SHARP ERROR BOUNDS FOR RITZ VECTORS
AND APPROXIMATE SINGULAR VECTORS

YUJI NAKATSUKASA

ABSTRACT. We derive sharp bounds for the accuracy of approximate eigenvectors (Ritz vectors) obtained by the Rayleigh-Ritz process for symmetric eigenvalue problems. Using information that is available or easy to estimate, our bounds improve the classical Davis-Kahan $\sin\theta$ theorem by a factor that can be arbitrarily large, and can give nontrivial information even when the $\sin\theta$ theorem suggests that a Ritz vector might have no accuracy at all. We also present extensions in three directions, deriving error bounds for invariant subspaces, singular vectors and subspaces computed by a (Petrov-Galerkin) projection SVD method, and eigenvectors of self-adjoint operators on a Hilbert space.

1. INTRODUCTION

It is well known that the eigenvector corresponding to a near-multiple eigenvalue is ill-conditioned. Specifically, the classical Davis-Kahan theory [2] implies that the condition number of eigenvectors of symmetric or Hermitian matrices is $1/\text{gap}$, where gap is the smallest distance between the particular eigenvalue and the other eigenvalues. For example, if $(\widehat{\lambda}, \widehat{x})$ with $\|\widehat{x}\| = 1$ is an approximation to an exact eigenpair (λ, x) of a symmetric matrix A with residual $\|r\| = \|A\widehat{x} - \widehat{\lambda}\widehat{x}\|$, then the Davis-Kahan $\sin\theta$ theorem gives the error bound for $\widehat{x}$ [2],[17, Ch. 11]:

$$(1.1) \qquad \sin\angle(x,\widehat{x}) \leq \frac{\|r\|}{\text{gap}_c},$$

where $\angle(x,\widehat{x}) = \text{acos}\frac{|\widehat{x}^T x|}{\|\widehat{x}\|\|x\|}$ and gap_c is the distance between $\widehat{\lambda}$ and the eigenvalues of A other than λ (where the subscript stands for "classical"). Here and throughout, $\|\cdot\|$ for vectors denotes the standard Euclidean norm. In view of the bound (1.1), it is commonly believed that if gap_c is smaller than the residual $\|r\|$, then we cannot guarantee any accuracy in the computed eigenvector $\widehat{x}$.

In this work, we partly challenge this belief. Namely, we examine the accuracy of eigenvectors obtained by the Rayleigh-Ritz process (R-R), the most widely-used process for computing partial (usually extremal) eigenpairs of large-scale symmetric/Hermitian matrices, and show that (1.1) can be improved—often significantly, and by a factor that can be arbitrarily large—using quantities that are readily available (or can be estimated cheaply) after the computation.

Of course, the classical Davis-Kahan bound is tight in general: In the absence of additional information other than $\|r\|$ and gap_c, we cannot improve (1.1), in that there exist examples for which the bound (1.1) is essentially tight. However, when $(\widehat{\lambda}, \widehat{x})$ is a computed approximate eigenpair (Ritz pair) obtained by R-R, there is usually abundant additional information available that (1.1) does not use: most importantly, the residual r is orthogonal to the trial subspace, which is rich in the eigenspace corresponding to not only λ but also eigenvalues close to $\widehat{\lambda}$. Moreover, since the trial subspace in R-R usually contains approximation to nearby eigenpairs (e.g., when looking for the smallest eigenvalues), a bound can be computed for Gap (which we call the "big Gap"), which is roughly the distance between the Ritz value $\widehat{\lambda}$ and eigenvalues not approximated by the Ritz values; see (2.3) for the precise definition. These are the crucial properties that allow us to improve the Davis-Kahan bound (1.1)—in other words, we take into account the matrix structure generated automatically by R-R to derive sharp bounds for the Ritz vector error.

Our results essentially show that up to a modest constant, the gap_c in (1.1) can be replaced by Gap, which is usually much wider, thus improving classical results. Another way to understand our results is via (structured) perturbation theory: while an eigenvector has condition number $1/\mathrm{gap}_c$ if a general perturbation is allowed, R-R imposes a structure in the perturbation that reduces the *structured* condition number to $1/\mathrm{Gap}$.

Qualitatively speaking, the fact that the accuracy of Ritz vectors depends on Gap rather than gap_c was pointed out by Ovtchinnikov [16, Thm. 4]. However, the bounds there involve quantities that are unavailable and difficult to estimate, such as the projector onto an exact eigenspace. Our bounds are easy to compute or estimate, using information that is available after a typical computation of an approximate eigenpair via the R-R process. Our bounds are also tight, in that they cannot be improved without additional information.

In addition, we extend the results in three ways. First, we obtain error bounds for invariant subspaces (spanned by more than one eigenvector) computed by R-R. This gives an answer to one of the open problems suggested in Davis-Kahan's classical paper. Second, we derive their SVD variants, establishing tight bounds for the quality of approximate singular vectors and singular subspaces associated with the largest singular values, obtained by a (Petrov-Galerkin) projection method. Finally, we generalize the error bounds to eigenvectors of self-adjoint operators on a Hilbert space.

Notation. $\lambda(A)$ denotes the spectrum (set of eigenvalues) of a symmetric matrix A. $\sigma(A) = \{\sigma_i(A)\}_{i=1}^{\min(m,n)}$ is the set of singular values of $A \in \mathbb{C}^{m \times n}$, where $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min}(A) = \sigma_{\min(m,n)}(A) \geq 0$. I_n denotes the $n \times n$ identity matrix. $Q_\perp \in \mathbb{C}^{n \times (n-k)}$ is the orthogonal complement of $Q \in \mathbb{C}^{n \times k}$. Quantities involved in the R-R process wear a hat (e.g., $\widehat{\lambda}, \widehat{x}, \widehat{X}$), and those with tildes are auxiliary objects for the analysis. Norms $\|\cdot\|$ without subscripts denote the spectral norm equal to the largest singular value, which for vectors are the Euclidean norm. $\|\|\cdot\|\|$ stands for inequalities that hold for any fixed unitarily invariant norm. Inequalities involving $\|\cdot\|_{2,F}$ hold for the spectral and Frobenius norms, but not necessarily for any unitarily invariant norm. We denote by $\mathcal{A}$ a self-adjoint operator on a Hilbert space, $\lambda(\mathcal{A})$ its spectrum, and $\|\mathcal{A}\|$ its spectral (operator) norm. We

drop the subscript i in $\widehat{\lambda}_i, \lambda_i$ when this can be done without causing confusion. We always normalize eigenvectors and Ritz vectors to have unit norm $\|x\| = \|\widehat{x}\| = 1$.

Unless otherwise stated, for definiteness we assume that the Ritz values $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$ approximate the smallest eigenvalues of A (and accordingly the Ritz values are arranged in increasing order $\widehat{\lambda}_1 \le \widehat{\lambda}_2 \le \cdots \le \widehat{\lambda}_k$). This is a typical situation in applications, and clearly the discussion covers the case where the largest eigenvalues are sought (if necessary by working with $-A$). A less common but still important case is when interior eigenvalues are desired, for example those lying in an interval (e.g., [13]). Our results are applicable to this case also; one subtlety here is that some care is needed in estimating Gap_i, since the Ritz values tend to contain outliers in this case.

2. Setup

2.1. **Big (good) Gap, small (bad) gap.** Let $A \in \mathbb{C}^{n \times n}$ be the (large) Hermitian matrix whose partial eigenvalues are sought, and let $Q \in \mathbb{C}^{n \times k}(n \ge k$, usually $n \gg k$) be a trial subspace with orthonormal columns $Q^*Q = I_k$ (obtained, e.g., via Lanczos, LOBPCG, Jacobi-Davidson, or the generalized Davidson method [1]). Following standard practice, for a matrix with orthonormal columns Q, we identify the matrix Q with its column space $\mathrm{Span}(Q)$. R-R obtains approximate eigenvalues (Ritz values) and eigenvectors (Ritz vectors) as follows:

(1) Compute the $k \times k$ matrix Q^*AQ.
(2) Compute the eigendecomposition $Q^*AQ = \Omega\widehat{\Lambda}\Omega^*$. $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k) = \mathrm{diag}(\widehat{\Lambda})$ are the Ritz values, and $\widehat{X} := [\widehat{x}_1, \ldots, \widehat{x}_k] = Q\Omega$ are the Ritz vectors.

The Ritz pairs $(\widehat{\lambda}_i, \widehat{x}_i)$ thus obtained satisfy $\widehat{x}_i \in \mathrm{span}(Q)$ for all i, and since $Q^*(AQ\Omega - Q\Omega\widehat{\Lambda}) = Q^*AQ\Omega - \Omega\widehat{\Lambda} = \Omega\widehat{\Lambda}\Omega^*\Omega - \Omega\widehat{\Lambda} = 0$ by construction, we have—crucially for this work—the orthogonality between Q and the residuals $A\widehat{x}_i - \widehat{\lambda}_i\widehat{x}_i \perp Q$ for every $1 \le i \le k$. Throughout we assume $k \ge 2$; indeed when $k = 1$ there is no room for improvement upon Davis-Kahan.

Underlying R-R is a matrix of particular structure: Let $Q_\perp$ be the orthogonal complement of Q, such that $[Q\ Q_\perp]$ is a square unitary matrix (and hence so is $[Q\Omega, Q_\perp]$), and consider the unitary transformation applied to A

$$(2.1) \quad \widetilde{A} := [Q\Omega, Q_\perp]^*A[Q\Omega, Q_\perp] = \begin{bmatrix} \widehat{\lambda}_1 & & & ---r_1^T--- \\ & \ddots & & \vdots \\ & & \widehat{\lambda}_k & ---r_k^T--- \\ & & & \\ \,\vert & & \,\vert & \\ r_1 & \cdots & r_k & A_3 \\ \,\vert & & \,\vert & \end{bmatrix}.$$

Here $R = (Q_\perp)^*AQ\Omega = [r_1, r_2, \ldots, r_k]$; we use the subscript 3 in A_3 because later we partition the $(1,1)$ block further into two pieces.

Suppose $(\lambda_i, \widetilde{x}_i)$ is an exact eigenpair of $\widetilde{A}$ such that $\widetilde{A}\widetilde{x}_i = \lambda_i\widetilde{x}_i$. Denote by $z_i \in \mathbb{C}^{n-k}$ the vector of the bottom $n-k$ elements of $\widetilde{x}_i$, and by w_i the ith element, and by $y_i \in \mathbb{C}^{k-1}$ the first k elements of $\widetilde{x}_i$, except the ith. For example when $i = 1$, we have $\widetilde{x}_1 = \begin{bmatrix} w_1 \\ y_1 \\ z_1 \end{bmatrix}$. Then since $x_i = [\widehat{X}, Q_\perp]\widetilde{x}_i$ is the corresponding eigenvector

of A, and $(\widehat{\lambda}_i, \widehat{x}_i)$ is a Ritz pair with $\widehat{x}_i = [\widehat{X}, Q_\perp]e_i$ where e_i is the ith column of identity I_n, it follows that $\cos\angle(x_i, \widehat{x}_i) = |e_i^T \widetilde{x}_i| = |w_i|$, and hence

$$(2.2) \qquad \sin\angle(x_i, \widehat{x}_i) = \sqrt{\|y_i\|^2 + \|z_i\|^2}.$$

This is a key fact in the forthcoming analysis.

Fundamental in this work is the distinction between the "big gap" Gap_i and the "small gap" gap_i, defined for $i = 1, \ldots, k$ by

$$(2.3) \qquad \mathrm{Gap}_i := \min|\lambda_i - \lambda(A_3)|, \qquad \mathrm{gap}_i := \min_{j \in \{1, \ldots, k\} \setminus i} |\lambda_i - \widehat{\lambda}_j|.$$

Intuitively, Gap_i measures the distance between the target λ_i and the undesired eigenvalues, whereas gap_i is that between λ_i and all the other eigenvalues,[1] including the desired ones (e.g., λ_2). For example when $R \to 0$, we have $\mathrm{Gap}_i \to \min|\lambda_i - \lambda_{k+1}|$; by contrast $\mathrm{gap}_i \to \min(|\lambda_i - \lambda_{i+1}|, |\lambda_i - \lambda_{i-1}|)$. Observe that $\mathrm{Gap}_i \geq \mathrm{gap}_i$, and we typically have $\mathrm{Gap}_i \gg \mathrm{gap}_i$. We illustrate this in Figure 2.1 for $i = 1$. Throughout the paper, it is helpful to consider the case $i = 1$, where the target eigenpair is the smallest one.[2]
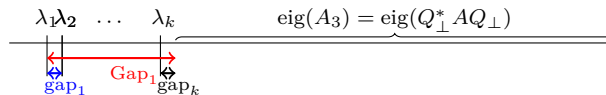

FIGURE 2.1. Illustration of a typical situation when the smallest eigenvalues are sought, and R is small enough so that $\widehat{\lambda}_i \approx \lambda_i$. While the small gap is $\mathrm{gap}_i = \min_{j \neq i} |\widehat{\lambda}_i - \lambda_j| \approx \min_{j \neq i} |\widehat{\lambda}_i - \widehat{\lambda}_j|$, the big Gap is much bigger $\mathrm{Gap}_i = \min|\lambda_i - \lambda(A_3)|$.

In addition to gap_i and Gap_i, some of the bounds we derive involve $|\lambda_i - \widehat{\lambda}_j|$ for a fixed $j \in \{1, \ldots, k\} \setminus i$. These lie between gap_i and Gap_i.

