STRUCTURAL CONVERGENCE RESULTS FOR APPROXIMATION OF DOMINANT SUBSPACES FROM BLOCK KRYLOV SPACES*

PETROS DRINEAS†, ILSE C. F. IPSEN‡, EUGENIA-MARIA KONTOPOULOU†, AND MALIK MAGDON-ISMAIL§

Abstract. This paper is concerned with approximating the dominant left singular vector space of a real matrix $\mathbf{A}$ of arbitrary dimension, from block Krylov spaces generated by the matrix $\mathbf{A}\mathbf{A}^T$ and the block vector $\mathbf{A}\mathbf{X}$. Two classes of results are presented. First are bounds on the distance, in the two- and Frobenius norms, between the Krylov space and the target space. The distance is expressed in terms of principal angles. Second are bounds for the low-rank approximation computed from the Krylov space compared to the best low-rank approximation, in the two- and Frobenius norms. For starting guesses $\mathbf{X}$ of full column-rank, the bounds depend on the tangent of the principal angles between $\mathbf{X}$ and the dominant right singular vector space of $\mathbf{A}$. The results presented here form the structural foundation for the analysis of randomized Krylov space methods. The innovative feature is a combination of traditional Lanczos convergence analysis with optimal approximations via least squares problems.

Key words. singular value decomposition, least squares, principal angles, gap-amplifying polynomials, random matrices

AMS subject classifications. 15A18, 15A42, 65F15, 65F20, 68W20

DOI. 10.1137/16M1091745

1. Introduction. Randomized methods for low-rank approximations from Krylov spaces are starting to emerge in the theoretical computer science community [34, 43]. This motivated us to produce a "proof of concept" for the approximation of dominant subspaces from Krylov spaces.

Low-rank versus subspace approximations. Our focus is the approximation of a dominant subspace of $\mathbf{A} \in \mathbb{R}^{m \times n}$. This is a different and harder problem than a low-rank approximation of $\mathbf{A}$. To wit, the objective of a low-rank approximation is a matrix $\mathbf{Z}$ with orthonormal columns that makes $\|\mathbf{A} - \mathbf{Z}\mathbf{Z}^T\mathbf{A}\|$ small in some unitarily invariant norm [21, 44]. In contrast, a subspace approximation aims at a space $\mathcal{K}$ that has a small angle with the dominant target space, which is the space spanned by the

†Department of Computer Science, Purdue University, West Lafayette, IN 47907 (pdrineas@purdue.edu, ekontopo@purdue.edu).

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (ipsen@ncsu.edu).

§Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180 (magdon@cs.rpi.edu).

singular vectors associated with the top k left singular vectors of $\mathbf{A}$.

For a dominant subspace to be well defined, the top k singular values must be separated by a gap from the remaining singular values of $\mathbf{A}$. In contrast, low-rank approximations can do without a singular value gap. Accuracy results for dominant subspace computations are automatically informative for low-rank approximations, but not vice versa. It is in this sense that dominant subspace approximations are harder.

This paper. We consider block Krylov space methods for computing dominant left singular vector spaces of general rectangular matrices, and we present structural, deterministic bounds on the quality of the subspaces for essentially general starting guesses. The innovative feature is a fusion of eigenvalue and singular value technology: We combine a traditional Lanczos convergence analysis [38] with optimal approximations via least squares problems [10, 11].

Our long-term goal is to put randomized Krylov space approximations on a firm numerical footing. However, at this preliminary first step, we make a few idealized assumptions:

1. The block Krylov spaces have maximal dimension.
2. The analysis assumes exact arithmetic and does not address the implementation of numerically stable recursions.

Future work will need to deal with the challenging issues of finite precision arithmetic and viable numerical implementations, including recursions, numerical stability, maintaining orthogonality, deflation, adaptation of block size, and restarting. Empirical evaluations will have to assess whether the bounds are tight enough to be informative in practice.

Overview. We start with a brief summary of our contributions (section 2), followed by a comparison to existing work (section 3). Auxiliary results (section 4) set the stage for the proof of the main theorems (sections 5–7 and Appendix A). We end the main part of the paper with a perspective on open problems (section 8).

2. Results. After setting the context (section 2.1), we give a brief summary of our bounds for the following: the distance between the Krylov space and the dominant left singular space (section 2.2); a particular dominant subspace approximation from the Krylov space (section 2.3); and the polynomials appearing in the approximation (section 2.4). We end this section with a discussion of options for bounding the distance between the initial guess and the dominant right singular vector space (section 2.5).

