Stable Algorithms for Large Sparse Eigenvalue Problems

By

CHAO-PING LIN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Applied Mathematics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Zhaojun Bai, Chair

James Bremer

Patrice Koehl

Committee in Charge

2021

Contents

List of Figures

List of Tables

Chao-Ping Lin
September 2021
Applied Mathematics

Stable Algorithms for Large Sparse Eigenvalue Problems

Abstract

In this dissertation, we consider the symmetric eigenvalue problem and the buckling eigenvalue problem. We study existing algorithms and propose stable variants for both eigenvalue problems.

We first analyze Hotelling's deflation for the symmetric eigenvalue problem $Ax = \lambda x$, where A is a symmetric matrix. Hotelling's deflation is a technique to displace computed eigenvalues of A. It is combined with an eigensolver to compute a partial eigendecomposition of A. Numerical stability of Hotelling's deflation is not well understood. In this dissertation, we derive computable upper bounds on the loss of orthogonality of computed eigenvectors and on the backward error norm of computed eigenpairs. From the upper bounds, we identify crucial quantities associated with the shifts and derive sufficient conditions for the backward stability of Hotelling's deflation. Based on these results, we propose a shift selection scheme for stabilizing Hotelling's deflation.

Next we consider the buckling eigenvalue problem $Kx = \lambda K_G x$, where the matrix K is positive semi-definite, the matrix K_G is indefinite, and the matrices K and K_G share a common nullspace. When K is positive definite, the shift-invert Lanczos method is a widely accepted method for the buckling eigenvalue problem. However, in our case, there are two issues associated with the method. First, the shift-invert operator $(K - \sigma K_G)^{-1}$ does not exist or is ill-conditioned. Second, the Lanczos vectors fall rapidly into the nullspace of K. The inner product induced by K leads to rapid growth of the Lanczos vectors in norm. The large norms introduce large round-off errors to the orthogonalization process, leading to loss of accuracy of compute solutions and even break down of the method. In this dissertation, we address these issues by generalizing the buckling spectral transformation to the singular pencil $K - \lambda K_G$ and regularizing the inner product to bound the norms of the Lanczos vectors. We propose a shift-invert Lanczos method for the buckling eigenvalue problem and develop a validation scheme using inertias.

Acknowledgments

First of all, I want to thank my advisor, Professor Zhaojun Bai, for guiding me over the past years. His knowledge, insight and patience are the biggest help and support for my graduate study. I also want to thank Professor Patrice Koehl and Professor James Bremer for serving on my qualify exam committee and dissertation committee, and Professor Naoki Saito and Professor Mark M. Rashid for serving on my qualifying exam committee.

This dissertation would not be complete without the help from many people. My special thanks go to Dr. Roger Grimes from LSTC and Dr. Ben-Shan Liao from Siemens PLM. Dr. Roger Grimes provides the industrial example for the shift-invert Lanczos method. Dr. Ben-Shan Liao shares his experience and provides some benchmark examples in industry. I also want to thank Professor Patrice Koehl, Professor Ding Lu, Professor Ren-Cang Li, and Professor Huiqing Xie for their contributions to the dissertation. Professor Patrice Koehl shares the data from biological analysis. Professor Ren-Cang Li shares his thought on Hotelling's deflation. Professor Ding Lu gives helpful suggestion leading to significant improvement of the analysis. The theoretical work by Professor Huiqing Xie is used to develop the shift-invert Lanczos method.

Beyond my graduate study, my parents and sister are always supportive for any decision I made. They accompany me to overcome many difficulties.

I had the fortune to meet some interesting people. They enrich my life during my graduate study. Especially, I would like to send my best wishes to Zhongruo Wang, Yunshen Zhou, Chengming Jiang, Luan Nyengen, Chih-Fan Chen, Jingyang Shu, Yanwen Luo, Xiaochen Liu, Jiawei Wang, Haotian Li, Chen Ji, Bohan Zhou, Yiqun Shao.

This work in this dissertation was supported in part by the NSF grant DMS-1913364.

Chapter 1

Introduction

1.1 Motivations

An algorithm follows a set of well-defined procedures to solve a mathematical problem. In finite precision arithmetic, however, behavior of algorithm could depart significantly from the theoretical expectation. Naive implementation would lead to instability of the algorithm. In this dissertation, we consider the symmetric eigenvalue problem and the buckling eigenvalue problem. We study existing algorithms and propose stable variants for both eigenvalue problems.

We first analyze Hotelling's deflation for the symmetric eigenvalue problem $Ax = \lambda x$, where A is a symmetric matrix. Hotelling's deflation is a technique to displace computed eigenvalues of A. It is combined with eigensolver to compute partial eigendecomposition of A. In the literature, numerical stability of Hotelling's deflation is not well understood. In the book [35, p. 585], Wilkinson comments that Hotelling's deflation has poor numerical stability. Parlett argues that, when deflating out the largest computed eigenvalue, the change to the smallest eigenvalue in magnitude would be at the same order of the round-off error incurred [26, Sec. 5.1]. Saad observes the loss of orthogonality of computed eigenvectors and proposes reorthogonalizing computed eigenvectors before applying deflation. Saad performs a backward stability analysis of this variant, and claims that the stability is determined by the angle between the computed eigenvector and the deflated subspace [27]. In this dissertation, we demonstrate that the stability of Hotelling's deflation is determined by *spectral gap* and *shift-gap ratio*. We propose a shift selection scheme for stabilizing Hotelling's deflation.

Next we consider the shift-invert Lanczos method for the buckling eigenvalue problem

$Kx = \lambda K_G$. The shift-invert Lanczos method is a Krylov subspace method with spectral transformation. It is a widely accepted method to compute solutions for structural analysis. The use of the inner product induced by a positive semidefinite matrix is common in practice. We refer to [16] and the references therein for further detail.

The buckling eigenvalue problem arises from a buckling analysis of structure . The problem has the properties that the matrix K is positive semi-definite, the matrix K_G is indefinite, and the matrices K and K_G share a common nullspace. The shift-invert Lanczos method is not applicable since (1) the shift-invert matrix $(K - \sigma K_G)^{-1}$ does not exist or is ill-conditioned and (2) the Lanczos vectors fall rapidly into the nullspace of K. The inner product induced by K leads to rapid growth of the Lanczos vectors in norm. The large norms of the Lanczos vectors introduce large round-off errors to the orthogonalization process, leading to loss of accuracy of computed solutions and even break down of the method [23, 25, 33].

In the past, the norms are controlled by restricting the Lanczos vectors to a proper subspace [23, 25]. A modified formula is also proposed to improve the Ritz vectors [25]. These techniques however can not stop the unbounded growth of the Lanczos vectors. In this dissertation, we develop alternative strategies to address these issues.

1.2 Contributions

The main contributions to the analysis of Hotelling's deflation are as follows.

1. We derive computable upper bounds on the loss of orthogonality of computed eigenvectors and on the backward error norm of computed eigenpairs.

2. We identify crucial quantities associated with the shifts and derive sufficient conditions for the backward stability of Hotelling's deflation.

3. We propose a shift selection scheme for stabilizing Hotelling's deflation.

Through numerical experiments, we demonstrate the sharpness of our bounds, and the effectiveness of our shift selection scheme.

Next we address the issues associated with the buckling eigenvalue problem $Kx = \lambda K_G x$. The main contributions on this part are as follows.

1. We generalize the buckling spectral transformation to the singular pencil $K - \lambda K_G$.

2. We regularize the inner product to bound the norms of Lanczos vectors.

We propose a shift-invert Lanczos method for the buckling eigenvalue problem and provide an implementation of the matrix-vector product based on the scheme in [3]. A validation scheme using inertias is developed. The efficacy of our method is demonstrated by a numerical example from industrial analysis.

1.3 Organization

This dissertation is organized as follows. In Chapter 2, we introduce backgrounds and theoretical tools. In particular, we introduce Lanczos method and several of its variants. In Chapter 3, we present analysis of Hotelling's deflation for the symmetric eigenvalue problem $Ax = \lambda x$. In Chapter 4, we propose a shift selection scheme for Hotelling's deflation and present the numerical results. In Chapter 5, we address the issues with the buckling eigenvalue problem $Kx = \lambda K_G x$. The efficacy of our strategy is demonstrated in Chapter 6.

1.4 Notation

Throughout this dissertation, capital letters are matrices and lower case letters are vectors or scalars. I_k is the k-by-k identity matrix. e_j is the jth column of the identity matrix I_k. X^{T} is the transpose of the matrix X and $X^{-\mathrm{T}}$ is the inverse of the transpose X^{T}. $\det(X)$ is the determinant of X. $\|\cdot\|_1$ and $\|\cdot\|_2$ are matrix or vector 1-norm and 2-norm, respectively. $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. $\mathcal{R}(X)$ is the range of X and $\mathcal{N}(X)$ is the nullspace of X. $\mathcal{S}_1 \oplus \mathcal{S}_2$ is the direct sum of two subspace $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}^{\perp}$ is the orthogonal complement to a subspace $\mathcal{S}$. $\mathcal{P}_{\mathcal{S}}$ is the orthogonal projection onto a subspace $\mathcal{S}$.

We also use the machine epsilon ε, which is the gap between 1 and the next largest floating point number. We use the Big O notation $O(\cdot)$ to state the order of magnitude. Other notations will be explained as used.

Chapter 2

Preliminaries

We first provide backgrounds for the rest of chapters.

2.1 Eigenvalue problems

2.1.1 Symmetric eigenvalue problem

The symmetric eigenvalue problem (SEP) is of the form

$$Ax = \lambda x, \tag{2.1}$$

where A is a n-by-n real-valued symmetric matrix. The eigenvalues of the symmetric matrix A are denoted by

$$\lambda(A) = \{\lambda \mid \det(A - \lambda I) = 0\}.$$

If $\lambda \in \lambda(A)$ and x is a nonzero vector satisfying (2.1), we say x is an eigenvector corresponding to the eigenvalue λ.

The symmetric eigenvalue problem (2.1) has n real eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n, \tag{2.2}$$

and there exist real eigenvectors $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{n \times n}$ satisfying

$$X^{\mathrm{T}} A X = \Lambda. \quad \text{and} \quad X^{\mathrm{T}} X = I_n,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ [8]. The ordering (2.2) of the eigenvalues λ_i is used throughout the rest of chapters.

2.1.2 Symmetric semidefinite generalized eigenvalue problem

The symmetric semidefinite generalized eigenvalue problem (SGEP) is of the form

$$Ax = \lambda Bx, \tag{2.3}$$

where A and B are n-by-n symmetric matrices and B is positive semidefinite. We call the matrix pencil $A - \lambda B \in \mathbb{R}^{n\times n}$ a symmetric semidefinite pencil. The eigenvalues of the symmetric semidefinite pencil $A - \lambda B$ are denoted by

$$\lambda(A, B) = \{\lambda \mid \det(A - \lambda B) = 0\}.$$

If $\lambda \in \lambda(A, B)$ and x is a nonzero vector satisfying (2.3), we say x is an eigenvector corresponding to the eigenvalue λ.

When the matrix B in (2.3) is a positive definite matrix, the generalized eigenvalue problem (2.3) is equivalent to the symmetric eigenvalue problem

$$L^{-1}AL^{-\mathrm{T}}y = \lambda y,$$

where $B = LL^{\mathrm{T}}$ is the Cholesky factorization and $y = L^{\mathrm{T}}x$. Therefore, it has n real eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n,$$

and there exist real eigenvectors $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{n\times n}$ satisfying

$$X^{\mathrm{T}}AX = \Lambda. \quad \text{and} \quad X^{\mathrm{T}}BX = I_n,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

When the matrix B in (2.3) is positive semidefinite, the following decomposition is derived by Fix and Heiberger [13] to analyze the generalized eigenvalue problem (2.3):

Lemma 2.1. *For a symmetric semi-definite pencil $A - \lambda B$, there exists a non-singular matrix $W \in \mathbb{R}^{n\times n}$ such that*

$$W^TAW = \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & A_{00} & A_{01} & A_{02} & \Theta & 0 \\ n_1 & A_{01}^T & A_{11} & A_{12} & & \\ n_2 & A_{02}^T & A_{12}^T & \Psi & & \\ n_0 & \Theta & & & 0 & \\ n_3 & 0 & & & & 0 \end{array} \quad \textit{and} \quad W^TBW = \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & I_{n_0} & & & & \\ n_1 & & I_{n_1} & & & \\ n_2 & & & 0 & & \\ n_0 & & & & 0 & \\ n_3 & & & & & 0 \end{array}, \tag{2.4}$$

where Ψ and Θ are non-singular, diagonal matrices with real diagonal entries.

From the decomposition (2.4), the finite eigenvalues of the generalized eigenvalue problem (2.3) are the eigenvalues of the symmetric eigenvalue problem

$$(A_{11} - A_{12}\Psi^{-1}A_{12}^{\mathrm{T}})z = \lambda z.$$

We further analyze the decomposition (2.4) in Chapter 5.

2.2 Lanczos method

Lanczos method [5] is a projection method for the symmetric eigenvalue problem (2.1). The method starts with a vector $v \in \mathbb{R}^n$ and builds up the Krylov subspace

$$\mathcal{K}^{j+1}(A, v) = \mathrm{span}\{v, Av, A^2 v, \ \dots\ , A^j v\}.$$

The symmetric eigenvalue problem (2.1) is projected onto the Krylov subspace $\mathcal{K}^{j+1}(A, v)$. At each iteration, an orthonormal basis $V_{j+1} = [v_1, v_2, \dots, v_{j+1}]$ of $\mathcal{K}^{j+1}(A, v)$ is computed by orthogonalizing the product Av_j against the basis V_j. The orthogonalization process generates the Rayleigh Ritz projection $T_j = V_j^{\mathrm{T}} A V_j$ satisfying

$$AV_j = V_j T_j + \beta_j v_{j+1} e_j^{\mathrm{T}}. \tag{2.5}$$

By construction, the projection matrix T_j is symmetric and tridiagonal. From (2.5), the basis vector v_{j+1} satisfies the three-term recurrence

$$\beta_j v_{j+1} = Av_j - \beta_{j-1} v_{j-1} - \alpha_j v_j, \tag{2.6}$$

where

$$T_j = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ & & \beta_{j-1} & \alpha_j \end{bmatrix}$$

The Lanczos method is summarized in Algorithm 2.1. To compute approximate eigenpairs (λ_i, x_i), we solve the reduced eigenvalue problem

$$T_j s_i = \theta_i s_i$$

by the QR algorithm [1] and set the Ritz pair $(\lambda_i, x_i) = (\theta_i, V_j s_i)$. From (2.5), the residual error of the Ritz pair (λ_i, x_i) can be estimated by

$$\|r_i\|_2 = \|Ax_i - \lambda_i x_i\|_2 = |\beta_j| \cdot |s_{ji}|, \quad s_{ji} = e_j^{\mathrm{T}} s_i.$$

Throughout the rest of chapters, the basis vectors v_i in the Lanczos method are referred to as the Lanczos vectors. The decomposition (2.5) of the symmetric matrix A is referred to as the Lanczos decomposition.

Algorithm 2.1 Lanczos method

```
$r \leftarrow v$, where $v$ is the starting vector
$v_0 \leftarrow 0$
$\beta_0 \leftarrow \|r\|_2$
for $j = 1, 2, \ldots$ do
    $v_j \leftarrow r/\beta_{j-1}$
    $r \leftarrow Av_j$
    $r \leftarrow r - \beta_{j-1}v_{j-1}$
    $\alpha_j \leftarrow v_j^T r$
    $r \leftarrow r - \alpha_j v_j$
    perform re-orthogonalization if necessary
    $\beta_j \leftarrow \|r\|_2$
    Compute the eigenvalue decomposition of $T_j$
    Check convergence
end for
Compute approximate eigenvectors of the converged eigenpairs
```

2.2.1 Loss of orthogonality

In the Lanczos method, the vector v_{j+1} is explicitly orthogonalized against the vector v_j. The orthogonality between the vectors v_{j+1} and v_i, $i < j$, is determined by the three-term recurrence (2.6). Round-off errors however enter the recurrence (2.6) in finite precision arithmetic. The equations satisfied by the computed Lanczos vectors v_i become

$$\beta_i v_{i+1} = Av_i - \beta_{i-1}v_{i-1} - \alpha_i v_i + f_i, \quad i = 1, \ \ldots \ , \ j, \tag{2.7}$$

where $f_i \in \mathbb{R}^n$ are the terms from the round-off errors.

From (2.7), for $i < j$, the inner products $\omega_{k,l} = v_k^{\mathrm{T}} v_l$ satisfy

$$\beta_j \omega_{i,j+1} = v_i^{\mathrm{T}} A v_j - \beta_{j-1}\omega_{i,j-1} - \alpha_j \omega_{i,j} + v_i^{\mathrm{T}} f_j \tag{2.8}$$

and

$$\beta_i \omega_{j,i+1} = v_j^{\mathrm{T}} A v_i - \beta_{i-1}\omega_{j,i-1} - \alpha_i \omega_{j,i} + v_j^{\mathrm{T}} f_i. \tag{2.9}$$

Substracting (2.8) from (2.9), and using the symmetry of A, we have the difference equation

$$\beta_j \omega_{i,j+1} = \beta_i \omega_{i+1,j} + (\alpha_i - \alpha_j)\omega_{i,j} + \beta_{i-1}\omega_{i-1,j} - \beta_{j-1}\omega_{i,j-1} + v_i^{\mathrm{T}} f_j - f_i^{\mathrm{T}} v_j \quad \text{for} \quad i < j. \tag{2.10}$$

The difference equation (2.10) describes how the loss of orthogonality evolves in the Lanczos method. From (2.10), the products $\omega_{i,j+1}$, $i < j$, is a weighted sum of the products $\omega_{i',j}$ and $\omega_{i',j-1}$ from the previous two Lanczos steps; the terms $v_i^{\mathrm{T}} f_j$ and $f_i^{\mathrm{T}} v_j$ from the round-off errors introduce perturbation to $\omega_{i,j+1}$. In exact arithmetic, we have the zero round-off errors $f_i = 0$ and $\omega_{i,i+1} = v_i^{\mathrm{T}} v_{i+1} = 0$. From the difference equation (2.10), $\omega_{i,j+1} = 0$ for all $i < j$. When the round-off errors are present, tiny products $\omega_{k,l}$ are introduced through the terms $\omega_{i,i+1}$, $v_i^{\mathrm{T}} f_j$ and $f_i^{\mathrm{T}} v_j$. The nonzero products $\omega_{k,l}$ are then propagated by the equation (2.10) in the subsequent iterations. Potential instability of the difference equation (2.10) leads to the amplification of the products $\omega_{i,j+1}$. The amplification results in the loss of orthogonality of the Lanczos vectors v_i [20, Sec. 10.6] [31].

Further analysis of the equations (2.7) leads to the well-known result by Paige [8, p. 379] [26, p. 295]

$$x_i^{\mathrm{T}} v_{j+1} = \frac{O(\varepsilon \|A\|_2)}{\beta_j s_{ji}}, \tag{2.11}$$

where $x_i = V_j s_i$ is the Ritz vector. The result (2.11) tells that the Lanczos vectors v_i are driven toward the converged Ritz vectors.

2.2.2 Reorthogonalization

We discuss commonly used techniques at step 10 of Algorithm 2.1 to maintain the orthogonality of the computed Lanczos vectors v_i:

- **Full reorthogonalization.** The vector r is explicitly orthogonalized against the vectors V_j at each iteration of the Lanczos method. The orthogonalization is done using the classical Gram-Schmidt

  $$r' \leftarrow r - V_j(V_j^{\mathrm{T}} r). \tag{2.12}$$

  The decrease in the norm $\|r\|_2$ after the orthogonalization is monitored. Additional orthogonalization by (2.12) is performed if $\|r'\|_2 < \frac{1}{\sqrt{2}}\|r\|_2$ [6].

- **Selective reorthogonalization.** From the difference equation (2.10), we compute an upper bound $\overline{\omega}_{k,l}$ on the loss of orthogonality $\omega_{k,l}$ by

$$\overline{\omega}_{i,j+1} = \frac{1}{\beta_j} \cdot \left(\beta_i \overline{\omega}_{i+1,j} + |\alpha_i - \alpha_j| \overline{\omega}_{i,j} + \beta_{i-1} \overline{\omega}_{i-1,j} + \beta_{j-1} \overline{\omega}_{i,j-1} + \overline{f}_j\right),$$

  where $\overline{f}_j$ is an estimate of the order of the terms $v_i^{\mathrm{T}} f_j - f_i^{\mathrm{T}} v_j$ [31]. We orthogonalize the vectors r and v_j against the vector v_i whenever $\overline{\omega}_{i,j+1} \geq \sqrt{\varepsilon}$ [30] [31]. We refer to [16] and the references therein for further details on the robust reorthogonalization technique.

2.3 Restarted Lanczos methods

In this section, we consider the Lanczos decomposition of a symmetric matrix A

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^{\mathrm{T}}. \tag{2.13}$$

We discuss techniques to re-compute the decomposition (2.13) with refined projection subspace.

2.3.1 Implicit restart

Implicit restart [32] first computes a new basis $V_{m+1}^{(1)}$ through QR iteration of the tridiagonal matrix T_m. The QR iteration is the transformation

$$T_m^{(1)} = Q_1^{\mathrm{T}} T_m Q_1,$$

where $T_m - \mu_1 I_m = Q_1 R_1$ is the QR decomposition and $\mu_1 \in \mathbb{R}$ is a shift. Implicit restart computes the new basis $V_m^{(1)}$ by the orthogonal transformation

$$V_m^{(1)} = V_m Q_1. \tag{2.14}$$

The orthogonal transformation in (2.14) serves to filter the starting vector v_1 by

$$v_1^{(1)} = \frac{(A - \mu_1 I_m) v_1}{\|(A - \mu_1 I_m) v_1\|_2}.$$

The new basis $V_m^{(1)}$ satisfies the equation

$$AV_m^{(1)} = V_m^{(1)} T_m^{(1)} + \beta_m v_{m+1} e_m^{\mathrm{T}} Q_1,$$

where the matrix $T_m^{(1)} = {V_m^{(1)}}^{\mathrm{T}} A V_m^{(1)}$ is the Rayleigh Ritz projection.