Recalling (2.1), the information clearly available after R-R are the Ritz pairs $(\widehat{\lambda}_i, \widehat{x}_i)$ for $i = 1, \ldots, k$ and the norms of the individual r_i, because they are equal to the residuals $\|r_i\| = \|A\widehat{x}_i - \widehat{\lambda}_i \widehat{x}_i\|$. In addition, one can reasonably expect that an estimate (or better yet, a lower bound) is available for Gap_i for each i, or at least for small i: when the smallest k eigenpairs are sought, the trial subspace Q—assuming it has been chosen appropriately by the algorithm used—is expected to be rich in the eigenspace corresponding to those eigenvalues. It then follows from standard eigenvalue perturbation theory that A_3 contains only eigenvalues that are roughly at least as large as $\lambda_{k+1}(A)$ (up to $\|R\|$, or indeed $\frac{\|R\|^2}{\mathrm{gap}}$ [12]). Therefore, although the exact value of $\lambda_{k+1}(A)$ is unknown, we can use the knowledge of the

---

[1]There is a subtle difference between $\mathrm{Gap}_i, \mathrm{gap}_i$ in (2.3) and gap_c in (1.1), in that the former is the difference between an exact desired eigenvalue and approximate undesired eigenvalues, whereas gap_c is the opposite. Our forthcoming analysis uses (2.3), and we return to this difference in Remark 5.4. Also note that we do not explicitly assume that $\mathrm{Gap}_i \geq \mathrm{gap}_i$; the bounds continue to hold without such assumptions, although when they do not hold, the bounds can be worse than Davis-Kahan.

[2]When the target eigenpairs are the smallest or largest ones, it also becomes easier to obtain lower bounds for gap_1 and Gap_i. For example, since the Courant-Fisher minimax theorem $\lambda_i \leq \widehat{\lambda}_i$, we have $\mathrm{gap}_1 \geq \min_{j \in \{1, \ldots, k\} \setminus 1} |\widehat{\lambda}_1 - \widehat{\lambda}_j|$, which involves only computed quantities. Similarly, we have $\mathrm{Gap}_i \geq \min|\widehat{\lambda}_i - \lambda(A_3)|$.

Ritz values $\widehat{\lambda}_i$ to estimate Gap_i, for example $\mathrm{Gap}_i \approx |\widehat{\lambda}_i - \widehat{\lambda}_{k+1}|$ or $\mathrm{Gap}_i \gtrsim |\widehat{\lambda}_i - \widehat{\lambda}_k|$; we use the latter, approximate lower bound in our experiments. Similarly, one can estimate gap_i for example as $\mathrm{gap}_i \approx \min_{j \neq i} |\widehat{\lambda}_i - \widehat{\lambda}_j|$.

In practice, an important feature of the residuals is that they are typically graded: $\|r_1\| \ll \|r_2\| \ll \cdots \ll \|r_k\|$. This is because the extremal eigenvalues converge much faster than interior ones; a fact deeply connected with polynomial (and rational) approximation theory [20, § 33]. We derive bounds (e.g., Theorem 4.1) that respect this property, and hence give sharp bounds in practical situations.

We note that previous bounds exist that involve the big Gap_i rather than gap_i; most notably (aside from Ovtchinnikov's result [16] mentioned in the introduction) Davis-Kahan's generalized $\sin\theta$ theorem where the angles between subspaces of different dimensions are bounded [2, Thm. 6.1]. In this case, however, (in addition to comparing, e.g., a vector and a subspace rather than two vectors) the numerator is replaced by the entire $\|R\|$ rather than the ith column $\|r_i\|$. The bounds we derive essentially show that, up to a small constant, (i) the small gap_i in (1.1) can be replaced by the big Gap_i, and (ii) the numerator is the ith column $\|r_i\|$. These combined give a massively improved error bound for $\widehat{x}_i$, especially for small values of i. The next section illustrates the first aspect, and the second will be covered in Section 4.

3. 2×2 PARTITIONING

We will derive three error bounds for Ritz vectors; the first, obtained in this section, is simple and vividly illustrates the roles of gap and Gap, but not sharp in a practical setting. In Section 4 we derive two more bounds that give better bounds in practice.

Here we consider a simplified 2×2 block partitioning of (2.1) where

$$(3.1) \qquad \widetilde{A}(= [Q\Omega, Q_\perp]^* A [Q\Omega, Q_\perp]) = \begin{bmatrix} \widehat{\Lambda}_1 & R^* \\ R & A_3 \end{bmatrix},$$

where $\widehat{\Lambda}_1 = \mathrm{diag}(\widehat{\lambda}_1, \widehat{\Lambda}_2) \in \mathbb{R}^{k \times k}$ and $\|R\| = \|A\widehat{X} - \widehat{X}\widehat{\Lambda}_1\|$ are the computed quantities. In other words, we do not distinguish the columns r_i of R but treat $\|R\|$ as a single residual term.

Below, we derive bounds for $\sin \angle(x_i, \widehat{x}_i)$ applicable to $i = 1, \ldots, k$. In our analysis, we assume that $\widehat{\lambda}_i$ is the $(1,1)$ element of $\widetilde{A}$. This simplifies the discussion and loses no generality as we can permute the leading $k \times k$ block of $\widetilde{A}$. Moreover, we drop the subscript i in the remainder of this section for simplicity.

Theorem 3.1. *Let A be a Hermitian matrix as in (2.1), for which $(\widehat{\lambda}, \widehat{x})$ is a Ritz pair with $\widehat{\lambda} = \widehat{\lambda}_1$. Let (λ, x) be an eigenvector of A, and let $\mathrm{Gap} = \min |\lambda - \lambda(A_3)|$ and $\mathrm{gap} = \min |\lambda - \lambda(\widehat{\Lambda}_2)|$. Then writing $R_2 = [r_2, r_3, \ldots, r_k]$, we have*

$$(3.2) \qquad \sin \angle(x, \widehat{x}) \leq \frac{\|R\|}{\mathrm{Gap}} \sqrt{1 + \frac{\|R_2\|^2}{\mathrm{gap}^2}} \quad \left(\leq \frac{\|R\|}{\mathrm{Gap}} \left(1 + \frac{\|R_2\|}{\mathrm{gap}}\right) \right).$$

Note that clearly $\|R_2\| \leq \|R\|$, so the result implies $\sin \angle(x, \widehat{x}) \leq \frac{\|R\|}{\mathrm{Gap}} \sqrt{1 + \frac{\|R\|^2}{\mathrm{gap}^2}}$ $\leq \frac{\|R\|}{\mathrm{Gap}} (1 + \frac{\|R\|}{\mathrm{gap}})$.

Proof. Let $\widetilde{x} = \begin{bmatrix} w \\ y \\ z \end{bmatrix}$ be an eigenvector of $\widetilde{A}$ as in (3.1) such that $\widetilde{A} \begin{bmatrix} w \\ y \\ z \end{bmatrix} = \lambda \begin{bmatrix} w \\ y \\ z \end{bmatrix}$ with $w \in \mathbb{C}, y \in \mathbb{C}^{k-1}$. Then since $\sin \angle(x, \widehat{x}) = \|[\begin{smallmatrix} y \\ z \end{smallmatrix}]\|$ from (2.2), the goal is to bound $\|y\|$ and $\|z\|$. The bottom part of $\widetilde{A}\widetilde{x} = \lambda\widetilde{x}$ gives

$$(\lambda I_{n-k} - A_3)z = R \begin{bmatrix} w \\ y \end{bmatrix},$$

from which we obtain

(3.3) $\|z\| \leq \|(\lambda I_{n-k} - A_3)^{-1}\| \left\| R \begin{bmatrix} w \\ y \end{bmatrix} \right\| \leq \|(\lambda I_{n-k} - A_3)^{-1}\| \|R\| = \dfrac{\|R\|}{\mathrm{Gap}}.$

Note that the denominator is Gap, not gap. We also note that the final bound is Davis-Kahan's generalized $\sin\theta$ theorem where subspaces of different sizes ($\widehat{x}$ and the $[x_1, \ldots, x_k]$) are compared (and when the perturbation is off-diagonal); in fact, we can also obtain $\|z\| \leq \frac{\|R\|}{\mathrm{Gap}} \|[\begin{smallmatrix} w \\ y \end{smallmatrix}]\|$, which is the generalized $\tan\theta$ theorem.

From the second block of $\widetilde{A}\widetilde{x} = \lambda\widetilde{x}$ we have

$$(\lambda I_{k-1} - \widehat{\Lambda}_2)y = R_2^* z,$$

and since $\|(\lambda I_{k-1} - \widehat{\Lambda}_2)y\| \geq \mathrm{gap}\, \|y\|$, we obtain the important bound

(3.4) $\|y\| \leq \dfrac{\|R_2\|}{\mathrm{gap}} \|z\|.$

Combining with (3.3) we obtain

(3.5) $\|y\| \leq \dfrac{\|R\|\|R_2\|}{\mathrm{gap} \cdot \mathrm{Gap}}.$

Therefore, we conclude that

$$\sin \angle(x, \widehat{x})^2 = \left\| \begin{bmatrix} y \\ z \end{bmatrix} \right\|^2 \leq \dfrac{\|R\|^2}{\mathrm{Gap}^2} \left(1 + \dfrac{\|R_2\|^2}{\mathrm{gap}^2} \right),$$

giving (3.2).   □

We make several remarks regarding the theorem.

Remark 3.1 (Qualitative behavior of bounds). Theorem 3.1 shows that

- if $\|R\| \leq \mathrm{gap}$, then $\sin \angle(x, \widehat{x}) \lesssim \frac{\|R\|}{\mathrm{Gap}}$,
- if $\|R\| \geq \mathrm{gap}$, then $\sin \angle(x, \widehat{x}) \lesssim \frac{\|R\|^2}{\mathrm{gap} \cdot \mathrm{Gap}}$.

Note how Davis-Kahan's bound is insufficient to explain these: When $\|R\| \leq \mathrm{gap}$, we improve the bound (1.1) by a factor $\mathrm{gap}\,/\,\mathrm{Gap}$, which is typically $\ll 1$. Moreover, when $\|R\| \geq \mathrm{gap}$, classical results suggest $\widehat{x}$ may have no accuracy at all. Nonetheless, Theorem 3.1 shows that there is still a nontrivial bound for $\sin \angle(x, \widehat{x})$ as long as $\|R\| \lesssim \sqrt{\mathrm{gap} \cdot \mathrm{Gap}}(\gg \mathrm{gap})$. These results are particularly relevant when only low-accuracy solutions are available, so that $\|R\|$ is much larger than working precision.

Remark 3.2 (Effect of finite precision arithmetic). Crucial in the above argument is that $\widehat{\Lambda}_1$ has zero off-diagonal elements. In practice in finite-precision arithmetic, the Rayleigh-Ritz process inevitably results in $\widehat{\Lambda}_1$ in (3.1) with off-diagonal elements that are $O(u)$ instead of 0, due to roundoff errors (assuming for simplicity $\|A\| = O(1)$). It is therefore important to address how they affect the bounds. As mentioned in the introduction, classical perturbation theory shows that these $O(u)$

terms will perturb $\widehat{x}$ by up to $O(u/\operatorname{gap})$. Since the off-diagonal $O(u)$ elements in $\widehat{\Lambda}_1$ indeed lie in the directions that perturb the eigenvector the most (we return to this in Section 4.2), to account for roundoff errors we will need to add the term $O(u/\operatorname{gap})$ to the bound (3.2). This remark becomes important especially when $\|R\|$ is small, so that $O(u/\operatorname{gap})$ is not negligible relative to $\frac{\|R\|}{\operatorname{Gap}}$. In other words, the folklore that eigenvectors cannot be computed with precision higher than u/gap is true; what we refute is the belief that the bound $\|R\|/\operatorname{gap}$ (or $\|r\|/\operatorname{gap}$) is sharp— our result shows that when $\|R\| > \operatorname{gap}$ but $\|R\| < \operatorname{Gap}$, Rayleigh-Ritz computes eigenvectors of much higher accuracy than $\|R\|/\operatorname{gap}$.

Remark 3.3 (Different partitionings). We can obtain different bounds depending on where to partition, that is, we can invoke the bound (3.2) by taking a $k \leftarrow k'$ for some $k' \leq k$. Each choice of k' gives a different bound, since each gives different values of $\|R\|$ and Gap (along with gap, though its dependence on k' is usually much less significant). If the computational cost is not a concern, one can compute all possible partitionings and take the smallest bound obtained. However, the bounds in Section 4 are often still better in practice.

Remark 3.4 (Proof via generalized Davis-Kahan and Saad). The result (3.2) can also be derived by combining (i) Saad's bound [18, Thm. 4.6], which bounds $\sin \angle(\widehat{x}_i, x_i)$ relative to $\sin \angle(Q, x_i)$, the angle between the desired eigenvector and the trial subspace, and (ii) the generalized Davis-Kahan $\sin \theta$ theorem [2], in which two subspaces of different dimensions are compared. Here we presented a first-principles derivation, as we use the same line of arguments to derive improved and generalized bounds in the forthcoming sections. Also noteworthy is Knyazev's paper [10], which generalizes Saad's bound to subspaces. He also shows that Ritz vectors contain quadratically small components in eigenvectors approximated by the other Ritz vectors. This is essentially captured in (3.4), which indicates $\|y\| = O(\|R\|^2)$ (absorbing the gaps in the constant). We revisit this phenomenon for subspaces in Section 5.

While we will not repeat them, Remarks 3.2 and 3.3 are relevant throughout the paper.