2.1. Setting. To approximate the dominant left singular vector subspace of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, given a starting guess $\mathbf{X} \in \mathbb{R}^{n \times s}$, we construct the Krylov space in[1] $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}\mathbf{X}$,

$$(2.1) \qquad \mathcal{K}_q \equiv \mathcal{K}_q(\mathbf{A}\mathbf{A}^T, \mathbf{A}\mathbf{X}) = \text{range} \begin{pmatrix} \mathbf{A}\mathbf{X} & (\mathbf{A}\mathbf{A}^T)\mathbf{A}\mathbf{X} & \cdots & (\mathbf{A}\mathbf{A}^T)^q\mathbf{A}\mathbf{X} \end{pmatrix}.$$

We assume maximal dimension, $\dim(\mathcal{K}_q) = (q+1)s$.

In contrast to [4, 5], the matrix $\mathbf{A}$ occurs not only in the powers $\mathbf{A}\mathbf{A}^T$ but also has a direct effect on the starting guess through $\mathbf{A}\mathbf{X}$. Furthermore, $\mathbf{X}$ is not required to have orthonormal columns and, at times, not even linearly independent columns.

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the *full* SVD of $\mathbf{A}$, so that $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$, and $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices. For a positive integer $1 \leq k < \text{rank}(\mathbf{A})$, identify

---

[1]The superscript T denotes the transpose, and $\| \cdot \|_2$ the two norm.

the dominant spaces by partitioning

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_k & \\ & \mathbf{\Sigma}_{k,\perp} \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{k,\perp} \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_k & \mathbf{V}_{k,\perp} \end{pmatrix},$$

where the diagonal matrix $\mathbf{\Sigma}_k$ contains the k largest singular values, hence is nonsingular. For the dominant subspaces to be well defined, the dominant k singular values of $\mathbf{A}$ must be strictly larger than the remaining ones, $1/\|\mathbf{\Sigma}_k^{-1}\|_2 > \|\mathbf{\Sigma}_{k,\perp}\|_2 > 0$.

2.2. Krylov space angles. We present bounds for the distance between the Krylov space $\mathcal{K}_q$ and the dominant left singular vector space range($\mathbf{U}_k$). Theorem 2.1 bounds the distance between $\mathcal{K}_q$ and the whole space, while Theorem 2.2 bounds the distance between $\mathcal{K}_q$ and an individual left singular vector. The distances are represented by principal angles, which are discussed in section 4.3; and they are bounded by generalized matrix functions in the sense of [3, 22], which are discussed in section 4.2.

Theorem 2.1 below is in the spirit of Rayleigh–Ritz bounds [9, 18]. It indicates how well the Krylov space $\mathcal{K}_q$ captures the targeted dominant left singular vector space range($\mathbf{U}_k$) in both the two- and Frobenius norms. Denote by $\mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k) \in \mathbb{R}^{k \times k}$ the diagonal matrix of principal angles between $\mathcal{K}_q$ and range($\mathbf{U}_k$), and by $\mathbf{\Theta}(\mathbf{X}, \mathbf{V}_k) \in \mathbb{R}^{k \times k}$ the diagonal matrix of the principal angles between range($\mathbf{X}$) and range($\mathbf{V}_k$).

THEOREM 2.1. *Let $\phi(x)$ be a polynomial of degree $2q + 1$ with odd powers only, such that $\phi(\mathbf{\Sigma}_k)$ is nonsingular. If* rank($\mathbf{V}_k^T \mathbf{X}$) = k, *then*[2]

$$\| \sin \mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k) \|_{2,F} \leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2 \, \|\phi(\mathbf{\Sigma}_k)^{-1}\|_2 \, \|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_{2,F}.$$

If, in addition, $\mathbf{X}$ has orthonormal or linearly independent columns, then

$$\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_{2,F} = \|\tan \mathbf{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_{2,F}$$

and

$$\| \sin \mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k) \|_{2,F} \leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2 \, \|\phi(\mathbf{\Sigma}_k)^{-1}\|_2 \, \|\tan \mathbf{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_{2,F}.$$

Proof. See section 5 for general and orthonormal $\mathbf{X}$, and Appendix A for $\mathbf{X}$ with linearly independent columns. □

Theorem 2.1 is reminiscent of the eigenvalue bounds [26, (2.18)] which contain a tangent on the left. The term $\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_{2,F}$ already appeared in previous analyses of randomized algorithms [15, 16, 17, 32], and bounds for it are discussed in section 2.5. If $\mathbf{X}$ is a random starting guess, such as a random sign matrix, a random Gaussian matrix[3], or a matrix with randomly chosen orthonormal columns, then state-of-the-art matrix concentration inequalities can be called upon.

In the special case where $\mathbf{X}$ has linearly independent columns the bounds admit a geometric interpretation: They depend on the tangents of angles between range($\mathbf{X}$) and the dominant right singular vector space range($\mathbf{V}_k$). The full-rank assumption for $\mathbf{V}_k^T \mathbf{X}$ means that the spaces range($\mathbf{V}_k$) and range($\mathbf{X}$) are sufficiently close, with all principal angles being less than $\pi/2$.

---

[2] The subscript F denotes the Frobenius norm.