Implicit restart then continues QR iterations on the matrix $T_m^{(1)}$ and updates the basis $V_m^{(1)}$. At i-th QR iteration, $i \geq 2$,

$$T_m^{(i)} = Q_i^{\mathrm{T}} T_m^{(i-1)} Q_i,$$

where $T_m^{(i-1)} - \mu_i I_m = Q_i R_i$ is the QR decomposition and $\mu_i \in \mathbb{R}$ is the shift, implicit restart updates the basis V_m by

$$V_m^{(i)} = V_m^{(i-1)} Q_i = V_m Q_1 Q_2 \dots Q_i.$$

After p QR iterations, $p < m$, the starting vector $v_1^{(p)}$ is

$$v_1^{(p)} = \frac{\prod_{i=1}^{p}(A - \mu_i I_m) v_1}{\| \prod_{i=1}^{p}(A - \mu_i I_m) v_1 \|_2}. \tag{2.15}$$

The new basis $V_m^{(p)}$ satisfies the equation

$$A V_m^{(p)} = V_m^{(p)} T_m^{(p)} + \beta_m v_{m+1} h^{\mathrm{T}}, \tag{2.16}$$

where the matrix

$$T_m^{(p)} = V_m^{(p)\mathrm{T}} A V_m^{(p)} = \begin{bmatrix} \alpha_1^{(p)} & \beta_1^{(p)} & & \\ \beta_1^{(p)} & \alpha_2^{(p)} & \ddots & \\ & \ddots & \ddots & \beta_{m-1}^{(p)} \\ & & \beta_{m-1}^{(p)} & \alpha_m^{(p)} \end{bmatrix}.$$

is the Rayleigh Ritz projection and $h^{\mathrm{T}} = e_m^{\mathrm{T}} Q_1 Q_2 \dots Q_p$.

To restart, implicit restart starts the Lanczos method with the filtered vector $v_1^{(p)}$ (2.15). We note that, in (2.16), each orthogonal matrix Q_i is upper Hessenberg, and the leading $m - p - 1$ entries of the row vector h^{T} are zero, i.e.,

$$h^{\mathrm{T}} = e_m^{\mathrm{T}} Q_1 Q_2 \dots Q_p = [0, \ \dots \ , 0, h_k, \ \dots \ , h_m], \quad k = m - p,$$

Therefore, from (2.16), the leading k basis vectors of $V_m^{(p)}$, denoted by V_k^+, satisfy

$$A V_k^+ = V_k^+ T_k^+ + (\beta_k^{(p)} v_{k+1}^{(p)} + \beta_m h_k v_{m+1}) e_k^{\mathrm{T}} := V_k^+ T_k^+ + \beta_k^+ v_{k+1}^+ e_k^{\mathrm{T}}$$

where T_k^+ is the leading k-by-k submatrix of the tridiagonal matrix $T_m^{(p)}$, and we can start with the $(k+1)$th step of the Lanczos method.

2.3.2 Thick restart

Thick restart [36] first solves the reduced eigenvalue problem

$$T_m s_i = \theta_i s_i.$$

The desired Ritz vectors $u_{i_1}, \ \dots \ , u_{i_k}$ are identified. Thick restart then starts the Lanczos method with the vectors

$$V_{k+1}^+ = [v_1^+, \ \dots \ , v_k^+, v_{k+1}^+] = [u_{i_1}, \ \dots \ , u_{i_k}, v_{m+1}] \, .$$

The starting vectors V_{k+1}^+ satisfy the equation

$$AV_k^+ = V_k^+ D_k + \beta_m v_{k+1}^+ h^{\mathrm{T}},$$

where $D_k = \mathrm{diag}(\theta_{i_1}, \ \ldots \ , \theta_{i_k})$, $h^{\mathrm{T}} = e_m^{\mathrm{T}} S_k$ and $S_k = [s_{i_1}, \ \ldots \ , s_{i_k}]$.

In the subsequent iterations, Lanczos method builds the projection subspace

$$\mathcal{K}_{j+1}^+ = \mathrm{span}\{u_{i_1}, \ldots, u_{i_k}, v_{m+1}, Av_{m+1}, \ldots, A^{j-k} v_{m+1}\}.$$

The orthonormal basis

$$V_{j+1}^+ = [v_1^+, \ \ldots \ , v_k^+, v_{k+1}^+, \ \cdots \ , v_{j+1}^+]$$

of $\mathcal{K}_{j+1}^+$ is computed by orthogonalizing the product Av_j^+ against the basis V_j^+. The orthogonalization process generates the Rayleigh Ritz projection $T_j^+ = V_j^{+\mathrm{T}} A V_j^+$ satisfying

$$AV_j^+ = V_j^+ T_j^+ + \beta_j^+ v_{j+1}^+ e_j^{\mathrm{T}}. \tag{2.17}$$

The projection matrix T_j^+ now is symmetric and has the banded structure,

$$T_j^+ = \begin{bmatrix} D_k & \beta_m h & & & & \\ \beta_m h^{\mathrm{T}} & \alpha_{k+1}^+ & \beta_{k+1}^+ & & & \\ & \beta_{k+1}^+ & \alpha_{k+2}^+ & \beta_{k+2}^+ & & \\ & & \ddots & \ddots & \ddots & \\ & & & \beta_{j-2}^+ & \alpha_{j-1}^+ & \beta_{j-1}^+ \\ & & & & \beta_{j-1}^+ & \alpha_j^+ \end{bmatrix}. \tag{2.18}$$

From (2.17) and (2.18), the Lanczos vector v_{j+1}^+ is computed by the recurrence

$$\beta_{j+1}^+ v_{j+1}^+ = \begin{cases} Av_j^+ - \alpha_j^+ v_j^+ - \beta_m V_{j-1}^+ h & \text{if} \quad j = k+1, \\ Av_j^+ - \alpha_j^+ v_j^+ - \beta_{j-1}^+ v_{j-1}^+ & \text{if} \quad j > k+1. \end{cases}$$

To compute approximate eigenpairs, we solve the reduced eigenvalue problem

$$T_j^+ s_i = \theta_i s_i$$

and set the Ritz pair $(\lambda_i = \theta_i, x_i = V_j^+ s_i)$. From (2.17), the residual error of the Ritz pair (λ_i, x_i) can be estimated by

$$\|r_i\|_2 = \|Ax_i - \lambda_i x_i\|_2 = |\beta_j| \cdot |s_{ji}|, \quad s_{ji} = e_j^{\mathrm{T}} s_i.$$

In finite precision arithmetic, we may perform full reorthogonalization to maintain the orthogonality of the computed Lanczos vectors [39]. A selective reorthogonalization scheme is developed in [36].

2.4 Shift-invert Lanczos method

Shift-invert Lanczos method [11, 16, 25] generalizes the Lanczos method to the SGEP (2.3). The method computes the eigenvalues λ and the associated eigenvectors x of (2.3) near a prescribed shift σ. It begins by converting (2.3) via a shift-invert spectral transformation into the equivalent eigenvalue problem

$$Cx \equiv (A - \sigma B)^{-1}Bx = \mu x, \quad \mu = \frac{1}{\lambda - \sigma}. \tag{2.19}$$

The spectral transformation (2.19) maps the eigenvalues λ near the shift σ into the extremal eigenvalues μ of C. Next, it is noted that C is symmetric with respect to the inner product induced by B. The Lanczos method is then run on C to compute the eigenpairs (μ, x) with the extremal eigenvalues μ.

The Lanczos method starts with a vector $v \in \mathbb{R}^n$ and builds the Krylov subspace $\mathcal{K}^{j+1}(C, v)$. An orthonormal basis V_{j+1} of $\mathcal{K}^{j+1}(C, v)$ is computed with the inner product induced by B. The orthogonalization process generates the projection $T_j = V_j^{\mathrm{T}} BCV_j$ satisfying

$$CV_j = V_j T_j + \beta_j v_{j+1} e_j^{\mathrm{T}} \tag{2.20}$$

with

$$T_j = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ & & \beta_{j-1} & \alpha_j \end{bmatrix}.$$

The Lanczos vector v_{j+1} satisfies the three-term recurrence

$$\beta_j v_{j+1} = Cv_j - \beta_{j-1} v_{j-1} - \alpha_j v_j.$$

The shift-invert Lanczos method is summarized in Algorithm 2.2. To compute approximate eigenpairs (λ_i, x_i) of the original problem (2.3), we solve the reduced eigenvalue problem

$$T_j s_i = \theta_i s_i$$

and set $(\lambda_i, x_i) = (\frac{1}{\theta_i} + \sigma, V_j s_i)$. From (2.20), the residue vector of (λ_i, x_i) is

$$r_i = Ax_i - \lambda_i Bx_i = -\frac{\beta_j s_{ji}}{\theta_i}(A - \sigma B)v_{j+1}, \quad s_{ji} = e_j^{\mathrm{T}} s_i.$$

Typically, the extremal eigenvalues of C are well separated. The eigenpairs (λ_i, x_i) near the shift σ converge after a few iterations of the shift-invert Lanczos method. At step 12 of Algorithm 2.2, to maintain the orthogonality of the Lanczos vector v_i, full reorthogonalization performs the classical Gram-Schmidt

$$r' \leftarrow r - V_j(V_j^{\mathrm{T}} Br).$$

Robust selective reorthogonalization schemes are developed in [16] [20, Sec. 10.6].

Algorithm 2.2 Shift-invert Lanczos method

```
 1: $r \leftarrow v$, where $v$ is the starting vector
 2: $v_0 \leftarrow 0$
 3: $p \leftarrow Br$
 4: $\beta_0 \leftarrow (p^T r)^{1/2}$
 5: for $j = 1, 2, \ldots$ do
 6:     $v_j \leftarrow r/\beta_{j-1}$
 7:     $r \leftarrow Cv_j$, where $C = (A - \sigma B)^{-1}B$
 8:     $r \leftarrow r - \beta_{j-1}v_{j-1}$
 9:     $p \leftarrow Br$
10:     $\alpha_j \leftarrow v_j^T p$
11:     $r \leftarrow r - \alpha_j v_j$
12:     perform re-orthogonalization if necessary
13:     $p \leftarrow Br$
14:     $\beta_j \leftarrow (p^T r)^{1/2}$
15:     Compute the eigenvalue decomposition of $T_j$
16:     Check convergence
17: end for
18: Compute approximate eigenvectors of the converged eigenpairs
```

2.5 Hotelling's deflation

Hotelling's deflation [19] is a technique to displace computed eigenvalues of symmetric matrix A. Given a computed eigenpair (λ, x) of A by an eigensolver, Hotelling's deflation displaces the eigenvalue λ by choosing a real shift σ and applying the low-rank update,

$$A' = A + \sigma \cdot xx^{\mathrm{T}}. \tag{2.21}$$

The low-rank update (2.21) displaces the computed eigenpair (λ, x) to an eigenpair $(\lambda + \sigma, x)$ of A' while the rest of the eigenpairs are not changed.

Hotelling's deflation is originally combined with the power method to compute the largest eigenvalues of A in magnitude [19] [26, Sec. 5.1] [28, p. 90] [35, p. 585]: we start with computing the largest eigenvalue λ and the associated eigenvector x of A by the power method. We apply Hotelling's deflation with the shift $\sigma = -\lambda$ to reveal the next largest eigenvalue λ' of A. Then we compute the eigenvalue λ' by applying the power method to the low-rank updated matrix A'.

Through the rest of chapters, Hotelling's deflation is referred to as explicit external deflation (EED).

2.6 Inertias of symmetric matrix

The inertias of a symmetric matrix A refer to the numbers $\nu_+(A)$, $\nu_0(A)$ and $\nu_-(A)$, where $\nu_+(A)$, $\nu_0(A)$ and $\nu_-(A)$ are the numbers of positive, zero and negative eigenvalues of A, respectively. Sylvester law states that the inertias are invariant under congruent transformation:

Theorem 2.1 (**Sylvester Law** [15, p. 448])**.** *If the matrix $A \in \mathbb{R}^{n\times n}$ is symmetric and the matrix $W \in \mathbb{R}^{n\times n}$ is non-singular, then the matrices A and $W^{\mathrm{T}}AW$ have the same inertias.*

The inertias of a symmetric matrix A are available from the LDL$^{\mathrm{T}}$ factorization of A. The inertias of A are equal to the inertias of the block diagonal matrix D by the Sylvester law [18, Sec. 11].

For the symmetric semidefinite generalized eigenvalue problem (2.3), the inertias can be used to count the number of eigenvalues in an interval. If some linear combination $\alpha A + \beta B$ is positive definite, the number of eigenvalues in an interval $[\sigma_l, \sigma_r]$ is equal to the difference [16]

$$\nu_-(A - \sigma_r B) - \nu_-(A - \sigma_l B). \tag{2.22}$$

The inertias in (2.22) can be computed by the LDL$^{\mathrm{T}}$ factorizations of the shifted matrices $A - \sigma_l B$ and $A - \sigma_r B$.

2.7 Backward error of computed eigenpairs

Let $(\widehat{\Lambda}, \widehat{X})$ be computed eigenpairs of a symmetric matrix A, and let U be the polar decomposition of $\widehat{X}$. We consider the symmetric backward errors for the SEP (2.1),

$$\mathcal{H}_U = \{\Delta \mid (A+\Delta)U = U\widehat{\Lambda}, \Delta = \Delta^{\mathrm{T}} \in \mathbb{R}^{n\times n}\}.$$

The following theorem by Sun [34] gives an upper bound on the norm $\min_{\Delta\in\mathcal{H}_U} \|\Delta\|_{\mathrm{F}}$.

Theorem 2.2. *The set $\mathcal{H}_U$ is non-empty and there exists a unique $\Delta_U \in \mathcal{H}_U$ such that*

$$\min_{\Delta\in\mathcal{H}_U} \|\Delta\|_{\mathrm{F}} = \|\Delta_U\|_{\mathrm{F}} \leq \sqrt{\|R\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathcal{R}(\widehat{X})}^{\perp} R\|_{\mathrm{F}}^2} \Big/ \sigma_{\min}(\widehat{X}),$$

where $R = A\widehat{X} - \widehat{X}\widehat{\Lambda}$ is the residual and $\mathcal{P}_{\mathcal{R}(\widehat{X})}^{\perp}$ is the orthogonal projection onto the orthogonal complement of $\mathcal{R}(\widehat{X})$.

For the SGEP (2.3), we consider the backward error of the computed eigenpair $(\widehat{\lambda}_i, \widehat{x}_i)$,

$$\eta(\widehat{\lambda}_i, \widehat{x}_i) := \min\left\{\epsilon \mid (A+\Delta A)\widehat{x}_i = \widehat{\lambda}_i(B+\Delta B)\widehat{x}_i,\ \|\Delta A\|_2 \leq \epsilon\|A\|,\ \|\Delta B\|_2 \leq \epsilon\|B\|\right\}, \quad (2.23)$$

where $\|A\|$ and $\|B\|$ are the norms of A and B, respectively. The following theorem, given by Frayssé and Toumazou [14], shows that the backward error $\eta(\widehat{\lambda}_i, \widehat{x}_i)$ can be computed from the residual norm $\|r_i\|_2 = \|A\widehat{x}_i - \widehat{\lambda}_i B\widehat{x}_i\|_2$.

Theorem 2.3. *The backward error defined in (2.23) is given by*

$$\eta(\widehat{\lambda}_i, \widehat{x}_i) = \frac{\|r_i\|_2}{(\|A\| + |\widehat{\lambda}_i| \cdot \|B\|)\|\widehat{x}_i\|_2}, \quad (2.24)$$

where $r_i = A\widehat{x}_i - \widehat{\lambda}_i B\widehat{x}_i$ is the residual vector.

When A and B are symmetric and $\widehat{\lambda}_i$ is real, we may consider the symmetric backward error

$$\eta_{\mathrm{T}}(\widehat{\lambda}_i, \widehat{x}_i) := \min\{\epsilon \mid (A+\Delta A)\widehat{x}_i = \widehat{\lambda}_i(B+\Delta B)\widehat{x}_i,\ \Delta A = \Delta A^{\mathrm{T}},\ \Delta B = \Delta B^{\mathrm{T}},$$
$$\|\Delta A\|_2 \leq \epsilon\|A\|,\ \|\Delta B\|_2 \leq \epsilon\|B\|\}.$$

The following theorem by Higham and Higham [17] analyzes the symmetric backward error $\eta_{\mathrm{T}}(\widehat{\lambda}_i, \widehat{x}_i)$.

Theorem 2.4. *If A and B are symmetric and $\widehat{\lambda}_i$ is real, we have $\eta_{\mathrm{T}}(\widehat{\lambda}_i, \widehat{x}_i) = \eta(\widehat{\lambda}_i, \widehat{x}_i)$.*

Chapter 3

Solving the symmetric eigenvalue problem with the EED

In this chapter, we analyze the EED for the symmetric eigenvalue problem (2.1). Without loss of generality, we consider an interval $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$ containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{n_e}$ at the lower end of the spectrum. We consider computing the partial eigendecomposition

$$AX_{n_e} = X_{n_e}\Lambda_{n_e}, \tag{3.1}$$

where $\Lambda_{n_e} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n_e})$, $X_{n_e} = [x_1, x_2, \ldots, x_{n_e}]$ and $X_{n_e}^{\mathrm{T}} X_{n_e} = I_{n_e}$.

To compute the partial eigendecomposition (3.1) with the EED, we compute the lowest eigenpair (λ, x) of A by an eigensolver such as TRLan [36]. We apply the EED with a shift $\sigma > \lambda_{\text{upper}} - \lambda$ to displace the eigenvalue λ to the higher end of the spectrum. Algorithm 3.1 summarizes the solution procedure. The solution procedure is referred to as the EED procedure.

In the literature, numerical stability of the EED procedure outlined in Algorithm 3.1 is not well understood. In [35, p. 585], Wilkinson comments that EED procedure is numerically unstable. Parlett argued that, when deflating out the largest computed eigenvalue, the change to the smallest eigenvalue in magnitude would be at the same order of the round-off error incurred [26, Sec. 5.1]. In [27], Saad observes the loss of orthogonality of computed eigenvectors and proposes reorthogonalizing computed eigenvectors before applying deflation. Saad performs a backward stability analysis of this variant, and claims that the stability is determined by the angle between the computed eigenvector and the deflated subspace.

In the following, we develop a backward stability analysis of the EED procedure. In

Algorithm 3.1 EED procedure

```
Input: (i) the symmetric matrix $A$. (ii) the interval $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$ at the lower end of the
    spectrum of $A$.
Output: $n_{ev}$ lowest eigenpairs of $A$.
$A_0 = A$;
compute the lowest eigenpair $(\lambda_1, x_1)$ of $A_0$ by EIGSOL;
for $j = 1, 2, \ldots$ do
    pick a shift $\sigma_j > \lambda_{\text{upper}} - \lambda_j$;
    $A_j = A_{j-1} + \sigma_j x_j x_j^{\mathrm{T}} = A + X_j \Sigma_j X_j^{\mathrm{T}}$;
    compute the lowest eigenpair $(\lambda_{j+1}, x_{j+1})$ of $A_j$ by EIGSOL;
    if $\lambda_{j+1} > \lambda_{\text{upper}}$, stop
end for
return the partial eigendecomposition (3.1) of $A$;
```

Section 3.1, we derive a governing equation of the EED procedure in finite precision arithmetic. In Section 3.2, we derive computable upper bounds on the loss of orthogonality of computed eigenvectors and on the symmetric backward error norm of computed eigenpairs. In Section 3.3, we identify the crucial quantities associated with the shifts and derive sufficient conditions for the backward stability of the EED procedure.

3.1 Governing equation of the EED procedure in finite precision arithmetic

We consider the eigensolver EIGSOL in Algorithm 3.1 to be a generic one. It could be TRLan [36] or ARPACK [22]. We only assume that EIGSOL can compute the lowest eigenpair $(\widehat{\lambda}, \widehat{x})$ of A with

$$A\widehat{x} = \widehat{\lambda}\widehat{x} + \eta,$$

where $\|\widehat{x}\|_2 = 1$ and the residual vector η satisfies

$$\|\eta\|_2 \leq tol \cdot \|A\|_2 \tag{3.2}$$

for a prescribed convergence tolerance tol.

The EED procedure starts with computing the lowest eigenpair $(\widehat{\lambda}_1, \widehat{x}_1)$ of A by

EIGSOL satisfying

$$A\widehat{x}_1 = \widehat{\lambda}_1\widehat{x}_1 + \eta_1,$$

where the residual vector η_1 satisfies (3.2). At the first EED step, we choose a shift σ_1 and define

$$\widehat{A}_1 \equiv A + \sigma_1\widehat{x}_1\widehat{x}_1^{\mathrm{T}}.$$

By choosing the shift $\sigma_1 > \lambda_{\text{upper}} - \widehat{\lambda}_1$, the lowest eigenpair of $\widehat{A}_1$ is an approximation of the second eigenpair (λ_2, x_2) of A. Subsequently, we use EIGSOL to compute the lowest eigenpair $(\widehat{\lambda}_2, \widehat{x}_2)$ of $\widehat{A}_1$ satisfying

$$\widehat{A}_1\widehat{x}_2 = \widehat{\lambda}_2\widehat{x}_2 + \eta_2,$$

where the residual vector η_2 satisfies (3.2). Meanwhile, expressing the computed eigenpair $(\widehat{\lambda}_1, \widehat{x}_1)$ in terms of $\widehat{A}_1$, we have

$$\widehat{A}_1\widehat{x}_1 = (\widehat{\lambda}_1 + \sigma_1)\widehat{x}_1 + \eta_1.$$

Proceeding to the second EED step, we choose a shift σ_2 and define

$$\widehat{A}_2 \equiv \widehat{A}_1 + \sigma_2\widehat{x}_2\widehat{x}_2^{\mathrm{T}} = A + \widehat{X}_2\Sigma_2\widehat{X}_2^{\mathrm{T}},$$

where $\widehat{X}_2 = [\widehat{x}_1, \widehat{x}_2]$ and $\Sigma_2 = \operatorname{diag}(\sigma_1, \sigma_2)$. By choosing the shift $\sigma_2 > \lambda_{\text{upper}} - \widehat{\lambda}_2$, the lowest eigenpair of $\widehat{A}_2$ is an approximation of the third eigenpair (λ_3, x_3) of A. Then we use EIGSOL again to compute the lowest eigenpair $(\widehat{\lambda}_3, \widehat{x}_3)$ of $\widehat{A}_2$ satisfying

$$\widehat{A}_2\widehat{x}_3 = \widehat{\lambda}_3\widehat{x}_3 + \eta_3,$$

where the residual vector η_3 satisfies (3.2). Meanwhile, expressing the computed eigenpairs $(\widehat{\lambda}_1, \widehat{x}_1)$ and $(\widehat{\lambda}_2, \widehat{x}_2)$ in terms of $\widehat{A}_2$, we have

$$\widehat{A}_2\widehat{X}_2 = \widehat{X}_2(\widehat{\Lambda}_2 + \Sigma_2) + \widehat{X}_2\Sigma_2\Phi_2 + E_2,$$

where $\widehat{\Lambda}_2 = \operatorname{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2)$, $E_2 = [\eta_1, \eta_2]$, and $\Phi_2 \in \mathbb{R}^{2\times 2}$ is the strictly lower triangular part of the matrix $\widehat{X}_2^{\mathrm{T}}\widehat{X}_2 - I_2$, i.e., $\Phi_2 + \Phi_2^{\mathrm{T}} = \widehat{X}_2^{\mathrm{T}}\widehat{X}_2 - I_2$.

