetilde{\sigma} = \widetilde{c}/\widetilde{s}$ and $\sigma_\star = c_\star/s_\star$; then

$$|\widetilde{s}^2 c_\star^2 - \widetilde{c}^2 s_\star^2| = |\widetilde{s}^2 - s_\star^2| = |c_\star^2 - \widetilde{c}^2| = \left|\frac{\widetilde{\sigma}^2}{1 + \widetilde{\sigma}^2} - \frac{\sigma_\star^2}{1 + \sigma_\star^2}\right|,$$

with the conventions $\infty/\infty = 1$ and $\infty - \infty = 0$.

The bound in Proposition 5.12 may be rather pessimistic, and we expect asymptotic convergence of order $\|\boldsymbol{r}\|^2$ due to the relation with the symmetric eigenvalue problem. It turns out that the desired result is easily generalized using the M-sine and the M^{-1}-norm. Specifically, let θ be defined as in (5.24) and define the residual norms

$$\rho(\boldsymbol{z}) = \|(L^{-1}NL^{-1} - \theta I)\boldsymbol{z}\| \quad \text{and} \quad \rho_M(\boldsymbol{w}) = \rho(L\boldsymbol{w}) = \|(N - \theta M)\boldsymbol{w}\|_{M^{-1}};$$

then we can immeadiately derive the following proposition.

Proposition 5.13 (Generalization of, e.g., [68, Thm. 11.7.1, Cor. 11.7.1])**.** *Suppose $\lambda_1 - \theta_1 < \theta_1 - \lambda_2$; then*

$$\frac{\rho_M(\boldsymbol{w}_1)}{\lambda_1 - \lambda_n} \leq \sin_M(\boldsymbol{w}_1, \boldsymbol{x}_1) \leq \tan_M(\boldsymbol{w}_1, \boldsymbol{x}_1) \leq \frac{\rho_M(\boldsymbol{w}_1)}{\theta_1 - \lambda_2}$$

and

$$\frac{\rho_M^2(\boldsymbol{w}_1)}{\lambda_1 - \lambda_n} \leq \lambda_1 - \theta_1 \leq \frac{\rho_M^2(\boldsymbol{w})}{\theta_1 - \lambda_2}.$$

Having the M^{-1}-norm for the residual instead of the M-norm might be surprising; however, the former is a natural choice in this context; see, e.g., [68, Ch. 15]. Moreover, Proposition 5.13 combined with the norm equivalence

$$\sigma_{\max}^{-1}(M)\,\|\boldsymbol{r}\|^2 \leq \|\boldsymbol{r}\|_{M^{-1}}^2 \leq \sigma_{\min}^{-1}(M)\,\|\boldsymbol{r}\|^2$$

implies that the converence of the generalized singular values must be of order $\|\boldsymbol{r}\|^2$. This result is verified in an example in the next section.

5.7 Numerical experiments

In this section we compare our new algorithms to JDGSVD and Zha's modified Lanczos algorithm by using tests similar to the examples found in [32] and Zha [95]. Additionally, we will apply Algorithm 5.1 and Algorithm 5.3 to general form Tikhonov regularization by approximating truncated GSVDs for several test problems. The first set of examples is detailed below.

Example 5.1. Let $A = CD$ and $B = SD$ be two $n \times n$ matrices, where

$$C = \operatorname{diag}(c_j), \quad c_j = (n - j + 1)/(2n), \quad S = \sqrt{I - C^2},$$
$$D = \operatorname{diag}(d_j), \quad d_j = \lceil j/(n/4) \rceil + r_j,$$

with r_j drawn from the standard uniform distribution on the open interval (0, 1).

Example 5.2. Let C and S be the same as in Example 5.1. Furthermore, let $A = UC\widetilde{D}W^*$ and $B = VS\widetilde{D}W^*$, where U, V, and W are random orthonormal matrices, and $\widetilde{D} = \operatorname{diag}(\widetilde{d}_j)$ with

$$\widetilde{d}_j = d_j - \min_{1\le j\le n} d_j + 10^{-\kappa}.$$

Three values for κ are considered, (a) $\kappa = 6$, (b) $\kappa = 9$, and (c) $\kappa = 12$.

Example 5.3. Let C and S be the same as in Example 5.1, and let $\widetilde{D}$ be the same as in Example 5.2. Let $\boldsymbol{f}$, $\boldsymbol{g}$, and $\boldsymbol{h}$ be random vectors on the unit $(n-1)$-sphere, and set

$$A = (I - 2\boldsymbol{f}\boldsymbol{f}^*)C\widetilde{D}(I - 2\boldsymbol{h}\boldsymbol{h}^*) \quad \text{and} \quad B = (I - 2\boldsymbol{g}\boldsymbol{g}^*)S\widetilde{D}(I - 2\boldsymbol{h}\boldsymbol{h}^*).$$

Note that $I - 2\boldsymbol{f}\boldsymbol{f}^*$, $I - 2\boldsymbol{g}\boldsymbol{g}^*$, and $I - 2\boldsymbol{h}\boldsymbol{h}^*$ are Householder reflections.

Example 5.4. Let

$$A = \texttt{sprand(n, n, 1e-1, 1)} \quad \text{and} \quad B = \texttt{sprand(n, n, 1e-1, 1e-2)},$$

where `sprand` is the MATLAB function with the same name.

Table 5.1: The median number of matrix-vector products the algorithms require for Examples 5.1–5.4 to compute an approximation satisfying (5.25). The tolerance $\tau = 10^{-3}$ was used for Zha's modified Lanczos algorithm, while $\tau = 10^{-6}$ was used for the remaining algorithms. The symbol – indicates a failure to converge up to the desired tolerance within the maximum number of iterations specified in the text, and the column **Cond** contains the condition numbers of $[A^T\ B^T]^T$.