[3] The elements of a random Gaussian matrix are independent and identically distributed normal random variables with mean zero and variance one.

Next, Theorem 2.2 bounds the distances between $\mathcal{K}_q$ and individual left singular vectors of $\mathbf{A}$. To this end, distinguish the k dominant singular values and associated left singular vectors,

$$\boldsymbol{\Sigma}_k = \mathrm{diag}\begin{pmatrix} \sigma_1 & \cdots & \sigma_k \end{pmatrix}, \qquad \mathbf{U}_k = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_k \end{pmatrix}.$$

THEOREM 2.2. *Let $\phi(x)$ be a polynomial of degree $2q+1$ with odd powers only, such that $\phi(\boldsymbol{\Sigma}_k)$ is nonsingular. If* $\mathrm{rank}(\mathbf{V}_k^T \mathbf{X}) = k$, *then*

$$|\sin \boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{u}_i)| \leq \frac{\|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2}{|\phi(\sigma_i)|} \, \|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_2, \qquad 1 \leq i \leq k.$$

If, in addition, $\mathbf{X}$ has orthonormal columns, then

$$|\sin \boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{u}_i)| \leq \frac{\|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2}{|\phi(\sigma_i)|} \, \|\tan \boldsymbol{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_2, \qquad 1 \leq i \leq k.$$

Proof. See section 6. $\square$

In the special case when $\mathbf{X}$ has orthonormal columns, the angle between a *single* left singular vector and $\mathcal{K}_q$ is bounded by *all* angles between $\mathbf{X}$ and the right singular vector space range$(\mathbf{V}_k)$.

2.3. Approximations from a Krylov space. The results here are motivated by work in the theoretical computer science community on randomized linear algebra [15]. There, a common objective is the best rank-k approximation to $\mathbf{A}$ with respect to a unitarily invariant norm,

$$\mathbf{A}_k \equiv \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^T.$$

The particular approximation $\hat{\mathbf{U}}_k$ computed by the proto-algorithm, Algorithm 2.1, guarantees a strong optimality property in the projection $\hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^T \mathbf{A}$: It is the best rank-$k$ approximation to $\mathbf{A}$ from $\mathcal{K}_q$ with respect to the Frobenius norm (see Lemma 7.1).

Algorithm 2.1 Proto-algorithm for a low-rank approximation of $\mathbf{A}$ from $\mathcal{K}_q$.

Require: $\mathbf{A} \in \mathbb{R}^{m \times n}$, starting guess $\mathbf{X} \in \mathbb{R}^{n \times s}$
        Target rank $k < \mathrm{rank}(\mathbf{A})$, provided $\sigma_k > \sigma_{k+1}$
        Block dimension $q \geq 1$ with $k \leq (q+1)s \leq m$
Ensure: $\hat{\mathbf{U}}_k \in \mathbb{R}^{m \times k}$ with orthonormal columns
  1: Set $\mathbf{K}_q = \begin{pmatrix} \mathbf{AX} & (\mathbf{AA}^T)\mathbf{AX} & \cdots & (\mathbf{AA}^T)^q \mathbf{AX} \end{pmatrix} \in \mathbb{R}^{m \times (q+1)s}$,
     and assume that $\mathrm{rank}(\mathbf{K}_q) = (q+1)s$.
  2: Compute an orthonormal basis $\mathbf{U}_K \in \mathbb{R}^{m \times (q+1)s}$ for range$(\mathbf{K}_q)$.
  3: Set $\mathbf{W} \equiv \mathbf{U}_K^T \mathbf{A} \in \mathbb{R}^{(q+1)s \times n}$, and assume $\mathrm{rank}(\mathbf{W}) \geq k$.
  4: Compute an orthonormal basis $\mathbf{U}_{W,k} \in \mathbb{R}^{(q+1)s \times k}$ for the k dominant left singular vectors of $\mathbf{W}$.
  5: Return $\hat{\mathbf{U}}_k = \mathbf{U}_K \mathbf{U}_{W,k} \in \mathbb{R}^{m \times k}$.

Theorem 2.3 presents a quality-of-approximation result for $\hat{\mathbf{U}}_k$. To this end we distinguish the orthonormal columns of $\hat{\mathbf{U}}_k = \begin{pmatrix} \hat{\mathbf{u}}_1 & \ldots & \hat{\mathbf{u}}_k \end{pmatrix} \in \mathbb{R}^{m \times k}$ and set

$$(2.2) \qquad \hat{\mathbf{U}}_i \equiv \begin{pmatrix} \hat{\mathbf{u}}_1 & \ldots & \hat{\mathbf{u}}_i \end{pmatrix} \in \mathbb{R}^{m \times i}, \qquad 1 \leq i \leq k,$$

and

$$\Delta \equiv \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2 \, \|\mathbf{V}_{k,\perp}^T \mathbf{X} \, (\mathbf{V}_k^T \mathbf{X})^\dagger\|_F.$$