In general, at the j-th EED step, we choose a shift σ_j and define

$$\widehat{A}_j \equiv \widehat{A}_{j-1} + \sigma_j\widehat{x}_j\widehat{x}_j^{\mathrm{T}} = A + \widehat{X}_j\Sigma_j\widehat{X}_j^{\mathrm{T}}, \tag{3.3}$$

where $\widehat{X}_j \equiv [\widehat{x}_1, \dots, \widehat{x}_j]$ and $\Sigma_j = \mathrm{diag}(\sigma_1, \dots, \sigma_j)$ with $\widehat{A}_0 \equiv A$. Then by choosing the shift $\sigma_j > \lambda_{\text{upper}} - \widehat{\lambda}_j$, the lowest eigenpair of $\widehat{A}_j$ is an approximation of the $(j+1)$-th eigenpair (λ_{j+1}, x_{j+1}) of A. We use EIGSOL to compute the lowest eigenpair $(\widehat{\lambda}_{j+1}, \widehat{x}_{j+1})$ of $\widehat{A}_j$ satisfying

$$\widehat{A}_j \widehat{x}_{j+1} = \widehat{\lambda}_{j+1} \widehat{x}_{j+1} + \eta_{j+1}, \tag{3.4}$$

where $\|\widehat{x}_{j+1}\|_2 = 1$ and the residual vector η_{j+1} satisfies (3.2), i.e.,

$$\|\eta_{j+1}\|_2 \leq tol \cdot \|A\|_2. \tag{3.5}$$

Meanwhile, for the computed eigenpairs $(\widehat{\lambda}_j, \widehat{x}_j)$ in terms of $\widehat{A}_j$, we have

$$\widehat{A}_j \widehat{x}_j = \widehat{A}_{j-1} \widehat{x}_j + \sigma_j \widehat{x}_j = (\lambda_j + \sigma_j) \widehat{x}_j + \eta_j, \tag{3.6}$$

and for the computed eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ with $1 \leq i \leq j-1$ in terms of $\widehat{A}_j$, we have

$$\begin{aligned}
\widehat{A}_j \widehat{x}_i &= \left(\widehat{A}_{j-1} + \sigma_j \widehat{x}_j \widehat{x}_j^{\mathrm{T}} \right) \widehat{x}_i \\
&= \left(\widehat{A}_{i-1} + \sigma_i \widehat{x}_i \widehat{x}_i^{\mathrm{T}} + \widehat{X}_{i+1:j} \Sigma_{i+1:j} \widehat{X}_{i+1:j}^{\mathrm{T}} \right) \widehat{x}_i \\
&= \left(\widehat{A}_{i-1} + \sigma_i \widehat{x}_i \widehat{x}_i^{\mathrm{T}} \right) \widehat{x}_i + \widehat{X}_{i+1:j} \Sigma_{i+1:j} \widehat{X}_{i+1:j}^{\mathrm{T}} \widehat{x}_i \\
&= \widehat{\lambda}_i \widehat{x}_i + \eta_i + \sigma_i \widehat{x}_i + \widehat{X}_{i+1:j} \Sigma_{i+1:j} \widehat{X}_{i+1:j}^{\mathrm{T}} \widehat{x}_i \\
&= (\widehat{\lambda}_i + \sigma_i) \widehat{x}_i + \widehat{X}_{i+1:j} \Sigma_{i+1:j} \widehat{X}_{i+1:j}^{\mathrm{T}} \widehat{x}_i + \eta_i \\
&= (\widehat{\lambda}_i + \sigma_i) \widehat{x}_i + \widehat{X}_j \Sigma_j \begin{bmatrix} 0 \\ \widehat{X}_{i+1:j}^{\mathrm{T}} \widehat{x}_i \end{bmatrix} + \eta_i,
\end{aligned} \tag{3.7}$$

where $\widehat{X}_{i+1:j} \equiv [\widehat{x}_{i+1}, \dots, \widehat{x}_j]$ and $\Sigma_{i+1:j} \equiv \mathrm{diag}(\sigma_{i+1}, \dots, \sigma_j)$.

Combining (3.6) and (3.7), we have

$$\widehat{A}_j \widehat{X}_j = \widehat{X}_j (\widehat{\Lambda}_j + \Sigma_j) + \widehat{X}_j \Sigma_j \Phi_j + E_j, \tag{3.8}$$

where $\widehat{\Lambda}_j = \mathrm{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_j)$, $E_j = [\eta_1, \dots, \eta_j]$, Φ_j is the strictly lower triangular part of the matrix $\widehat{X}_j^{\mathrm{T}} \widehat{X}_j - I_j$ and $\Phi_j + \Phi_j^{\mathrm{T}} = \widehat{X}_j^{\mathrm{T}} \widehat{X}_j - I_j$. Eqs. (3.4) and (3.8) are referred to as *the governing equations of the EED procedure in finite precision arithmetic.*

Now we introduce the following two quantities associated with the shifts $\sigma_1, \dots, \sigma_j$ for a j-step EED:

- the *spectral gap* of $\widehat{A}_j$, defined as the separation between the computed eigenvalues and the shifted ones:

$$\gamma_j \equiv \min_{\lambda \in \mathcal{I}_{j+1}, \theta \in \mathcal{J}_j} |\lambda - \theta| > 0, \tag{3.9}$$

where $\mathcal{I}_{j+1} \equiv \{\widehat{\lambda}_1, \dots, \widehat{\lambda}_j, \widehat{\lambda}_{j+1}\}$, the set of computed eigenvalues, and $\mathcal{J}_j \equiv \{\widehat{\lambda}_1 + \sigma_1, \dots, \widehat{\lambda}_j + \sigma_j\}$, the set of computed eigenvalues after shifting (see Figure 3.1 for an illustration);

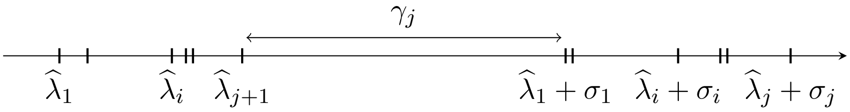

γ_j
$\widehat{\lambda}_1$
$\widehat{\lambda}_i$
$\widehat{\lambda}_{j+1}$
$\widehat{\lambda}_1 + \sigma_1$
$\widehat{\lambda}_i + \sigma_i$
$\widehat{\lambda}_j + \sigma_j$

Figure 3.1: Illustration of the spectral gap γ_j

- the *shift-gap ratio*, defined as the ratio of the largest shift to the spectral gap γ_j:

$$\tau_j \equiv \frac{1}{\gamma_j} \cdot \max_{1 \le i \le j} |\sigma_i|. \tag{3.10}$$

We will see that γ_j and τ_j are crucial quantities to characterize the backward stability of the EED procedure.

3.2 Backward stability analysis

In this section, we derive upper bounds on the following two quantities measuring the accuracy of the computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$:

- the loss of orthogonality of the computed eigenvectors $\widehat{X}_{j+1}$,

$$\omega_{j+1} \equiv \|\widehat{X}_{j+1}^{\mathrm{T}} \widehat{X}_{j+1} - I_{j+1}\|_{\mathrm{F}}, \tag{3.11}$$

- the symmetric backward error norm of the computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$,

$$\delta_{j+1} \equiv \min_{\Delta \in \mathcal{H}_{U_{j+1}}} \|\Delta\|_{\mathrm{F}}, \tag{3.12}$$

where $\mathcal{H}_{U_{j+1}}$ is the set of the symmetric backward errors for the orthonormal basis U_{j+1} from the polar decomposition of $\widehat{X}_{j+1}$, namely,

$$\mathcal{H}_{U_{j+1}} \equiv \left\{ \Delta \mid (A + \Delta) U_{j+1} = U_{j+1} \widehat{\Lambda}_{j+1},\ \Delta = \Delta^{\mathrm{T}} \in \mathbb{R}^{n \times n} \right\}. \tag{3.13}$$

For a prescribed tolerance *tol* of the stopping criterion (3.5) for an eigensolver EIGSOL, the EED procedure is considered to be *backward stable* if

$$\omega_{j+1} = O(tol) \tag{3.14}$$

and

$$\delta_{j+1} = O(tol \cdot \|A\|_2), \tag{3.15}$$

where the constants in the big-O notations are low-degree polynomials in the number j of the EED steps.

3.2.1 Loss of orthogonality

We first prove the following lemma to reveal the structure of the orthogonality between the computed eigenvectors $\widehat{X}_{j+1}$.

Lemma 3.1. *By the governing equations* (3.4) *and* (3.8) *of j steps of EED, if $\tau_j \omega_j < \sqrt{2}$, then for $i = 1, 2, \ldots, j$, the matrices $\Gamma_i \equiv \widehat{\Lambda}_i + \Sigma_i - \widehat{\lambda}_{i+1} I_i$ and $I_i + \Phi_i^{\mathrm{T}} \Sigma_i \Gamma_i^{-1}$ are non-singular, and*

$$\widehat{X}_i^{\mathrm{T}} \widehat{x}_{i+1} = \Gamma_i^{-1} \left(I_i + \Phi_i^{\mathrm{T}} \Sigma_i \Gamma_i^{-1} \right)^{-1} \left[\widehat{X}_i^{\mathrm{T}} \eta_{i+1} - E_i^{\mathrm{T}} \widehat{x}_{i+1} \right]. \tag{3.16}$$

Furthermore,

(i) $\|\Gamma_i^{-1}\|_2 \leq \gamma_j^{-1}$,

(ii) $\|(I_i + \Phi_i^{\mathrm{T}} \Sigma_i \Gamma_i^{-1})^{-1}\|_2 \leq (1 - \tau_j \omega_j / \sqrt{2})^{-1}$,

where γ_j and τ_j are the spectral gap and the shift-gap ratio defined in (3.9) *and* (3.10)*, respectively, and ω_j is the loss of the orthogonality defined in* (3.11)*.*

Proof. By the governing equations (3.4) and (3.8) of j steps of EED, for $1 \leq i \leq j$, we have

$$\widehat{X}_i^{\mathrm{T}} \widehat{A}_i \widehat{x}_{i+1} = \widehat{\lambda}_{i+1} \widehat{X}_i^{\mathrm{T}} \widehat{x}_{i+1} + \widehat{X}_i^{\mathrm{T}} \eta_{i+1}$$

and

$$\widehat{X}_i^{\mathrm{T}} \widehat{A}_i \widehat{x}_{i+1} = (\widehat{\Lambda}_i + \Sigma_i) \widehat{X}_i^{\mathrm{T}} \widehat{x}_{i+1} + \Phi_i^{\mathrm{T}} \Sigma_i \widehat{X}_i^{\mathrm{T}} \widehat{x}_{i+1} + E_i^{\mathrm{T}} \widehat{x}_{i+1}.$$

Consequently,

$$(\Gamma_i + \Phi_i^{\mathrm{T}} \Sigma_i) \widehat{X}_i^{\mathrm{T}} \widehat{x}_{i+1} = \widehat{X}_i^{\mathrm{T}} \eta_{i+1} - E_i^{\mathrm{T}} \widehat{x}_{i+1}, \tag{3.17}$$

where $\Gamma_i = \widehat{\Lambda}_i + \Sigma_i - \widehat{\lambda}_{i+1} I_i$ is a diagonal matrix.

By the definition (3.9) of the spectral gap γ_j, the minimal singular value

$$\sigma_{\min}(\Gamma_i) = \min_{1 \leq k \leq i} |\widehat{\lambda}_k + \sigma_k - \widehat{\lambda}_{i+1}| \geq \gamma_j > 0.$$

Hence the matrix Γ_i is non-singular and the bound (i) holds.

Since Φ_i is the strictly lower triangular part of the matrix $\widehat{X}_i^{\mathrm{T}}\widehat{X}_i - I_i$,

$$\|\Phi_i^{\mathrm{T}}\|_2 \le \|\Phi_i^{\mathrm{T}}\|_{\mathrm{F}} = \frac{\omega_i}{\sqrt{2}} \le \frac{\omega_j}{\sqrt{2}}.$$

Consequently,

$$\|\Phi_i^{\mathrm{T}}\Sigma_i\Gamma_i^{-1}\|_2 \le \|\Phi_i^{\mathrm{T}}\|_2\|\Gamma_i^{-1}\|_2\|\Sigma_i\|_2 \le \frac{\omega_j}{\sqrt{2}} \cdot \gamma_j^{-1} \cdot \|\Sigma_j\|_2 = \frac{\omega_j}{\sqrt{2}} \cdot \tau_j < 1, \tag{3.18}$$

where for the last inequality, we use the assumption $\tau_j\omega_j < \sqrt{2}$. By (3.18), the matrix $I_i + \Phi_i^{\mathrm{T}}\Sigma_i\Gamma_i^{-1}$ is non-singular and the bound (ii) holds due to $\|(I+G)^{-1}\|_2 \le (1-\|G\|_2)^{-1}$ for any G with $\|G\|_2 < 1$.

Since both matrices Γ_i and $I_i + \Phi_i^{\mathrm{T}}\Sigma_i\Gamma_i^{-1}$ are invertible, the identity (3.16) follows from (3.17). □

Next we exploit the structure of the product $\widehat{X}_i^{\mathrm{T}}\widehat{x}_{i+1}$ to derive a computable upper bound on the loss of orthogonality ω_{j+1} of the computed eigenvectors $\widehat{X}_{j+1}$.

Theorem 3.1. *By the governing equations (3.4) and (3.8) of j steps of EED, if $\tau_j\omega_j < \sqrt{2}$, then the loss of orthogonality ω_{j+1} of the computed eigenvectors $\widehat{X}_{j+1}$ defined in (3.11) satisfies*

$$\omega_{j+1} \le 2\,\frac{c_j}{\gamma_j}\left(1 + 2\frac{c_j}{\gamma_j}\|E_{j+1}\|_{\mathrm{F}}\right)\|E_{j+1}\|_{\mathrm{F}}, \tag{3.19}$$

where $c_j = (1-\tau_j\omega_j/\sqrt{2})^{-1}$, and γ_j and τ_j are the spectral gap and the shift-gap ratio defined in (3.9) and (3.10), respectively.

Proof. By the definition (3.11), we have

$$\omega_{j+1}^2 = 2 \cdot \left\|\Phi_{j+1}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 = 2 \cdot \sum_{i=1}^{j} \|\widehat{X}_i^{\mathrm{T}}\widehat{x}_{i+1}\|_2^2. \tag{3.20}$$

Recall Lemma 3.1 that, for any $1 \le i \le j$,

$$\|\widehat{X}_i^{\mathrm{T}}\widehat{x}_{i+1}\|_2 \le \frac{c_j}{\gamma_j} \cdot \|\widehat{X}_i^{\mathrm{T}}\eta_{i+1} - E_i^{\mathrm{T}}\widehat{x}_{i+1}\|_2.$$

Hence we can derive from (3.20) that

$$\begin{aligned}\omega_{j+1}^2 &\le \frac{2c_j^2}{\gamma_j^2} \cdot \sum_{i=1}^{j} \|\widehat{X}_i^{\mathrm{T}}\eta_{i+1} - E_i^{\mathrm{T}}\widehat{x}_{i+1}\|_2^2 \\ &= \frac{2c_j^2}{\gamma_j^2} \cdot \frac{1}{2}\|\widehat{X}_{j+1}^{\mathrm{T}}E_{j+1} - E_{j+1}^{\mathrm{T}}\widehat{X}_{j+1}\|_{\mathrm{F}}^2 \\ &\le \frac{2c_j^2}{\gamma_j^2} \cdot 2\|\widehat{X}_{j+1}^{\mathrm{T}}E_{j+1}\|_{\mathrm{F}}^2 \;\le\; \frac{4c_j^2}{\gamma_j^2} \cdot \|\widehat{X}_{j+1}^{\mathrm{T}}\|_2^2\|E_{j+1}\|_{\mathrm{F}}^2.\end{aligned}$$

Since

$$\|\widehat{X}_{j+1}^{\mathrm{T}}\|_2^2 = \|\widehat{X}_{j+1}^{\mathrm{T}}\widehat{X}_{j+1}\|_2 \le \|I_{j+1}\|_2 + \|\widehat{X}_{j+1}^{\mathrm{T}}\widehat{X}_{j+1} - I_{j+1}\|_2 \le 1 + \omega_{j+1},$$

we arrive at

$$\omega_{j+1}^2 \le \frac{4c_j^2}{\gamma_j^2} \cdot (1+\omega_{j+1}) \cdot \|E_{j+1}\|_{\mathrm{F}}^2. \tag{3.21}$$

Letting $t = \omega_{j+1}/\chi_{j+1}$, where $\chi_{j+1} = 2c_j\|E_{j+1}\|_{\mathrm{F}}/\gamma_j$, then the inequality (3.21) is recast as

$$t^2 - \chi_{j+1}t - 1 \le 0. \tag{3.22}$$

By the fact that the quadratic polynomial in (3.22) is concave, we conclude that

$$t \le \frac{1}{2} \cdot \left(\chi_{j+1} + \sqrt{4 + \chi_{j+1}^2}\right) \le \frac{1}{2} \cdot (\chi_{j+1} + 2 + \chi_{j+1}) \le 1 + \chi_{j+1}.$$

This proves the upper bound in (3.19). □

3.2.2 Symmetric backward error norm

We derive a computable upper bound on the symmetric backward error norm δ_{j+1} of computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$ of A defined in (3.12). First, the following lemma gives an upper bound on the norm of the residue for $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$:

$$R_{j+1} \equiv A\widehat{X}_{j+1} - \widehat{X}_{j+1}\widehat{\Lambda}_{j+1}. \tag{3.23}$$

Lemma 3.2. *By the governing equations (3.4) and (3.8) of j steps of EED, if $\tau_j\omega_j < \sqrt{2}$, then for the computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$ of A, the Frobenius norm of the residual R_{j+1} defined in (3.23) satisfies*

$$\|R_{j+1}\|_{\mathrm{F}} \le \left(1 + \sqrt{2}c_j\tau_j(1+\omega_{j+1})\right)\|E_{j+1}\|_{\mathrm{F}}, \tag{3.24}$$

where $c_j = (1 - \tau_j\omega_j/\sqrt{2})^{-1}$, and γ_j and τ_j are the spectral gap and the shift-gap ratio defined in (3.9) and (3.10), respectively.

Proof. From the governing equation (3.8) of the EED procedure after $j+1$ steps, we have

$$\widehat{A}_{j+1}\widehat{X}_{j+1} = \widehat{X}_{j+1}(\widehat{\Lambda}_{j+1} + \Sigma_{j+1}) + \widehat{X}_{j+1}\Sigma_{j+1}\Phi_{j+1} + E_{j+1}. \tag{3.25}$$

On the other hand, by the definition (3.3) of $\widehat{A}_{j+1}$, we have

$$\begin{aligned}\widehat{A}_{j+1}\widehat{X}_{j+1} &= A\widehat{X}_{j+1} + \widehat{X}_{j+1}\Sigma_{j+1}\widehat{X}_{j+1}^{\mathrm{T}}\widehat{X}_{j+1} \\ &= A\widehat{X}_{j+1} + \widehat{X}_{j+1}\Sigma_{j+1}(\Phi_{j+1} + I_{j+1} + \Phi_{j+1}^{\mathrm{T}}).\end{aligned} \tag{3.26}$$

Combining (3.25) and (3.26), we obtain the residual

$$R_{j+1} = A\widehat{X}_{j+1} - \widehat{X}_{j+1}\widehat{\Lambda}_{j+1} = E_{j+1} - \widehat{X}_{j+1}\Sigma_{j+1}\Phi_{j+1}^{\mathrm{T}}. \tag{3.27}$$

Consequently, the norm of the residual R_{j+1} is bounded by

$$\|R_{j+1}\|_{\mathrm{F}} \leq \|E_{j+1}\|_{\mathrm{F}} + \|\widehat{X}_{j+1}\|_2\|\Sigma_{j+1}\Phi_{j+1}^{\mathrm{T}}\|_{\mathrm{F}}. \tag{3.28}$$

Note that $\Sigma_{j+1} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{j+1})$ and Φ_{j+1}^{T} is the strictly upper triangular part of the matrix $\widehat{X}_{j+1}^{\mathrm{T}}\widehat{X}_{j+1} - I_{j+1}$, and we have

$$\begin{aligned}\|\widehat{X}_{j+1}\|_2\|\Sigma_{j+1}\Phi_{j+1}^{\mathrm{T}}\|_{\mathrm{F}} &\leq \|\widehat{X}_{j+1}\|_2\|\Sigma_j\|_2\|\Phi_{j+1}^{\mathrm{T}}\|_{\mathrm{F}} \\ &\leq \|\widehat{X}_{j+1}\|_2\|\Sigma_j\|_2 \cdot \frac{1}{\sqrt{2}}\omega_{j+1} \\ &\leq \frac{1}{\sqrt{2}}\|\Sigma_j\|_2\sqrt{(1+\omega_{j+1})\omega_{j+1}^2},\end{aligned} \tag{3.29}$$

where, for the third inequality, we again use the fact that $\|\widehat{X}_{j+1}\|_2 \leq \sqrt{1+\omega_{j+1}}$ by the definition of the loss of orthogonality ω_{j+1}. Left-multiplying (3.21) by $1+\omega_{j+1}$, we know that

$$(1+\omega_{j+1})\omega_{j+1}^2 \leq \frac{4c_j^2}{\gamma_j^2} \cdot (1+\omega_{j+1})^2 \cdot \|E_{j+1}\|_{\mathrm{F}}^2.$$

Plugging into (3.29), we obtain

$$\begin{aligned}\|\widehat{X}_{j+1}\|_2\|\Sigma_{j+1}\Phi_{j+1}^{\mathrm{T}}\|_{\mathrm{F}} &\leq \sqrt{2}\|\Sigma_j\|_2 c_j\gamma_j^{-1} \cdot (1+\omega_{j+1}) \cdot \|E_{j+1}\|_{\mathrm{F}} \\ &= \sqrt{2}c_j\tau_j \cdot (1+\omega_{j+1}) \cdot \|E_{j+1}\|_{\mathrm{F}}.\end{aligned}$$

Combine with (3.28) and we arrive at the upper bound (3.24) of $\|R_{j+1}\|_{\mathrm{F}}$. □

From Lemma 3.2 and Theorem 2.2, we have the following computable upper bound on the symmetric backward error norm δ_{j+1} of the computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$ of A.