3.1. **Experiments.** To illustrate Theorem 3.1, we conduct the following experiment; throughout, all experiments were carried out in MATLAB version R2017a using IEEE double precision arithmetic with unit roundoff $\approx 1.1 \times 10^{-16}$. Let

$$A = \begin{bmatrix} \widehat{\Lambda}_1 & R^* \\ R & A_3 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

where $n = 10$ (the precise size of n is insignificant), $\widehat{\Lambda}_1 = -\begin{bmatrix} 1+\operatorname{gap} & \\ & 1 \end{bmatrix}$ and $A_3 \succeq 0$, so that Gap ≥ 1. We take $R \in \mathbb{R}^{k \times 2}$ to be randomly generated matrices using MATLAB's `randn` function, scaled so that $\|R\|$ is fixed to a value 10^{-i} for $i = 0, \ldots, 15$. For each i, we generate 100 such matrices R, and find the largest value of $\sin \angle(x, \widehat{x})$ from the 100 runs (note that $\widehat{x} = [1, 0, \ldots, 0]^T$ by construction). These are shown as "observed" in Figure 3.1, along with (i) the classical bound $\|R\|/\operatorname{gap}$, (ii) the new bound (3.2), (iii) the bound u/gap, in view of Remark 3.2.

In view of Remark 3.2, $\sin\angle(x,\widehat{x})$ is bounded by the maximum of (3.2) and (a small multiple of) u/gap. Of course, we always have the trivial bound $\sin\angle(x,\widehat{x}) \leq 1$, so putting these together, we have the following bound in finite-precision arithmetic:

$$(3.6) \qquad \sin\angle(x,\widehat{x}) \leq \min\left(1, \max\left(O(\frac{u}{\operatorname{gap}}), \frac{\|R\|}{\operatorname{Gap}}\sqrt{1+\frac{\|R\|^2}{\operatorname{gap}^2}}\right)\right).$$

We observe in Figure 3.1 that this is indeed the case, and the new bound (3.2) gives remarkably sharp bounds for the observed values of $\sin\angle(x,\widehat{x})$ (when it is not dominated by u/gap, and gives a nontrivial bound ≤ 1). This is despite the fact that we are plotting the looser bound $\frac{\|R\|}{\operatorname{Gap}}\sqrt{1+\frac{\|R\|^2}{\operatorname{gap}^2}}$ with $\|R_2\|$ in (3.2) replaced by $\|R\|$, as using $\|R_2\|$ makes the bound depend on the particular random instance of R.

As discussed above, the new bound (3.2) has two asymptotic behaviors: $\approx \|R\|/\operatorname{Gap}$ when $\|R\| \leq \operatorname{gap}$, and $\approx \|R\|^2/(\operatorname{gap}\cdot\operatorname{Gap})$ when $\|R\| \geq \operatorname{gap}$. This can be seen in the plots, as the change of slope in the new bound around $\|R\| \approx \operatorname{gap}$. From the plots with $\operatorname{gap} = 10^{-3}$ and 10^{-5}, we see that this transition also reflects the observed values of $\sin\angle(x,\widehat{x})$ quite accurately. In all cases, the classical Davis-Kahan bound $\|R\|/\operatorname{gap}$ tends to be severe overestimates (and $\|r\|/\operatorname{gap}$ as in (1.1) is not much different), and the new bound can provide nontrivial information (bound smaller than 1) even when the Davis-Kahan bound is useless with $\|R\|/\operatorname{gap} > 1$, and the difference between Davis-Kahan and the new bound widens when gap is small.

4. Improved error bounds for Ritz vectors

The above experiments illustrate the sharpness of the bound (3.2) given the information $\|R\| = \|[r_1, \ldots, r_k]\|$ and $\widehat{\lambda}_i$, along with $\min(\operatorname{eig}(A_3))$. When applied in practice, however, we find that the bound (3.2) is usually a severe overestimate, as we illustrate in Section 4.1. The reason is that it does not distinguish r_1 from r_k (say), while typically we have $\|r_1\| \ll \|r_k\|$, reflecting the difference in speed with which each Ritz pair converges, typically the extremal ones converging first.

As noted in Section 2.1, after R-R one also has information on the individual norms $\|r_i\| = \|A\widehat{x}_i - \widehat{\lambda}_i\widehat{x}_i\|$. Here we derive bounds that are essentially sharp using all the information. We shall show that if $\|r_i\|$ are sufficiently small, then $\sin\angle(x,\widehat{x}) \lesssim \frac{\|r_1\|}{\operatorname{Gap}_1}$. This is usually a massive improvement over (3.2), and essentially sharp: we cannot improve the bound below $\frac{\|r_1\|}{\operatorname{Gap}_1}$. The argument is similar to Theorem 3.1 but with more elaborate manipulations. The strategy is the same: bound $\|y\|$ in terms of $\|z\|$, and use this to bound $\left\|\begin{bmatrix} y \\ z \end{bmatrix}\right\|$.

Theorem 4.1. *In the setting of Theorem* 3.1,

- *If* $\operatorname{Gap} > \frac{\|R_2\|^2}{\operatorname{gap}}$, *then*

$$(4.1) \qquad \sin\angle(x,\widehat{x}) \leq \frac{\|r_1\|}{\operatorname{Gap} - \frac{\|R_2\|^2}{\operatorname{gap}}}\sqrt{1+\frac{\|R_2\|^2}{\operatorname{gap}^2}}.$$
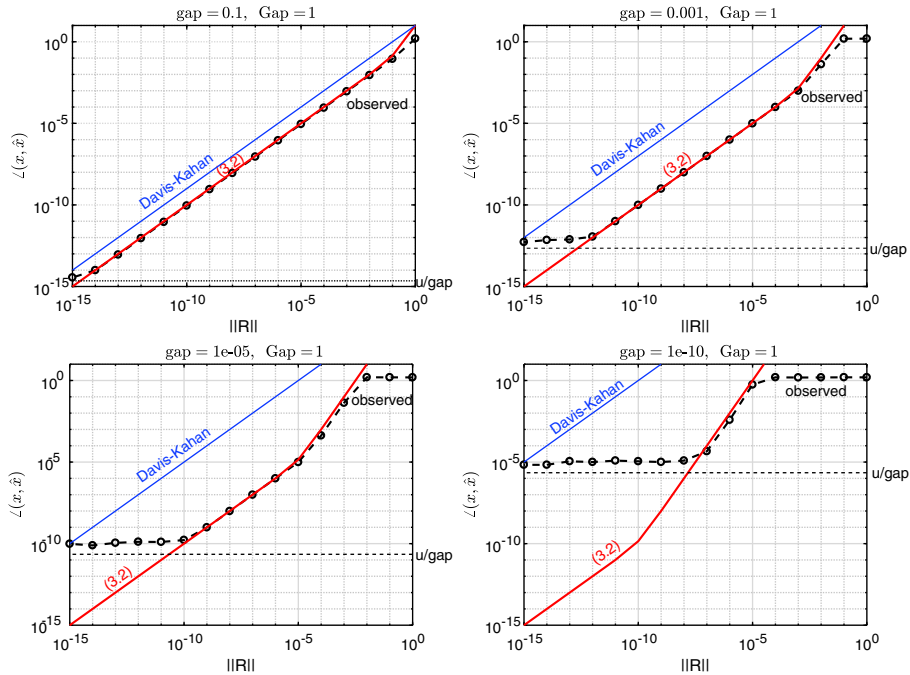
FIGURE 3.1. Illustration of our bound (3.2) (red; with R_2 replaced by R), varying gap (upper-left: gap $= 10^{-1}$, upper-right: gap $= 10^{-3}$, lower-left: gap $= 10^{-5}$, lower-right: gap $= 10^{-10}$). Observe how sharp (3.2) is, relative to the classical Davis-Kahan bound $\|R\|/$ gap. When $\|R\| \leq u/$ gap, the bound in finite-precision arithmetic would be the maximum between the new bound and $u/$ gap (dashed black, constant line); see (3.6).

- *If* $\mathrm{Gap} > \sum_{i=2}^{k} \frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}$, *then*

$$(4.2) \qquad \sin \angle(x, \widehat{x}) \leq \frac{\|r_1\|}{\mathrm{Gap} - \sum_{i=2}^{k} \frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}} \sqrt{1 + \left(\sum_{i=2}^{k} \frac{\|r_i\|}{|\lambda - \widehat{\lambda}_i|} \right)^2}.$$

Proof. We first prove (4.1). The main idea is to improve the bound (3.3) on $\|z\|$. As before we have $(\lambda I_{k-1} - \widehat{\Lambda}_2)y = R_2^* z$, so $\|y\| \leq \frac{\|R_2\|\|z\|}{\mathrm{gap}}$. We also have

$$(\lambda I_{n-k} - A_3)z = [r_1 \ R_2] \begin{bmatrix} w \\ y \end{bmatrix}.$$

This gives $(\lambda I_{n-k} - A_3)z - R_2 y = r_1 w$, hence $\|(\lambda I_{n-k} - A_3)z\| - \|R_2 y\| \leq \|r_1 w\|$. Using $\|y\| \leq \frac{\|R_2\|\|z\|}{\mathrm{gap}}$ we obtain

$$\|(\lambda I_{n-k} - A_3)z\| - \frac{\|R_2\|^2 \|z\|}{\mathrm{gap}} \leq \|r_1 w\|.$$

Noting that $\sigma_{\min}(\lambda I_{n-k} - A_3) = \text{Gap}$, we have $\|(\lambda I_{n-k} - A_3)z\| \geq \text{Gap}\,\|z\|$, hence

$$(\text{Gap} - \frac{\|R_2\|^2}{\text{gap}})\|z\| \leq \|r_1 w\|.$$

Using the assumption $\text{Gap} > \frac{\|R_2\|^2}{\text{gap}}$ and the trivial bound $\|w\| \leq 1$ we obtain

$$\|z\| \leq \frac{\|r_1\|}{\text{Gap} - \frac{\|R_2\|^2}{\text{gap}}}.$$

This together with $\|y\| \leq \frac{\|R_2\|\|z\|}{\text{gap}}$ yields

$$\sin\angle(x,\widehat{x}) = \left\|\begin{bmatrix} y \\ z \end{bmatrix}\right\| \leq \frac{\|r_1\|}{\text{Gap} - \frac{\|R_2\|^2}{\text{gap}}}\sqrt{1 + \frac{\|R_2\|^2}{\text{gap}^2}},$$

giving (4.1).

The remaining task is to prove (4.2). The idea is to improve the bound (3.4) on $\|y\|$, or rather its individual entries, using

$$(\lambda I_{k-1} - \widehat{\Lambda}_2)y = R_2^* z.$$

Writing $y = [y_2, \ldots, y_k]^T$, the ith ($i = 2, \ldots, k$) element gives $(\lambda - \widehat{\lambda}_i)y_i = r_i^* z$, hence

$$(4.3) \qquad |y_i| = \frac{|r_i^* z|}{|\lambda - \widehat{\lambda}_i|} \leq \frac{\|r_i\|\|z\|}{|\lambda - \widehat{\lambda}_i|}, \quad i = 2, \ldots, k.$$

We also have

$$(\lambda I_{n-k} - A_3)z = [r_1, r_2, \ldots, r_k]\begin{bmatrix} w \\ y_2 \\ \vdots \\ y_k \end{bmatrix}.$$

This gives $(\lambda I_{n-k} - A_3)z - R_2 y = r_1 w$, and

$$(4.4) \qquad \|(\lambda I_{n-k} - A_3)z\| - \sum_{i=2}^{k}\|r_i y_i\| \leq \|r_1 w\|,$$

so using (4.3) we obtain

$$\|(\lambda I_{n-k} - A_3)z\| - \sum_{i=2}^{k}\frac{\|r_i\|^2\|z\|}{|\lambda - \widehat{\lambda}_i|} \leq \|r_1 w\|.$$

Again using $\|(\lambda I_{n-k} - A_3)z\| \geq \text{Gap}\,\|z\|$, we therefore obtain

$$(\text{Gap} - \sum_{i=2}^{k}\frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|})\|z\| \leq \|r_1 w\|.$$

Hence, using the assumption $\text{Gap} > \sum_{i=2}^{k}\frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}$ and the trivial bound $\|w\| \leq 1$ we obtain

$$\|z\| \leq \frac{\|r_1\|}{\text{Gap} - \sum_{i=2}^{k}\frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}}.$$

The fact $\sin\angle(x,\widehat{x}) = \left\|\begin{bmatrix} y \\ z \end{bmatrix}\right\|$ together with (4.3) completes the proof of (4.2). $\quad\square$

Note that since the bounds $\|y\| \leq \frac{\|R_2\|\|z\|}{\text{gap}}$ and (4.3) are both valid, in both bounds (4.1) and (4.2), the term with the square root can be replaced with the minimum, that is, $\sqrt{1 + \min\left(\frac{\|R_2\|^2}{\text{gap}^2}, \left(\sum_{i=2}^{k} \frac{\|r_i\|}{|\lambda - \widehat{\lambda}_i|}\right)^2\right)}$. This applies also to the bounds to follow, but for brevity we do not repeat this remark.

We also note that the bounds (4.1) and (4.2) are not comparable. The bound (4.1) involves the small gap, which (4.2) avoids to some extent by using the individual residuals $\|r_i\|$; however, the heavy use of triangular inequalities in the bound (4.4) suggests (4.1) can still be a significant overestimate. The "sharpest" bound one can obtain would be via directly bounding the norm $\|y\| = \|(\lambda I_{k-1} - \widehat{\Lambda}_2)^{-1}R_2^* z\|$. Nonetheless, experiments suggest (4.2) is often a good bound, as we illustrate now.

4.1. **Experiments.** We illustrate Theorem 4.1 with experiments more practical than Section 3.1. We let $A \in \mathbb{R}^{1000 \times 1000}$ be the classical tridiagonal matrix with 2 on the diagonal and -1 on the super- and sub-diagonals. This is a 1D Laplacian matrix, obtained by finite difference discretization. We then run the LOBPCG algorithm [11] to compute the smallest eigenpair with a random initial guess, working with a $k = 50$-dimensional subspace.