THEOREM 2.3. *Let $\phi(x)$ be a polynomial of degree $2q+1$ with odd powers only, such that $\phi(\boldsymbol{\Sigma}_k)$ is nonsingular and $\phi(\sigma_i) \geq \sigma_i$ for $1 \leq i \leq k$. If $\mathrm{rank}(\mathbf{V}_k^T \mathbf{X}) = k$, then for $1 \leq i \leq k$,*

$$(2.3) \qquad \|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_F \leq \|\mathbf{A} - \mathbf{A}_i\|_F + \Delta,$$

$$(2.4) \qquad \|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_2 \leq \|\mathbf{A} - \mathbf{A}_i\|_2 + \Delta,$$

$$(2.5) \qquad \sigma_i - \Delta \leq \|\hat{\mathbf{u}}_i^T \mathbf{A}\|_2 \leq \sigma_i.$$

If, in addition, $\mathbf{X}$ has orthonormal columns, then

$$\Delta = \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2 \, \|\tan \boldsymbol{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_F.$$

Proof. See section 7.   □

Bounds of the form (2.5) were already proposed in [34, Theorem 1] as a finer, *vectorwise*, way to capture the quality of approximations to individual left singular vectors of $\mathbf{A}$. Empirical evidence [34] suggests that error metrics of the form (2.3) and (2.4) indicate the quality of the *aggregate* approximation and are therefore coarser than (2.5).

2.4. Judicious choice of polynomials. We show the existence of and present bounds for the polynomials in Theorems 2.1, 2.2, and 2.3. The strict inequality $\mathrm{rank}(\mathbf{A}) > k$ in Algorithm 2.1 allows us to express the relative singular gap as

$$(2.6) \qquad \frac{\sigma_k - \sigma_{k+1}}{\sigma_{k+1}} \geq \gamma > 0,$$

which is equivalent to $\sigma_k \geq (1+\gamma)\sigma_{k+1} > 0$.

LEMMA 2.4. *If (2.6) holds, then there exists a polynomial $\phi(x)$ of degree $2q+1$ with odd powers only, such that $\phi(\sigma_i) \geq \sigma_i > 0$ for $1 \leq i \leq k$, and*

$$|\phi(\sigma_i)| \leq \frac{4\sigma_{k+1}}{2^{(2q+1)\,\min\{\sqrt{\gamma},1\}}}, \qquad i \geq k+1.$$

Hence

$$\|\phi(\boldsymbol{\Sigma}_k)^{-1}\|_2 \leq \sigma_k^{-1} \qquad and \qquad \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2 \leq \frac{4\sigma_{k+1}}{2^{(2q+1)\,\min\{\sqrt{\gamma},1\}}}.$$

Proof. See section 4.5.   □

We apply Lemma 2.4 to the previous results, first for the special case when $\mathbf{X}$ has linearly independent columns. Abbreviate

$$\Gamma(\boldsymbol{\Theta}, \gamma, q) \equiv 4\,\frac{\|\tan \boldsymbol{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_2}{2^{(2q+1)\,\min\{\sqrt{\gamma},1\}}}.$$

To keep things short, we consider only the two-norm bound for Theorem 2.1.

COROLLARY 2.5. *Let (2.6) hold, and let $\mathrm{rank}(\mathbf{V}_k^T \mathbf{X}) = k$. If $\mathbf{X}$ has orthonormal columns, then*

$$\|\sin \boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_2 \leq \Gamma(\boldsymbol{\Theta}, \gamma, q)\,\frac{\sigma_{k+1}}{\sigma_k} \leq \frac{\Gamma(\boldsymbol{\Theta}, \gamma, q)}{1+\gamma}$$

and

$$|\sin \boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{u}_i)| \leq \Gamma(\boldsymbol{\Theta}, \gamma, q)\,\frac{\sigma_{k+1}}{\sigma_i}, \qquad 1 \leq i \leq k.$$

Proof. Apply Lemma 2.4 to Theorems 2.1 and 2.2.   □

COROLLARY 2.6. *Let* (2.6) *hold, and let* $\mathrm{rank}(\mathbf{V}_k^T\mathbf{X}) = k$. *If* $\mathbf{X}$ *has orthonormal columns, then Theorem* 2.3 *holds with*

$$\Delta \leq \Gamma(\mathbf{\Theta}, \gamma, q)\,\sigma_{k+1},$$

so that for $1 \leq i \leq k$

$$\|\mathbf{A} - \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T\mathbf{A}\|_F \leq \|\mathbf{A} - \mathbf{A}_i\|_F + \Gamma(\mathbf{\Theta}, \gamma, q)\,\sigma_{k+1},$$
$$\|\mathbf{A} - \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T\mathbf{A}\|_2 \leq \|\mathbf{A} - \mathbf{A}_i\|_2 + \Gamma(\mathbf{\Theta}, \gamma, q)\,\sigma_{k+1},$$
$$\sigma_i - \Gamma(\mathbf{\Theta}, \gamma, q)\,\sigma_{k+1} \leq \|\hat{\mathbf{u}}_i^T\mathbf{A}\|_2 \leq \sigma_i.$$