Theorem 3.2. *By the governing equations (3.4) and (3.8) of j steps of EED, if $\tau_j\omega_j < \sqrt{2}$ and $\omega_{j+1} < 1$, then the symmetric backward error norm δ_{j+1} of the computed eigenpairs $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$ of A defined in (3.12) has the following upper bound*

$$\delta_{j+1} \leq \sqrt{2}\left(\frac{1+c_j\tau_j(1+\omega_{j+1})}{\sqrt{1-\omega_{j+1}}}\right)\|E_{j+1}\|_{\mathrm{F}}, \tag{3.30}$$

where $c_j = (1-\tau_j\omega_j/\sqrt{2})^{-1}$, and γ_j and τ_j are the spectral gap and the shift-gap ratio defined in (3.9) and (3.10), respectively.

Proof. For the computed eigenvectors $\widehat{X}_{j+1}$ of A, let U_{j+1} be the orthonormal basis from the polar decomposition of $\widehat{X}_{j+1}$, and the set $\mathcal{H}_{U_{j+1}}$ be defined in (3.13). It follows from the definition (3.12) and Theorem 2.2 that

$$\delta_{j+1} = \min_{\Delta\in\mathcal{H}_{U_{j+1}}} \|\Delta\|_{\mathrm{F}} \le \frac{1}{\sigma_{\min}(\widehat{X}_{j+1})}\sqrt{\|R_{j+1}\|_{\mathrm{F}}^2 + \|\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}R_{j+1}\|_{\mathrm{F}}^2}, \tag{3.31}$$

where R_{j+1} is the residual of $(\widehat{\Lambda}_{j+1}, \widehat{X}_{j+1})$ defined in (3.23), and $\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}$ is the orthogonal projection onto the orthogonal complement of the subspace $\mathcal{R}(\widehat{X}_{j+1})$, i.e.,

$$\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})} = I - \widehat{X}_{j+1}(\widehat{X}_{j+1}^T\widehat{X}_{j+1})^{-1}\widehat{X}_{j+1}^T.$$

By the equation (3.27), $\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}R_{j+1} = \mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}E_{j+1}$. Hence we have

$$\|\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}R_{j+1}\|_{\mathrm{F}} = \|\mathcal{P}^{\perp}_{\mathcal{R}(\widehat{X}_{j+1})}E_{j+1}\|_{\mathrm{F}} \le \|E_{j+1}\|_{\mathrm{F}}. \tag{3.32}$$

On the other hand, by the definition (3.11) of ω_{j+1} and the assumption $\omega_{j+1} < 1$, we have

$$|\sigma_{\min}^2(\widehat{X}_{j+1}) - 1| \le \|\widehat{X}_{j+1}^{\mathrm{T}}\widehat{X}_{j+1} - I_{j+1}\|_2 \le \omega_{j+1} < 1,$$

which implies the following lower bound of the singular value

$$\sigma_{\min}(\widehat{X}_{j+1}) \ge \sqrt{1-\omega_{j+1}}. \tag{3.33}$$

Plug (3.32) and (3.33) into (3.31) and recall the upper bound of $\|R_{j+1}\|_{\mathrm{F}}$ in Lemma 3.2, and then we obtain

$$\begin{aligned}\delta_{j+1} &\le \sqrt{1 + \left(1 + \sqrt{2}c_j\tau_j\cdot(1+\omega_{j+1})\right)^2}\cdot\frac{\|E_{j+1}\|_{\mathrm{F}}}{\sqrt{1-\omega_{j+1}}}\\ &\le \sqrt{2}\left(\frac{1+c_j\tau_j(1+\omega_{j+1})}{\sqrt{1-\omega_{j+1}}}\right)\|E_{j+1}\|_{\mathrm{F}},\end{aligned}$$

where the second inequality is due to $1+(1+\sqrt{2}a)^2 \le 2(1+2a+a^2) = 2(1+a)^2$. This completes the proof. □

3.3 Conditions for the backward stability

In Theorems 3.1 and 3.2, the upper bounds (3.19) and (3.30) for ω_{j+1} and δ_{j+1} involve the quantity ω_j from the previous EED step. In this section, under a mild assumption, we derive explicit upper bounds for ω_{j+1} and δ_{j+1}, and then reveal conditions for the backward stability of the EED procedure.

Lemma 3.3. *Consider j steps of EED governed by Eqs.* (3.4) *and* (3.8)*. Assume*

$$\tau_j \frac{\|A\|_2}{\gamma_j} \cdot 4\sqrt{j+1} \cdot tol < 0.1. \tag{3.34}$$

Then

(i) it holds that

$$\tau_i\omega_i < 0.11 \quad \text{and} \quad c_i = (1 - \tau_i\omega_i/\sqrt{2})^{-1} < 2 \quad \text{for } i = 1, 2, \dots, j; \tag{3.35}$$

(ii) the loss of orthogonality ω_{j+1} is bounded by

$$\omega_{j+1} \leq \left(\frac{\|A\|_2}{\gamma_j} \cdot 5\sqrt{j+1}\right) \cdot tol; \tag{3.36}$$

(iii) the backward error norm δ_{j+1} is bounded by

$$\delta_{j+1} \leq \left(\tau_j \cdot 5\sqrt{j+1}\right) \cdot tol \cdot \|A\|_2. \tag{3.37}$$

Proof. First observe that by the definitions (3.9) and (3.10), γ_i is monotonically decreasing with the index i and $\tau_i \geq 1$ is monotonically increasing with i. Therefore, the assumption (3.34) implies the inequalities

$$\frac{\|A\|_2}{\gamma_i} \cdot 4\sqrt{i+1} \cdot tol < 0.1 \quad \text{and} \quad \tau_i \frac{\|A\|_2}{\gamma_{i-1}} \cdot 4\sqrt{i} \cdot tol < 0.1 \quad \text{for all } i \leq j. \tag{3.38}$$

Since the stopping criterion (3.5) of EIGSOL implies

$$\|E_i\|_{\mathrm{F}} = \|[\eta_1, \dots, \eta_i]\|_{\mathrm{F}} \leq \sqrt{i} \cdot tol \cdot \|A\|_2, \tag{3.39}$$

inequalities (3.38) leads to

$$\frac{4}{\gamma_i} \cdot \|E_{i+1}\|_{\mathrm{F}} < 0.1 \quad \text{and} \quad \tau_i \cdot \frac{4}{\gamma_{i-1}} \cdot \|E_i\|_{\mathrm{F}} < 0.1 \quad \text{for all } i \leq j. \tag{3.40}$$

(i) We prove the inequality (3.35) by induction. To begin with, recall that $\|\widehat{x}_1\|_2 = 1$, which implies $\omega_1 = \|\widehat{x}_1^T\widehat{x}_1 - 1\|_F = 0$, $\tau_1\omega_1 = 0 < 0.11$, and $c_1 = 1 < 2$. Hence (3.35) holds for $i = 1$. Now, for $2 \leq i \leq j$, assume that $\tau_{i-1}\omega_{i-1} < 0.11$ and $c_{i-1} < 2$. Since $\tau_{i-1}\omega_{i-1} < 0.11$, we can apply Theorem 3.1 and derive from (3.19) that

$$\tau_i\omega_i \leq \tau_i \cdot \frac{2c_{i-1}}{\gamma_{i-1}}\|E_i\|_{\mathrm{F}} \cdot \left(1 + \frac{2c_{i-1}}{\gamma_{i-1}}\|E_i\|_{\mathrm{F}}\right) < 0.1 \cdot (1 + 0.1) = 0.11, \tag{3.41}$$

where the last inequality of (3.41) is by $2c_{i-1} < 4$ and (3.40). This implies immediately

$$c_i = (1 - \tau_i\omega_i/\sqrt{2})^{-1} \leq (1 - 0.11/\sqrt{2})^{-1} < 2.$$

Therefore, (3.35) follows by induction.

(ii) Since we have $\tau_j\omega_j < 0.11$ and $c_j < 2$ by (3.35), we can apply Theorem 3.1 and derive from (3.19) that

$$\omega_{j+1} \leq \frac{2c_j}{\gamma_j} \cdot \|E_{j+1}\|_{\mathrm{F}} \cdot \left(1 + \frac{2c_j}{\gamma_j}\|E_{j+1}\|_{\mathrm{F}}\right) \leq \frac{4}{\gamma_j} \cdot \|E_{j+1}\|_{\mathrm{F}} \cdot (1 + 0.1)\,, \tag{3.42}$$

where in the second inequality we used $2c_j < 4$ and the first inequality in (3.40). Recall the error bound of $\|E_{j+1}\|_{\mathrm{F}}$ from (3.39) and we obtain (3.36).

(iii) We have $\tau_j\omega_j < 0.11$ and $c_j < 2$ by (3.35). It also follows from (3.42) and (3.40) that $\omega_{j+1} < 0.11$. Therefore, we can apply Theorem 3.2 and derive from (3.30) that

$$\delta_{j+1} \leq \sqrt{2}\left(\frac{1 + c_j\tau_j(1+\omega_{j+1})}{\sqrt{1-\omega_{j+1}}}\right)\|E_{j+1}\|_{\mathrm{F}} \leq \sqrt{2}\left(\frac{1 + 2\tau_j(1+0.11)}{\sqrt{1-0.11}}\right) \cdot \|E_{j+1}\|_{\mathrm{F}},$$

where in second inequality we used $0 \leq \omega_{j+1} < 0.11$. Since $\tau_j \geq 1$ by definition (3.10), we can relax the leading constant as $\sqrt{2}(1.06 + 2.36 \cdot \tau_j) \leq \sqrt{2}(3.42 \cdot \tau_j) < 5\tau_j$. Recall the error bound of $\|E_{j+1}\|_{\mathrm{F}}$ from (3.39) and we prove (3.37). □

By the error bounds (3.36) and (3.37) in Lemma 3.3, we can see that the quantities $\gamma_j^{-1}\|A\|_2$ and τ_j play important roles for the stability of the EED procedure. A sufficient condition to achieve the backward stability (3.14) and (3.15) is given by $\gamma_j^{-1}\|A\|_2 = O(1)$ and $\tau_j = O(1)$. In summary, we have the following theorem for the backward stability of the EED procedure.

Theorem 3.3. *Under the assumptions of the residual norm $\|\eta_i\|_2$ of EIGSOL satisfying* (3.2) *and the inequality* (3.34)*, the backward stability of the EED procedure, in the sense of* (3.14) *and* (3.15)*, is guaranteed if the shifts $\sigma_1, \ldots, \sigma_j$ are dynamically chosen such that*

$$\gamma_j^{-1}\|A\|_2 = O(1) \quad \textit{and} \quad \tau_j = O(1). \tag{3.43}$$

We note that when the shifts σ_j are dynamically chosen such that the conditions (3.43) are satisfied, the assumption of the inequality (3.34) is indeed mild.

Remark 3.1. From the upper bound (3.36) of the loss of orthogonality ω_{j+1}, we see that if the spectral gap γ_j is too small, i.e., $\gamma_j \ll \|A\|_2$, then ω_{j+1} could be amplified by a factor of $\gamma_j^{-1}\|A\|_2$. On the other hand, from the upper bound (3.37) of the symmetric backward error norm δ_{j+1}, we see that when τ_j is too large, i.e., $\tau_j \gg 1$, δ_{j+1} could be amplified by a factor of τ_j. We will demonstrate these observations in the numerical experiments in next chapter.

Chapter 4

Stabilizing the EED procedure

In this chapter, we propose a shift selection scheme for stabilizing the EED procedure discussed in Chapter 3 and present numerical results.

4.1 Shift selection scheme

We consider the following choice of the shift at the j-th EED step,

$$\sigma_j = \mu - \widehat{\lambda}_j, \tag{4.1}$$

where $\mu \in \mathbb{R}$ is a parameter with $\mu > \lambda_{\text{upper}}$. Recall that $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$ is an interval at the lower end of the spectrum. The shift selection scheme (4.1) has been used in several previous works, although without elaboration on the choice of the parameter μ [24, 37] [26, Sec. 5.1]. We discuss how to choose the parameter μ such that the sufficient conditions (3.43) for the backward stability of the EED procedure can hold.

The shift selection scheme (4.1) implies that the spectral gap γ_j in (3.9) satisfies

$$\gamma_j = \min_{\theta \in \mathcal{I}_{j+1},\ \lambda \in \mathcal{J}_j} |\lambda - \theta| = \min_{1 \le i \le j+1} |\mu - \widehat{\lambda}_i|, \tag{4.2}$$

where $\mathcal{I}_{j+1} = \{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_j, \widehat{\lambda}_{j+1}\}$ and $\mathcal{J}_j = \{\widehat{\lambda}_1 + \sigma_1, \ldots, \widehat{\lambda}_j + \sigma_j\} = \{\mu\}$. On the other hand, it also implies that the shift-gap ratio τ_j in (3.10) satisfies

$$\tau_j = \frac{1}{\gamma_j} \cdot \max_{1 \le i \le j} |\sigma_i| = \frac{\max_{1 \le i \le j} |\mu - \widehat{\lambda}_i|}{\min_{1 \le i \le j+1} |\mu - \widehat{\lambda}_i|}. \tag{4.3}$$

Now recall that $\mu > \lambda_{\text{upper}}$ and the computed eigenvalues $\widehat{\lambda}_i \in [\lambda_{\text{low}}, \lambda_{\text{upper}}]$, for $i = 1, 2, \ldots, j+1$, so we have

$$\mu - \lambda_{\text{upper}} \le \min_{1 \le i \le j+1} |\mu - \widehat{\lambda}_i| \le \max_{1 \le i \le j+1} |\mu - \widehat{\lambda}_i| \le \mu - \lambda_{\text{low}}.$$

Hence, (4.2) and (4.3) lead to

$$\gamma_g \leq \gamma_j \leq \gamma_g \, \tau_g \quad \text{and} \quad \tau_j \leq \tau_g, \tag{4.4}$$

where

$$\gamma_g \equiv \mu - \lambda_{\text{upper}} \quad \text{and} \quad \tau_g \equiv \frac{\mu - \lambda_{\text{low}}}{\mu - \lambda_{\text{upper}}}.$$

Now we focus on the choice of the parameter μ such that the quantities γ_g and τ_g satisfy the conditions (3.43). Let us consider a frequently encountered case in practice where the width of the interval $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$ satisfies

$$\lambda_{\text{upper}} - \lambda_{\text{low}} \leq \frac{1}{2}\|A\|_2.$$

Then by setting

$$\mu = \widehat{\lambda}_1 + \|A\|_2,$$

we have

$$\frac{1}{2}\|A\|_2 \leq \gamma_g = \left(1 - \frac{\lambda_{\text{upper}} - \widehat{\lambda}_1}{\|A\|_2}\right)\|A\|_2 \leq \|A\|_2$$

and

$$\tau_g = 1 + \frac{\lambda_{\text{upper}} - \lambda_{\text{low}}}{\gamma_g} \leq 2.$$

Consequently, by (4.4), the spectral gap γ_j and the shift-gap ratio τ_j satisfy the desired conditions (3.43).

In summary, assuming that $\lambda_{\text{upper}} - \lambda_{\text{low}} \leq \frac{1}{2}\|A\|_2$. we recommend the use of the shift selection scheme at the j-th EED,

$$\sigma_j = \mu - \widehat{\lambda}_j \quad \text{with} \quad \mu = \widehat{\lambda}_1 + \|A\|_2, \tag{4.5}$$

to compute the eigenvalues in the interval $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$.

4.2 Algorithm

We summarize the EED procedure with the shift selection scheme (4.5) in Algorithm 4.1

Algorithm 4.1 EED procedure with the shift selection scheme (4.5)

```
Input: (i) the symmetric matrix $A$. (ii) the interval $\mathcal{I} = [\lambda_{\text{low}}, \lambda_{\text{upper}}]$ at the lower end of the
    spectrum of $A$. (iii) the relative tolerance $tol$ in (3.2) for EIGSOL.
Output: the approximate eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ of $A$ in the interval $\mathcal{I}$.
 1: $\widehat{A}_0 = A$;
 2: use EIGSOL to compute the lowest eigenpair $(\widehat{\lambda}_1, \widehat{x}_1)$ of $\widehat{A}_0$ and an estimate Anorm of $\|A\|_2$;
 3: $\mu = \widehat{\lambda}_1 + \texttt{Anorm}$;
 4: for $j = 1, 2, \ldots$ do
 5:     $\sigma_j = \mu - \widehat{\lambda}_j$;
 6:     $\widehat{A}_j = \widehat{A}_{j-1} + \sigma_j \widehat{x}_j \widehat{x}_j^{\mathrm{T}} = A + \widehat{X}_j \Sigma_j \widehat{X}_j^{\mathrm{T}}$;
 7:     compute the lowest eigenpair $(\widehat{\lambda}_{j+1}, \widehat{x}_{j+1})$ of $\widehat{A}_j$ by EIGSOL;
 8:     check if all the eigenpairs in the interval $\mathcal{I}$ have been computed;
 9: end for
10: return the approximate eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ in the interval $\mathcal{I}$;
```

A few remarks are in order:

- In practice, we never need to form the matrix $\widehat{A}_j$ at step 6 explicitly. We can assume that the only operation that is required by EIGSOL is the matrix-vector product $y := \widehat{A}_j x$.

- At step 7, the computation of the lowest eigenpair $(\widehat{\lambda}_{j+1}, \widehat{x}_{j+1})$ of $\widehat{A}_j$ can be accelerated by warm starting the EIGSOL with the lowest unconverged Ritz vectors of $\widehat{A}_{j-1}$. This is possible for iterative eigensolver such as TRLan [36].

- At step 8, an ideal validation method is to use the inertias of the shifted matrix $A - \lambda_{\text{upper}} I$. However, computation of the inertias could be a prohibitive cost for large matrices. An empirical validation is to monitor the lowest eigenvalue $\widehat{\lambda}_{j+1}$ of $\widehat{A}_j$. All eigenpairs in the interval $\mathcal{I}$ are considered to be found when $\widehat{\lambda}_{j+1}$ is outside the interval $\mathcal{I}$.

4.3 Numerical results

In this section, we first use synthetic examples to verify the sharpness of the upper bounds (3.19) and (3.30) on the loss of orthogonality and the symmetric backward error norm of the EED procedure under the choice (4.5) of the shifts σ_j. We present the cases the shifts σ_j may lead to numerical instability of the EED procedure. Then we demonstrate the numerical stability of the EED procedure for a set of large sparse symmetric matrices arising from applications.

We use TRLan as the eigensolver in Algorithm 4.1. TRLan is a C implementation of the thick-restart Lanczos method with adaptive sizes of the projection subspace [36,38,39]. The convergence criterion of an approximate eigenpair $(\widehat{\lambda}_{j+1}, \widehat{x}_{j+1})$ is the residual norm satisfying

$$\|\eta_{j+1}\|_2 = \|\widehat{A}_j \widehat{x}_{j+1} - \widehat{\lambda}_{j+1} \widehat{x}_{j+1}\|_2 < tol \cdot \texttt{Anorm},$$

where tol is a user-specified tolerance and `Anorm` is a 2-norm estimate of A computed by TRLan. The starting vector is a random vector.

Example 4.1. In this example, we demonstrate the sharpness of the upper bounds (3.19) and (3.30) on the loss of orthogonality and the symmetric backward error norm with the choice (4.5) of the shifts σ_j.

We consider a diagonal matrix A with diagonal elements

$$a_{kk} = \begin{cases} \frac{1}{2} d_k, & \text{if } 1 \leq k \leq n/2, \\ \frac{1}{2}(1 + d_{k-n/2}), & \text{if } n/2 < k \leq n, \end{cases}$$

where $d_k = 10^{-5(1-\frac{k-1}{n/2-1})}$ and the matrix size $n = 500$. The spectrum range of A is $(0, 1]$. The eigenvalues of A are clustered around 0 and 0.5. We are interested in computing the $n_e = 65$ eigenvalues in the interval $\mathcal{I} = [0,\ 10^{-4}]$. The computed 2-norm of A is $\texttt{Anorm} = 1.00$.

To closely observe the convergence, TRLan is modified so that the convergence test is performed at each Lanczos iteration. The maximal dimension m of the projection subspace is set to be 40.

Numerical results of the EED procedure for computing all the n_e eigenvalues in the interval $\mathcal{I}$ are summarized in Table 4.1, where the 4th column is the loss of orthogonality ω_{n_e}, the 5th column is the upper bound (3.19) of ω_{n_e}, the 6th column is the norm of the residue R_{n_e} (3.23), and the 7th column is the upper bound (3.30) of δ_{n_e}.