Alg		**Zha**		**JDGSVD**		**GDGSVD**		**MDGSVD**	
Ex	**Cond**	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$
5.1	4.97e+00	3390	–	1524	6188	580	3072	502	730
5.2a	3.99e+06	19082	–	2008	5396	992	2326	1054	622
5.2b	3.99e+09	19082	–	2008	5396	998	2312	1036	628
5.2c	3.99e+12	19082	–	2008	5374	998	2312	1030	622
5.3a	3.99e+06	17810	–	1964	5418	996	2318	1048	616
5.3b	3.99e+09	17810	–	1964	5418	996	2288	1036	616
5.3c	3.99e+12	17810	–	1964	5418	996	2288	1048	628
5.4	1.41e+00	–	1262	–	–	2334	244	2314	240

We generate the matrices from Examples 5.1–5.4 for $n = 1000$, allowing us to verify the results. For Algorithm 5.1 and Algorithm 5.3 we set the minimum

dimension to 10, the maximum dimension to 30, and the maximum number of restarts to 100. For JDGSVD we use the same minimum and maximum dimensions in combination with a maximum of 10 and 1000 inner and outer iterations, respectively. Furthermore, we let JDGSVD use standard extraction to find the largest generalized singular value, and refined extraction to find the smallest generalized singular value. We have implemented Zha's modified Lanczos algorithm with LSQR, and let LSQR use the tolerance 10^{-12} and a maximum of $\lceil 10\sqrt{n} \rceil = 320$ iterations. The maximum number of outer-iterations for the modified Lanczos algorithm is 100.

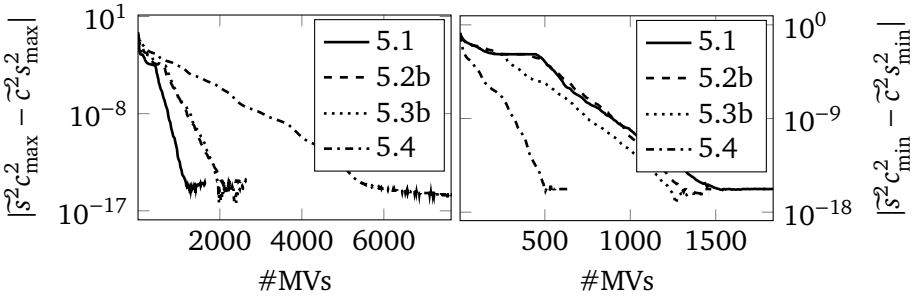


Figure 5.1: The convergence history of MDGSVD as the errors from (5.25) compared to the number of matrix-vector products, with results for the largest (left) and smallest (right) generalized singular pairs.

We run each test with 500 different starting vectors, and record the number of matrix-vector products required until an approximate generalized singular pair $(\widetilde{c}, \widetilde{s})$ satisfies

$$|\widetilde{s}^2 c_{\max}^2 - \widetilde{c}^2 s_{\max}^2| < \tau \quad \text{or} \quad |\widetilde{s}^2 c_{\min}^2 - \widetilde{c}^2 s_{\min}^2| < \tau, \tag{5.25}$$

where we use $\tau = 10^{-3}$ for Zha's modified Lanczos algorithm and $\tau = 10^{-6}$ for the remaining algorithms. The median results are shown in Table 5.1. We notice that the convergence of Zha's method is markedly slower here than in [95]. Additional testing has indicated that the difference is caused by the larger choice of n, which in turn decreases the gap between the generalized singular pairs. JDGSVD does not require accurate solutions from the inner iterations and is significantly faster, but fails to converge to a sufficiently accurate solution in the last example. Compared to JDGSVD, GDGSVD approximately reduces the number of matrix-vector multiplications by a factor of 2 for $\sigma_{\max}$ and by a factor of 2 to 2.4 for $\sigma_{\min}$, and has no problem finding a solution for the last example. MDGSVD performs only slightly worse than GDGSVD for the largest generalized singular pairs on average, but uses approximately 4 times fewer MVs than GDGSVD for the smallest generalized singular pairs in almost all tests.

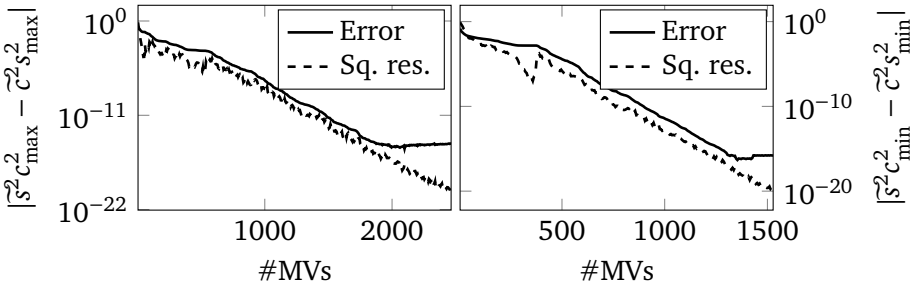


Figure 5.2: The errors of the largest (left) and smallest (right) generalized singular pairs approximations compared to the square of the relative residual norm in the right-hand side of (5.8). The results are for Example 5.2a and MDGSVD.

Figure 5.1 shows the convergence of MDGSVD. The monotone behavior and asymptotic linear convergence of the method are clearly visible. We can also see that the asymptotic convergence is significantly better than the worst-case bound from Proposition 5.4. Figure 5.2 shows a comparison between the relative residual norm (5.8) and the convergence of the generalized singular pairs for Example 5.2a. The results for the other examples are similar, and are therefore omitted. Although the graphs belonging to the smallest generalized singular pairs suggest temporary misconvergence, the comparison still demonstrates that (5.8) is an asymptotically suitable indicator for the convergence of the generalized singular pairs. Moreover, the convergence of the generalized singular pairs appears to be quadratic in the residual norm.