Proof. Apply Lemma 2.4 to Theorem 2.3.   □

To achieve an additive error of $\Gamma(\mathbf{\Theta}, \gamma, q) \leq \epsilon$, set q to be the smallest integer that exceeds

$$(2.7) \qquad q \geq \frac{1}{2\min\left\{\sqrt{\gamma}, 1\right\}}\left(\log_2 4\|\tan\mathbf{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_2 - \log_2\epsilon\right).$$

Thus, as the singular value gap γ decreases, the dimension of the space $\mathcal{K}_q$ increases. More specifically, q increases logarithmically with higher target accuracy ϵ and increasing distance of $\mathbf{X}$ from the dominant right singular vector space of $\mathbf{A}$.

If $\mathbf{X}$ is rank deficient, then Corollaries 2.5 and 2.6 still hold with

$$\Gamma(\mathbf{\Theta}, \gamma, q) = 4\,\frac{\|\mathbf{V}_{k,\perp}^T\mathbf{X}(\mathbf{V}_k^T\mathbf{X})^\dagger\|_2}{2^{(2q+1)\,\min\left\{\sqrt{\gamma}, 1\right\}}}.$$

2.5. The initial guess. It remains to bound $\|\mathbf{V}_{k,\perp}^T\mathbf{X}(\mathbf{V}_k^T\mathbf{X})^\dagger\|_{2,F}$. The simplest way might be strong submultiplicativity,

$$\|\mathbf{V}_{k,\perp}^T\mathbf{X}(\mathbf{V}_k^T\mathbf{X})^\dagger\|_{2,F} \leq \|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_{2,F}\,\|(\mathbf{V}_k^T\mathbf{X})^\dagger\|_2 = \frac{\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_{2,F}}{\sigma_k(\mathbf{V}_k^T\mathbf{X})},$$

followed by separate bounds for the individual factors.

Ideally, the starting guess $\mathbf{X}$ should be close to $\mathrm{range}(\mathbf{V}_k)$ and far away from $\mathrm{range}(\mathbf{V}_{k,\perp})$, so that $\sigma_k(\mathbf{V}_k^T\mathbf{X})$ is large and $\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_{2,F}$ is small. The assumption $\sigma_k(\mathbf{V}_k^T\mathbf{X}) > 0$ is critical for our results, hence a necessary condition for the user-specified matrix $\mathbf{X} \in \mathbb{R}^{n \times s}$ is $\mathrm{rank}(\mathbf{X}) \geq k$, while trying to keep the column dimension $s \geq k$ small.

If $\mathbf{X}$ is a random Gaussian, then $\sigma_k(\mathbf{V}_k^T\mathbf{X})$ is bounded away from zero with high probability even for $s = k$. However, there are many other choices for $\mathbf{X}$ that come with lower bounds for $\sigma_k(\mathbf{V}_k^T\mathbf{X})$. They include random sign matrices [1, 30], the fast randomized Hadamard transform [2, 39], the subsampled randomized Hadamard transform [17, 42], the fast randomized discrete cosine transform [36], and input sparsity time embeddings [13, 33, 35].

In contrast, keeping $\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_F$ small is relatively easy. For typical random matrices $\mathbf{X}$, one can show that, with high probability,

$$\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_{2,F} \leq c\,\|\mathbf{V}_{k,\perp}\|_F \leq c\sqrt{n},$$

where c is a small constant.

For instance, if $s = 1$ and $\mathbf{X}$ is a Gaussian column vector, then

$$(2.8) \qquad \mathbf{E}\left[\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_F^2\right] = \|\mathbf{V}_{k,\perp}^T\|_F^2 \le n.$$

Markov's inequality guarantees that, with probability at least .9,

$$\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_F \le \sqrt{10n}.$$

Essentially all randomized embedding matrices satisfy variants of (2.8), and we expect the iteration count q in (2.7) to be logarithmic in n.

From a numerical point of view, a starting guess $\mathbf{X}$ with orthonormal columns is preferable. Thus one could pick a random matrix $\mathbf{X}$ and apply a thin QR decomposition $\mathbf{X} = \mathbf{QR}$. However, this significantly complicates the derivation of bounds for $\|\mathbf{V}_{k,\perp}^T\mathbf{X}\|_{2,F}$ and $\|(\mathbf{V}_k^T\mathbf{X})^\dagger\|_2$, as most matrix concentration inequalities apply only to the original random matrix $\mathbf{X}$, not to its orthonormal basis $\mathbf{Q}$. For instance, if $\mathbf{X}$ is a random matrix whose entries are ± 1 with equal probability, then $\mathbf{Q}$ does not inherit this property. Fortunately, the subsampled Hadamard transform [17, 42] is one of a few random matrices with orthonormal columns, hence amenable to application of matrix concentration inequalities.