Figure 4.1 (left) shows the convergence of $\sin \angle(\widehat{x}_1, x_1)$ along with four bounds: Davis-Kahan's $\sin \angle(\widehat{x}_1, x_1) \leq \frac{\|r_1\|}{\text{gap}}$, (4.1), (4.2), and (3.2) from the previous section. Some data are missing for (4.1) and (4.2) in the early steps as they violated the assumption Gap $> \frac{\|R_2\|^2}{\text{gap}}$ or Gap $> \sum_{i=2}^{k} \frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}$; note that these assumptions can be checked inexpensively. To estimate Gap and gap we used the available quantities Gap $\gtrsim \min|\widehat{\lambda}_1 - \widehat{\lambda}_k|$, gap $\approx \min|\widehat{\lambda}_1 - \widehat{\lambda}_2|$; the plots look nearly identical if the exact values are used.
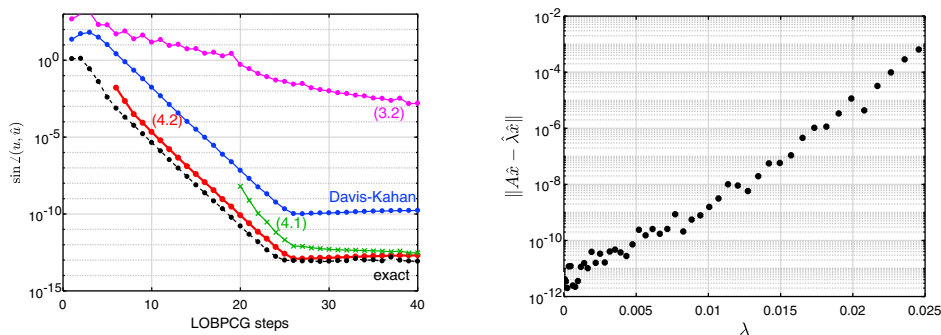
FIGURE 4.1. Left: convergence of $\sin \angle(\widehat{x}_1, x_1)$ (shown as exact), and its bounds (4.1), (4.2) and the Davis-Kahan bound (1.1). Right: scatterplot of λ vs. residuals $\|r_i\| = \|A\widehat{x}_i - \widehat{\lambda}_i\widehat{x}_i\|$ for $i = 1, 2, \ldots, k = 50$, after 20 LOBPCG iterations. Note how $\|r_i\|$ are graded $\|r_i\| \ll \|r_j\|$ for $i \ll j$.

We make several observations. First, (4.2) gave sharp bounds for $\sin \angle(\widehat{x}_1, x_1)$ when applicable. For example after eight LOBPCG iterations, Davis-Kahan's $\sin \theta$ theorem gives bounds > 1, suggesting $\widehat{x}_1$ may have no accuracy at all. Nonetheless, (4.2) correctly shows that it has at least accuracy $\lesssim 10^{-3}$. Second, the bound (3.2) is poor throughout, because it takes the entire residual matrix norm $\|R\|$ in the numerator, without respecting the fact that the residuals $\|r_i\| = \|A\widehat{x}_i - \widehat{\lambda}_i\widehat{x}_i\|$ are typically graded and hence $\|r_1\| \ll \|R\|$, as illustrated in Figure 4.1 (right). Finally, the asymptotic behavior of the bounds as

$\|R\| \to 0$ (many LOBPCG steps) are also in stark contrast. This is because up to first order in $\|R\|$, (4.1) and (4.2) are $\frac{\|r_1\|}{\text{Gap}}$, whereas Davis-Kahan involves the smaller gap $\frac{\|r_1\|}{\text{gap}}$.

4.2. **Structured condition number.** Here we interpret Theorem 4.1 from the standpoint of perturbation theory. Namely, we regard R in (2.1) as a perturbation to the block diagonal matrix $\widetilde{A}_0 := \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k, A_3)$ having eigenvectors $e_1, \ldots, e_k$ (besides others), the first k canonical vectors. Examining $\angle(x_i, \widehat{x}_i)$ is equivalent to examining how much the eigenvector e_i of $\widetilde{A}_0$ gets perturbed by R.

In the opening we mentioned that the condition number of an eigenvector is $1/\text{gap}_i$. That is, there exists a perturbation E such that $\widetilde{A}_0 + E$ has an eigenvector $\widehat{e}_i$ with

$$(4.5) \qquad \sin \angle(e_i, \widehat{e}_i) = \frac{\|E\|}{\text{gap}_i} + O(\|E\|^2).$$

Yet, the two bounds in Theorem 4.1 show that writing $R := \widetilde{A} - \widetilde{A}_0$ (slightly and harmlessly abusing notation), we have

$$(4.6) \qquad \sin \angle(e_i, \widehat{e}_i) = \frac{\|r_i\|}{\text{Gap}_i} + O(\|R\|^2).$$

Note the two changes, both potentially significant: first, gap is replaced by Gap. Second, the norm of the entire perturbation $\|E\|$ is replaced by the individual $\|r_i\|$, the perturbation only in the ith column of $\widetilde{A}_0$.

An explanation of this effect can be made via structured perturbation analysis. In R-R, the perturbation R in (2.1) is highly structured in two ways: the nonzero pattern, and the grading of $\|r_i\|$. For example, the perturbation E that would perturb the eigenvector e_1 the most is the $(1,2)$ and $(2,1)$ elements in $\widetilde{A}_0$, as they connect the eigenvalues λ_1 and λ_2, resulting in the (unstructured) condition number $1/|\lambda_1 - \lambda_2| = 1/\text{gap}_1$. However, these are forced to be zero by the R-R construction. Within the structured perturbation allowed in R-R, e_1 is perturbed most by the $(k+1, 1)$ and $(1, k+1)$ elements, assuming for the moment A_3 is diagonalized. These elements connect the eigenvalues λ_1 and $\min(\lambda(A_3)) \approx \lambda_{k+1}$, resulting in the structured condition number $1/\text{Gap}_1$. Regarding the grading of $\|r_i\|$, the r_j $(j \neq i)$ terms have no effect on $\widehat{e}_i$ up to $O(\|r_j\|^2)$, making r_i the only term that affects the leading term in (4.6).

5. Bounds for invariant subspaces

We now turn to bounding errors for invariant subspaces spanned by more than one eigenvector. Besides being the natural object in many applications, it is sometimes necessary to resort to subspaces instead of individual eigenvectors, when multiple or near-multiple eigenvalues are present. For example, if gap $= O(u)$, none of the above bounds would be useful, as the $O(\frac{u}{\text{gap}})$ term in (3.6) due to roundoff errors is always present. Below we derive bounds that give useful information in such cases.

We briefly recall the definition of angles between subspaces. The angles $\{\theta_i\}_{i=1}^{k_1}$ between two subspaces spanned by $X \in \mathbb{C}^{n \times k_1}, Y \in \mathbb{C}^{n \times k_1}$ with orthonormal columns are defined by $\theta_i = \text{acos}(\sigma_i(X^*Y))$ and denoted by $\angle(X, Y)$; they are known as the canonical angles or principal angles [5, Thm. 6.4.3]. Equivalently, we have $\sin \theta_i = \sigma_i(X_\perp^* Y)$ (as can be verified, e.g., via the CS decomposition [5, Thm. 2.5.2]), which is what we use below (and used above for $k_1 = 1$ to obtain (2.2)).

To clarify the situation, rewrite (2.1) as

$$(5.1) \qquad \widetilde{A} := [\widehat{X}_1 \ \widehat{X}_2 \ \widehat{X}_3]^* A [\widehat{X}_1 \ \widehat{X}_2 \ \widehat{X}_3] = \begin{bmatrix} \widehat{\Lambda}_1 & 0 & R_1^* \\ 0 & \widehat{\Lambda}_2 & R_2^* \\ R_1 & R_2 & A_3 \end{bmatrix},$$

where $[\widehat{X}_1 \; \widehat{X}_2 \; \widehat{X}_3]$ is an orthogonal matrix, with $[\widehat{X}_1 \; \widehat{X}_2] = Q\Omega$, $\widehat{X}_1 \in \mathbb{C}^{n \times k_1}, \widehat{X}_2 \in \mathbb{C}^{n \times (k-k_1)}$, and $\widehat{X}_3 \in \mathbb{C}^{n \times (n-k)}$. Our goal is to bound $\||\sin \angle (\widehat{X}_1, X_1)\||$ from above, where $X_1 \in \mathbb{C}^{n \times k_1}$ is a matrix of k_1 exact eigenvectors of A, i.e., $AX_1 = X_1 \Lambda_1$. Defining $\widetilde{X}_1 = [\widehat{X}_1 \; \widehat{X}_2 \; \widehat{X}_3]^* X_1$, we have $\widetilde{A}\widetilde{X}_1 = \widetilde{X}_1 \Lambda_1$, so the columns of $\widetilde{X}_1$ are eigenvectors of $\widetilde{A}$. With the partitioning $\widetilde{X}_1 = \begin{bmatrix} W \\ Y \\ Z \end{bmatrix}$ with $W \in \mathbb{C}^{k_1 \times k_1}, Y \in \mathbb{C}^{(k-k_1) \times k_1}, Z \in \mathbb{C}^{(n-k) \times k_1}$, it therefore follows that

$$\left\| \left| \sin \angle (\widehat{X}_1, X_1) \right| \right\| = \left\| \left| [\widehat{X}_2 \; \widehat{X}_3]^* X_1 \right| \right\| = \left\| \left| \begin{bmatrix} Y \\ Z \end{bmatrix} \right| \right\|.$$

This extends (2.2), and is a key identity in the forthcoming analysis.

Sometimes we deal with the angles between subspaces of different dimensions, say $[\widehat{X}_1 \; \widehat{X}_2] \in \mathbb{C}^{n \times k}$ and $X_1 \in \mathbb{C}^{n \times k_1}$ with $k_1 \leq k$. In this case the angles are defined via $\sin \theta_i = \sigma_i(X_1^*([\widehat{X}_1 \; \widehat{X}_2]_\perp))$ for $i = 1, \ldots, k_1$.

Here is the extension of the previous bounds to invariant subspaces. Note that gap and Gap are redefined; we use the same notation as they reduce to the same values when $k_1 = 1$.

Theorem 5.1. *Let $A, \widetilde{A}$ be as in (5.1), with $(\widehat{\Lambda}_1, \widehat{X}_1)$ being k_1 Ritz pairs. Let (Λ_1, X_1) be a set of k_1 exact eigenpairs $AX_1 = X_1 \Lambda_1$. Let $\mathrm{Gap} = \min |\lambda(\Lambda_1) - \lambda(A_3)|$ and $\mathrm{gap} = \min |\lambda(\Lambda_1) - \lambda(\widehat{\Lambda}_2)|$. Then writing*

$$R = [R_1 \; R_2] := [R_1 \; r_{k_1+1}, \ldots, r_k] \in \mathbb{C}^{n \times k}$$

where $R_1 \in \mathbb{C}^{n \times k_1}$, we have

$$(5.2) \quad \left\| \left| \sin \angle (X, \widehat{X}) \right| \right\| \leq \frac{\||R\||}{\mathrm{Gap}} \left(1 + \frac{\|R_2\|}{\mathrm{gap}} \right), \qquad \| \sin \angle (X, \widehat{X}) \|_{2,F} \leq \frac{\|R\|_{2,F}}{\mathrm{Gap}} \sqrt{1 + \frac{\|R_2\|^2}{\mathrm{gap}^2}}.$$

Moreover, if $\mathrm{Gap} > \frac{\|R_2\|^2}{\mathrm{gap}}$, then

$$(5.3) \quad \left\| \left| \sin \angle (X, \widehat{X}) \right| \right\|_{2,F} \leq \frac{\|R_1\|_{2,F}}{\mathrm{Gap} - \frac{\|R_2\|^2}{\mathrm{gap}}} \sqrt{1 + \frac{\|R_2\|^2}{\mathrm{gap}^2}},$$

and if $\mathrm{Gap} > \sum_{i=k_1+1}^{k} \frac{\|r_i\|^2}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|}$, then

$$(5.4) \quad \left\| \left| \sin \angle (X, \widehat{X}) \right| \right\|_{2,F} \leq \frac{\|R_1\|_{2,F}}{\mathrm{Gap} - \sum_{i=k_1+1}^{k} \frac{\|r_i\|_2^2}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|}} \sqrt{1 + \left(\sum_{i=k_1+1}^{k} \frac{\|r_i\|}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|} \right)^2}.$$

Proof. The proof mimics that of Theorem 4.1, extending the discussion from vectors to subspaces. Let $\widetilde{X}_1 = [\widehat{X}_1, \widehat{X}_2, \widehat{X}_3] X_1 = \begin{bmatrix} W \\ Y \\ Z \end{bmatrix} \in \mathbb{C}^{n \times k_1}$ be an invariant subspace of $\widetilde{A}$ such that

$$(5.5) \quad \widetilde{A} \begin{bmatrix} W \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} W \\ Y \\ Z \end{bmatrix} \Lambda_1.$$

Then the bottom part of the equation gives

$$(5.6) \quad Z\Lambda_1 - A_3 Z = [R_1 \; R_2] \begin{bmatrix} W \\ Y \end{bmatrix} = R \begin{bmatrix} W \\ Y \end{bmatrix}.$$

Using a well-known bound for Sylvester's equations (e.g., [14, Lem. 2], [19, Ch. V]), along with the fact $\min(\lambda(\Lambda_1) - \lambda(A_3)) = \mathrm{Gap}$, we obtain

$$(5.7) \quad \||Z\|| \leq \frac{\||R\|| \left\| \left| \begin{bmatrix} W \\ Y \end{bmatrix} \right| \right\|}{\mathrm{Gap}} \leq \frac{\||R\||}{\mathrm{Gap}},$$

where for the last inequality we used the fact $\||XY\|| \leq \||X\|| \||Y\||$ [7, Cor. 3.5.10]. As in (3.3), this is the generalized Davis-Kahan $\sin\theta$ theorem.