From Table 4.1, we observe that with the choice (4.5) of the shifts σ_j, $\gamma_g^{-1}\texttt{Anorm} \approx 1$ and $\tau_g \approx 1$. Therefore, the conditions (3.43) of the spectral gap γ_j and the shift-gap ratio τ_j for the backward stability are satisfied. Consequently, the loss of orthogonality of the computed eigenvectors is $\omega_{n_e} = O(tol)$ and the symmetric backward error norm of the computed eigenpairs $(\widehat{\Lambda}_{n_e}, \widehat{X}_{n_e})$ is $\delta_{n_e} = O(tol \cdot \texttt{Anorm})$. In addition, we observe that the upper bounds (3.19) and (3.30) of ω_{n_e} and δ_{n_e} are tight within an order of magnitude.

Table 4.1: Numerical stability of the EED procedure for different tolerances *tol* (Example 4.1). The eigensolver is the TRLan.

tol	μ	γ_g	ω_{n_e}	bound (3.19)	$\|R_{n_e}\|_\mathrm{F}$	bound (3.30)
10^{-6}	1.00	1.00	$2.37 \cdot 10^{-6}$	$1.59 \cdot 10^{-5}$	$7.87 \cdot 10^{-6}$	$2.24 \cdot 10^{-5}$
10^{-8}	1.00	1.00	$1.78 \cdot 10^{-8}$	$1.58 \cdot 10^{-7}$	$7.95 \cdot 10^{-8}$	$2.24 \cdot 10^{-7}$
10^{-10}	1.00	1.00	$1.82 \cdot 10^{-10}$	$1.58 \cdot 10^{-9}$	$7.94 \cdot 10^{-10}$	$2.24 \cdot 10^{-9}$

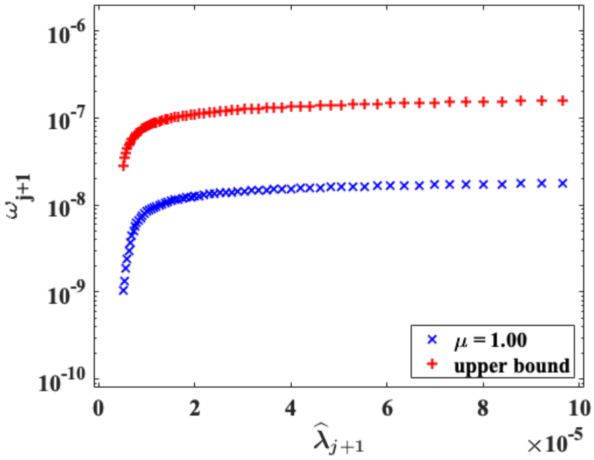


Figure 4.1: The loss of orthogonality ω_{j+1} and the upper bound (3.19) of ω_{j+1} against the computed eigenvalues $\widehat{\lambda}_{j+1}$ for $2 \leq j+1 \leq n_e$, $tol = 10^{-8}$ (Example 4.1).

Example 4.2. In this example, we illustrate that improperly chosen shifts σ_j may lead to instability of the EED procedure.

We consider the same diagonal matrix A as in Example 4.1. The combination of TRLan and EED is used to compute the $n_e = 65$ eigenvalues in the interval $\mathcal{I} = [0,\, 10^{-4}]$. Let us set the shifts $\sigma_j = \mu - \widehat{\lambda}_j$ with $\mu = 2 \cdot 10^{-4}$, which is much smaller than the recommended value of $\mu = \widehat{\lambda}_1 + \|A\|_2 \approx 1.00$. Numerical results are summarized in Table 4.2, where the tolerance $tol = 10^{-8}$ for TRLan. We observe that $\gamma_j = O(\gamma_g) \ll$ `Anorm`, and the loss of orthogonality of the computed eigenvectors is indeed amplified by a factor of $\gamma_g^{-1} \cdot$ `Anorm`. We note that since $\tau_j = O(1)$, the symmetric backward error norms $\delta_{n_e} = O(tol \cdot$ `Anorm`$)$.

Now we flip the sign of the diagonal elements of A defined in Example 4.1, and set $n = 200$. We compute $n_e = 74$ eigenvalues in the interval $\mathcal{I} = [-1.0, -0.5001]$ using the EED procedure. The computed 2-norm of A is `Anorm` $= 1.00$.

Table 4.2: Instability of TRLan with EED when the spectral gaps $\gamma_j = O(\gamma_g)$ are too small.

tol	μ	γ_g	ω_{n_e}	bound (3.19)	$\|R_{n_e}\|_\mathrm{F}$	bound (3.30)
10^{-6}	$2 \cdot 10^{-4}$	10^{-4}	$8.26 \cdot 10^{-3}$	$1.79 \cdot 10^{-1}$	$8.00 \cdot 10^{-6}$	$3.29 \cdot 10^{-5}$
10^{-8}	$2 \cdot 10^{-4}$	10^{-4}	$8.28 \cdot 10^{-5}$	$1.53 \cdot 10^{-3}$	$7.96 \cdot 10^{-8}$	$3.22 \cdot 10^{-7}$
10^{-10}	$2 \cdot 10^{-4}$	10^{-4}	$8.27 \cdot 10^{-7}$	$1.52 \cdot 10^{-5}$	$7.95 \cdot 10^{-10}$	$3.22 \cdot 10^{-9}$

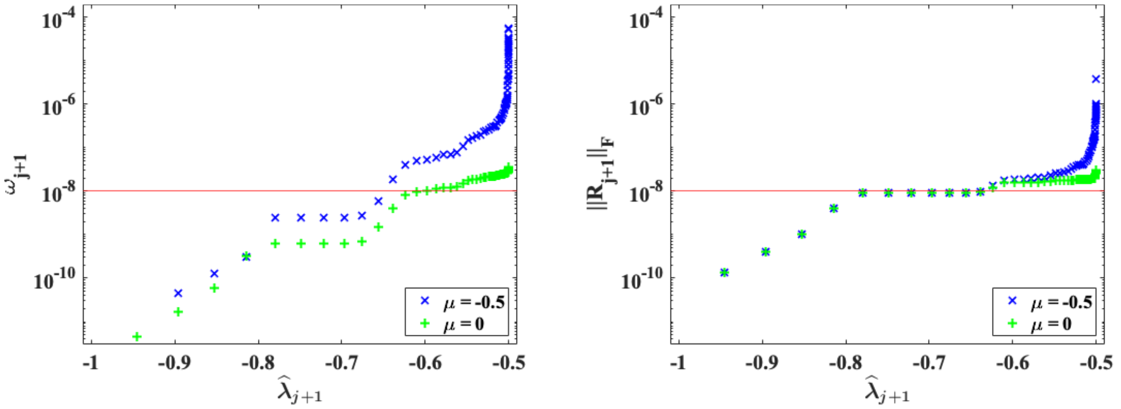


Figure 4.2: The loss of orthogonality ω_{j+1} (left) and the residual norm $\|R_{j+1}\|_{\mathrm{F}}$ (right) against the computed eigenvalues $\widehat{\lambda}_{j+1}$ for $2 \leq j+1 \leq n_e$. The red lines are *tol* (left) and $tol \cdot \texttt{Anorm}$ (right).

Instead of the choice (4.5) for the shifts σ_j, we set $\sigma_j = \mu - \widehat{\lambda}_j$ with $\mu = -0.5$. The blue ×-lines in Figure 4.2 are the loss of orthogonality and the residual norms for the computed eigenpairs $(\widehat{\lambda}_{j+1}, \widehat{x}_{j+1})$ for $2 \leq j+1 \leq n_e$. We observe that for the first 6 computed eigenvalues in the subinterval $[-1.0, -0.75]$ of $\mathcal{I}$, since the spectral gap $\gamma_j \geq 0.25$ and the shift-gap ratio $\tau_j \leq 2$, the computed eigenpairs are backward stable with $\omega_6 = 2.48 \cdot 10^{-9} = O(tol)$ and $\|R_6\|_{\mathrm{F}} = 9.05 \cdot 10^{-9} = O(tol \cdot \texttt{Anorm})$. However, for the computed eigenvalues in the subinterval $[-0.75, -0.5001]$ of $\mathcal{I}$, the computed eigenpairs are not backward stable due to the facts that the spectral gaps γ_j become small, $\gamma_j \approx 1.03 \cdot 10^{-4}$, and the shift-gap ratios τ_j grows up to $\tau_j \approx 4.86 \cdot 10^3$. Consequently, the loss of orthogonality ω_{n_e} and the residual norm $\|R_{n_e}\|_{\mathrm{F}}$ are increased by a factor of up to 10^3, respectively. The stability are restored if the shifts are chosen according to the recommendation (4.5) as shown by the green +-lines in Figure 4.2.

Example 4.3. In this example, we demonstrate the numerical stability of the EED procedure for a set of large sparse symmetric matrices from applications.

The statistics of the matrices are summarized in Table 4.3, where n is the size of the matrix, nnz is the number of nonzero entries of the matrix, $[\lambda_{\min}, \lambda_{\max}]$ is the spectrum range, and n_e is the number of eigenvalues in the interval $\mathcal{I} = [\lambda_{\mathrm{low}}, \lambda_{\mathrm{upper}}]$. The quantities n_e are calculated by computing the inertias of the shifted matrices $A - \lambda_{\mathrm{upper}} I$. `Laplacian` is the negative 2D Laplacian on a 200-by-200 grids with Dirichlet boundary condition [21]. `worms20` is the graph Laplacian worms20_10NN in machine learning datasets [7]. `SiO, Si34H36, Ge87H76` and `Ge99H100` are Hamiltonian matrices from PARSEC collection [7].

We run TRLan with a maximal number m of Lanczos vectors to compute the lowest

Table 4.3: Statistics of the test matrices.

matrix	n	nnz	$[\lambda_{\min}, \lambda_{\max}]$	$[\lambda_{\text{low}}, \lambda_{\text{upper}}]$	n_e
`Laplacian`	$40,000$	$199,200$	$[0, 7.9995]$	$[0, 0.07]$	205
`worms20`	$20,055$	$260,881$	$[0, 6.0450]$	$[0, 0.05]$	289
`SiO`	$33,401$	$1,317,655$	$[-1.6745, 84.3139]$	$[-1.7, 2.0]$	182
`Si34H36`	$97,569$	$5,156,379$	$[-1.1586, 42.9396]$	$[-1.2, 0.4]$	310
`Ge87H76`	$112,985$	$7,892,195$	$[-1.214, 32.764]$	$[-1.3, -0.0053]$	318
`Ge99H100`	$112,985$	$8,451,395$	$[-1.226, 32.703]$	$[-1.3, -0.0096]$	372

eigenpairs of the matrix $\widehat{A}_j$. The convergence test is performed at each restart of TRLan. All the converged eigenvalues in the interval $\mathcal{I}$ are shifted by EED. Meanwhile, we also keep a maximal number m_0 of the lowest unconverged eigenvectors as the starting vectors of TRLan for the matrix $\widehat{A}_{j+1}$. All the eigenvalues in $\mathcal{I}$ are assumed to be computed when the lowest converged eigenvalue is outside the interval $\mathcal{I}$. This combination of TRLan and EED is referred to as TRLED [1].

TRLED is compiled using the `icc` compiler (version 2021.1) with the optimization flag `-O2`, and linked to BLAS and LAPACK available in Intel Math Kernel Library (version 2021.1.1). The experiments are conducted on a MacBook with 1.6 GHz Intel Core i5 CPU and 8GB of RAM.

For numerical experiments, we set the maximal number of Lanczos vectors $m = 150$. When starting TRLED for $\widehat{A}_{j+1}$, the maximal number of the starting vectors is $m_0 = 75$. The convergence tolerance for the residual norm was set to $tol = 10^{-8}$ as a common practice for solving large scale eigenvalue problems with double precision [30].

Numerical results of TRLED are summarized in Table 4.4, where the 2nd column is the number $\widehat{n}_e$ of the computed eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ in the interval $\mathcal{I}$, the 3rd column is the number $j_{\max}$ of steps of EED performed, the 4th column is the loss of orthogonality $\omega_{\widehat{n}_e}$, and the 5th column is the relative residual norm $\|R_{\widehat{n}_e}\|_{\text{F}}/\texttt{Anorm}$ of the computed eigenpairs $(\widehat{\Lambda}_{n_e}, \widehat{X}_{n_e})$. From the quantities n_e in Table 4.3 and $\widehat{n}_e$ in Table 4.4, we see that for all test matrices the eigenvalues in the prescribed intervals $\mathcal{I}$ are successfully computed with the desired backward stability.

The left plot of Figure 4.3 is a profile of the number of converged eigenvalues at each external deflation of a total of 74 EEDs for the matrix `Ge99H100`. The right plot of Figure 4.3 shows the relative residual norms of all 372 computed eigenpairs in the interval. We observe

[1] https://github.com/cplin722/trleed

Table 4.4: Numerical results of TRLED.

matrix	$\widehat{n}_e$	$j_{\max}$	$\omega_{\widehat{n}_e}$	$\|R_{\widehat{n}_e}\|_{\mathrm{F}}/\texttt{Anorm}$	CPU time (sec.)	
					TRLED	TRLan
`Laplacian`	205	60	$1.93 \cdot 10^{-8}$	$6.33 \cdot 10^{-8}$	66.5	86.0
`worms20`	289	86	$2.63 \cdot 10^{-8}$	$7.24 \cdot 10^{-8}$	57.3	74.8
`SiO`	182	41	$2.33 \cdot 10^{-8}$	$4.71 \cdot 10^{-8}$	42.4	47.1
`Si34H36`	310	72	$3.41 \cdot 10^{-8}$	$7.50 \cdot 10^{-8}$	309.9	310.4
`Ge87H76`	318	66	$4.08 \cdot 10^{-8}$	$8.50 \cdot 10^{-8}$	388.7	421.0
`Ge99H100`	372	74	$3.65 \cdot 10^{-8}$	$7.63 \cdot 10^{-8}$	501.1	533.4

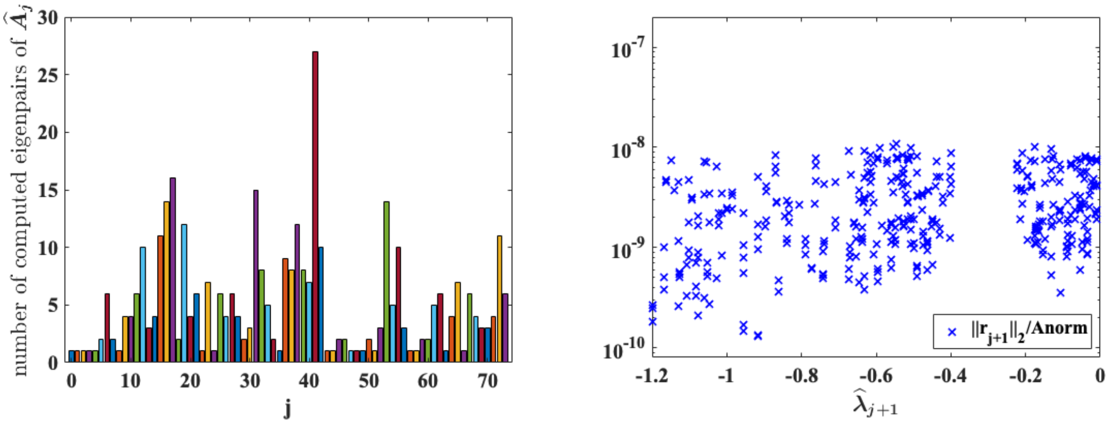


Figure 4.3: The number of deflated eigenpairs at each EED for the matrix `Ge99H100` (left). The relative residual norms of 372 computed eigenpairs (right).

that a large number of converged eigenvalues are deflated and shifted away at some EED steps.

To examine whether the multiple explicit external deflations lead to a significant increase in execution time, in the 6th and 7th columns of Table 4.4, we record the CPU time of TRLED and TRLan for computing all eigenvalues in the same intervals. For TRLan, we set the maximal number of Lanczos vectors to $n_e + 150$. The restart scheme with `restart=1` is used. TRLan is compiled and executed under the same setting as TRLED. We observe comparable execution time of TRLED and TRLan.

Chapter 5

Shift-invert Lanczos method for the buckling eigenvalue problem

In this chapter, we consider the buckling eigenvalue problem

$$Kx = \lambda K_G x, \tag{5.1}$$

where K and K_G are $n \times n$ sparse symmetric matrices, K is positive semidefinite and K_G is indefinite, and the matrices K and K_G share a common nullspace $\mathcal{Z}_c$. We consider computing a few nonzero finite eigenvalues around a prescribed shift $\sigma \neq 0$ and the associated eigenvectors x perpendicular to the common nullspace $\mathcal{Z}_c$. We assume that a basis $Z \equiv [Z_N \ Z_C]$ of the nullspace of K and a basis Z_C of the common nullspace $\mathcal{Z}_c$ of K and K_G are available, and the pencil $K - \lambda K_G$ is simultaneouly diagonalizable.

When the matrix K in (5.1) is positive definite, the shift-invert Lanczos method, introduced in Section 2.4, is a widely accepted method to compute solutions of (5.1) near a prescribed shift σ [16]: the buckling eigenvalue problem (5.1) is first converted via a buckling spectral transformation into the equivalent eigenvalue problem

$$Cx = (K - \sigma K_G)^{-1} Kx = \mu x, \quad \mu = \frac{\lambda}{\lambda - \sigma}. \tag{5.2}$$

The solutions of (5.2) are computed by the Lanczos method with the inner product induced by K.

For our buckling eigenvalue problem (5.1), however,

- the matrices K and K_G share a common nullspace $\mathcal{Z}_c$. The shift-invert matrix $(K - \sigma K_G)^{-1}$ does not exist or is extremely ill-conditioned.

- The Lanczos vectors fall rapidly into the nullspace $\mathcal{N}(K)$. The inner product induced by K leads to rapid growth of the Lanczos vectors in norm. The large norms of the Lanczos vectors introduce large round-off errors to the orthogonalization process, leading to loss of accuracy of computed solutions and even break down of the method [23, 25, 33].

In the past, the norms are controlled by restricting the Lacnzos vectors to a proper subspace. In [25], the authors note that the subspace $\mathcal{R}(C)$ is both a complimentary subspace of $\mathcal{N}(K)$ and an invariant subspace of C. It is proposed to restrict the Lanczos vectors v_i by picking a starting vector v in the subspace $\mathcal{R}(C)$. This technique may fail since the fall is caused by the instability of the difference equation (2.10) [25]. In [23], Meerbergen proposes to project the Lanczos vectors v_i back onto the subspace $\mathcal{R}(C)$ through implicit restart. From the governing equation

$$CV_j = V_jT_j + \beta_j v_{j+1}e_j^{\mathrm{T}} + F_j, \tag{5.3}$$

Meerbergen notes that, with the QR decomposition $T_j = Q_1R_1$, the new basis $V_j^{(1)} = V_jQ_1$ by the implicit restart satisfies the equation

$$V_j^{(1)} + \beta_j v_{j+1}e_j^{\mathrm{T}}R_1^{-1} + F_jR_1^{-1} = CV_jR_1^{-1}.$$

Applying implicit restart will effectively put the Lanczos vectors $v_1, \ldots, v_{j-1}$ back onto the subspace $\mathcal{R}(C)$. From the governing equation (5.3), the following inexpensive formula is also proposed to improve the Ritz vector [25],

$$x_i = V_{j+1}w_i, \quad w_i = \frac{1}{\theta_i}\begin{bmatrix} T_js_i \\ \beta_{ji} \end{bmatrix}. \tag{5.4}$$

It is observed that the modified formula (5.4) will effectively put the Ritz vector x_i back onto the subspace $\mathcal{R}(C)$.

In this chapter, we propose alternative strategies to address these issues. We first derive a canonical form of the pencil $K - \lambda K_G$ in Section 5.1. We then convert (5.1) into an equivalent ordinary eigenvalue problem $Cx = \mu x$ by generalizing the buckling spectral transformation (5.2) in Section 5.2. In Section 5.3, we construct a positive definite matrix M by applying low-rank updating to the matrix K. We show that the matrix C is symmetric with respect to the inner product induced by M. In Section 5.4, we propose a shift-invert Lanczos method for the buckling eigenvalue problem (5.1) and provide an implementation of the matrix-vector product $u = Cv$. A validation scheme using inertias is developed in Section 5.5.

5.1 Canonical form

We start with a canonical form of the pencil $K - \lambda K_G$. For the compactness of presentation, we interchange the roles of K and K_G in (5.1) and consider the reversal of the pencil $K - \lambda K_G$, i.e., $K_G - \lambda^\# K$.

Theorem 5.1. *For the pencil $K_G - \lambda^\# K$, there exists a non-singular matrix $W \in \mathbb{R}^{n \times n}$ such that*

$$W^T K_G W = \begin{array}{c|ccc} & n_1 & n_2 & n_3 \\ \hline n_1 & \Lambda_1^\# & & \\ n_2 & & \Lambda_2^\# & \\ n_3 & & & 0 \end{array} \quad \textit{and} \quad W^T K W = \begin{array}{c|ccc} & n_1 & n_2 & n_3 \\ \hline n_1 & I_{n_1} & & \\ n_2 & & 0 & \\ n_3 & & & 0 \end{array}, \tag{5.5}$$

where $\Lambda_1^\#$ and $\Lambda_2^\#$ are diagonal matrices with real diagonal entries, and $\Lambda_2^\#$ is non-singular. Furthermore, by conformally partitioning $W = [W_1, W_2, W_3]$, we have

$$W_3^T W_1 = 0 \quad \textit{and} \quad W_3^T W_2 = 0, \tag{5.6}$$

Proof. see Appendix A. □

By the canonical form (5.5), we immediately know that (i) the columns of W_3 span the common nullspace $\mathcal{Z}_c$ of K and K_G, and the columns of $[W_1 \;\; W_2]$ span the orthogonal complement to $\mathcal{Z}_c$, i.e., $\mathcal{Z}_c^\perp$; (ii) the columns of W_1 are eigenvectors associated with real finite eigenvalues $(\Lambda_1^\#, I_{n_1})$ of the pencil $K_G - \lambda^\# K$ and are perpendicular to $\mathcal{Z}_c$; (iii) The columns of W_2 are eigenvectors associated with an infinite eigenvalue $(\Lambda_2^\#, 0)$ of the pencil $K_G - \lambda^\# K$ and are perpendicular to $\mathcal{Z}_c$; (iv) For $x \in \mathcal{Z}_c$, $(\lambda^\#, x)$ is an eigenpair of the pencil $K_G - \lambda^\# K$ for any $\lambda^\# \in \mathbb{C}$.