3. Comparison to existing work. Our work on subspace computations is motivated by a recent probabilistic approach for low-rank approximations via block Krylov spaces [34].

Randomized methods. Analyses of numerical methods that compute dominant subspaces and eigenvectors from randomized starting vectors date back at least to the 1980s. They include the power method and inverse iteration [14, 24] and information theoretic analyses of Lanczos methods [27, 28].

Current analyses in theoretical computer science focus on low-rank approximations [21, 44], rather than subspace computations, and as such tend not to produce bounds for the accuracy of subspaces such as those in section 2.2.

A popular approach towards low-rank approximation is subspace iteration, which makes use of only the last iterate $(\mathbf{AA}^T)^q\mathbf{AX}$ [21, 44]. Then came block Krylov methods, which exploit all of the iterates $(\mathbf{AA}^T)^j\mathbf{AX}$, $0 \le j \le q$. The analysis in [34] relies on generalized matrix functions [3, 22], but is limited to Gaussian random matrices for starting guesses $\mathbf{X} \in \mathbb{R}^{n \times k}$, and to Chebyshev polynomials for ϕ. The eponymous *gap-dependent bound* [34, Theorem 13] requires a gap between the kth and $(k + 1)$st singular values, and can be considered a special case of Theorem 2.3. However, [34, Theorems 10, 11, and 12] do not require a singular value gap such as (2.6). Such *gap-independent bounds* are informative for low-rank approximations, but not for computations of specific subspaces, as explained in section 1.

Close on the heels of [34] is [43], with a focus on gap-independent bounds of the type (2.4) and random Gaussian starting guesses [43, Theorem 3.1]. The proof techniques in [43] resemble ours and leverage our prior work; see Lemma 7.1 and [10].

Traditional, deterministic methods. Although non-numerical in nature as well, our results are nevertheless guided in spirit by foundational work on eigenvalue and invariant subspace computations, including the standard Lanczos convergence analysis [38, section 6.6], a geometric view of Krylov space methods [6, 7], block Lanczos methods [29, 37], and Rayleigh–Ritz bounds [9, 18], but also by Krylov space methods for singular value problems [4, 5].

A more detailed comparison, though, seems elusive due to differences in both the computational problem and the algorithm. The analyses in [6, 7] target vector

rather than block methods, for eigenvalues and invariant subspaces of non-Hermitian matrices, with a concern for restarting. The block methods in [29, 37] are Lanczos methods for Hermitian eigenvalue problems, and the analyses exploit the (block) tridiagonal structure resulting from recursions. Although singular value problems are considered in [4] and in [5] with block methods, the Krylov spaces are different and the focus is on algorithmic issues of augmenting and restarting the Lanczos process, rather than subspace distances. For the solution of ill-posed least squares problems via LSQR, [25] analyzes the accuracy of a regularized solution by bounding the sine between $\mathcal{K}_j(\mathbf{A}^T\mathbf{A}, \mathbf{A}^T\mathbf{v})$ and a dominant right singular vector space; however, all singular values must be distinct.

In the context of low-rank approximations, [40] proposes a Lanczos bidiagonalization with one-sided reorthogonalization.

In contrast, the context of this paper is singular vector spaces for general matrices of any dimension; and an algorithm that is not tied to a particular recursion and, due to steps 3–5, is not a straightforward Krylov method. Furthermore, the key feature of our analyses is a least squares approach [10, 11] that assures the quality of the approximation.

4. Auxiliary results. We review submultiplicative inequalities for norms, the matrix Pythagoras theorem, and solutions of multiple right-hand side least squares problems in the two norm (section 4.1). We also present the matrix functions to represent elements in the Krylov space $\mathcal{K}_q$ (section 4.2), review angles between subspaces (section 4.3), and introduce gap-amplifying polynomials (section 4.4).

4.1. Norm inequalities, Pythagoras, and least squares. We make frequent use of the strong submultiplicativity of the Frobenius norm [23, p. 211]. For matrices $\mathbf{Y}_1 \in \mathbb{R}^{m\times k}$ and $\mathbf{Y}_2 \in \mathbb{R}^{k\times n}$,

$$\|\mathbf{Y}_1\mathbf{Y}_2\|_F \leq \|\mathbf{Y}_1\|_2\|\mathbf{Y}_2\|_F,$$
$$\|\mathbf{Y}_1\mathbf{Y}_2\|_F \leq \|\mathbf{Y}_1\|_F\|\mathbf{Y}_2\|_2.$$

If, in addition, $\mathrm{rank}(\mathbf{Y}_1) = \mathrm{rank}(\mathbf{Y}_2) = k$, then [8, Theorem 2.2.3]