From the second block of (5.5) we have

$$(5.8) \qquad\qquad Y\Lambda_1 - \widehat{\Lambda}_2 Y = R_2^* Z,$$

hence again from the Sylvester equation bound

$$(5.9) \qquad\qquad \||Y\|| \leq \frac{\|R_2\| \||Z\||}{\text{gap}}.$$

Together with (5.7) we obtain

$$(5.10) \qquad\qquad \||Y\|| \leq \frac{\||R\|| \|R_2\|}{\text{gap} \cdot \text{Gap}}.$$

Therefore, we conclude that

$$\left\||\sin\angle(X,\widehat{X})\right\|| = \left\|\left|\begin{bmatrix} Y \\ Z \end{bmatrix}\right\|\right| \leq \frac{\||R\||}{\text{Gap}}\left(1 + \frac{\|R_2\|}{\text{gap}}\right),$$

the first inequality in (5.2). For the spectral and Frobenius norms, using the stronger inequality

$$(5.11) \qquad\qquad \left\|\begin{bmatrix} A \\ B \end{bmatrix}\right\|_{2,F} \leq \sqrt{\|A\|_{2,F}^2 + \|B\|_{2,F}^2},$$

we obtain the second result in (5.2).

We next prove (5.3). From (5.6) we obtain $\||Z\Lambda_1 - A_3 Z\|| - \|R_2 Y\| \leq \||R_1 W\||$, hence using (5.9) we have

$$\||Z\Lambda_1 - A_3 Z\|| - \frac{\|R_2\|^2 \||Z\||}{\text{gap}} \leq \||R_1 W\||.$$

Again using the Sylvester equation bound $\||Z\Lambda_1 - A_3 Z\|| \geq \text{Gap}\,\||Z\||$, we therefore obtain

$$\left(\text{Gap} - \frac{\|R_2\|^2}{\text{gap}}\right)\||Z\|| \leq \||R_1 W\||.$$

Hence using the assumption $\text{Gap} > \frac{\|R_2\|^2}{\text{gap}}$ and the trivial bound $\|W\| \leq 1$ along with $\||R_1 W\|| \leq \||R_1\|| \|W\|$, we obtain

$$\||Z\|| \leq \frac{\||R_1\||}{\text{Gap} - \frac{\|R_2\|^2}{\text{gap}}}.$$

Finally, using (5.11) again we obtain

$$\left\||\sin\angle(X,\widehat{X})\right\||_{2,F} = \left\|\left|\begin{bmatrix} Y \\ Z \end{bmatrix}\right\|\right|_{2,F} \leq \frac{\|R_1\|_{2,F}}{\text{Gap} - \frac{\|R_2\|^2}{\text{gap}}}\sqrt{1 + \frac{\|R_2\|^2}{\text{gap}^2}},$$

giving (5.3).

It remains to establish (5.4). Taking the ith row of (5.8) gives $y_i\Lambda_1 - \widehat{\lambda}_{k_1+i}y_i = r_{k_1+i}^* Z$ for $i = 1, \ldots, k - k_1$, where y_i is the ith row of Y. Hence

$$(5.12) \qquad\qquad \||y_i\|| \leq \frac{\|r_{k_1+i}\| \||Z\||}{\min|\lambda(\Lambda_1) - \widehat{\lambda}_{k_1+i}|}, \quad i = 1, \ldots, k - k_1.$$

We also have $Z\Lambda_1 - A_3 Z = [R_1, r_{k_1+1}, \ldots, r_k] \begin{bmatrix} W \\ y_1 \\ \vdots \\ y_{k-k_1} \end{bmatrix}$. This gives $(Z\Lambda_1 - A_3 Z) -$

$[r_{k_1+1}, \ldots, r_k] \begin{bmatrix} y_1 \\ \vdots \\ y_{k-k_1} \end{bmatrix} = R_1 W$, so using (5.12) we obtain

$$(5.13) \qquad \|\|Z\Lambda_1 - A_3 Z\|\| - \sum_{i=k_1+1}^{k} \frac{\|r_i\|^2 \|\|Z\|\|}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|} \leq \|\|R_1 W\|\|.$$

Since $\|\|Z\Lambda_1 - A_3 Z\|\| \geq \|\|Z\|\|/\operatorname{Gap}$ as before, this gives

$$(\operatorname{Gap} - \sum_{i=k_1+1}^{k} \frac{\|r_i\|^2}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|})\|\|Z\|\| \leq \|\|R_1 W\|\| \leq \|\|R_1\|\|\|W\|.$$

Hence using the assumption $\operatorname{Gap} > \sum_{i=k_1+1}^{k} \frac{\|r_i\|^2}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|}$ and the trivial bound $\|W\| \leq 1$ we obtain

$$\|\|Z\|\| \leq \frac{\|\|R_1\|\|}{\operatorname{Gap} - \sum_{i=k_1+1}^{k} \frac{\|r_i\|^2}{\min |\lambda(\Lambda_1) - \widehat{\lambda}_i|}}.$$

We use the fact $\left\|\left\| \sin \angle(X, \widehat{X}) \right\|\right\| = \left\|\left\| \begin{bmatrix} Y \\ Z \end{bmatrix} \right\|\right\|$ together with (5.12) to complete the proof of (5.4), again using (5.11). $\qquad \square$

Four remarks are in order.

Remark 5.1 (Vector vs. subspace bounds). The $\| \cdot \|_{2,F}$ bounds in Theorem 5.1 reduce to the vector bounds in the previous sections by taking $k_1 = 1$. Thus they can be regarded as proper generalizations.

Remark 5.2 (Bounds for unitarily invariant norms). Just like the two bounds in (5.2), (5.3) and (5.4) have their counterparts applicable to any unitarily invariant norm, obtained by replacing the final factors of the form $\sqrt{1+b^2}$ by $1 + b$. The proof is identical up to the very end, where we use the bound $\left\|\left\| \begin{bmatrix} A \\ B \end{bmatrix} \right\|\right\| \leq \|\|A\|\| + \|\|B\|\|$ instead of (5.11).

Remark 5.3 (Question 10.3 by Davis-Kahan). At the end of their landmark paper, Davis and Kahan [2] suggest four open problems. Among them, Question 10.2 asks for an extension of their theorems to the case where $\mathbb{C}^n$ is split into a pair of three (instead of two) subspaces in two ways, X_1, X_2, X_3 (exact eigenspaces) and $\widehat{X}_1, \widehat{X}_2, \widehat{X}_3$ (approximate ones). Namely, using information such as the Ritz values and residuals, can one bound the subspace angles? We argue that the above results give an answer—the setting in (5.1) is precisely in this form, and Theorem 5.2 gives sharp bounds for $\|\| \sin \angle(X_1, \widehat{X}_1)\|\|$.

Remark 5.4 (On the definition of gap). As mentioned in footnote 1, there is a difference in what is measured between our gap, Gap in (2.3) and gap_c in the classical treatments [2],[17, Ch. 11]. For example, (5.2) reduces to $\frac{\|r_1\|}{\operatorname{Gap}}$ when $k_1 = k = 1$ (hence R_2 is empty), but Gap here is the difference between λ (an exact desired eigenvalue) and $\lambda(A_3)$ (approximations to undesired eigenvalues), and the same can be said of gap in (2.3). By contrast, gap_c in (1.1) is the distance between the approximate desired eigenvalue $\widehat{\lambda}$ and exact undesired eigenvalues. It turns out that for (1.1), both gaps are applicable—indeed, we can obtain (1.1) from Theorem 5.1, takinig $k_1 = k = n - 1$: note that $\|\sin \angle(x, \widehat{x})\| = \|\widehat{x}^T X_\perp\| = \|\sin \min \angle(X_\perp, \widehat{X}_\perp)\|$ (this can be verified, e.g., via the CS decomposition [5, Thm. 2.5.2]; here $[x, X_\perp]$ and $[\widehat{x}, \widehat{X}_\perp]$ are orthogonal), and invoke (5.2) taking $X \leftarrow X_\perp$, $\widehat{X} \leftarrow \widehat{X}_\perp$, and Λ_2 empty. Then Gap in (5.2) becomes

gap in (1.1), and since the residuals are related by $R = r_1^T$ their norms are the same $\|R\| = \|r_1\|$, so (5.2) reduces precisely to (1.1). In other words, the bound (1.1) holds regardless of which definition of gap is used. However, the analysis in this paper uses gap, Gap in (2.3), and we have not proved that our theorems hold with gap_c.

Remark 5.5 (Proof techniques). The reader might have also noticed that the proofs above are basically repeated applications of well-known norm inequalities in matrix analysis. One might then wonder, why do they appear to give stronger results than previous ones? The answer appears to lie in (2.1)—the simple but crucial unitary transformation from A to $\widetilde{A}$ that simplifies the task to bounding $\|Y\|, \|Z\|$, as in (2.2). By contrast, most classical results work with A and start from the residual equation $A\widehat{X}_1 - \widehat{X}_1\widehat{\Lambda}_1 = R$ and derive bounds on $\angle(X_1, \widehat{X}_1)$: for example the Davis-Kahan $\sin\theta$ theorem can be obtained essentially by left-multiplying X_3^T, taking X_2 empty. Knyazev [10, Sec. 4] employed ingenious techniques to obtain (among others) essentially (5.10). Obtaining "sharper" bounds like (5.13) in a similar manner appears to be highly challenging. Once we reformulate the problem as in (2.1)–(2.2), the derivation becomes significantly simpler (in the author's opinion).

6. SVD

We now present an SVD analogue of Theorem 5.1, deriving bounds for the accuracy of singular vectors and singular subspaces obtained by a Petrov-Galerkin projection method. Such methods proceed as follows: project A onto lower-dimensional trial subspaces spanned by $\widehat{U} \in \mathbb{C}^{m \times k_m}, \widehat{V} \in \mathbb{C}^{n \times k_n}$ having orthonormal columns (for how to choose $\widehat{U}, \widehat{V}$ see, e.g., [1, 6, 21]), compute the SVD of the small $k_m \times k_n$ matrix $\widehat{U}^* A \widehat{V} = \widetilde{U}\widehat{\Sigma}\widetilde{V}^*$ and obtain an approximate economical SVD as $A \approx (\widehat{U}\widetilde{U})\widehat{\Sigma}(\widehat{V}\widetilde{V})^*$, which is of rank $\leq \min(k_m, k_n)$. Some of the columns of $\widehat{U}\widetilde{U}$ and $\widehat{V}\widetilde{V}$ then approximate the exact left and right singular vectors of A. Our goal is to quantify their accuracy. We focus on the most frequently encountered case where an approximate SVD is sought, that is, the leading singular vectors are being approximated.

Theorem 6.1. *Let $A \in \mathbb{C}^{m \times n}$ with $m \geq n$, of the form*

$$(6.1) \qquad [\widehat{U}_1 \ \widehat{U}_2 \ \widehat{U}_3]^* A [\widehat{V}_1 \ \widehat{V}_2 \ \widehat{V}_3] = \begin{bmatrix} \widehat{\Sigma}_1 & 0 & R_1 \\ 0 & \widehat{\Sigma}_2 & R_2 \\ S_1 & S_2 & A_3 \end{bmatrix} =: \widetilde{A},$$

where $[\widehat{U}_1 \ \widehat{U}_2 \ \widehat{U}_3]$ and $[\widehat{V}_1 \ \widehat{V}_2 \ \widehat{V}_3]$ are square unitary, and $\widehat{\Sigma}_1 \in \mathbb{R}^{k_1 \times k_1}$, $\widehat{\Sigma}_2 \in \mathbb{R}^{(k_m - k_1) \times (k_n - k_1)}$ with $\begin{bmatrix} \widehat{\Sigma}_1 \\ \widehat{\Sigma}_2 \end{bmatrix}$ equal to $\begin{bmatrix} \mathrm{diag}(\widehat{\sigma}_1, \widehat{\sigma}_2, \ldots, \widehat{\sigma}_k) \\ 0 \end{bmatrix}$ if $k_m \geq k_n = k$, and $\begin{bmatrix} \mathrm{diag}(\widehat{\sigma}_1, \widehat{\sigma}_2, \ldots, \widehat{\sigma}_k) \ 0_{k \times (k_n - k)} \end{bmatrix}$ if $k = k_m < k_n$. Let (Σ_1, U_1, V_1) be the set of k_1 leading singular triplets of A. Define $\mathrm{Gap} = \min(\sigma(\widehat{\Sigma}_1) - \sigma(A_3))$ and $\mathrm{gap} = \sigma_{\min}(\Sigma_1) - \|\widehat{\Sigma}_2\|$, and suppose that $\mathrm{Gap}, \mathrm{gap} > 0$. Write $[S_1 \ S_2] = S$, $\begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = R$ and for brevity define $\|\Theta\| := \max(\| \sin\angle(U_1, \widehat{U}_1)\|, \| \sin\angle(V_1, \widehat{V}_1)\|)$. Then we have

$$(6.2) \qquad \|\Theta\| \leq \frac{\max(\|R\|, \|S\|)}{\mathrm{Gap}} \left(1 + \frac{\max(\|R_2\|, \|S_2\|)}{\mathrm{gap}}\right).$$

Moreover, provided that $\mathrm{Gap} > \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}$, we have

$$(6.3) \qquad \|\Theta\| \leq \frac{\max(\|S_1\|, \|R_1\|)}{\mathrm{Gap} - \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}} \left(1 + \frac{\max(\|R_2\|, \|S_2\|)}{\mathrm{gap}}\right).$$

Finally, define $k' := \max(k_m - k_1, k_n - k_1)$, and denote by r_{2i}^T the ith row of R_{21} and by s_{2i} the ith column of S_{21} (setting $r_{2i} = 0$ for $i > k_m - k_1$ and $s_{2i} = 0$ for $i > k_n - k_1$, and

$\widehat{\sigma}_{k_1+i} = 0$ *for* $i > \min(k_m, k_n) - k_1)$. *If* $\mathrm{Gap} > \sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)^2}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}}$, *then*

$$(6.4) \qquad \|\|\Theta\|\| \leq \frac{\max(\|\|S_1\|\|, \|\|R_1\|\|)}{\mathrm{Gap} - \sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)^2}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}}} \left(1 + \sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}} \right).$$

Though not displayed for brevity, slightly improved bounds for $\|\cdot\|_{2,F}$ analogous to those in Theorem 5.1 are available for each bound above. The derivation is again the same, using the inequality $\left\| \begin{bmatrix} A \\ B \end{bmatrix} \right\|_{2,F} \leq \sqrt{\|A\|_{2,F}^2 + \|B\|_{2,F}^2}$.