5.2 Generalized buckling spectral transformation

Mathematically, a generalized buckling spectral transformation of the singular pencil $K - \lambda K_G$ is to replace the inverse in (5.2) by the pseudo-inverse and leads to the ordinary eigenvalue problem

$$Cx = \mu x \quad \text{with} \quad C = (K - \sigma K_G)^\dagger K, \tag{5.7}$$

where $(K - \sigma K_G)^\dagger$ is the pseudo-inverse of the singular matrix $K - \sigma K_G$ [15, p. 290]. Note that the non-zero real shift σ cannot be an eigenvalue of the pencil $K - \lambda K_G$.

We now present the relationship of non-trivial eigenpairs between the original generalized eigenvalue problem (5.1) and the ordinary eigenvalue problem (5.7). We first use the canonical form (5.5) to derive an eigenvalue decomposition of C and provide the eigenvalue and eigenvector relations between C and $K_G - \lambda^{\#} K$.

Lemma 5.1. *With the canonical form (5.5) in Theorem 5.1, an eigenvalue decomposition of the matrix C defined in (5.7) is given by*

$$CW = W \begin{bmatrix} (I_{n_1} - \sigma\Lambda_1^{\#})^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix}. \tag{5.8}$$

Proof. Recall that, since the matrix $K - \sigma K_G$ is symmetric,

$$\mathcal{R}(K - \sigma K_G) = \mathcal{N}(K - \sigma K_G)^{\perp} = \mathcal{Z}_c^{\perp}. \tag{5.9}$$

In addition, by the condition (5.6) in the canonical form (5.5), we have

$$\mathcal{R}(W_1) \oplus \mathcal{R}(W_2) = \mathcal{R}(W_3)^{\perp} = \mathcal{Z}_c^{\perp}. \tag{5.10}$$

Therefore, from (5.9) and (5.10),

$$\mathcal{R}(K - \sigma K_G) = \mathcal{R}(W_1) \oplus \mathcal{R}(W_2) = \mathcal{R}(W_3)^{\perp} = \mathcal{Z}_c^{\perp}. \tag{5.11}$$

Now note that, from the canonical form (5.5),

$$W^T K W = \begin{bmatrix} I_{n_1} & & \\ & 0 & \\ & & 0 \end{bmatrix} \quad \text{and} \quad W^T(K - \sigma K_G)W = \begin{bmatrix} I_{n_1} - \sigma\Lambda_1^{\#} & & \\ & -\sigma\Lambda_2^{\#} & \\ & & 0 \end{bmatrix}.$$

Therefore, we have

$$W^T K W = \begin{bmatrix} I_{n_1} & & \\ & 0 & \\ & & 0 \end{bmatrix} = W^T(K - \sigma K_G)W \begin{bmatrix} (I_{n_1} - \sigma\Lambda_1^{\#})^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix}. \tag{5.12}$$

Left multiplying (5.12) by $(K - \sigma K_G)^{\dagger} W^{-T}$,

$$(K - \sigma K_G)^{\dagger} K W = (K - \sigma K_G)^{\dagger}(K - \sigma K_G) W \begin{bmatrix} (I_{n_1} - \sigma\Lambda_1^{\#})^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix}. \tag{5.13}$$

The pseudo-inverse $(K - \sigma K_G)^{\dagger}$ satisfies the Moore-Penrose conditions [15, p. 290], which give

$$(K - \sigma K_G)^{\dagger}(K - \sigma K_G) = \mathcal{P}_{\mathcal{R}((K-\sigma K_G)^T)} = \mathcal{P}_{\mathcal{R}(K-\sigma K_G)}, \tag{5.14}$$

namely $(K - \sigma K_G)^{\dagger}(K - \sigma K_G)$ is an orthogonal projection onto $\mathcal{R}((K - \sigma K_G)^T) = \mathcal{R}(K - \sigma K_G)$. Therefore, from (5.11) and (5.14),

$$(K - \sigma K_G)^{\dagger}(K - \sigma K_G)W = W \begin{bmatrix} I_{n_1} & & \\ & I_{n_2} & \\ & & 0 \end{bmatrix}. \tag{5.15}$$

From Eqs. (5.13) and (5.15), we have the eigenvalue decomposition (5.8) of C.

□

Lemma 5.2. *The matrix C defined in (5.7) has the following properties:*

(i) $(\lambda^{\#}, x)$ is an eigenpair of $K_G - \lambda^{\#}K$ with non-zero finite $\lambda^{\#}$ and $x \in \mathcal{Z}_c^{\perp}$ if and only if (μ, x) is an eigenpair of C with $\mu \neq 0$ and $\mu \neq 1$ and $x \in \mathcal{Z}_c^{\perp}$, where $\mu = \frac{1}{1-\sigma\lambda^{\#}}$.

(ii) $(\lambda^{\#}, x)$ is an eigenpair of $K_G - \lambda^{\#}K$ with $\lambda^{\#} = 0$ and $x \in \mathcal{Z}_c^{\perp}$ if and only if (μ, x) is an eigenpair of C with $\mu = 1$ and $x \in \mathcal{Z}_c^{\perp}$.

(iii) $(\lambda^{\#}, x)$ is an eigenpair of $K_G - \lambda^{\#}K$ with $|\lambda^{\#}| = \infty$ and $x \in \mathcal{Z}_c^{\perp}$ if and only if (μ, x) is an eigenpair of C with $\mu = 0$ and $x \in \mathcal{Z}_c^{\perp}$.

(iv) If $x \in \mathcal{Z}_c$, $Cx = 0$.

Proof. The lemma can be proved by comparing the eigenvalue decomposition (5.8) of C with the canonical form (5.5) of $K_G - \lambda^{\#}K$. Specifically, for (i) and (ii), recall that each column of W_1 is an eigenvector associated with a real, finite eigenvalue $\lambda^{\#}$ of the pencil $K_G - \lambda^{\#}K$ and the eigenvector is perpendicular to the common nullspace $\mathcal{Z}_c$. From (5.8), each column of W_1 is now an eigenvector associated with a non-zero, finite eigenvalue $\mu = (1 - \sigma\lambda^{\#})^{-1}$ of the eigenproblem (5.7).

To show (iii), recall that each column of W_2 is an eigenvector associated with an infinite eigenvalue of the pencil $K_G - \lambda^{\#}K$ and the eigenvector is perpendicular to the common nullspace $\mathcal{Z}_c$. From (5.8), each column of W_2 is now an eigenvector associated with zero eigenvalue of the eigenproblem (5.7).

Finally, for (iv), the common nullspace $\mathcal{Z}_c$ is spanned by the columns of W_3 and, from (5.8), we know that $Cx = 0$ if $x \in \mathcal{Z}_c$. □

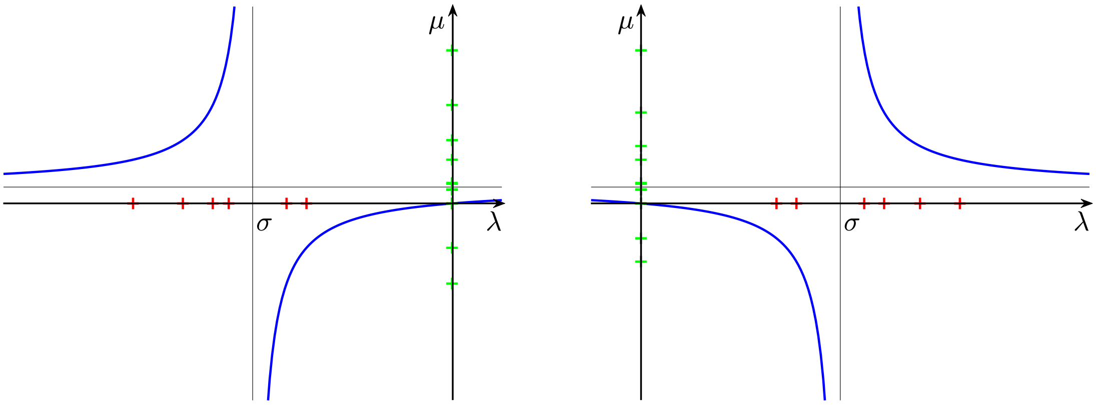


Figure 5.1: Buckling spectral transformation with $\sigma < 0$ (left) and $\sigma > 0$ (right).

The following theorem provides the relationship of non-trivial eigenpairs between the original generalized eigenvalue problem (5.1) and the ordinary eigenvalue problem (5.7).

Theorem 5.2. *(λ, x) is an eigenpair of the pencil $K - \lambda K_G$ with non-zero finite eigenvalue λ and $x \in \mathcal{Z}_c^\perp$ if and only if (μ, x) is an eigenpair of the matrix C in (5.7) with $\mu \neq 0$ and $\mu \neq 1$ and $x \in \mathcal{Z}_c^\perp$, where $\mu = \frac{\lambda}{\lambda - \sigma}$ and $\sigma \neq 0$.*

Proof. Note that (λ, x) is an eigenpair of $K - \lambda K_G$ with non-zero finite eigenvalue λ and $x \in \mathcal{Z}_c^\perp$ if and only if $(\lambda^\#, x)$ is an eigenpair of $K_G - \lambda^\# K$ with non-zero finite eigenvalue $\lambda^\# = \lambda^{-1}$ and $x \in \mathcal{Z}_c^\perp$. Also, from Lemma 5.2 (i), we know that $(\lambda^\#, x)$ is an eigenpair of $K_G - \lambda^\# K$ with non-zero finite eigenvalue $\lambda^\#$ and $x \in \mathcal{Z}_c^\perp$ if and only if (μ, x) is an eigenpair of the eigenvalue problem $Cx = \mu x$ with $\mu = \frac{1}{1 - \sigma\lambda^\#}$, $\mu \neq 0$ and $\mu \neq 1$, and $x \in \mathcal{Z}_c^\perp$. Therefore, (λ, x) is an eigenpair of the pencil $K - \lambda K_G$ with non-zero finite eigenvalue λ and $x \in \mathcal{Z}_c^\perp$ if and only if (μ, x) is an eigenpair of the eigenvalue problem $Cx = \mu x$ with $\mu = \frac{\lambda}{\lambda - \sigma}$, $\mu \neq 0$ and $\mu \neq 1$, and $x \in \mathcal{Z}_c^\perp$. □

By Theorem 5.2, near the shift σ, the eigenpairs (λ, x) of $K - \lambda K_G$ with non-zero finite eigenvalues λ and $x \in \mathcal{Z}_c^\perp$ are transformed into eigenpairs (μ, x) of C with non-zero eigenvalues μ, which typically are well-separated, and those away from the shift σ are transformed into clustered eigenpairs (μ, x) of C near unity as shown in Figure 5.1. We note that the eigenpairs (μ, x) with $\mu = 0$ or $\mu = 1$ are not the ones of interest. The eigenpairs $(1, x)$ correspond to eigenpairs of $K - \lambda K_G$ with infinite eigenvalues and the eigenpairs $(0, x)$ correspond to eigenpairs of $K - \lambda K_G$ with $x \in \mathcal{N}(K)$. □

5.3 Regularized inner product

In this section, we introduce a positive definite matrix M from a low-rank updating of K, and then show that the matrix C in the generalized buckling spectral transformation (5.7) is symmetric with respect to the inner product induced by M.

Theorem 5.3. *Let C be defined in (5.7). Let $Z = [Z_N \; Z_C]$ span the nullspace $\mathcal{N}(K)$ and Z_C span the common nullspace $\mathcal{Z}_c$ of K and K_G. Define*

$$M = K + (K_G Z_N) H_N (K_G Z_N)^T + Z_C H_C Z_C^T, \tag{5.16}$$

where H_N and H_C are arbitrary positive definite matrices. Then

- *(i) the matrix M is positive definite,*
- *(ii) the matrix C is symmetric with respect to the inner product induced by M.*

Proof. By the canonical form (5.5), we have

$$\mathcal{N}(K) = \mathcal{R}(W_2) \oplus \mathcal{R}(W_3) = \mathcal{R}(Z_N) \oplus \mathcal{R}(Z_C) \quad \text{and} \quad \mathcal{Z}_c = \mathcal{R}(W_3) = \mathcal{R}(Z_C),$$

and

$$\begin{bmatrix} Z_N & Z_C \end{bmatrix} = \begin{bmatrix} W_2 & W_3 \end{bmatrix} \begin{bmatrix} R_{22} & O \\ R_{32} & R_{33} \end{bmatrix}$$

for some matrices $R_{22} \in \mathbb{R}^{n_2 \times n_2}$, $R_{32} \in \mathbb{R}^{n_3 \times n_2}$, $R_{33} \in \mathbb{R}^{n_3 \times n_3}$, and R_{22} and R_{33} are nonsingular. Therefore,

$$W^T K_G Z_N = W^T K_G (W_2 R_{22} + W_3 R_{32}) = W^T K_G W_2 R_{22} = \begin{bmatrix} 0 \\ \Lambda_2^{\#} R_{22} \\ 0 \end{bmatrix}.$$

Since the basis W satisfies the condition (5.6),

$$W^T Z_C = W^T W_3 R_{33} = \begin{bmatrix} 0 \\ 0 \\ (W_3^T W_3) R_{33} \end{bmatrix}.$$

Therefore,

$$W^T M W = W^T \big(K + (K_G Z_N) H_N (K_G Z_N)^T + Z_C H_C Z_C^T\big) W = \begin{bmatrix} I_{n_1} & & \\ & \widehat{H}_N & \\ & & \widehat{H}_C \end{bmatrix}, \tag{5.17}$$

where

$$\widehat{H}_N = \Lambda_2^{\#} R_{22} H_N R_{22}^T \Lambda_2^{\#} \quad \text{and} \quad \widehat{H}_C = (W_3^T W_3) R_{33} H_C R_{33}^T (W_3^T W_3).$$

To prove that M is positive definite, we show that both $\widehat{H}_N$ and $\widehat{H}_C$ are positive definite. For the matrix $\widehat{H}_N$, we note that the matrix H_N is positve definite and the matrix R_{22} is non-singular. Also, from Theorem 5.1, the diagonal matrix $\Lambda_2^{\#}$ is non-singular. Therefore, the matrix $\widehat{H}_N$ is positive definite. For the matrix $\widehat{H}_C$, we note that the matrix H_C is positive definite and the matrix R_{33} is non-singular. Also, since the matrix W_3 is of full rank, the symmetric matrix $W_3^T W_3$ is non-singular. Therefore, the matrix $\widehat{H}_C$ is also positive definite. This proves (i).

To prove (ii), by the eigenvalue decomposition (5.8) of C and (5.17), we have

$$W^T MCW = W^T MWW^{-1}CW = \begin{bmatrix} (I_{n_1} - \sigma\Lambda_1^{\#})^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix}.$$

Therefore, the matrix MC is symmetric, which means that the matrix C is symmetric with respect to the inner product induced by M. □

5.4 Shift-invert Lanczos method

By Theorem 5.3, the matrix C in (5.7) is symmetric with respect to the inner product induced by the positive definite matrix M in (5.16). It naturally leads that to solve the buckling eigenvalue problem (5.1), we can use the Lanczos method on the matrix C with the inner product induced by M. This new strategy is also referred to as the shift-invert Lanczos method and outlined in Algorithm 5.1.

We provide an implementation of the matrix-vector product $u = Cv$ at line 7 of Algorithm 5.1. We first show that the matrix-vector product $u = Cv = (K - \sigma K_G)^{\dagger} Kv$ is connected with the solution of a consistent singular linear system with constraint.

Theorem 5.4. *Given $v \in \mathbb{R}^n$, the vector*

$$u = (K - \sigma K_G)^{\dagger} Kv \tag{5.18}$$

is the unique solution of the consistent singular linear system

$$(K - \sigma K_G)u = Kv \tag{5.19}$$

Algorithm 5.1 Shift-invert Lanczos method for the buckling eigenvalue problem (5.1)

```
$r \leftarrow v$, where $v$ is the starting vector
$v_0 \leftarrow 0$
$p \leftarrow Mr$, where $M = K + (K_G Z_N) H_N (K_G Z_N)^T + Z_C H_C Z_C^T$
$\beta_0 \leftarrow (p^T r)^{1/2}$
for $j = 1, 2, \ldots$ do
    $v_j \leftarrow r/\beta_{j-1}$
    $r \leftarrow C v_j$, where $C = (K - \sigma K_G)^\dagger K$
    $r \leftarrow r - \beta_{j-1} v_{j-1}$
    $p \leftarrow Mr$
    $\alpha_j \leftarrow v_j^T p$
    $r \leftarrow r - \alpha_j v_j$
    perform re-orthogonalization if necessary
    $p \leftarrow Mr$
    $\beta_j \leftarrow (p^T r)^{1/2}$
    Compute the eigenvalue decomposition of $T_j$
    Check convergence
end for
Compute approximate eigenvectors of the converged eigenpairs
```

with the constraint

$$Z_C^T u = 0, \tag{5.20}$$

where Z_C is a basis of the common nullspace of K and K_G.

Proof. First note that since both K and $K - \sigma K_G$ are symmetric, we have

$$\mathcal{R}(K) = \mathcal{N}(K)^\perp \quad \text{and} \quad \mathcal{R}(K - \sigma K_G) = \mathcal{N}(K - \sigma K_G)^\perp = \mathcal{Z}_c^\perp \tag{5.21}$$

and

$$\mathcal{Z}_c = \mathcal{N}(K - \sigma K_G) \subset \mathcal{N}(K). \tag{5.22}$$

Therefore from (5.21) and (5.22),

$$Kv \in \mathcal{R}(K) \subset \mathcal{R}(K - \sigma K_G),$$

which implies that the linear system (5.19) is consistent. From (5.18),

$$(K-\sigma K_G)u = (K-\sigma K_G)(K-\sigma K_G)^{\dagger}Kv = \mathcal{P}_{\mathcal{R}(K-\sigma K_G)}Kv = Kv, \tag{5.23}$$

where $\mathcal{P}_{\mathcal{R}(K-\sigma K_G)}$ is an orthogonal projection onto $\mathcal{R}(K-\sigma K_G)$ (by the Moore-Penrose conditions [15, p. 290]). This means that u is a solution of the consistent singular linear system (5.19).

On the other hand, from (5.18) and (5.23),

$$u = (K-\sigma K_G)^{\dagger}Kv = (K-\sigma K_G)^{\dagger}(K-\sigma K_G)u = \mathcal{P}_{\mathcal{R}((K-\sigma K_G)^T)}u = \mathcal{P}_{\mathcal{R}(K-\sigma K_G)}u.$$

Since $\mathcal{R}(K-\sigma K_G) = \mathcal{Z}_c^{\perp}$, it implies that u is perpendicular to the common nullspace $\mathcal{Z}_c$, which is also the nullspace $\mathcal{N}(K-\sigma K_G)$.

The uniqueness can be shown as follows. Given two solutions u_1 and u_2 to (5.19), the difference u_1-u_2 would satisfy $(K-\sigma K_G)(u_1-u_2)=0$, which implies $u_1-u_2 \in \mathcal{Z}_c$. However, since both solutions satisfy the constraint (5.20), $Z_C^T(u_1-u_2)=0$. Therefore, $u_1-u_2=0$. □

We now present method to compute the matrix-vector product $u = Cv$. First, we have the following theorem to extract a non-singular submatrix of $K-\sigma K_G$ by exploiting the basis Z_C.

Theorem 5.5. *Let $Z_C \in \mathbb{R}^{n\times n_3}$ be a basis of $\mathcal{N}(K-\sigma K_G)$ and $P \in \mathbb{R}^{n\times n}$ be a permutation matrix such that $P^TZ_C \equiv \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$, and $Y_2 \in \mathbb{R}^{n_3\times n_3}$ is non-singular. Define*

$$S = P^T(K-\sigma K_G)P \quad \textit{and} \quad S = \begin{array}{c c} & \begin{array}{cc} n-n_3 & n_3 \end{array} \\ \begin{array}{c} n-n_3 \\ n_3 \end{array} & \begin{bmatrix} S_{11}^{\sigma} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \end{array}. \tag{5.24}$$

Then

(1) the submatrix $S_{11}^{\sigma} \in \mathbb{R}^{(n-n_3)\times(n-n_3)}$ is non-singular,

(2) $\nu_+(S_{11}^{\sigma}) = \nu_+(K-\sigma K_G)$ and $\nu_-(S_{11}^{\sigma}) = \nu_-(K-\sigma K_G)$, where $\nu_+(X)$ and $\nu_-(X)$ denote the numbers of positive and negative eigenvalues of the symmetric matrix X, respectively.

Proof. Let

$$E = \begin{array}{c c} & \begin{array}{cc} n-n_3 & n_3 \end{array} \\ \begin{array}{c} n-n_3 \\ n_3 \end{array} & \begin{bmatrix} I_{n-n_3} & Y_1 \\ 0 & Y_2 \end{bmatrix} \end{array} \in \mathbb{R}^{n\times n}.$$

The matrix E is non-singular since Y_2 is non-singular. By the congruence transformation, we have

$$E^T SE = E^T P^T (K - \sigma K_G) PE = E^T \begin{bmatrix} S_{11}^\sigma & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} E = \begin{array}{c c} & \begin{array}{cc} n-n_3 & n_3 \end{array} \\ \begin{array}{c} n-n_3 \\ n_3 \end{array} & \begin{bmatrix} S_{11}^\sigma & 0 \\ 0 & 0 \end{bmatrix} \end{array}. \tag{5.25}$$

Sylvester's law [15, p. 448] tells that the matrices $K - \sigma K_G$ and $E^T SE$ have the same inertias. In particular, from (5.25), we know that

$$\nu_+(K - \sigma K_G) = \nu_+(S_{11}^\sigma), \quad \nu_-(K - \sigma K_G) = \nu_-(S_{11}^\sigma),$$

and

$$\nu_0(K - \sigma K_G) = \nu_0(S_{11}^\sigma) + n_3 \tag{5.26}$$

But $\nu_0(K - \sigma K_G) = \dim(\mathcal{N}(K - \sigma K_G)) = n_3$. Therefore, from (5.26), $\nu_0(S_{11}^\sigma) = 0$ and S_{11}^σ is non-singular. □

Theorem 5.5 was inspired by [2, Theorem 2.2] where the authors consider solving a consistent semi-definite linear systems $Ax = b$ from the electromagnetic applications [3]. The matrix A, generated from the finite element modeling, is positive semi-definite and an explicit basis of the nullspace of A is available. This explicit basis of the nullspace is then used to identify a non-singular part of A and a solution of the linear system can be computed from it. Although in the generalized eigenvalue probem (5.1), the matrix $K - \sigma K_G$ is indefinite, we found that the strategy developed in [2] can be generalized to the system (5.19) and (5.20).