(4.1)
$$(\mathbf{Y}_1\mathbf{Y}_2)^\dagger = \mathbf{Y}_2^\dagger\mathbf{Y}_1^\dagger.$$

LEMMA 4.1 (matrix Pythagoras). *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$. *If* $\mathbf{A}^T\mathbf{B} = \mathbf{0}$, *then*

$$\|\mathbf{A} + \mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2.$$

Proof. From $\mathbf{A}^T\mathbf{B} = \mathbf{0}$ and the linearity of the trace follows

$$\|\mathbf{A} + \mathbf{B}\|_F^2 = \mathrm{trace}\left(\mathbf{A}^T\mathbf{A} + \mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}\right)$$
$$= \mathrm{trace}(\mathbf{A}^T\mathbf{A}) + \mathrm{trace}(\mathbf{B}^T\mathbf{B}) = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2. \qquad \square$$

References to optimal solutions for multiple right-hand side least squares problems in the two norm are hard to find. Lemma 4.2 below is easy to prove in the Frobenius norm. An elaborate proof for Schatten p norms can be found in [31, Theorems 3.2 and 3.3]. For the sake of completeness, we present a straightforward proof for the two norm.

LEMMA 4.2. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times p}$. *Then*[4] $\mathbf{A}^\dagger\mathbf{B}$ *is a solution of*

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}} \|\mathbf{B} - \mathbf{A}\mathbf{X}\|_2$$

---

[4]The superscript † denotes the Moore–Penrose inverse.

with least squares residual $\|(\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)\mathbf{B}\|_2 = \min_{\mathbf{X} \in \mathbb{R}^{n \times p}} \|\mathbf{B} - \mathbf{A}\mathbf{X}\|_2$.

Proof. Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be a thin SVD, and let $\begin{pmatrix} \mathbf{U} & \mathbf{U}_\perp \end{pmatrix} \in \mathbb{R}^{m \times m}$ be an orthogonal matrix. Any $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfies

$$\begin{aligned} \|\mathbf{B} - \mathbf{A}\mathbf{X}\|_2^2 &= \|\mathbf{U}\mathbf{U}^T(\mathbf{B} - \mathbf{A}\mathbf{X}) + \mathbf{U}_\perp \mathbf{U}_\perp^T (\mathbf{B} - \mathbf{A}\mathbf{X})\|_2^2 \\ &= \|\mathbf{U}(\mathbf{U}^T\mathbf{B} - \boldsymbol{\Sigma}\mathbf{V}^T\mathbf{X}) + \mathbf{U}_\perp \mathbf{U}_\perp^T\mathbf{B}\|_2^2 \\ &= \|\mathbf{U}\mathbf{T}_1 + \mathbf{U}_\perp \mathbf{T}_2\|_2^2, \end{aligned}$$

where $\mathbf{T}_1 \equiv \mathbf{U}^T\mathbf{B} - \boldsymbol{\Sigma}\mathbf{V}^T\mathbf{X}$ and $\mathbf{T}_2 \equiv \mathbf{U}_\perp^T\mathbf{B}$. Let $\mathbf{y}_{opt} \in \mathbb{R}^p$ with $\|\mathbf{y}_{opt}\|_2 = 1$ satisfy $\|\mathbf{U}_\perp \mathbf{T}_2 \mathbf{y}_{opt}\|_2^2 = \|\mathbf{U}_\perp \mathbf{T}_2\|_2^2$. The vector Pythagoras theorem implies

$$\begin{aligned} \|\mathbf{B} - \mathbf{A}\mathbf{X}\|_2^2 &\geq \|(\mathbf{U}\mathbf{T}_1 + \mathbf{U}_\perp \mathbf{T}_2)\mathbf{y}_{opt}\|_2^2 \\ &= \|\mathbf{U}\mathbf{T}_1 \mathbf{y}_{opt}\|_2^2 + \|\mathbf{U}_\perp \mathbf{T}_2 \mathbf{y}_{opt}\|_2^2 \geq \|\mathbf{U}_\perp \mathbf{T}_2 \mathbf{y}_{opt}\|_2^2. \end{aligned}$$

Combining all of the above gives

$$\|\mathbf{B} - \mathbf{A}\mathbf{X}\|_2^2 \geq \|(\mathbf{U}\mathbf{T}_1 + \mathbf{U}_\perp \mathbf{T}_2)\mathbf{y}_{opt}\|_2^2 \geq \|\mathbf{U}_\perp \mathbf{T}_2\|_2^2 = \|\mathbf{U}_\perp \mathbf{U}_\perp^T \mathbf{B}\|_2^2.$$

This lower bound is achieved by $\mathbf{X}_{opt} = \mathbf{A}^\dagger \mathbf{B}$,

$$\|\mathbf{B} - \mathbf{A}\mathbf{X}_{opt}\|_2^2 = \|\mathbf{B} - \mathbf{A}\mathbf{A}^\dagger\mathbf{B}\|_2^2 = \|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{B}\|_2^2 = \|\mathbf{U}_\perp \mathbf{U}_\perp^T\mathbf{B}\|_2^2. \qquad \square$$