Example 5.5. Given a large, sparse, and ill-conditioned matrix A, consider the problem of reconstructing exact data $\boldsymbol{x}_\star$ from measured data $\boldsymbol{b} = A\boldsymbol{x}_\star + \boldsymbol{e}$, where $\boldsymbol{e}$ is a noise vector. A regularized solution may be determined with general form Tikhonov regularization by computing

$$\boldsymbol{x}_\mu = \operatorname*{argmin}_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \mu\|B\boldsymbol{x}\|^2$$

for some operator B with $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\boldsymbol{0}\}$, and some parameter $\mu > 0$. For the purpose of this example, we take several $n \times n$ matrices A and length n solution vectors $\boldsymbol{x}_\star$ from Regularization Tools [27], and for B we use the $(n-1) \times n$ finite difference operator

$$B = \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix}.$$

The entries of the noise vectors $\boldsymbol{e}$ are independently drawn from the standard normal distribution, after which the vector $\boldsymbol{e}$ is scaled such that $\epsilon = \mathbb{E}[\|\boldsymbol{e}\|] = 0.01\|\boldsymbol{b}\|$. We select the parameters μ such that $\|A\boldsymbol{x}_{\boldsymbol{\mu}} - \boldsymbol{b}\| = \eta\epsilon$, where $\eta = 1 + 3.090232/\sqrt{2n}$ so that $\|\boldsymbol{e}\| \leq \eta\epsilon$ with probability 0.999; see also (4.17).

Consider Example 5.5, where we can write $\boldsymbol{x}_{\boldsymbol{\mu}}$ as

$$\boldsymbol{x}_{\boldsymbol{\mu}} = X(\Sigma_A^*\Sigma_A + \mu\Sigma_B^*\Sigma_B)^{-1}\Sigma_A^*U^*\boldsymbol{b} = \sum_{i=1}^{n} \frac{c_i}{c_i^2 + \mu s_i^2}\boldsymbol{x}_i\boldsymbol{u}_i^*\boldsymbol{b}.$$

For large-scale problems with rapidly decaying c_i and multiple right-hand sides $\boldsymbol{b}$, it may attractive to approximate the truncated GSVD and compute the above summation only for a few of the largest generalized singular pairs and their corresponding generalized singular vectors. In particular, we use our GDGSVD and MDGSVD methods to approximate the truncated GSVD consisting of the 15 largest generalized singular pairs and vectors. We use minimum and maximum dimensions 15 and 45, respectively, and a maximum of 100 restarts. We deflate or terminate when the right-hand side of (5.8) is less than 10^{-6}, and seed the search space with the nullspace of B spanned by the vector $(1, \ldots, 1)^T$. We consider two different cases. In the first case, we deflate the seeded vector and terminate as soon as the relative residual for the second largest generalized singular pair is sufficiently small. In the second case we deflate the seeded vector plus four additional vectors, and terminate when the relative residual corresponding to the sixth largest generalized singular pair is less than 10^{-6}. We use the approximated truncated GSVDs to compute $\boldsymbol{x}_{\boldsymbol{\mu}}$, and compare it with the solution obtained with the exact truncated GSVD.

The experiments are repeated with 1000 different initial vectors and noise vectors, and the median results are reported in Table 5.2 and Table 5.3. Test problems Deriv2-{1,2,3} all use the same matrix A, but have different right-hand sides and solutions; the same is true for Gravity-{1,2,3}. Test problems Heat-{1,5} have the same solutions, but different A and $\boldsymbol{b}$. The tables show a reduction in the required number of matrix-vector products for multidirectional subspace expansion, with reduction factors approximately between 1.25 to 2.15 or better in the majority of cases. However, the reduced number of matrix-vector products may come at the cost of an increased relative error in the reconstructed solution and an increased angle between the exact and approximated generalized singular vector $\boldsymbol{x}_2$, although not consistently.

Table 5.2: Truncated GSVD tests where only the nullspace of B is deflated, and the iterations are terminated when the relative residual for the second largest generalized singular pair after (1, 0) is sufficiently small. The columns **Rank** and **Eff. cond** contain the numerical rank and effective condition number of A; and $\mathbf{sin}\,(x_2, \widetilde{x}_2)$ is a measure for the error in the approximation of the generalized singular vector corresponding to the second largest generalized singular pair.

Alg			**GDGSVD**			**MDGSVD**		
Ex	**Rank**	**Eff. cond**	$\mathbf{sin}\,(x_2, \widetilde{x}_2)$	**Rel. Err.**	**#MV**	$\mathbf{sin}\,(x_2, \widetilde{x}_2)$	**Rel. Err.**	**#MV**
Baart	13	5.30e + 12	1.25e − 5	2.25e − 5	1632	3.40e − 6	2.08e − 5	72
Deriv2-1	1024	1.27e + 06	1.51e − 5	9.50e − 5	4228	7.92e − 6	2.75e − 5	2074
Deriv2-2	1024	1.27e + 06	1.51e − 5	8.61e − 5	4228	7.92e − 6	9.90e − 6	2074
Deriv2-3	1024	1.27e + 06	1.51e − 5	1.91e − 3	4228	7.92e − 6	1.22e − 4	2074
Foxgood	30	3.88e + 12	2.90e − 5	1.15e − 5	4148	4.59e − 5	2.31e − 5	2830
Gravity-1	45	5.80e + 12	1.56e − 5	2.52e − 4	3764	1.06e − 5	9.97e − 4	1750
Gravity-2	45	5.80e + 12	1.56e − 5	5.88e − 4	3764	1.06e − 5	9.85e − 4	1750
Gravity-3	45	5.80e + 12	1.56e − 5	3.04e − 4	3764	1.06e − 5	3.77e − 4	1750
Heat-1	587	6.18e + 12	3.52e − 5	4.20e − 2	4976	9.48e − 6	7.16e − 2	802
Heat-5	1022	1.27e + 03	1.18e − 5	5.73e − 2	5036	7.44e − 6	1.28e − 1	616
Phillips	1024	2.90e + 10	1.25e − 5	5.82e − 3	4188	7.87e − 6	1.58e − 3	1762
Shaw	20	4.32e + 12	2.98e − 5	2.24e − 2	3644	1.32e − 5	7.75e − 3	2308
Wing	8	1.01e + 12	1.61e − 5	1.39e − 5	4492	4.55e − 5	1.44e − 4	3064

Table 5.3: Truncated GSVD tests and results similar to Table 5.2, but in this case with the approximation of the five largest generalized singular pairs after the pair (1, 0) corresponding to the nullspace of the regularization operator.