Proof. Let $\left(\Sigma_1, \widetilde{U}_1, \widetilde{V}_1 \right)$ be a set of exact singular triplets of $\tilde{A}$, i.e., $\tilde{A}\widetilde{V}_1 = \widetilde{U}_1 \Sigma_1$ and $\widetilde{U}_1^* \tilde{A} = \Sigma_1 \widetilde{V}_1^*$. Write $\widetilde{V}_1 = \begin{bmatrix} \widetilde{V}_{11} \\ \widetilde{V}_{21} \\ \widetilde{V}_{31} \end{bmatrix}, \widetilde{U}_1 = \begin{bmatrix} \widetilde{U}_{11} \\ \widetilde{U}_{21} \\ \widetilde{U}_{31} \end{bmatrix}$, so that

$$(6.5) \qquad \begin{bmatrix} \widehat{\Sigma}_1 & 0 & R_1 \\ 0 & \widehat{\Sigma}_2 & R_2 \\ S_1 & S_2 & A_3 \end{bmatrix} \begin{bmatrix} \widetilde{V}_{11} \\ \widetilde{V}_{21} \\ \widetilde{V}_{31} \end{bmatrix} = \begin{bmatrix} \widetilde{U}_{11} \\ \widetilde{U}_{21} \\ \widetilde{U}_{31} \end{bmatrix} \Sigma_1$$

and

$$(6.6) \qquad \begin{bmatrix} \widetilde{U}_{11}^* & \widetilde{U}_{21}^* & \widetilde{U}_{31}^* \end{bmatrix} \begin{bmatrix} \widehat{\Sigma}_1 & 0 & R_1 \\ 0 & \widehat{\Sigma}_2 & R_2 \\ S_1 & S_2 & A_3 \end{bmatrix} = \Sigma_1 \begin{bmatrix} \widetilde{V}_{11}^* & \widetilde{V}_{21}^* & \widetilde{V}_{31}^* \end{bmatrix}.$$

As in the previous sections, we have the crucial identities

$$(6.7) \qquad \left\|\| \sin\angle(U_1, \widehat{U}_1) \|\right\| = \left\|\| \begin{bmatrix} \widetilde{U}_{21} \\ \widetilde{U}_{31} \end{bmatrix} \|\right\|, \quad \left\|\| \sin\angle(V_1, \widehat{V}_1) \|\right\| = \left\|\| \begin{bmatrix} \widetilde{V}_{21} \\ \widetilde{V}_{31} \end{bmatrix} \|\right\|.$$

To prove the theorem we first bound $\|\|\widetilde{U}_{21}\|\|$ with respect to $\|\|\widetilde{U}_{31}\|\|$, and similarly bound $\|\|\widetilde{V}_{21}\|\|$ with respect to $\|\|\widetilde{V}_{31}\|\|$.

From the second block of (6.5) we obtain

$$(6.8) \qquad \widehat{\Sigma}_2 \widetilde{V}_{21} + R_2 \widetilde{V}_{31} = \widetilde{U}_{21} \Sigma_1,$$

and the second block of (6.6) gives

$$(6.9) \qquad \widetilde{U}_{21}^* \widehat{\Sigma}_2 + \widetilde{U}_{31}^* S_2 = \Sigma_1 \widetilde{V}_{21}^*.$$

Taking norms and using the triangular inequality and the fact $\sigma_{\min}(X)\|\|Y\|\| \leq \|\|XY\|\| \leq \|\|X\|\|\|\|Y\|\|$ (the lower bound holds if $X \in \mathbb{C}^{m \times n}, m \geq n$) in (6.8) and (6.9), we obtain

$$(6.10) \qquad \begin{aligned} \left\|\|\widetilde{U}_{21}\|\right\| \sigma_{\min}(\Sigma_1) - \left\|\|\widetilde{V}_{21}\|\right\| \|\widehat{\Sigma}_2\| &\leq \left\|\|R_2\widetilde{V}_{31}\|\right\|, \\ \left\|\|\widetilde{V}_{21}\|\right\| \sigma_{\min}(\Sigma_1) - \left\|\|\widetilde{U}_{21}\|\right\| \|\widehat{\Sigma}_2\| &\leq \left\|\|\widetilde{U}_{31}^* S_2\|\right\|. \end{aligned}$$

By adding the first inequality times $\sigma_{\min}(\Sigma_1)$ and the second inequality times $\|\widehat{\Sigma}_2\|$, we eliminate the $\|\|\widetilde{V}_{21}\|\|$ term, and recalling the assumption $\sigma_{\min}(\Sigma_1) > \|\widehat{\Sigma}_2\|$ we obtain

$$\left\|\|\widetilde{U}_{21}\|\right\| \leq \frac{\sigma_{\min}(\Sigma_1)\|\|R_2\widetilde{V}_{31}\|\| + \|\widehat{\Sigma}_2\|\|\|\widetilde{U}_{31}^* S_2\|\|}{(\sigma_{\min}(\Sigma_1))^2 - \|\widehat{\Sigma}_2\|^2}.$$

Eliminating $\|\|\widetilde{U}_{21}\|\|$ from (6.10) similarly yields

$$\left\|\|\widetilde{V}_{21}\|\right\| \leq \frac{\sigma_{\min}(\Sigma_1)\|\|\widetilde{U}_{31}^* S_2\|\| + \|\widehat{\Sigma}_2\|\|\|R_2\widetilde{V}_{31}\|\|}{(\sigma_{\min}(\Sigma_1))^2 - \|\widehat{\Sigma}_2\|^2}.$$

Combining these two inequalities we obtain

$$\max(\left\|\left|\widetilde{U}_{21}\right|\right\|, \left\|\left|\widetilde{V}_{21}\right|\right\|) \leq \frac{\max(\||\widetilde{U}_{31}^* S_2\||, \||R_2 \widetilde{V}_{31}\||)}{\sigma_{\min}(\Sigma_1) - \|\widehat{\Sigma}_2\|}$$

$$\leq \frac{\max(\||\widetilde{U}_{31}\||, \||\widetilde{V}_{31}\||) \max(\|R_2\|, \|S_2\|)}{\sigma_{\min}(\Sigma_1) - \|\widehat{\Sigma}_2\|}$$

$$(6.11) \qquad = \frac{\max(\|R_2\|, \|S_2\|)}{\mathrm{gap}} \max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|).$$

Together with (6.7) it follows that

$$\max(\left\|\left|\sin \angle(U_1, \widehat{U}_1)\right|\right\|, \left\|\left|\sin \angle(V_1, \widehat{V}_1)\right|\right\|) = \max\left(\left\|\left|\begin{bmatrix} \widetilde{U}_{21} \\ \widetilde{U}_{31} \end{bmatrix}\right|\right\|, \left\|\left|\begin{bmatrix} \widetilde{V}_{21} \\ \widetilde{V}_{31} \end{bmatrix}\right|\right\|\right)$$

$$\leq \max(\||\widetilde{U}_{21}\|| + \||\widetilde{U}_{31}\||, \||\widetilde{V}_{21}\|| + \||\widetilde{V}_{31}\||)$$

$$(6.12) \qquad \leq (1 + \frac{\max(\|R_2\|, \|S_2\|)}{\mathrm{gap}}) \max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|).$$

The remaining task is to bound $\max(\||\widetilde{U}_{31}\||, \||\widetilde{V}_{31}\||)$. The bottom block of (6.5) gives

$$(6.13) \qquad S_1 \widetilde{V}_{11} + S_2 \widetilde{V}_{21} + A_3 \widetilde{V}_{31} = \widetilde{U}_{31} \Sigma_1.$$

Hence recalling that $[S_1 \ S_2] = S$ we have

$$(6.14) \qquad \sigma_{\min}(\Sigma_1)\left\|\left|\widetilde{U}_{31}\right|\right\| \leq \left\|\left|S\right|\right\| + \|A_3\|\left\|\left|\widetilde{V}_{31}\right|\right\|.$$

Similarly, from the last block of (6.6)

$$(6.15) \qquad \widetilde{U}_{11}^* R_1 + \widetilde{U}_{21}^* R_2 + \widetilde{U}_{31}^* A_3 = \Sigma_1 \widetilde{V}_{31}^*,$$

we obtain

$$(6.16) \qquad \sigma_{\min}(\Sigma_1)\left\|\left|\widetilde{V}_{31}\right|\right\| \leq \left\|\left|R\right|\right\| + \|A_3\|\left\|\left|\widetilde{U}_{31}\right|\right\|.$$

We multiply (6.14) by $\sigma_{\min}(\Sigma_1)$ and (6.16) by $\|A_3\|$, and add them to eliminate the $\||\widetilde{V}_{31}\||$ terms, to obtain

$$(\sigma_{\min}(\Sigma_1)^2 - \|A_3\|^2)\left\|\left|\widetilde{U}_{31}\right|\right\| \leq \|A_3\|\left\|\left|R\right|\right\| + \sigma_{\min}(\Sigma_1)\left\|\left|S\right|\right\|$$

$$\leq (\sigma_{\min}(\Sigma_1) + \|A_3\|) \max(\left\|\left|R\right|\right\|, \left\|\left|S\right|\right\|).$$

Hence by the assumption $\mathrm{Gap} = \sigma_{\min}(\Sigma_1) - \|A_3\| > 0$, we have

$$\left\|\left|\widetilde{U}_{31}\right|\right\| \leq \frac{\max(\left\|\left|R\right|\right\|, \left\|\left|S\right|\right\|)}{\mathrm{Gap}}.$$

Eliminating the $\||\widetilde{U}_{31}\||$ terms from (6.14) and (6.16) yields the same bound for $\||\widetilde{V}_{31}\||$, hence

$$\max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|) \leq \frac{\max(\left\|\left|R\right|\right\|, \left\|\left|S\right|\right\|)}{\mathrm{Gap}}.$$

Combine this with (6.11) and (6.7) to obtain (6.2).

We next prove (6.3). From (6.13) we also obtain

$$(6.17) \qquad \sigma_{\min}(\Sigma_1)\left\|\left|\widetilde{U}_{31}\right|\right\| \leq \left\|\left|S_1\right|\right\| + \|S_2\|\left\|\left|\widetilde{V}_{21}\right|\right\| + \|A_3\|\left\|\left|\widetilde{V}_{31}\right|\right\|,$$

and from (6.15),

$$(6.18) \qquad \sigma_{\min}(\Sigma_1)\left\|\left|\widetilde{V}_{31}\right|\right\| \leq \left\|\left|R_1\right|\right\| + \left\|\left|\widetilde{U}_{21}\right|\right\|\|R_2\| + \left\|\left|\widetilde{U}_{31}\right|\right\|\|A_3\|.$$

Again eliminate the $\|\widetilde{V}_{31}\|$ terms by multiplying (6.17) by $\sigma_{\min}(\Sigma_1)$ and (6.18) by $\|A_3\|$, and adding them:

$$
\begin{aligned}
(\sigma_{\min}(\Sigma_1)^2 &- \|A_3\|^2)\left\|\left\|\widetilde{U}_{31}\right\|\right\| \\
&\leq \sigma_{\min}(\Sigma_1)(\left\|\left\|S_1\right\|\right\| + \|S_2\|\left\|\left\|\widetilde{V}_{21}\right\|\right\|) + \|A_3\|(\left\|\left\|R_1\right\|\right\| + \|R_2\|\left\|\left\|\widetilde{U}_{21}\right\|\right\|) \\
&\leq (\sigma_{\min}(\Sigma_1) + \|A_3\|)(\max(\left\|\left\|S_1\right\|\right\|, \left\|\left\|R_1\right\|\right\|) \\
&\quad + \max(\|S_2\|, \|R_2\|)\max(\left\|\left\|\widetilde{U}_{21}\right\|\right\|, \left\|\left\|\widetilde{V}_{21}\right\|\right\|)).
\end{aligned}
$$

Therefore, using (6.11) we obtain

$$
\begin{aligned}
\left\|\left\|\widetilde{U}_{31}\right\|\right\| &\leq \frac{\max(\left\|\left\|S_1\right\|\right\|, \left\|\left\|R_1\right\|\right\|) + \max(\|S_2\|, \|R_2\|)\max\left(\left\|\left\|\widetilde{U}_{21}\right\|\right\|, \left\|\left\|\widetilde{V}_{21}\right\|\right\|\right)}{\sigma_{\min}(\Sigma_1) - \|A_3\|} \\
&\leq \frac{1}{\mathrm{Gap}}\left(\max(\left\|\left\|S_1\right\|\right\|, \left\|\left\|R_1\right\|\right\|) + \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}\max\left(\left\|\left\|\widetilde{U}_{31}\right\|\right\|, \left\|\left\|\widetilde{V}_{31}\right\|\right\|\right)\right).
\end{aligned}
$$

As before, eliminating $\left\|\left\|\widetilde{U}_{31}\right\|\right\|$ from (6.17) and (6.18) yields the same bound for $\left\|\left\|\widetilde{V}_{31}\right\|\right\|$, hence

$$
\begin{aligned}
\max(&\left\|\left\|\widetilde{U}_{31}\right\|\right\|, \left\|\left\|\widetilde{V}_{31}\right\|\right\|) \\
&\leq \frac{1}{\mathrm{Gap}}\left(\max(\left\|\left\|S_1\right\|\right\|, \left\|\left\|R_1\right\|\right\|) + \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}\max(\left\|\left\|\widetilde{U}_{31}\right\|\right\|, \left\|\left\|\widetilde{V}_{31}\right\|\right\|)\right).
\end{aligned}
$$

Therefore using the assumption $\mathrm{Gap} > \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}$ we obtain

$$
\max(\left\|\left\|\widetilde{U}_{31}\right\|\right\|, \left\|\left\|\widetilde{V}_{31}\right\|\right\|) \leq \frac{\max(\left\|\left\|S_1\right\|\right\|, \left\|\left\|R_1\right\|\right\|)}{\mathrm{Gap} - \frac{\max(\|S_2\|, \|R_2\|)^2}{\mathrm{gap}}}.
$$

Together with (6.12) we obtain (6.3).