By Theorem 5.5, the method to solve (5.19), i.e., compute the matrix-vector product $u = Cv = (K - \sigma K_G)^\dagger K v$, can be described in two steps:

1. Find a solution u_p of the consistent singular linear system (5.19).

2. Compute $u = \mathcal{P}_{\mathcal{R}(K-\sigma K_G)} u_p$ to satisfy the constraint (5.20), where $\mathcal{P}_{\mathcal{R}(K-\sigma K_G)}$ is an orthogonal projection onto $\mathcal{R}(K - \sigma K_G)$.

Specifically, in Step 1, find the permutation matrix P as described in Theorem 5.5, and rewrite (5.19) in the partitioned form (5.24):

$$\begin{bmatrix} S_{11}^\sigma & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \in \mathcal{R}(S), \tag{5.27}$$

where

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \equiv P^T u \quad \text{and} \quad \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \equiv P^T K v.$$

Since S_{11}^{σ} is non-singular, S_{11}^{σ} is of full rank and the leading $n - n_3$ columns of S are linearly independent. On the other hand, we know that $\operatorname{rank}(S) = \operatorname{rank}(K - \sigma K_G) = n - n_3$. Therefore, the leading $n - n_3$ columns of S is a basis of $\mathcal{R}(S)$, and there is a solution w_p of (5.27) with $w_2 = 0$. Direct substitution gives

$$w_p = \begin{bmatrix} (S_{11}^{\sigma})^{-1} c_1 \\ 0 \end{bmatrix},$$

where the inverse $(S_{11}^{\sigma})^{-1}$ can be computed using the sparse $\mathrm{LDL}^{\mathrm{T}}$factorization of S_{11}^{σ} [4, 10]. A solution u_p of (5.19) is then given by

$$u_p = P \begin{bmatrix} (S_{11}^{\sigma})^{-1} c_1 \\ 0 \end{bmatrix}.$$

In Step 2, since Z_C is a basis of $\mathcal{N}(K - \sigma K_G)$, which is the orthogonal complement to $\mathcal{R}(K - \sigma K_G)$, the vector u can be computed by the projection

$$u = \mathcal{P}_{\mathcal{R}(K - \sigma K_G)} u_p = (I - Z_C (Z_C^T Z_C)^{-1} Z_C^T) u_p.$$

If Z_C is an orthonormal basis, then

$$u = \mathcal{P}_{\mathcal{R}(K - \sigma K_G)} u_p = (I - Z_C Z_C^T) u_p.$$

5.5 Eigenvalue counting

In this section, as a validation scheme, we discuss a way to count the number of eigenvalues in a given interval. In the following, $\nu_+(A)$ and $\nu_-(A)$ denote the number of positive and negative eigenvalues of a symmetric matrix A, respectively. $n(\alpha, \beta)$ and $n^{\#}(\alpha, \beta)$ denote the numbers of eigenvalues of the pencil $K - \lambda K_G$ and the reversed pencil $K_G - \lambda^{\#} K$ in an interval (α, β), respectively.

First, we consider the following lemma.

Lemma 5.3. *Let $Z = [Z_N \ Z_C]$ span the nullspace $\mathcal{N}(K)$ and Z_C span the common nullspace $\mathcal{Z}_c$ of K and K_G, then* *(i) for* $\alpha < 0$*,* $n(\alpha, 0) = \nu_-(K - \alpha K_G) - \nu_-(Z_N^T K_G Z_N)$*,*

(ii) for $\alpha > 0$*,* $n(0, \alpha) = \nu_-(K - \alpha K_G) - \nu_+(Z_N^T K_G Z_N)$*.*

In addition, the matrix $Z_N^T K_G Z_N$ *is non-singular.*

Proof. The proof is based on the following two facts: (1) (λ, x) is an eigenpair of the pencil $K - \lambda K_G$ with non-zero finite eigenvalue λ and $x \in \mathcal{Z}_c^\perp$ if and only if $(\lambda^\#, x)$ is an eigenpair of the pencil $K_G - \lambda^\# K$ with non-zero finite eigenvalue $\lambda^\# = \lambda^{-1}$ and $x \in \mathcal{Z}_c^\perp$. (2) By the canonical form (5.5), we have

$$W^T(K_G - \frac{1}{\alpha}K)W = \begin{bmatrix} \Lambda_1^\# - \frac{1}{\alpha}I_{n_1} & & \\ & \Lambda_2^\# & \\ & & 0 \end{bmatrix}.$$

Consequently, by Sylvester's law, we have

$$\begin{aligned} \nu_-(K_G - \frac{1}{\alpha}K) &= \nu_-(\Lambda_1^\# - \frac{1}{\alpha}I_{n_1}) + \nu_-(\Lambda_2^\#), \\ \nu_+(K_G - \frac{1}{\alpha}K) &= \nu_+(\Lambda_1^\# - \frac{1}{\alpha}I_{n_1}) + \nu_+(\Lambda_2^\#). \end{aligned}$$

Now, for (i), since $\alpha < 0$,

$$\begin{aligned} n(\alpha, 0) &= n^\#(-\infty, \frac{1}{\alpha}) \\ &= \nu_-(\Lambda_1^\# - \frac{1}{\alpha}I_{n_1}) \\ &= \nu_-(K_G - \frac{1}{\alpha}K) - \nu_-(\Lambda_2^\#) \\ &= \nu_-(K - \alpha K_G) - \nu_-(\Lambda_2^\#). \end{aligned} \tag{5.28}$$

For (ii), since $\alpha > 0$,

$$\begin{aligned} n(0, \alpha) &= n^\#(\frac{1}{\alpha}, +\infty) \\ &= \nu_+(\Lambda_1^\# - \frac{1}{\alpha}I_{n_1}) \\ &= \nu_+(K_G - \frac{1}{\alpha}K) - \nu_+(\Lambda_2^\#) \\ &= \nu_-(K - \alpha K_G) - \nu_+(\Lambda_2^\#). \end{aligned} \tag{5.29}$$

On the other hand, by the canonical form (5.5), we have

$$\mathcal{N}(K) = \mathcal{R}(Z_N) \oplus \mathcal{R}(Z_C) = \mathcal{R}(W_2) \oplus \mathcal{R}(W_3) \quad \text{and} \quad \mathcal{Z}_c = \mathcal{R}(Z_C) = \mathcal{R}(W_3),$$

and

$$Z_N = W_2 R_{22} + W_3 R_{32},$$

where $R_{22} \in \mathbb{R}^{n_2 \times n_2}$, $R_{32} \in \mathbb{R}^{n_3 \times n_2}$ and R_{22} is non-singular. Also, we know that $W_2^T K_G W_2 = \Lambda_2^{\#}$. Therefore,

$$Z_N^T K_G Z_N = R_{22}^T (W_2^T K_G W_2) R_{22} = R_{22}^T \Lambda_2^{\#} R_{22}.$$

This implies that the matrix $Z_N^T K_G Z_N$ is non-singular, and by Sylvester's law, we have

$$\nu_-(\Lambda_2^{\#}) = \nu_-(Z_N^T K_G Z_N) \quad \text{and} \quad \nu_+(\Lambda_2^{\#}) = \nu_+(Z_N^T K_G Z_N). \tag{5.30}$$

The lemma is an immediate consequence of (5.28), (5.29) and (5.30). □

Lemma 5.3 establishes the relation between the number of eigenvalues in the interval $(\alpha, 0)$ or $(0, \alpha)$ and the inertia $\nu_-(K - \alpha K_G)$. Below, we discuss how to express the inertia $\nu_-(K - \alpha K_G)$ in terms of the submatrix S_{11}^{α} in (5.24).

Lemma 5.4. *In terms of the submatrix S_{11}^{α} in (5.24),*

$$\nu_-(K - \alpha K_G) = \nu_-(S_{11}^{\alpha}). \tag{5.31}$$

Proof. The equality (5.31) immediately follows from Theorem 5.5. □

Combining Lemmas 5.3 and 5.4, we have the following theorem which provides a computational approach to count the number of eigenvalues of $K - \lambda K_G$ using the inertias of S_{11}^{α}.

Theorem 5.6. *In terms of the submatrix S_{11}^{α} in (5.24), we have*

(*i*) $n(\alpha, 0) = \nu_-(S_{11}^{\alpha}) - \nu_-(Z_N^T K_G Z_N)$, *if* $\alpha < 0$.

(*ii*) $n(0, \alpha) = \nu_-(S_{11}^{\sigma}) - \nu_+(Z_N^T K_G Z_N)$, *if* $\alpha > 0$.

In practice, the inertia $\nu_-(S_{11}^{\alpha})$ is a by-product of the sparse LDL$^{\mathrm{T}}$factorizations of the submatrix S_{11}^{α} [18, p. 214]. The inertias $\nu_-(Z_N^T K_G Z_N)$ and $\nu_+(Z_N^T K_G Z_N)$ can be easily computed since the size of $Z_N^T K_G Z_N$ is small in buckling analysis.

Chapter 6

Numerical Results

In this chapter, we begin with a synthetic example to illustrate the growth of the norms of Lanczos vectors with K-inner product, and the consequence of the growth, as discussed in Chapter 5. Then we demonstrate the efficacy of the proposed shift-invert Lanczos method for an example arising in industrial buckling analysis of structures.

Algorithm 5.1 is implemented in MATLAB [1]. The accuracy of a computed eigenpair $(\widehat{\lambda}_i, \widehat{x}_i)$ of the generalized eigenvalue problem (5.1) is measured by the relative residual norm

$$\eta(\widehat{\lambda}_i, \widehat{x}_i) \equiv \frac{\|K\widehat{x}_i - \widehat{\lambda}_i K_G \widehat{x}_i\|_2}{(\|K\|_1 + |\widehat{\lambda}_i| \|K_G\|_1)\|\widehat{x}_i\|_2}.$$

The Euclidean angle $\theta_i = \angle(\widehat{x}_i, \mathcal{Z}_c)$ is computed for checking if $\widehat{x}_i$ is perpendicular to the common nullspace $\mathcal{Z}_c$ of K and K_G [14, 17].

Example 6.1. Let us consider the following matrix pair (K, K_G) similar to the ones constructed by Meerbergen [23] and Stewart [33]:

$$K = Q\Lambda Q^T \in \mathbb{R}^{n\times n} \quad \text{and} \quad K_G = Q\Phi Q^T \in \mathbb{R}^{n\times n},$$

where $Q \in \mathbb{R}^{n\times n}$ is a random orthogonal matrix, $\Lambda \in \mathbb{R}^{n\times n}$ and $\Phi \in \mathbb{R}^{n\times n}$ are diagonal matrices with diagonal elements

$$\Lambda_{kk} = \begin{cases} k, & \text{if} \quad 1 \le k \le n-m \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \quad \Phi_{kk} = (-1)^k, \quad 1 \le k \le n.$$

By construction, K is positive semi-definite and K_G is indefinite, and the pencil $K - \lambda K_G$ is regular. The last m columns of Q form a basis of the nullspace $\mathcal{N}(K)$. For $1 \le k \le n-m$, the

[1] https://github.com/cplin722/bucklingEigs

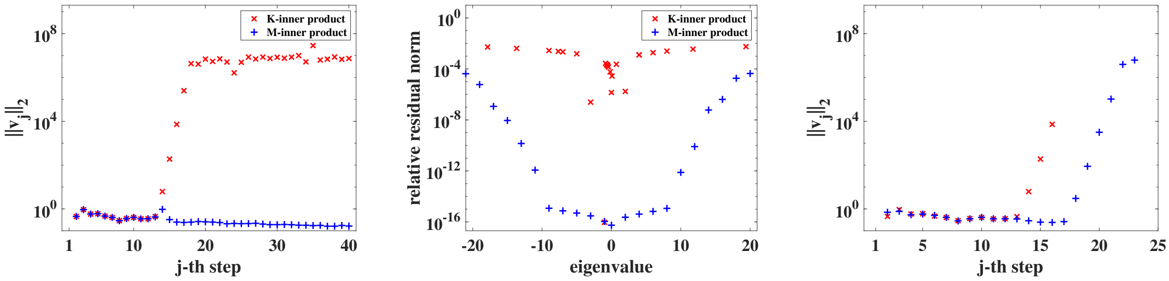


Figure 6.1: Left: the 2-norms of the Lanczos vectors v_j. Middle: the relative residual norms of the approximate eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$. Right: the 2-norms of the Lanczos vectors v_j with (+) and without (x) implicit restart.

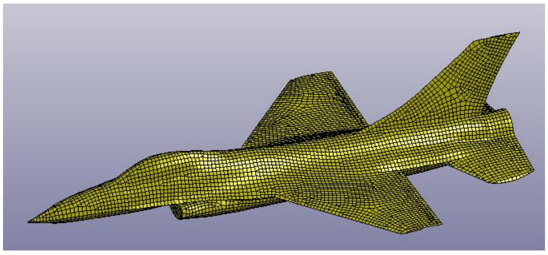
k-th column of Q is an eigenvector and the associated eigenvalue is $\lambda_k = (-1)^k \cdot k$. The zero eigenvalue of $C \equiv (K - \sigma K_G)^{-1} K$ is a well-separated eigenvalue, and the associated eigenspace is also the nullspace of K. We use the MATLAB function `ldl` to compute the LDL$^{\mathrm{T}}$factorization of the shifted matrix $K - \sigma K_G$.

For numerical experiments, we take $n = 500$ and $m = 1$. We use the buckling spectral transformation (5.2) with the shift $\sigma = -0.6$. We run the Lanczos method with K-inner product, and the starting vector Cx_0 with $x_0 = [1, \ \ldots, \ 1]^T$. The approximate eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ of (5.1) are computed by $(\widehat{\lambda}_i, \widehat{x}_i) = (\frac{\sigma \widehat{\mu}_i}{\widehat{\mu}_i - 1}, \widehat{x}_i)$.

The left plot of Figure 6.1 shows the 2-norms of 40 Lanczos vectors v_j. As observed by Meerbergen [23] and Stewart [33], the 2-norms of Lanczos vectors v_j grow rapidly. Consequently, as shown in the middle plot of Figure 6.1, the accuracy of approximate eigenpairs $(\widehat{\lambda}_i, \widehat{x}_i)$ deteriorates. In contrast, when we replace the K-inner product by the positive definite M-inner product with $H_N = I_m$, we observe that the 2-norms of the Lanczos vectors are well bounded. Multiple eigenvalues near the shift σ are computed with the relative residual norms around the machine precision.

In [23], Meerbergen proposed to control the norms of the Lanczos vectors by applying implicit restart. We experimented with the schemes with and without the implicit restart. The results are shown in the right plot of Figure 6.1. We observe that the 2-norms of the Lanczos vectors still grow rapidly.

Example 6.2. This is an example from the buckling analysis of a finite element model of an airplane shown in Figure 6.2. The size of the pencil $K - \lambda K_G$ is $n = 67,512$. The stiffness matrix K is positive semi-definite and the dimension of the nullspace $\mathcal{N}(K)$ is known to be 6, which corresponds to the 6 rigid body modes [12]. The geometric stiffness matrix K_G is symmetric



i	$d_i/\|K_G\|_1$	$\frac{\|Kz_i\|_2}{\|K\|_1\|z_i\|_2}$	$\frac{\|K_Gz_i\|_2}{\|K_G\|_1\|z_i\|_2}$
1	$6.90 \cdot 10^{-5}$	$2.74 \cdot 10^{-16}$	$6.78 \cdot 10^{-5}$
2	$3.25 \cdot 10^{-5}$	$4.88 \cdot 10^{-16}$	$9.06 \cdot 10^{-6}$
3	$2.32 \cdot 10^{-5}$	$4.71 \cdot 10^{-16}$	$1.19 \cdot 10^{-5}$
4	$7.32 \cdot 10^{-16}$	$2.68 \cdot 10^{-17}$	$5.01 \cdot 10^{-18}$
5	$1.26 \cdot 10^{-16}$	$1.90 \cdot 10^{-17}$	$4.89 \cdot 10^{-18}$
6	$7.81 \cdot 10^{-18}$	$2.37 \cdot 10^{-17}$	$5.00 \cdot 10^{-18}$

Figure 6.2: Left: Finite element model of an airplane. Right: Accuracy of the bases for the nullspace of K and common nullspace of K and K_G. The second column shows the singular values d_i of K_GY with Y being an orthonormal basis of $\mathcal{N}(K)$. The third and fourth columns show the accuracy of the basis $Z = [Z_N \ Z_C] = [z_1 \ z_2 \ \dots \ z_6]$.

but indefinite. The basis Z of $\mathcal{N}(K)$ is computed by $Z = [-(K_{11}^{-1}K_{12})^T \ I_6]^T$ [12], where $[K_{11} \ K_{12}] \in \mathbb{R}^{(n-6)\times n}$ is the leading block rows of K. The dimension of the common nullspace $\mathcal{Z}_c$ of K and K_G is 3, which can be easily computed from the basis Z, see [15, Theorem 6.4.1]. The accuracy of the bases is shown in the table in Figure 6.2. We are interested in computing the nonzero eigenvalues of the pencil $K - \lambda K_G$ in an interval around zero and the associated eigenvectors perpendicular to the common nullspace $\mathcal{Z}_c$.

We use the method to compute the matrix-vector product $u = Cv$ described in Section 5.4. We determine the permutation matrix P by maximizing the number of non-zero entries in the last n_3 columns of S in (5.24). The MATLAB function `ldl`, which uses MA57 [9] for real sparse matrices, is used to compute the sparse LDL$^{\mathrm{T}}$factorization of the submatrix S_{11}^{σ}. The pivot tolerance $\tau = 0.1$ is used to control the numerical stability of the factorization [9]. In defining the positive definite matrix M, we form the product K_GZ_N and normalize each column of the matrices K_GZ_N and Z_C. The condition number of K_GZ_N after the normalization is $\kappa_2(K_GZ_N) = 1.03$. Then we set the matrices $H_N = \omega I_{n_2}$ and $H_C = \omega I_{n_3}$, $\omega = \|K\|_1$, to balance the matrix M [29]. The starting vector of the Lanczos procedure is $v = Cx_0$ with x_0 being a random vector [25].

To monitor the progress of the shift-invert Lanczos method, an approximate eigenpair $(\widehat{\mu}_i, \widehat{x}_i)$ computed from an eigenpair $(\widehat{\mu}_i, \widehat{s}_i)$ of the reduced matrix T_j is considered to have converged if the following two conditions are satisfied:

$$|\widehat{\mu}_i| \geq tol \quad \text{and} \quad \frac{|\sigma|}{(\widehat{\mu}_i - 1)^2}|\beta_j||e_j^T\widehat{s}_i| < tol,$$

where the first condition excludes the zero eigenvalues and the second condition bounds the error of the computed eigenvalue $\widehat{\lambda}_i = \frac{\sigma\widehat{\mu}_i}{\widehat{\mu}_i - 1}$ with the prescribed tolerance tol (see [11, 16]

Table 6.1: Results of 12 computed eigenvalues in the interval $(-8, 0)$ after 38 steps of the Lanczos method with the shift $\sigma = -4.0$. $\|\widehat{X}^T M \widehat{X} - I_{12}\|_F = 4.75 \cdot 10^{-12}$ with $\widehat{X} \equiv [\widehat{x}_1 \ \ldots \ \widehat{x}_{12}]$.

i	$\widehat{\lambda}_i$	$\eta(\widehat{\lambda}_i, \widehat{x}_i)$	$\cos\angle(\widehat{x}_i, \mathcal{Z}_c)$
1	-2.716598	$1.48 \cdot 10^{-17}$	$8.52 \cdot 10^{-17}$
2	-2.883589	$1.73 \cdot 10^{-17}$	$8.27 \cdot 10^{-17}$
3	-3.292700	$1.37 \cdot 10^{-17}$	$4.84 \cdot 10^{-18}$
4	-3.378406	$1.01 \cdot 10^{-17}$	$2.38 \cdot 10^{-17}$
5	-5.754628	$2.72 \cdot 10^{-17}$	$4.04 \cdot 10^{-17}$
6	-5.854071	$2.92 \cdot 10^{-17}$	$3.47 \cdot 10^{-17}$
7	-6.089281	$3.14 \cdot 10^{-17}$	$2.47 \cdot 10^{-17}$
8	-6.228974	$2.67 \cdot 10^{-17}$	$6.24 \cdot 10^{-17}$
9	-6.784766	$5.33 \cdot 10^{-16}$	$4.93 \cdot 10^{-17}$
10	-6.886759	$2.57 \cdot 10^{-15}$	$7.67 \cdot 10^{-18}$
11	-7.561377	$1.88 \cdot 10^{-12}$	$1.31 \cdot 10^{-16}$
12	-7.745144	$3.83 \cdot 10^{-12}$	$4.87 \cdot 10^{-17}$

and [26, p. 357]). In this numerical example, we experiment with the tolerance $tol = 10^{-6}$.

We now show the numerical results for computing nonzero eigenvalues of the pencil $K - \lambda K_G$ and corresponding eigenvectors perpendicular to the common nullspace $\mathcal{Z}_c$ in the interval $(-8, 8)$. First, let us consider the left-half interval $(-8, 0)$. With the shift $\sigma = -4.0$, the shift-invert Lanczos method (Algorithm 5.1) computed 12 eigenvalues to the machine precision in the interval $(-8, 0)$ at 38-th iteration. The accuracy of the computed eigenpairs $(\widehat{\lambda}_i = \frac{\sigma\widehat{\mu}_i}{\widehat{\mu}_i - 1}, \widehat{x}_i)$ is shown in Table 6.1. To validate the number of eigenvalues in the interval $(-8, 0)$, we use the counting scheme described in Section 5.5. Using the inertias of the submatrix S_{11}^{α} with $\alpha = -8$ and Theorem 5.5, we have

$$n(-8, 0) = \nu_-(S_{11}^{\alpha}) - \nu_-(Z_N^T K_G Z_N) = 15 - 3 = 12.$$

This matches the number of eigenvalues found in the interval.