Alg	GDGSVD			MDGSVD		
Ex	$\sin(x_2, \widetilde{x}_2)$	Rel. Err.	#MV	$\sin(x_2, \widetilde{x}_2)$	Rel. Err.	#MV
Baart	1.82e − 6	3.19e − 6	1996	2.61e − 8	1.51e − 7	74
Deriv2-1	8.03e − 6	8.99e − 6	6088	6.54e − 6	1.08e − 5	3604
Deriv2-2	8.03e − 6	3.52e − 6	6088	6.54e − 6	4.03e − 6	3604
Deriv2-3	8.03e − 6	6.25e − 5	6088	6.54e − 6	4.25e − 5	3604
Foxgood	6.91e − 6	3.36e − 6	6808	1.07e − 5	5.20e − 6	5485
Gravity-1	1.93e − 6	1.14e − 5	5600	4.85e − 6	4.10e − 5	4012
Gravity-2	1.93e − 6	3.11e − 5	5600	4.85e − 6	3.50e − 5	4012
Gravity-3	1.93e − 6	8.39e − 6	5600	4.85e − 6	1.86e − 5	4012
Heat-1	2.70e − 6	2.82e − 2	7520	5.14e − 6	4.74e − 2	1948
Heat-5	7.92e − 6	4.63e − 2	6676	2.92e − 6	2.48e − 2	1804
Phillips	4.74e − 6	3.49e − 4	5912	2.30e − 6	1.63e − 4	3574
Shaw	1.91e − 6	6.51e − 5	5772	2.67e − 6	1.80e − 4	5620
Wing	8.33e − 6	4.16e − 6	5292	1.44e − 5	7.26e − 6	4618

5.8 Conclusion

We have discussed two iterative methods for the computation of a few extremal generalized singular values and vectors. The first method can be seen as a generalized Davidson-type method, and the second as a further generalization. Specifically, the second method uses multidirectional subspace expansion combined with a truncation phase to find improved search directions, while ensuring moderate subspace growth. Both methods allow for a natural and straightforward thick restart. We have also derived two different methods for the deflation of generalized singular values and vectors.

We have characterized the locally optimal search directions and expansion vectors in both the generalized Davidson method and the multidirectional method. Note that these search directions generally cannot be computed during the iterations. The inability to compute these optimal search directions motivates multidirectional subspace expansion and its reliance on the extraction process, as well as the removal of low-quality search directions. We have argued that our methods can still achieve (asymptotic) linear convergence and have provided asymptotic bounds on the rate of convergence. Additionally, we have shown that the convergence of both methods is monotonic, and have concluded the theoreti-

cal analysis by developing Rayleigh–Ritz theory and generalizing known results for the Hermitian eigenvalue problem to the Hermitian positive definite generalized eigenvalue problem that corresponds to the GSVD.

The theoretical convergence behavior is supported by our numerical experiments. Moreover, the numerical experiments demonstrate that our generalized Davidson-type method is competitive with existing methods, and suitable for approximating the truncated GSVD of matrix pairs with rapidly decaying generalized singular values. Significant additional performance improvements may be obtained by our new multidirectional method.

Chapter 6

Conclusion

We have seen in Chapter 2 that matrix balancing may be useful for generating high-quality field of value based spectral inclusion regions. Furthermore, these inclusion regions can be approximated efficiently when balancing is combined with projections onto Krylov subspaces. The combination of Krylov subspaces and balancing leads naturally to our new "Krylov and balance" (K+B) approach, which is computationally cheap and typically yields excellent results. Another benefit is that the K+B approach is matrix free; however, there are disadvantages as well. Most notably, the K+B approach cannot be used to compute the diagonal scaling matrix for the original full-size matrix and may, in rare situations, compute inclusion regions that are too small. Possible future research may include balancing and field of values based inclusion regions for the generalized eigenvalue problem, quadratic eigenvalue problem, and the polynomial eigenvalue problem; see, e.g., [33, 54].

In Chapter 3 we have presented a two-sided Krylov–Schur method as a natural generalization of the one-sided Krylov–Schur approach by Stewart, and as a more stable alternative to the two-sided Lanczos algorithm. In addition to the advantages and disadvantages mentioned in Section 3.9, we would like to state that there is no benefit in using two-sided Krylov–Schur for Hermitian matrices. Furthermore, we do not expect any major advantages of two-sided Krylov–Schur over one-sided Krylov–Schur when only computing exterior eigenvalues. Instead, our two-sided method shows its strengths primarily in applications where left and right eigenvectors or eigenspaces are required simultaneously. It could be useful and interesting if future research provides us with more insight when two-sided Krylov–Schur may be expected to yield better results or performance than its one-sided counterpart.

We have introduced a new method in Chapter 4 for large-scale Tikhonov regularization that combines a new multidirectional subspace expansion with optional

truncation in order to produce a higher quality search space. The multidirectional expansion generates a richer search space, whereas the truncation ensures moderate growth. In addition, we have discussed a straightforward parameter choice for multiparameter regularization, that satisfies the discrepancy principle and is based on easy to compute derivatives. Although it is not clear how to extend the parameter selection from this chapter to non-smooth regularization terms, research in automatic parameter selection for more general regularization terms could prove valuable.

In Chapter 5 we have derived two competitive methods, for computing extremal generalized singular values and vectors, as well as for approximating the truncated generalized singular value decomposition of matrix pairs with rapidly decaying generalized singular values. The first method can be seen as a generalized Davidson-type method, while the second method builds upon the multidirectional subspace expansion and truncation from the previous chapter. The idea is again to find improved search directions in each iteration with multidirectional subspace expansion, while ensuring moderate subspace growth with a fast truncation phase. Both of our new methods allow for natural and straightforward thick restarts, which are essential parts of Algorithms 5.1 and 5.3. Numerical experiments suggest that the latter two algorithms have the potential to become "methods of choice", although the current lack of preconditioning might be a disadvantage. Future work may include adapting the multidirectional subspace expansion and truncation to different matrix decompositions.