The remaining task is to establish (6.4). For this, we revisit (6.8), (6.9), and now bound the individual $\widetilde{U}_{21i}$, the ith row of $\widetilde{U}_{21}$. We obtain

$$
\begin{aligned}
\left\|\left\|\widetilde{U}_{21i}\right\|\right\|\sigma_{\min}(\Sigma_1) - \left\|\left\|\widetilde{V}_{21i}\right\|\right\|\widehat{\sigma}_{k_1+i} &\leq \left\|\left\|r_{2i}^T\widetilde{V}_{31}\right\|\right\|, \\
\left\|\left\|\widetilde{V}_{21i}\right\|\right\|\sigma_{\min}(\Sigma_1) - \left\|\left\|\widetilde{U}_{21i}\right\|\right\|\widehat{\sigma}_{k_1+i} &\leq \left\|\left\|\widetilde{U}_{31}^*s_{2i}\right\|\right\|.
\end{aligned}
$$

Eliminating $\widetilde{V}_{21i}$ and $\widetilde{U}_{21i}$ as before gives

$$
\left\|\left\|\widetilde{U}_{21i}\right\|\right\| \leq \frac{\sigma_{\min}(\Sigma_1)\left\|\left\|r_{2i}^T\widetilde{V}_{31}\right\|\right\| + \widehat{\sigma}_{k_1+i}\left\|\left\|\widetilde{U}_{31}^*s_{2i}\right\|\right\|}{(\sigma_{\min}(\Sigma_1))^2 - \widehat{\sigma}_{k_1+i}^2},
$$

$$
\left\|\left\|\widetilde{V}_{21i}\right\|\right\| \leq \frac{\sigma_{\min}(\Sigma_1)\left\|\left\|\widetilde{U}_{31}^*s_{2i}\right\|\right\| + \widehat{\sigma}_{k_1+i}\left\|\left\|r_{2i}^T\widetilde{V}_{31}\right\|\right\|}{(\sigma_{\min}(\Sigma_1))^2 - \widehat{\sigma}_{k_1+i}^2}.
$$

Note that when $k_m \neq k_n$, $(\widetilde{U}_{21i}, r_{2i})$ or $(\widetilde{V}_{21i}, s_{2i})$ is empty for large i; by taking $\widehat{\sigma}_{k_1+i} = 0$ for such i the argument carries over. We therefore have for every i

$$
(6.19) \qquad \max(\left\|\left\|\widetilde{U}_{21i}\right\|\right\|, \left\|\left\|\widetilde{V}_{21i}\right\|\right\|) \leq \frac{\max(\|s_{2i}\|, \|r_{2i}\|)\max(\left\|\left\|\widetilde{U}_{31}\right\|\right\|, \left\|\left\|\widetilde{V}_{31}\right\|\right\|)}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}}.
$$

From (6.13) we also obtain

$$
(6.20) \qquad \sigma_{\min}(\Sigma_1)\left\|\left\|\widetilde{U}_{31}\right\|\right\| \leq \|A_3\|\left\|\left\|\widetilde{V}_{31}\right\|\right\| + \left\|\left\|S_1\right\|\right\| + \sum_{i=1}^{k_n-k_1}\|s_{2i}\|\left\|\left\|\widetilde{V}_{21i}\right\|\right\|,
$$

and from (6.15),

$$(6.21) \qquad \sigma_{\min}(\Sigma_1)\left\|\left|\widetilde{V}_{31}\right|\right\| \leq \left\|\left|\widetilde{U}_{31}\right|\right\|\|A_3\| + \left\|\left|R_1\right|\right\| + \sum_{i=1}^{k_m-k_1} \left\|\left|\widetilde{U}_{21i}\right|\right\|\|r_{2i}\|.$$

Hence eliminating $\left\|\left|\widetilde{U}_{31}\right|\right\|$ and then $\left\|\left|\widetilde{V}_{31}\right|\right\|$, and using (6.19) gives

$$\max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|)$$

$$\leq \frac{1}{\mathrm{Gap}}\left(\max(\left\|\left|S_1\right|\right\|, \left\|\left|R_1\right|\right\|) + \sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)^2}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}} \max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|)\right).$$

Thus by the assumption $\mathrm{Gap} > \dfrac{\sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)^2}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}}}{}$ we have

$$\max(\left\|\left|\widetilde{U}_{31}\right|\right\|, \left\|\left|\widetilde{V}_{31}\right|\right\|) \leq \frac{\max(\left\|\left|S_1\right|\right\|, \left\|\left|R_1\right|\right\|)}{\mathrm{Gap} - \frac{\sum_{i=1}^{k'} \frac{\max(\|r_{2i}\|, \|s_{2i}\|)^2}{\sigma_{\min}(\Sigma_1) - \widehat{\sigma}_{k_1+i}}}{}}.$$

Finally, the bound (6.4) follows from combining this with (6.19) and (6.7). □

We note that R or S in the above theorem is allowed to be empty, as in the case where a one-sided projection is employed. This includes the popular randomized SVD algorithm [6]. We make two more remarks.

Remark 6.1 (Other approaches for the SVD). A standard approach to extending results in symmetric eigenvalue problems to the SVD is to use the Jordan-Wielandt matrix, for example as in [12, Sec. 3]. As pointed out in [15], this has the slight downside of introducing spurious eigenvalues at 0. Moreover, the results via Jordan-Wielandt we obtained were less clean and looser than Theorem 6.1. Another approach is to work with the Gram matrix A^*A, but this unnecessarily squares the singular values and modifies gap and Gap. For these reasons, we have chosen to work directly with the SVD equations.

Remark 6.2 (Proof of (6.2) via Wedin and [15]). As in Remark 3.4, a proof for (6.2) can be given by combining Wedin's result (the SVD analogue of Davis-Kahan) and [15] (SVD analogue of Saad's result). The sharper bounds (6.2) and (6.3) cannot be obtained this way.

7. Eigenvectors of a self-adjoint operator

So far we have specialized to finite-dimensional matrices as the analysis is elementary and the situation is more transparent. In this final section, as in [10, 16], we extend the discussion to the infinite-dimensional case, where the matrix is generalized to a self-adjoint operator $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ on a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$. Unlike the previous studies, which assumed the operators are bounded, our discussion allows $\mathcal{A}$ to be unbounded, thus is applicable for example to differential operators $\mathcal{A}u = u''$; in this case, we assume that $\mathcal{A}$ is densely defined, as is customary.

Let Q be a subspace of $\mathcal{H}$, which is of finite dimension k with orthonormal basis $q_1, \ldots, q_k$. In the Rayleigh-Ritz process for $\mathcal{A}$, we compute the $k \times k$ matrix A_1 with (i,j) element $\langle q_i, \mathcal{A}q_j \rangle$ and its eigenvalue decomposition $A_1 = \Omega \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k)\Omega^*$ to obtain the Ritz values $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$ and Ritz vectors $[q_1, \ldots, q_k]\Omega$. Denote by $Q_1, Q_2 \in \mathcal{H}$ the resulting Ritz subspaces corresponding to disjoint sets of eigenvalues of A_1 (we have $Q = Q_1 \oplus Q_2$), and let Q_3 be the (infinite-dimensional) orthogonal complement of Q such that $\mathcal{H} = Q_1 \oplus Q_2 \oplus Q_3$ is an orthogonal direct sum.

For simplicity, we treat the case where Q_1 is one-dimensional (subspace versions can be obtained, generalizing Section 5). That is, let $(\widehat{\lambda}, \widehat{u})$ be a Ritz pair with $\widehat{u} = q_1$, and suppose that $\mathcal{A}u = \lambda u$; note that this is an assumption, as a self-adjoint operator may

not have any eigenvalue (e.g., [8, Ch. 9]), although the spectrum is always nonempty. The goal is to bound $\sin\angle(\widehat{u}, u)$.

Denote by $\mathcal{P}_i$ the orthogonal projectors onto each subspace Q_i. We define $\mathcal{A}_{ij} := \mathcal{P}_i \mathcal{A} \mathcal{P}_j$. Then the R-R process forces $\mathcal{A}_{12} = 0, \mathcal{A}_{21} = 0$. Note that $\mathcal{A}_{13}^* = \mathcal{A}_{31}, \mathcal{A}_{23}^* = \mathcal{A}_{32}$ (where $*$ denotes the adjoint of the operators), and these terms represent the residuals, hence we write $\|R_1\| = \|R_{31}\|$ and $\|R\| = \|R_{31} + \mathcal{A}_{32}\|$. Also define $\|R_2\| = \|\mathcal{A}_{32}\| (= \|\mathcal{A}_{23}\|)$, and $\|r_i\| = \|\mathcal{A}_{32} \mathcal{P}_{2,i}\| (= \|\mathcal{P}_{2,i} \mathcal{A}_{23}\|)$ for $i = 2, \ldots, k$, where $\mathcal{P}_{2,i}$ is the one-dimensional projection onto the ith Ritz vector. The quantities gap and Gap are defined by gap $= \min|\widehat{\lambda} - \lambda(\mathcal{A}_{22})|$, Gap $= \min|\widehat{\lambda} - \lambda(\mathcal{A}_{33})|$, in which $\lambda(\mathcal{A}_{ii})$ denotes the spectrum of the restriction of $\mathcal{A}_{ii}$ to Q_i.

Theorem 7.1. *Under the above assumptions and notation,*

$$(7.1) \qquad \sin\angle(u, \widehat{u}) \leq \frac{\|R\|}{\mathrm{Gap}} \sqrt{1 + \frac{\|R_2\|^2}{\mathrm{gap}^2}} \quad \left(\leq \frac{\|R\|}{\mathrm{Gap}}(1 + \frac{\|R_2\|}{\mathrm{gap}}) \right).$$

Moreover, if $\mathrm{Gap} > \frac{\|R_2\|^2}{\mathrm{gap}}$, *then*

$$(7.2) \qquad \sin\angle(u, \widehat{u}) \leq \frac{\|R_1\|}{\mathrm{Gap} - \frac{\|R_2\|^2}{\mathrm{gap}}} \sqrt{1 + \frac{\|R_2\|^2}{\mathrm{gap}^2}},$$

and if $\mathrm{Gap} > \sum_{i=2}^{k} \frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}$, *then*

$$(7.3) \qquad \sin\angle(u, \widehat{u}) \leq \frac{\|R_1\|}{\mathrm{Gap} - \sum_{i=2}^{k} \frac{\|r_i\|^2}{|\lambda - \widehat{\lambda}_i|}} \sqrt{1 + \left(\sum_{i=2}^{k} \frac{\|r_i\|}{|\lambda - \widehat{\lambda}_i|} \right)^2}.$$

Proof. Writing $u = \sum_{i=1}^{3} u_i$ with $u_i \in Q_i$, the Q_i-component of $\mathcal{A}u = \lambda u$ each implies

$$(7.4a) \qquad \lambda u_1 = \mathcal{A}_{11} u_1 + \mathcal{A}_{13} u_3,$$

$$(7.4b) \qquad \lambda u_2 = \mathcal{A}_{22} u_2 + \mathcal{A}_{23} u_3,$$

$$(7.4c) \qquad \lambda u_3 = \mathcal{A}_{31} u_1 + \mathcal{A}_{32} u_2 + \mathcal{A}_{33} u_3.$$

Our goal is to bound $\sin\angle(u, \widehat{u}) = \sqrt{\|u_2\|^2 + \|u_3\|^2}$.

We first derive (7.1), an analogue of Theorem 3.1. By (7.4c), we have

$$\|(\mathcal{A}_{33} - \lambda)u_3\| = \|\mathcal{A}_{31}u_1 + \mathcal{A}_{32}u_2\| = \|(\mathcal{A}_{31} + \mathcal{A}_{32})u\| \leq \|\mathcal{A}_{31} + \mathcal{A}_{32}\| = \|R\|.$$

Together with the fact $\|(\mathcal{A}_{33} - \lambda)u_3\| \geq \mathrm{Gap}\,\|u_3\|$ ([9, § V.3.5]; to see this, note that $v = (\mathcal{A}_{33} - \lambda)u_3$ implies $(\mathcal{A}_{33} - \lambda)^{-1}v = u_3$, hence $\|u_3\| \leq \|(\mathcal{A}_{33} - \lambda)^{-1}\|\|v\| \leq \mathrm{Gap}\,\|v\|$), we obtain

$$(7.5) \qquad \|u_3\| \leq \frac{\|R\|}{\mathrm{Gap}}.$$

Now (7.4b) gives $(\mathcal{A}_{22} - \lambda)u_2 = -\mathcal{A}_{23}u_3$. Since $\|\mathcal{A}_{23}\| = \|\mathcal{A}_{23}^*\| = \|\mathcal{A}_{32}\| = \|R_2\|$, and $\|(\mathcal{A}_{22} - \lambda)u_2\| \geq \mathrm{gap}\,\|u_2\|$, we thus have $\|u_2\| \leq \frac{\|R_2\|}{\mathrm{gap}}\|u_3\|$. Using this and (7.5), we obtain (7.1).