Next let us consider the right-half interval $(0, 8)$. In this case, we use the shift $\sigma = 4.0$. By the shift-invert Lanczos method (Algorithm 5.1), we found 13 eigenvalues to the machine precision in the interval $(0, 8)$ at 44-th iteration. The accuracy of the computed eigenpairs $(\widehat{\lambda}_i = \frac{\sigma\widehat{\mu}_i}{\widehat{\mu}_i - 1}, \widehat{x}_i)$ are shown in Table 6.2. To validate the number of eigenvalues in the interval $(0, 8)$, we again use the counting scheme described in Section 5.5. Using the inertias of the submatrix S_{11}^{α} with $\alpha = 8$ and Theorem 5.5, we have

$$n(0, 8) = \nu_-(S_{11}^{\alpha}) - \nu_+(Z_N^T K_G Z_N) = 13 - 0 = 13.$$

This also matches the number of computed eigenvalues in the interval.

Table 6.2: Results of 13 computed eigenvalues in the interval $(0, 8)$ after 44 steps of the Lanczos method with the shift $\sigma = 4.0$. $\|\widehat{X}^T M \widehat{X} - I_{13}\|_F = 1.79 \cdot 10^{-11}$ with $\widehat{X} \equiv [\widehat{x}_1 \ \ldots \ \widehat{x}_{13}]$.

i	$\widehat{\lambda}_i$	$\eta(\widehat{\lambda}_i, \widehat{x}_i)$	$\cos\angle(\widehat{x}_i, \mathcal{Z}_c)$
1	2.967043	$3.80 \cdot 10^{-17}$	$1.10 \cdot 10^{-16}$
2	3.025965	$2.96 \cdot 10^{-17}$	$3.39 \cdot 10^{-17}$
3	3.917831	$1.71 \cdot 10^{-17}$	$7.71 \cdot 10^{-17}$
4	4.008941	$1.61 \cdot 10^{-17}$	$7.13 \cdot 10^{-17}$
5	4.591063	$2.43 \cdot 10^{-17}$	$4.29 \cdot 10^{-17}$
6	4.662575	$2.64 \cdot 10^{-17}$	$2.47 \cdot 10^{-17}$
7	5.699271	$5.24 \cdot 10^{-17}$	$7.45 \cdot 10^{-17}$
8	5.725937	$7.44 \cdot 10^{-17}$	$1.38 \cdot 10^{-17}$
9	6.465175	$7.40 \cdot 10^{-16}$	$1.14 \cdot 10^{-16}$
10	6.598173	$7.96 \cdot 10^{-15}$	$2.18 \cdot 10^{-16}$
11	7.285975	$4.45 \cdot 10^{-15}$	$3.32 \cdot 10^{-16}$
12	7.626265	$2.41 \cdot 10^{-14}$	$1.39 \cdot 10^{-15}$
13	7.880296	$1.24 \cdot 10^{-12}$	$3.71 \cdot 10^{-14}$

Chapter 7

Concluding remarks

In the first part of this dissertation, we analyzed the EED procedure for the symmetric eigenvalue problem $Ax = \lambda x$. In the analysis, two crucial quantities associated with the shifts are identified: the spectral gap γ_j and the shift-gap ratio τ_j. The following sufficient conditions for the backward stability of the EED procedure are derived:

$$\gamma_j^{-1}\|A\|_2 = O(1) \quad \text{and} \quad \tau_j = O(1).$$

A shift selection scheme is proposed to satisfy the conditions. We conclude that the EED procedure is backward stable with a proper choice of the shifts.

Future work include generalization of the EED procedure to the SGEP (2.3) and spectrum analysis of the deflated matrix $\widehat{A}_j$. Improving existing eigensolvers with the EED procedure is also a subject worth studying.

In the second part of this dissertation, we study the buckling eigenvalue problem $Kx = \lambda K_G x$, and address the issues associated with the shift-invert Lanczos method. In this part, the buckling spectral transformation is generalized to the singular pencil $K - \lambda K_G$, and a regularization scheme is proposed for the inner product. An implementation of the matrix-vector is provided and a validation scheme using inertias is proposed. For the industrial example, we found that our method can successfully compute the eigenvalues, and the associated eigenvectors, in an interval. For this part, one future direction is to study the choice of the matrices H_N and H_C for the optimal conditioning of M.

Appendix A

Canonical form of a symmetric semi-definite pencil $A - \lambda B$

We give a constructive derivation of a canonical form of a symmetric semi-definite pencil $A - \lambda B$, namely A is symmetric and B is symmetric semi-positive definite.

Theorem A.1. *For a symmetric semi-definite pencil $A - \lambda B$, there exists a non-singular matrix $W \in \mathbb{R}^{n \times n}$ such that*

$$W^T A W = \begin{array}{c} \\ {\scriptstyle 2n_0} \\ {\scriptstyle n_1} \\ {\scriptstyle n_2} \\ {\scriptstyle n_3} \end{array} \!\!\begin{array}{c} \begin{array}{cccc} {\scriptstyle 2n_0} & {\scriptstyle n_1} & {\scriptstyle n_2} & {\scriptstyle n_3} \end{array} \\ \left[\begin{array}{cccc} S & & & \\ & \Lambda & & \\ & & \Psi & \\ & & & 0 \end{array}\right] \end{array} \quad \text{and} \quad W^T B W = \begin{array}{c} \\ {\scriptstyle 2n_0} \\ {\scriptstyle n_1} \\ {\scriptstyle n_2} \\ {\scriptstyle n_3} \end{array} \!\!\begin{array}{c} \begin{array}{cccc} {\scriptstyle 2n_0} & {\scriptstyle n_1} & {\scriptstyle n_2} & {\scriptstyle n_3} \end{array} \\ \left[\begin{array}{cccc} \Omega & & & \\ & I_{n_1} & & \\ & & 0 & \\ & & & 0 \end{array}\right] \end{array}, \qquad (A.1)$$

where

$$S \equiv I_{n_0} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \Omega \equiv I_{n_0} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

Λ and Ψ are diagonal matrices with real diagonal entries, and Ψ is non-singular. Moreover, we have

$$n_0 = \dim(\mathcal{N}(B)) - n_2 - n_3,$$

$$n_1 = \operatorname{rank}(B) - n_0,$$

$$n_2 = \operatorname{rank}(\mathcal{P}_{\mathcal{N}(B)} A \mathcal{P}_{\mathcal{N}(B)}),$$

$$n_3 = \dim(\mathcal{N}(A) \cap \mathcal{N}(B)),$$

where $\mathcal{P}_{\mathcal{N}(B)}$ is the orthogonal projection onto $\mathcal{N}(B)$.

We first introduce the following lemma due to Fix and Heiberger [13], also see [26, Sec. 15.5].

Lemma A.1. *For the symmetric semi-definite pencil $A-\lambda B$, there exists a non-singular matrix $W \in \mathbb{R}^{n\times n}$ such that*

$$W^TAW = \begin{array}{c} \\ n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \left[\begin{array}{ccccc} A_{00} & A_{01} & A_{02} & \Theta & 0 \\ A_{01}^T & A_{11} & A_{12} & & \\ A_{02}^T & A_{12}^T & \Psi & & \\ \Theta & & & 0 & \\ 0 & & & & 0 \end{array}\right] \end{array} \quad \text{and} \quad W^TBW = \begin{array}{c} \\ n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \left[\begin{array}{ccccc} I_{n_0} & & & & \\ & I_{n_1} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{array}\right] \end{array},$$

where Ψ and Θ are non-singular, diagonal matrices with real diagonal entries.

Proof of Theorem A.1. By Lemma 2.1, there exists a non-singular matrix $W_0 \in \mathbb{R}^{n\times n}$ such that

$$A^{(1)} \equiv W_0^TAW_0 = \begin{array}{c} \\ n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \left[\begin{array}{ccccc} A_{00} & A_{01} & A_{02} & \Theta & 0 \\ A_{01}^T & A_{11} & A_{12} & & \\ A_{02}^T & A_{12}^T & \Psi & & \\ \Theta & & & 0 & \\ 0 & & & & 0 \end{array}\right] \end{array}$$

and

$$B^{(1)} \equiv W_0^TBW_0 = \begin{array}{c} \\ n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \left[\begin{array}{ccccc} I_{n_0} & & & & \\ & I_{n_1} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{array}\right] \end{array},$$

where Ψ and Θ are non-singular, diagonal matrices with real diagonal entries.

Let

$$W_1 \equiv \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & I_{n_0} & & & & \\ n_1 & & I_{n_1} & & & \\ n_2 & & & I_{n_2} & & \\ n_0 & -\Theta^{-1}A_{00}/2 & -\Theta^{-1}A_{01} & -\Theta^{-1}A_{02} & I_{n_0} & \\ n_3 & & & & & I_{n_3} \end{array},$$

then

$$A^{(2)} \equiv W_1^T A^{(1)} W_1 = \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & 0 & & & \Theta & \\ n_1 & & A_{11} & A_{12} & & \\ n_2 & & A_{12}^T & \Psi & & \\ n_0 & \Theta & & & 0 & \\ n_3 & & & & & 0 \end{array}$$

and

$$B^{(2)} \equiv W_1^T B^{(1)} W_1 = \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & I_{n_0} & & & & \\ n_1 & & I_{n_1} & & & \\ n_2 & & & 0 & & \\ n_0 & & & & 0 & \\ n_3 & & & & & 0 \end{array}.$$

Next let

$$W_2 \equiv \begin{array}{c|ccccc|} & n_0 & n_1 & n_2 & n_0 & n_3 \\ n_0 & I_{n_0} & & & & \\ n_1 & & I_{n_1} & & & \\ n_2 & & -\Psi^{-1}A_{12}^T & I_{n_2} & & \\ n_0 & & & & I_{n_0} & \\ n_3 & & & & & I_{n_3} \end{array},$$

then

$$A^{(3)} \equiv W_2^T A^{(2)} W_2 = \begin{array}{cc} & \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \begin{array}{c} n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} & \begin{bmatrix} 0 & & & \Theta & \\ & C & & & \\ & & \Psi & & \\ \Theta & & & 0 & \\ & & & & 0 \end{bmatrix} \end{array}$$

and

$$B^{(3)} \equiv W_2^T B^{(2)} W_2 = \begin{array}{cc} & \begin{array}{ccccc} n_0 & n_1 & n_2 & n_0 & n_3 \end{array} \\ \begin{array}{c} n_0 \\ n_1 \\ n_2 \\ n_0 \\ n_3 \end{array} & \begin{bmatrix} I_{n_0} & & & & \\ & I_{n_1} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \end{array},$$

where $C \in \mathbb{R}^{n_1 \times n_1}$ is symmetric and $C = A_{11} - A_{12}\Psi^{-1}A_{12}^T$.

Define the permutation matrix

$$P_3 \equiv \begin{bmatrix} I_{n_0} & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{n_1} & 0 & 0 \\ 0 & 0 & 0 & I_{n_2} & 0 \\ 0 & I_{n_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{n_3} \end{bmatrix},$$

then

$$A^{(4)} \equiv P_3^T A^{(3)} P_3 = \begin{array}{cc} & \begin{array}{ccccc} n_0 & n_0 & n_1 & n_2 & n_3 \end{array} \\ \begin{array}{c} n_0 \\ n_0 \\ n_1 \\ n_2 \\ n_3 \end{array} & \begin{bmatrix} & \Theta & & & \\ \Theta & & & & \\ & & C & & \\ & & & \Psi & \\ & & & & 0 \end{bmatrix} \end{array}$$

and

$$B^{(4)} \equiv P_3^T B^{(3)} P_3 = \begin{array}{c} \\ n_0 \\ n_0 \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccccc} n_0 & n_0 & n_1 & n_2 & n_3 \end{array} \\ \left[\begin{array}{ccccc} I_{n_0} & & & & \\ & 0 & & & \\ & & I_{n_1} & & \\ & & & 0 & \\ & & & & 0 \end{array}\right] \end{array}.$$

Since $C \in \mathbb{R}^{n_1 \times n_1}$ is symmetric, it admits the eigen-decomposition

$$C = Q \Lambda Q^T,$$

where $Q \in \mathbb{R}^{n_1 \times n_1}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{n_1 \times n_1}$ is a diagonal matrix. Applying the congruent transformation associated with $W_4 \equiv \operatorname{diag}(I_{n_0}, \Theta^{-1}, Q, I_{n_2}, I_{n_3})$, we have

$$A^{(5)} \equiv W_4^T A^{(4)} W_4 = \begin{array}{c} \\ n_0 \\ n_0 \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccccc} n_0 & n_0 & n_1 & n_2 & n_3 \end{array} \\ \left[\begin{array}{ccccc} & I_{n_0} & & & \\ I_{n_0} & & & & \\ & & \Lambda & & \\ & & & \Psi & \\ & & & & 0 \end{array}\right] \end{array}$$

and

$$B^{(5)} \equiv W_4^T B^{(4)} W_4 = \begin{array}{c} \\ n_0 \\ n_0 \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccccc} n_0 & n_0 & n_1 & n_2 & n_3 \end{array} \\ \left[\begin{array}{ccccc} I_{n_0} & & & & \\ & 0 & & & \\ & & I_{n_1} & & \\ & & & 0 & \\ & & & & 0 \end{array}\right] \end{array}.$$

Last, define the permutation matrix $P_5 \equiv \operatorname{diag}(E, I_{n_1}, I_{n_2}, I_{n_3})$, where the matrix

$E \equiv [e_1\ e_{n_0+1}\ e_2 \ldots\ e_{2n_0}]$, and we have the canonical form in (A.1)

$$A^{(6)} \equiv P_5^T A^{(5)} P_5 = \begin{array}{c|cccc|} & 2n_0 & n_1 & n_2 & n_3 \\ \hline 2n_0 & S & & & \\ n_1 & & \Lambda & & \\ n_2 & & & \Psi & \\ n_3 & & & & 0 \\ \hline \end{array}$$

and

$$B^{(6)} \equiv P_5^T B^{(5)} P_5 = \begin{array}{c|cccc|} & 2n_0 & n_1 & n_2 & n_3 \\ \hline 2n_0 & \Omega & & & \\ n_1 & & I_{n_1} & & \\ n_2 & & & 0 & \\ n_3 & & & & 0 \\ \hline \end{array},$$

where

$$S \equiv I_{n_0} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \Omega \equiv I_{n_0} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The canonical form (A.1) is obtained with $W \equiv W_0 W_1 W_2 P_3 W_4 P_5$.

Now we interpret the dimension of each block matrix. From the canonical form of B in Eq. (A.1), we can infer that $n_0 = \dim(\mathcal{N}(B)) - n_2 - n_3$ and $n_1 = \operatorname{rank}(B) - n_0$. Also, $n_3 = \dim(\mathcal{N}(A) \cap \mathcal{N}(B))$. To interpret n_2, let $Z \in \mathbb{R}^{n \times (n_0+n_2+n_3)}$ be the basis of $\mathcal{N}(B)$ consisting of the columns of W and consider the QR decomposition of $Z = QR$. Since Q is an orthonormal basis of $\mathcal{N}(B)$, $\operatorname{rank}(\mathcal{P}_{\mathcal{N}(B)} A \mathcal{P}_{\mathcal{N}(B)}) = \operatorname{rank}(Q^T A Q)$. By the Sylvester's law, $\operatorname{rank}(Q^T A Q) = \operatorname{rank}(Z^T A Z)$. But, from the canonical form (A.1), $Z^T A Z = \operatorname{diag}(0_{n_0}, \Psi, 0_{n_3})$ and $\operatorname{rank}(Z^T A Z) = n_2$. Therefore, $n_2 = \operatorname{rank}(\mathcal{P}_{\mathcal{N}(B)} A \mathcal{P}_{\mathcal{N}(B)})$. □

Corollary A.1. *The symmetric semi-definite pencil $A - \lambda B$ is simultaneously diagonalizable if and only if $n_0 = 0$. In this case, we have the canonical form*

$$W^T A W = \begin{array}{c|ccc|} & n_1 & n_2 & n_3 \\ \hline n_1 & \Lambda & & \\ n_2 & & \Psi & \\ n_3 & & & 0 \\ \hline \end{array} \quad \textit{and} \quad W^T B W = \begin{array}{c|ccc|} & n_1 & n_2 & n_3 \\ \hline n_1 & I_{n_1} & & \\ n_2 & & 0 & \\ n_3 & & & 0 \\ \hline \end{array},$$

Proof. From the pairs (S, Ω) and $(\Psi, 0)$ in Eq. (A.1), we note that the algebraic and geometric multiplicity of the infinite eigenvalues are $2n_0 + n_2$ and $n_0 + n_2$, respectively. Therefore, the symmetric semi-definite pencil $A - \lambda B$ is simultaneously diagonalizable if and only if $n_0 = 0$. □

References

[1] E. C. Anderson, Z. Bai, C. H. Bischof, L. S. Blackford, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. Hammarling, A. M. McKenney, and D. C. Sorensen. *LAPACK Users' Guide (3rd Edition).* SIAM, Philadelphia, PA, 1999.

[2] P. Arbenz and Z. Drmac. On positive semidefinite matrices with known null space. *SIAM Journal on Matrix Analysis and Applications*, 24(1):132–149, 2002.

[3] P. Arbenz, R. Geus, and S. Adam. Solving Maxwell eigenvalue problems for accelerating cavities. *Physical Review Special Topics - Accelerators and Beams*, 4(2):022001, 2001.

[4] C. Ashcraft, R. G. Grimes, and J. G. Lewis. Accurate symmetric indefinite linear equation solvers. *SIAM Journal on Matrix Analysis and Applications*, 20(2):513–561, 1998.

[5] Z. Bai, J. W. Demmel, J. J. Dongarra, A. Ruhe, and H. A. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide.* SIAM, Philadelphia, PA, 2000.

[6] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart. Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization. *Mathematics of Computation*, 30(136):772–795, 1976.

[7] T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1–25, 2011.

[8] J. W. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

[9] I. S. Duff. MA57—a code for the solution of sparse symmetric definite and indefinite systems. *ACM Transactions on Mathematical Software (TOMS)*, 30(2):118–144, 2004.

[10] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct Methods for Sparse Matrices (2nd Edition).* Oxford, New York, NY, 2017.

[11] T. Ericsson and A. Ruhe. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Mathematics of Computation*, 35(152):1251–1268, 1980.

[12] C. Farhat and M. Géradin. On the general solution by a direct method of a large-scale singular system of linear equations: application to the analysis of floating structures. *International Journal for Numerical Methods in Engineering*, 41(4):675–696, 1998.

[13] G. Fix and R. Heiberger. An algorithm for the ill-conditioned generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 9(1):78–88, 1972.

[14] V. Frayssé and V. Toumazou. A note on the normwise perturbation theory for the regular generalized eigenproblem. *Numerical Linear Algebra with Applications*, 5(1):1–10, 1998.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations (4th Edition).* Johns Hopkins University Press, Baltimore, MD, 2013.

[16] R. G. Grimes, J. G. Lewis, and H. D. Simon. A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 15(1):228–272, 1994.

[17] D. J. Higham and N. J. Higham. Structured backward error and condition of generalized eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 20(2):493–512, 1998.

[18] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, PA, 2002.

[19] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[20] T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis.* Dover Publications, Mineola, NY, 2000.

[21] A. V. Knyazev. Laplacian in 1D, 2D, or 3D (https://www.mathworks.com/matlabcentral/fileexchange/27279-laplacian-in-1d-2d-or-3d). MATLAB Central File Exchange, 2015. Retrieved March 24, 2020.

[22] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, PA, 1998.

[23] K. Meerbergen. The Lanczos method with semi-definite inner product. *BIT Numerical Mathematics*, 41(5):1069–1078, 2001.

[24] J. H. Money and Q. Ye. Algorithm 845: EIGIFP: a MATLAB program for solving large symmetric generalized eigenvalue problems. *ACM Transactions on Mathematical Software (TOMS)*, 31(2):270–279, 2005.

[25] B. Nour-Omid, B. N. Parlett, T. Ericsson, and P. S. Jensen. How to implement the spectral transformation. *Mathematics of Computation*, 48(178):663–673, 1987.

[26] B. N. Parlett. *The Symmetric Eigenvalue Problem.* SIAM, Philadelphia, PA, 1998.

[27] Y. Saad. Numerical solution of large nonsymmetric eigenvalue problems. *Computer Physics Communications*, 53(1):71–90, 1989.

[28] Y. Saad. *Numerical Methods for Large Eigenvalue Problems (2nd Edition).* SIAM, Philadelphia, PA, 2011.

[29] J. A. Sifuentes, Z. Gimbutas, and L. F. Greengard. Randomized methods for rank-deficient linear systems. *Electronic Transactions on Numerical Analysis*, 44:177–188, 2015.

[30] H. D. Simon. Analysis of the symmetric Lanczos algorithm with reorthogonalization methods. *Linear Algebra and its Applications*, 61:101–131, 1984.

[31] H. D. Simon. The Lanczos algorithm with partial reorthogonalization. *Mathematics of Computation*, 42(165):115–142, 1984.

[32] D. C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13(1):357–385, 1992.

[33] G. W. Stewart. On the semidifinite B-Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1458–1468, 2009.

[34] J.-G. Sun. A note on backward perturbations for the Hermitian eigenvalue problem. *BIT Numerical Mathematics*, 35(3):385–393, 1995.

[35] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, UK, 1965.

[36] K. Wu and H. D. Simon. Thick-restart Lanczos method for large symmetric eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 22(2):602–616, 2000.

[37] I. Yamazaki, Z. Bai, D. Lu, and J. J. Dongarra. Matrix powers kernels for thick-restart Lanczos with explicit external deflation. In *2019 IEEE International Parallel Distributed Processing Symposium (IPDPS)*, pages 472–481. IEEE, 2019.

[38] I. Yamazaki, Z. Bai, H. D. Simon, L.-W. Wang, and K. Wu. Adaptive projection subspace dimension for the thick-restart Lanczos method. *ACM Transactions on Mathematical Software (TOMS)*, 37(3):27, 2010.

[39] I. Yamazaki, K. Wu, and H. D. Simon. nu-TRLan User Guide version 1.0. Technical report, LBNL 1288E, 2008.