Bibliography

[1] Z. Bai. *The CSD, GSVD, their applications and computations*. IMA Preprint Series 958. University of Minnesota, 1992.

[2] Z. Bai, D. Day, J. Demmel, and J. Dongarra. *A test matrix collection for non-Hermitian eigenvalue problems*. Technical report. University of Tennessee, 1996.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. A. van der Vorst, eds. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[4] M. Belge, M. E. Kilmer, and E. L. Miller. *Efficient determination of multiple regularization parameters in a generalized L-curve framework*. Inverse Problems 18.4 (2002), pp. 1161–1183.

[5] I. Bendixson. *Sur les racines d'une équation fondamentale*. Acta Math. 25.1 (1902), pp. 359–365.

[6] J. Berns–Müller and A. Spence. *Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems*. SIAM J. Matrix Anal. Appl. 28.4 (2006), pp. 1069–1082.

[7] T. Betcke. *Optimal scaling of generalized and polynomial eigenvalue problems*. SIAM J. Matrix Anal. Appl. 30.4 (2008), pp. 1320–1338.

[8] T. Betcke. *The generalized singular value decomposition and the method of particular solutions*. SIAM J. Sci. Comput. 30.3 (2008), pp. 1278–1295.

[9] T. Braconnier, F. Chatelin, and J. C. Dunyach. *Highly nonnormal eigenproblems in the aeronautical industry*. Japan J. Ind. Appl. Math. 12.1 (1995), pp. 123–136.

[10] T. Braconnier, F. Chatelin, and V. Fraysse. *The influence of large nonnormality on the quality of convergence of iterative methods in linear algebra*. Preprint. CERFACS, 1994.

[11] T. Braconnier and N. J. Higham. *Computing the field of values and pseudospectra using the Lanczos method with continuation*. BIT 36.3 (1996), pp. 422–440.

[12] T. Braconnier, R. A. McCoy, and V. Toumazou. *Using the field of values for pseudospectra generation*. Preprint. CERFACS, 1997.

[13] C. Brezinski, M. Redivo-Zaglia, G. Rodriguez, and S. Seatzu. *Multi-parameter regularization techniques for ill-conditioned linear systems*. Numer. Math. 94.2 (2003), pp. 203–228.

[14] D. Calvetti and L. Reichel. *Tikhonov regularization of large linear problems*. BIT 43.2 (2003), pp. 263–283.

[15] S. Chaturantabut and D. C. Sorensen. *Nonlinear model reduction via discrete empirical interpolation*. SIAM J. Sci. Comput. 32.5 (2010), pp. 2737–2764.

[16] T.-Y. Chen and J. W. Demmel. *Balancing sparse matrices for computing eigenvalues*. Linear Algebra Appl. 309.1–3 (2000), pp. 261–287.

[17] J. K. Cullum and T. Zhang. *Two-sided Arnoldi and nonsymmetric Lanczos algorithms*. SIAM J. Matrix Anal. Appl. 24.2 (2002), pp. 303–319.

[18] R. J. A. David and D. S. Watkins. *An inexact Krylov–Schur algorithm for the unitary eigenvalue problem*. Linear Algebra Appl. 429.5-6 (2008), pp. 1213–1228.

[19] L. Elsner and M. Paardekooper. *On measures of nonnormality of matrices*. Linear Algebra Appl. 92 (1987), pp. 107–123.

[20] D. Fong and M. A. Saunders. *LSMR: an iterative algorithm for sparse least-squares problems*. SIAM J. Sci. Comput. 33.5 (2011), pp. 2950–2971.

[21] M. Fornasier, V. Naumova, and S. V. Pereverzyev. *Parameter choice strategies for multipenalty regularization*. SIAM J. Numer. Anal. 52.4 (2014), pp. 1770–1794.

[22] S. Gazzola and P. Novati. *Multi-parameter Arnoldi–Tikhonov methods*. Electron. Trans. Numer. Anal. 40 (2013), pp. 452–475.

[23] S. Gazzola and L. Reichel. *A new framework for multi-parameter regularization*. BIT 56.3 (2016), pp. 919–949.

[24] G. H. Golub and C. F. Van Loan. *Matrix Computations*. 4th ed. Johns Hopkins University Press, 2013.

[25] C. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman Publishing, Boston, 1984.

[26] K. E. Gustafson and D. K. M. Rao. *Numerical Range: The Field of Values of Linear Operators and Matrices*. Springer–Verlag New York, Inc., 1997.

[27] P. C. Hansen. *Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems*. Numer. Algorithms 6 (1994), pp. 1–35.

[28] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, Philadelpia, PA, 1998.

[29] P. Henrici. *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*. Numer. Math. 4.1 (1962), pp. 24–40.

[30] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. 2nd ed. SIAM, Philadelpia, PA, 2002.

[31] M. Hochbruck and M. E. Hochstenbach. *Subspace extraction for matrix functions*. Preprint. 2005.

[32] M. E. Hochstenbach. *A Jacobi–Davidson type method for the generalized singular value problem*. Linear Algebra Appl. 431 (2009), pp. 471–487.

[33] M. E. Hochstenbach. *Fields of values and inclusion regions for matrix pencils*. Electron. Trans. Numer. Anal. 38 (2011), pp. 98–112.

[34] M. E. Hochstenbach and L. Reichel. *An iterative method for Tikhonov regularization with a general linear regularization operator*. J. Integral Equations Appl. 22.3 (2010), pp. 465–482.

[35] M. E. Hochstenbach, L. Reichel, and X. Yu. *A Golub–Kahan-type reduction method for matrix pairs*. J. Sci. Comput. (2015), pp. 1–23.

[36] M. E. Hochstenbach, D. A. Singer, and P. F. Zachlin. *Eigenvalue inclusion regions from inverses of shifted matrices*. Linear Algebra Appl. 429.10 (2008), pp. 2481–2496.

[37] M. E. Hochstenbach, D. A. Singer, and P. F. Zachlin. *Numerical approximation of the field of values of the inverse of a large matrix*. Textos Mat. Sér. B 44 (2013), pp. 59–71.