We now turn to (7.3); the proof of (7.2) is similar and omitted. As in Theorem 5.1, the idea is to improve the estimate of $\|u_2\|$ using (7.4b). Projecting it onto $\mathcal{P}_{2,i}$ gives $\mathcal{P}_{2,i}(\mathcal{A}_{22} - \lambda)u_2 + \mathcal{P}_{2,i}\mathcal{A}_{23}u_3 = 0$ for $i = 2, \ldots, k$, and by assumption $\mathcal{P}_{2,i}\mathcal{A}_{22} = \widehat{\lambda}_i \mathcal{P}_{2,i}$, so $(\widehat{\lambda}_i - \lambda)\mathcal{P}_{2,i}u_2 + \mathcal{P}_{2,i}\mathcal{A}_{23}u_3 = 0$, hence

$$(7.6) \qquad \|\mathcal{P}_{2,i}u_2\| \leq \frac{\|\mathcal{P}_{2,i}\mathcal{A}_{23}u_3\|}{|\widehat{\lambda}_i - \lambda|} \leq \frac{\|\mathcal{P}_{2,i}\mathcal{A}_{23}\|\|u_3\|}{|\widehat{\lambda}_i - \lambda|} = \frac{\|r_i\|\|u_3\|}{|\widehat{\lambda}_i - \lambda|},$$

where we used $\|r_i\| = \|\mathcal{P}_{2,i}\mathcal{A}_{23}\|$ for the final equality. The inequality (7.6) holds for $i = 2, \ldots, k$.

Now since $\mathcal{A}_{32} = \mathcal{A}_{32}\mathcal{P}_2 = \sum_{i=2}^{k}\mathcal{A}_{32}\mathcal{P}_{2,i}$, we can rewrite (7.4c) as $\mathcal{A}_{31}u_1 + \sum_{i=2}^{k}\mathcal{A}_{32}\mathcal{P}_{2,i}u_2$ $= -(\mathcal{A}_{33} - \lambda)u_3$. Therefore

$$\|(\mathcal{A}_{33} - \lambda)u_3\| \leq \|\mathcal{A}_{31}u_1\| + \sum_{i=2}^{k}\|\mathcal{A}_{32}\mathcal{P}_{2,i}u_2\| \leq \|R_1\| + \sum_{i=2}^{k}\|\mathcal{A}_{32}\mathcal{P}_{2,i}\|\|\mathcal{P}_{2,i}u_2\|$$

$$\leq \|R_1\| + \sum_{i=2}^{k}\frac{\|r_i\|^2\|u_3\|}{|\widehat{\lambda}_i - \lambda|},$$

where we used (7.6) and $\|r_i\| = \|\mathcal{A}_{32}\mathcal{P}_{2,i}\|$. Together with $\|(\mathcal{A}_{33} - \lambda)u_3\| \geq \mathrm{Gap}\,\|u_3\|$ we obtain

$$(7.7) \qquad \|u_3\| \leq \frac{\|R_1\|}{\mathrm{Gap}\left(1 - \sum_{i=2}^{k}\frac{\|r_i\|^2}{|\widehat{\lambda}_i - \lambda|}\right)}.$$

Finally, (7.6) together with $\|u_2\|^2 = \sum_{i=2}^{k}\|\mathcal{P}_{2,i}u_2\|^2$ gives

$$\|u_2\|^2 \leq \left(\sum_{i=2}^{k}\frac{\|r_i\|}{|\widehat{\lambda}_i - \lambda|}\right)^2\|u_3\|^2,$$

so

$$\sin\angle(u,\widetilde{u}) = \sqrt{\|u_2\|^2 + \|u_3\|^2} \leq \frac{\|R_1\|}{\mathrm{Gap}\left(1 - \sum_{i=2}^{k}\frac{\|r_i\|^2}{|\widehat{\lambda}_i - \lambda|}\right)}\sqrt{1 + \left(\sum_{i=2}^{k}\frac{\|r_i\|}{|\widehat{\lambda}_i - \lambda|}\right)^2},$$

completing the proof of (7.3). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

7.1. Experiments: Sturm-Liouville eigenvalue problem.
We illustrate Theorem 7.1 with a simple Sturm-Liouville eigenvalue problem (e.g., [4, § 3.5])

$$(7.8) \qquad \mathcal{A}u = u'' = \lambda u, \qquad u'(0) = \alpha u(0), \quad u'(\pi) = \beta u(\pi), \qquad u \in \mathcal{H} = H^2(0,\pi).$$

$\mathcal{A}$ is an unbounded self-adjoint operator, with a full set of (infinitely many) orthonormal eigenfunctions. Here we take $\alpha = 1, \beta = -1$. The exact eigenvalues are $\lambda_i = -\nu_i^2$, where ν_i are the solutions for $\tan\pi\nu = 2\nu/(\nu^2 - 1)$, with corresponding eigenfunction $\nu_i\cos\nu_i x + \alpha\sin\nu_i x$ [4, § 3.5]. We attempt to compute the eigenpairs with the smoothest eigenfunctions, i.e., eigenpairs closest to 0. To do this, a natural idea is to take low-degree polynomials. We take the trial subspace to be the k-dimensional subspace of polynomials p of degree up to $k + 1$ that satisfy the two boundary conditions $p'(0) = \alpha p(0)$ and $p'(\pi) = \beta p(\pi)$. Figure 7.1 (left) shows the basis functions obtained in this way, for $k = 7$. Such computations can be done conveniently using Chebfun [3].

Having defined the subspace Q, we can perform R-R to obtain the Ritz vectors (which are functions in $\mathcal{H}$ here), along with the Ritz values.

Figure 7.1 shows the convergence of $\angle(u,\widehat{u})$ to the eigenfunction u for the smallest eigenpair and its bounds, analogous to Figure 4.1. As in that experiment, our bound (7.2) gives tighter bounds for the actual error, although here Davis-Kahan also performs well, since gap is not very small.

Finally, in Figure 7.2 we illustrate the behavior of the residual function $\mathcal{A}\widehat{u} - \widehat{\lambda}\widehat{u}$ as k varies. Note that $\widehat{u}$ is determined up to a sign flip ± 1; here we chose $\widehat{u}(1) > 0$. We make two observations. First, evidently the norm $\|\mathcal{A}\widehat{u} - \widehat{\lambda}\widehat{u}\|$ decays rapidly as k increases, essentially like the right plot in Figure 7.1. The second and more interesting observation is that the residuals appear to become more and more oscillatory (nonsmooth) as k grows. This is a typical phenomenon, and can be explained as follows. As emphasized repeatedly in this paper, R-R forces the residual to be orthogonal to Q, which contains the "smoothest" functions. Consequently, in the Legendre expansion of the residual $\mathcal{A}\widehat{u} - \widehat{\lambda}\widehat{u} = \sum_{i=0}^{\infty}c_iP_i(x)$,
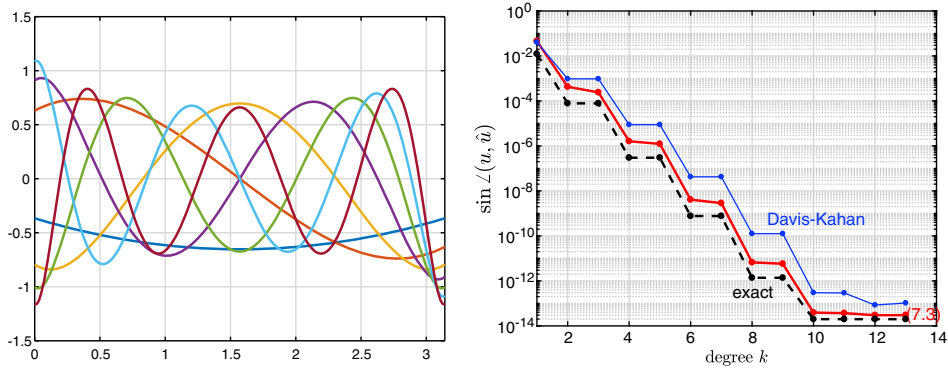
FIGURE 7.1. Left: Basis functions for projection subspace Q, satisfying $u'(0) = u(0), u'(\pi) = -u(\pi)$. Right: Convergence of $\angle(u, \widehat{u})$ and its bounds.

$|c_i|$ are small for $i < k$; they are bounded roughly by $\|u_2\|$, which is $O(\|\mathcal{A}\widehat{u} - \widehat{\lambda}\widehat{u}\|^2)$ by (7.6). This also reflects the main result in [10]; recall Remark 3.4. By growing k, the residual becomes orthogonal to more and more of these smoothest functions, and therefore becomes more oscillatory.
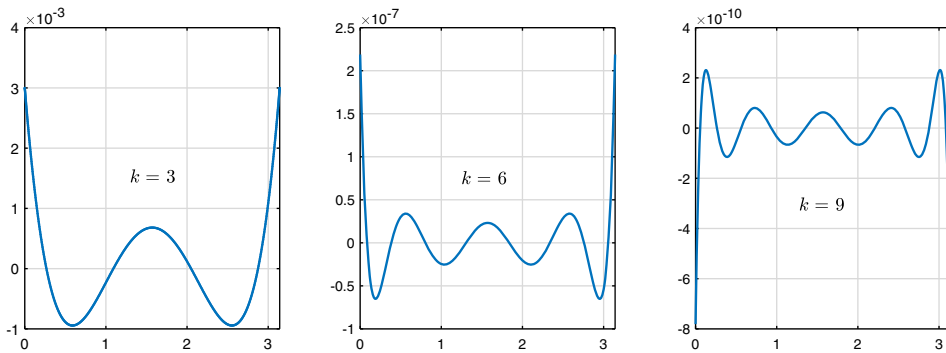
FIGURE 7.2. Residual function $\mathcal{A}\widehat{u} - \widehat{\lambda}\widehat{u}$ for $k = 3, 6$ and 9.

ACKNOWLEDGMENTS

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst (eds.), *Templates for the Solution of Algebraic Eigenvalue Problems*, A practical guide, Software, Environments, and Tools, vol. 11, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. MR1792141

[2] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III*, SIAM J. Numer. Anal. **7** (1970), 1–46, DOI 10.1137/0707001. MR264450

[3] T. A. Driscoll, N. Hale, and L. N. Trefethen. *Chebfun Guide*. Pafnuty Publications, 2014.

[4] G. B. Folland, *Fourier Analysis and its Applications*, The Wadsworth & Brooks/Cole Mathematics Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1992. MR1145236

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 2013. MR3024913

[6] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev. **53** (2011), no. 2, 217–288, DOI 10.1137/090771806. MR2806637

[7] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1991. MR1091716

[8] J. K. Hunter and B. Nachtergaele, *Applied Analysis*, World Scientific Publishing Co., Inc., River Edge, NJ, 2001. MR1829589

[9] T. Kato, *Perturbation Theory for Linear Operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition. MR1335452

[10] A. V. Knyazev, *New estimates for Ritz vectors*, Math. Comp. **66** (1997), no. 219, 985–995, DOI 10.1090/S0025-5718-97-00855-7. MR1415802

[11] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput. **23** (2001), no. 2, 517–541, DOI 10.1137/S1064827500366124. Copper Mountain Conference (2000). MR1861263

[12] C.-K. Li and R.-C. Li, *A note on eigenvalues of perturbed Hermitian matrices*, Linear Algebra Appl. **395** (2005), 183–190, DOI 10.1016/j.laa.2004.08.026. MR2112884

[13] R. Li, Y. Xi, E. Vecharynski, C. Yang, and Y. Saad, *A thick-restart Lanczos algorithm with polynomial filtering for Hermitian eigenvalue problems*, SIAM J. Sci. Comput. **38** (2016), no. 4, A2512–A2534, DOI 10.1137/15M1054493. MR3537014

[14] R.-C. Li and L.-H. Zhang, *Convergence of the block Lanczos method for eigenvalue clusters*, Numer. Math. **131** (2015), no. 1, 83–113, DOI 10.1007/s00211-014-0681-6. MR3383329

[15] Y. Nakatsukasa, *Accuracy of singular vectors obtained by projection-based SVD methods*, BIT **57** (2017), no. 4, 1137–1152, DOI 10.1007/s10543-017-0665-x. MR3736004

[16] E. Ovtchinnikov, *Cluster robust error estimates for the Rayleigh-Ritz approximation. I. Estimates for invariant subspaces*, Linear Algebra Appl. **415** (2006), no. 1, 167–187, DOI 10.1016/j.laa.2005.06.040. MR2214751

[17] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Applied Mathematics, vol. 20, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original. MR1490034

[18] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Classics in Applied Mathematics, vol. 66, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Revised edition of the 1992 original [1177405]. MR3396212

[19] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, Inc., Boston, MA, 1990. MR1061154

[20] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. MR1444820

[21] L. Wu, E. Romero, and A. Stathopoulos, *PRIMME_SVDS: a high-performance preconditioned SVD solver for accurate large-scale computations*, SIAM J. Sci. Comput. **39** (2017), no. 5, S248–S271, DOI 10.1137/16M1082214. MR3716557

Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom
Email address: nakatsukasa@maths.ox.ac.uk