[38] M. E. Hochstenbach and I. N. Zwaan. *Matrix balancing for field of values type inclusion regions* (2013). Submitted.

[39] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

[40] P. Howland, M. Jeon, and H. Park. *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*. SIAM J. Matrix Anal. Appl. 25.1 (2003), pp. 165–179.

[41] K. Ito, B. Jin, and T. Takeuchi. *Multi-parameter Tikhonov regularization*. arXiv:1102.1173v2 [math.NA]. Preprint. Mar. 2011.

[42] I. M. Jaimoukha and E. M. Kasenally. *Implicitly restarted Krylov subspace methods for stable partial realizations*. SIAM J. Matrix Anal. Appl. 18.3 (1997), pp. 633–652.

[43] Z. Jia and G. W. Stewart. *An analysis of the Rayleigh–Ritz method for approximating eigenspaces*. Math. Comp. 70.234 (2000), pp. 637–647.

[44] C. R. Johnson. *Functional characterizations of the field of values and the convex hull of the spectrum*. Proc. Amer. Math. Soc. 61.2 (1976), pp. 201–204.

[45] C. R. Johnson. *Numerical determination of the field of values of a general complex matrix*. SIAM J. Numer. Anal. 15.3 (1978), pp. 595–602.

[46] W. Kahan, B. N. Parlett, and E. Jiang. *Residual bounds on approximate eigensystems of nonnormal matrices*. SIAM J. Numer. Anal. 19.3 (1982), pp. 470–484.

[47] M. E. Kilmer, P. C. Hansen, and M. I. Español. *A projection-based approach to general-form Tikhonov regularization*. SIAM J. Sci. Comput. 29.1 (2007), pp. 315–330.

[48] A. Knyazev. "Preconditioned Eigensolvers." In *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Ed. by Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. A. van der Vorst. Section 11.3. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[49] A. V. Knyazev. *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*. SIAM J. Sci. Comput. 23.2 (2001), pp. 517–541.

[50] D. Kressner. *Numerical Methods and Software for General and Structured Eigenvalue Problems*. PhD thesis. Technische Universität Berlin, 2004.

[51] D. Kressner. *A periodic Krylov–Schur algorithm for large matrix products*. Numer. Math. 103.3 (2006), pp. 461–483.

[52] K. Kunisch and T. Pock. *A Bilevel optimization approach for parameter learning in variational models*. SIAM J. Imaging Sci. 6.2 (2013), pp. 938–983.

[53] J. Lampe, L. Reichel, and H. Voss. *Large-scale Tikhonov regularization via reduction by orthogonal projection*. Linear Algebra Appl. 436.8 (2012), pp. 2845–2865.

[54] D. Lemonnier and P. Van Dooren. *Balancing regular matrix pencils*. SIAM J. Matrix Anal. Appl. 28.1 (2006), pp. 253–263.

[55] R.-C. Li and Q. Ye. *A Krylov subspace method for quadratic matrix polynomials with applications to constrained least squares problems*. SIAM J. Matrix Anal. Appl. 25.2 (2003), pp. 405–528.

[56] S. Lu and S. V. Pereverzyev. *Multi-parameter regularization and its numerical realization*. Numer. Math. 118.1 (2011), pp. 1–31.

[57] S. Lu, S. V. Pereverzyev, Y. Shao, and U. Tautenhahn. *Discrepancy curves for multi-parameter regularization*. J. Inverse Ill-Posed Probl. 18.6 (2010), pp. 655–676.

[58] T. A. Manteuffel. *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*. Numer. Math. 31.2 (1978), pp. 183–208.

[59] T. A. Manteuffel and G. Starke. *On hybrid iterative methods for nonsymmetric systems of linear equations*. Numer. Math. 73.4 (1996), pp. 489–506.

[60] *The Matrix Market*. `math.nist.gov/MatrixMarket`, a repository for test matrices.

[61] K. Meerbergen and R. Morgan. "Inexact Methods." In *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Ed. by Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. A. van der Vorst. Section 11.2. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[62] S. Noschese and L. Reichel. *Inverse subspace problems with applications*. Numer. Linear Algebra Appl. 21 (2014), pp. 589–603.

[63] E. E. Osborne. *On pre-conditioning of matrices*. J. ACM 7.4 (1960), pp. 338–345.

[64] E. Ovtchinnikov. *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems I: the preconditioning aspect*. SIAM J. Numer. Anal. 41.1 (2003), pp. 258–271.

[65] E. Ovtchinnikov. *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems II: the subspace acceleration*. SIAM J. Numer. Anal. 41.1 (2003), pp. 271–286.

[66] C. C. Paige and M. A. Saunders. *Towards a generalized singular value decomposition*. SIAM J. Numer. Anal. 18.3 (1981), pp. 398–405.

[67] C. C. Paige and M. A. Saunders. *LSQR: An algorithm for sparse linear equations and sparse least squares*. ACM Trans. Math. Softw. 8.1 (1982), pp. 43–71.

[68] B. N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelpia, PA, 1998.

[69] B. N. Parlett and C. Reinsch. *Balancing a matrix for calculation of eigenvalues and eigenvectors*. Numer. Math. 13.4 (1969), pp. 293–304.

[70] P. Perona and J. Malik. *Scale-space and edge detection using anisotropic diffusion*. IEEE Trans. Pattern Anal. Mach. Intell. 12.7 (1990), pp. 629–639.

[71] L. Reichel, F. Sgallari, and Q. Ye. *Tikhonov regularization based on generalized Krylov subspace methods*. Appl. Numer. Math. 62.9 (2012), pp. 1215–1228.

[72] L. Reichel and X. Yu. *Matrix decompositions for Tikhonov regularization*. Electron. Trans. Numer. Anal. 43 (2015), pp. 223–243.

[73] L. Reichel and X. Yu. *Tikhonov regularization via flexible Arnoldi reduction*. BIT 55.4 (2015), pp. 1145–1168.

[74] J. Rommes and N. Martins. *Computing large-scale system eigenvalues most sensitive to parameter changes, with applications to power system small-signal stability*. IEEE Trans. Power Syst. 32.2 (2008), pp. 434–442.

[75] A. Ruhe. *On the closeness of eigenvalues and singular values for almost normal matrices*. Linear Algebra Appl. 11.1 (1975), pp. 87–93.

[76] A. Ruhe. "The two-sided Arnoldi algorithm for nonsymmetric eigenvalue problems." In *Matrix Pencils*. Ed. by B. Kågström and A. Ruhe. Springer, Berlin, Heidelberg, 1983, pp. 104–120.

[77] A. Ruhe. *Closest normal matrix finally found!* BIT 27.4 (1987), pp. 585–598.

[78] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. 2nd ed. SIAM, Philadelpia, PA, 2011.

[79] G. L. G. Sleijpen, J. van den Eshof, and P. Smit. *Optimal a priori error bounds for the Rayleigh–Ritz method*. Math. Comp. 72.242 (2002), pp. 677–684.

[80] G. L. G. Sleijpen, H. A. van der Vorst, and E. Meijerink. *Efficient expansion of subspaces in the Jacobi–Davidson method for standard and generalized eigenproblems*. Electron. Trans. Numer. Anal. 7 (1998), pp. 75–89.

[81] G. W. Stewart. *A Krylov–Schur algorithm for large eigenproblems*. SIAM J. Matrix Anal. Appl. 23.3 (2001/02), pp. 601–614.

[82] G. W. Stewart. *Matrix Algorithms II: Eigensystems*. SIAM, Philadelpia, PA, 2001.

[83] G. W. Stewart. *A generalization of Saad's theorem on Rayleigh–Ritz approximations*. Linear Algebra Appl. 327.1–3 (2001), pp. 115–119.

[84] G. W. Stewart. *Addendum to: "A Krylov–Schur algorithm for large eigenproblems"*. SIAM J. Matrix Anal. Appl. 24.2 (2002), pp. 599–601.

[85] G. W. Stewart. *On the numerical analysis of oblique projectors*. SIAM J. Matrix Anal. Appl. 32.1 (2011), pp. 309–348.

[86] M. Stoll. *A Krylov–Schur approach to the truncated SVD*. Linear Algebra Appl. 436.8 (2012), pp. 2795–2806.

[87] D. Szyld. *The many proofs of an identity on the norm of oblique projections*. Numer. Algorithms 42.3 (2006), pp. 309–323.

[88] F. Tisseur. *Backward error and condition of polynomial eigenvalue problems*. Linear Algebra Appl. 309.1-3 (2000), pp. 339–361.

[89] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra*. Princeton University Press, Princeton, NJ, 2005, pp. xviii+606.

[90] H. A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*. Vol. 13. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2003, pp. xiv+221.

[91] D. S. Watkins. *A case where balancing is harmful*. Electron. Trans. Numer. Anal. 23 (2006), pp. 1–4.

[92] H. Wielandt. *On eigenvalues of sums of normal matrices*. Pacific J. Math 5 (1955), pp. 633–638.

[93] T. G. Wright and L. N. Trefethen. *Large-scale computation of pseudospectra using ARPACK and eigs*. SIAM J. Sci. Comput. 23.2 (2001), pp. 591–605.

[94] Q. Ye. *Optimal expansion of subspaces for eigenvector approximations*. Linear Algebra Appl. 428.4 (2008), pp. 911–918.

[95] H. Zha. *Computing the generalized singular values/vectors of large space or structured matrix pairs*. Numer. Math. 72 (1996), pp. 391–417.

[96] Y. Zhou and Y. Saad. *Block Krylov–Schur method for large symmetric eigenvalue problems*. Numer. Algorithms 47.4 (2008), pp. 341–359.

[97] I. N. Zwaan and M. E. Hochstenbach. *Generalized Davidson and multidirectional-type methods for the generalized singular value decomposition* (2017). Preprint.

[98] I. N. Zwaan and M. E. Hochstenbach. *Krylov-Schur type restarts for the two-sided Arnoldi method*. SIAM J. Matrix Anal. Appl. 38.2 (2017), pp. 297–321.

[99] I. N. Zwaan and M. E. Hochstenbach. *Multidirectional subspace expansion for one-parameter and multiparameter Tikhonov regularization*. J. Sci. Comput. 70.3 (2017), pp. 990–1009.

Index

Summary

Generalized Krylov methods for large-scale matrix problems

This dissertation concerns the development of new Krylov subspace methods for large-scale (generalized) matrix problems. Of particular interest are standard eigenvalue problems, generalized singular value problems, and general form regularization problems, which commonly emerge in various fields of natural and applied sciences. The focus is on applications where the matrices are large enough that using direct methods is no longer feasible, and structured enough to facilitate fast matrix-vector products; that is, applications suitable for Krylov methods.

Four main subjects are considered in this dissertation: matrix balancing for field of value type inclusion regions, two-sided Krylov–Schur restarts, multidirectional subspace expansion for generalized and multiparameter Tikhonov regularization, and multidirectional subspace expansion for computing generalized singular values and vectors.

The field of values of a matrix is convex, guaranteed to contain all eigenvalues, and its boundary is often tight around the eigenvalues and can be approximated efficiently. Therefore, using the field of values as an inclusion region may be an attractive alternative in an exploratory phase to computing eigenvalues. However, occasionally the numerical radius of a matrix is much larger than its spectral radius, which makes the field of values meaningless as an inclusion region. We show that in this case, the quality of the field of values as an inclusion region may often be improved by balancing the matrix. Balancing is an existing technique designed to decrease the disparity between row and column norms through a carefully constructed diagonal similarity transform. Several interesting connections with the nonnormality of matrices are investigated and emphasized. Moreover, we propose a new, simple, and fast balancing methodology for computing spectral inclusion regions, where the Hessenberg matrix resulting from the Arnoldi process is balanced and used to approximate the field of values. The effectiveness of the method is demonstrated with numerical experiments.

Next, we derive two-sided Krylov–Schur as an extension of the Krylov–Schur restarting method to the two-sided Arnoldi method for large-scale nonnormal matrices. This extension allows for the simultaneous approximation of left and right eigenvectors, and thus eigenvalue condition numbers, while working exclusively with orthonormal bases. Specifically, two-sided Krylov–Schur maintains orthonormal bases for a separate left and right Krylov subspace, and applies only orthonormal transformations to these bases during the restarts. We may therefore expect better numerical stability compared to unsymmetric Lanczos if the method is carefully implemented. We describe algorithms for both standard Ritz extraction and harmonic Ritz extraction, and present several error bounds. Numerical examples where we compute the least sensitive eigenvalues or use the left and right shift-invariant subspace bases to approximate pseudospectra illustrate the usefulness of two-sided Krylov–Schur.

Generating high-quality search spaces for generalized Tikhonov regularization and multiparameter Tikhonov regularization may be challenging. In the latter case, selecting suitable regularization parameters may also be challenging. We introduce a new method for large-scale multiparameter Tikhonov regularization with general regularization operators. The method works by repeatedly extending the search space in multiple directions, similar to generalized Krylov, and subsequently removing the less promising directions to ensure moderate growth of the search space. Moreover, we propose a discrepancy principle based parameter selection strategy related to perturbation results. Numerical experiments are performed to test the algorithms.

Finally, we describe two subspace algorithms for computing extremal generalized singular values and vectors, that are also suitable for approximating truncated generalized singular value decompositions. The first algorithm can be seen as a restarted generalized Davidson algorithm, and the second algorithm improves upon the first with multidirectional subspace expansion. This multidirectional subspace expansion is also followed by a truncation step, although it is slightly different from the previous one. Furthermore, we provide additional insight into the multidirectional subspace expansion technique with several interesting theoretical observations, and generalize numerous error bounds for the symmetric eigenvalue problem to the generalized singular value problem.

Curriculum Vitae

Ian Zwaan was born on 22 December 1989 in Fredericksburg, VA, the United States of America. He grew up in the Netherlands and graduated from high school in 2007. He subsequently studied both Mathematics and Computer Science at the University of Utrecht and obtained two Bachelor degrees *cum laude* in early 2011. He completed his master in Mathematical Sciences *cum laude* in 2013 with the thesis *Compressed Sensing accelerated radial acquisitions for dynamic Magnetic Resonance Imaging*.

In February 2012 Ian started his PhD project at Eindhoven University of Technology under the supervision of dr. Michiel Hochstenbach and prof.dr.ir Barry Koren. The results obtained from his research are presented in this dissertation. The PhD position was funded by NWO within the project “Innovative methods for large matrix problems”.

List of publications

Preprints

1. I. N. Zwaan and M. E. Hochstenbach *Generalized Davidson and multidirectional-type methods for the generalized singular value decomposition*.
2. M. E. Hochstenbach and I. N. Zwaan *Matrix balancing for field of values type inclusion regions*.

Refereed journal papers

3. I. N. Zwaan and M. E. Hochstenbach *Krylov-Schur type restarts for the two-sided Arnoldi method*. SIAM J. Matrix Anal. Appl. 38.2 (2017), pp. 297–321.
4. I. N. Zwaan and M. E. Hochstenbach *Multidirectional subspace expansion for one-parameter and multiparameter Tikhonov regularization*. J. Sci. Comput. 70.3 (2017), pp. 990–1009.

Non-refereed proceedings papers

5. L. Gonzales, K. Huijssen, R. Jha, T. van Leeuwen, A. Sbrizi, W. van Valenberg, I. N. Zwaan, E. de Weerdt. *Patient-adaptive compressed sensing for MRI*. Proceedings of the 106th European Study Group Mathematics with Industry, 2015.
6. N. Banagaaya, N. Budko, H. van Doorn, G. Khimshiashvili, R. Klooster, PJ. Lindenbergh, J. Verdijck, F. Vermolen, and I. N. Zwaan. *The Mathematics of French Fries*. Proceedings of the 98th European Study Group Mathematics with Industry, 2014.

Popular mathematics

7. T van Leeuwen, A. Sbrizzi, I. N. and Zwaan. *Beeldreconstructie in MRI met compressed sensing*. NAW, 5/17.2 (June 2016), pp. 114-118.

Acknowledgments

There are many people to whom I owe my gratitude for being able become a Ph.D. researcher. Thank you all for your support and positive influence. First, I would like to thank the many teachers I have had over the years for their education and for providing me with the necessary mathematical knowledge. I would like to thank Gerard Sleijpen in particular, not just for teaching me about the fine art of numerical mathematics and numerical linear algebra, but also for his supervision of my master thesis, and for for bringing me into contact with a researcher in Eindhoven who had just received an NWO vidi grant and was looking for two Ph.D. students.

Michiel, I am honored to have had you as my mentor, and I am grateful for your contribution to my education. You effected both personal and intellectual growth, and inspired me for the future. Your advice and encouragement, as well as your faith and confidence in me, has been invaluable. I am also thankful to Barry Koren for his well judged advice and guidance when necessary.

My sincere appreciation to prof.dr. Joost Batenburg, dr.ir. Martin van Gijzen, prof. Per Christian Hansen, prof. Karl Meerbergen, and prof.dr. Siep Weiland for refereeing this dissertation and for being part of my defense committee. I would also like to thank the Netherlands Organisation for Scientific Research (NWO) for the financial support.

My gratitude extends to Tristan van Leeuwen and Alessandro Sbrizzi. We have had some interesting discussions and wrote an nice article together. I sincerely hope that we can continue to collaborate in the future.

I am grateful for all CASA members and the great atmosphere they help create. In particular, I appreciate the discussions and conversations I have had with my office mates Sarah Gaaf, Xiulei Cao, Sangye Lungten, and René Beltman. Furthermore, I wish to thank Enna van Dijk who could always help me with the "administrative details" I knew little about.

Last, but certainly not least, it is impossible to truly express how much I appreciate the care and support of my parents and my brothers, without whom I certainly could not have written this dissertation.