4.2. Matrix functions for the Krylov space. The elements of the Krylov space $\mathcal{K}_q$ in (2.1) can be expressed in terms of matrices $\hat{\phi}(\mathbf{A}\mathbf{A}^T)\mathbf{A}\mathbf{X} \in \mathbb{R}^{m \times s}$, where $\hat{\phi}$ is a polynomial of degree q. From the point of view of singular values, though, we need a higher degree polynomial,

$$\hat{\phi}(\mathbf{A}\mathbf{A}^T)\,\mathbf{A}\mathbf{X} = \mathbf{U}\,\hat{\phi}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T)\boldsymbol{\Sigma}\,\mathbf{V}^T\mathbf{X} = \mathbf{U}\,\phi(\boldsymbol{\Sigma})\,\mathbf{V}^T\mathbf{X}.$$

Here ϕ is a polynomial of degree $2q + 1$ with odd powers only and represents a *generalized matrix function* [3, 22]. Since

$$\boldsymbol{\Sigma} = \mathrm{diag}\begin{pmatrix} \sigma_1 & \cdots & \sigma_{\min\{m,n\}} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

is rectangular, the polynomial ϕ is applied to the diagonal elements of $\boldsymbol{\Sigma}$ only and returns a diagonal matrix of the same dimension,

$$\phi(\boldsymbol{\Sigma}) \equiv \mathrm{diag}\begin{pmatrix} \phi(\sigma_1) & \cdots & \phi(\sigma_{\min\{m,n\}}) \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

With this, we denote elements in $\mathcal{K}_q$ by

(4.2) $$\boldsymbol{\Phi} \equiv \mathbf{U}\phi(\boldsymbol{\Sigma})\mathbf{V}^T\mathbf{X} \ \in \mathbb{R}^{m \times s}.$$

Clearly,

(4.3) $$\mathrm{range}(\boldsymbol{\Phi}) \subset \mathcal{K}_q.$$

The assumption $\dim(\mathcal{K}_q) = (q+1)s \leq m$ in step 1 of Algorithm 2.1 guarantees that $\mathbf{U}_K$ has exactly $(q+1)s$ orthonormal columns, which form a basis for $\mathcal{K}_q$.

4.3. Angles between subspaces. Let $\mathbf{Q} \in \mathbb{R}^{n \times s}$ and $\mathbf{W}_k \in \mathbb{R}^{n \times k}$, with $k \leq s$, be matrices with orthonormal columns. Hence, the singular values $\sigma_j(\mathbf{W}_k^T \mathbf{Q})$ lie between zero and one, and we can write $\sigma_j(\mathbf{W}_k^T \mathbf{Q}) = \cos \theta_j$, $1 \leq j \leq k$. The *principal* or *canonical angles* between range($\mathbf{Q}$) and range($\mathbf{W}_k$) are [19, section 6.4.3]

$$0 \leq \theta_1 \leq \cdots \leq \theta_k \leq \pi/2,$$

where $\theta_j = \cos^{-1}(\sigma_j(\mathbf{W}_k^T \mathbf{Q}))$. Following [41, Definition I.5.3], we define the diagonal matrix of principal angles between the subspaces spanned by the columns of $\mathbf{Q}$ and the columns of $\mathbf{W}_k$:

$$\boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \equiv \mathrm{diag}\begin{pmatrix} \theta_1 & \cdots & \theta_k \end{pmatrix}.$$

Hence the singular values of $\mathbf{W}_k^T \mathbf{Q}$ are the diagonal elements of $\cos \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k)$. From [19, section 6.4.3] and [41, section I.5.3] follows that the distance between range($\mathbf{W}_k$) and range($\mathbf{Q}$) in the two- and Frobenius norms, respectively, equals

$$(4.4) \qquad \| \sin \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_{2,F} = \|(\mathbf{I} - \mathbf{W}_k \mathbf{W}_k^T)\mathbf{Q}\|_{2,F}.$$

In particular, $\| \sin \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_2 = \sin \theta_k$, so the two-norm distance is determined by the largest principal angle.

Assume that range($\mathbf{W}_k$) and range($\mathbf{Q}$) are sufficiently close, so that the largest angle $\theta_k < \pi/2$. This is equivalent to $\cos \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k)$ being nonsingular, and rank($\mathbf{W}_k^T \mathbf{Q}$) $= k$. Then [45, section 3] implies that the tangents of the principal angles satisfy

$$\| \tan \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_{2,F} = \| \sin \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k)(\cos \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k))^\dagger \|_{2,F}$$

$$(4.5) \qquad\qquad\qquad = \left\| (\mathbf{I} - \mathbf{W}_k \mathbf{W}_k^T)\mathbf{Q} \left(\mathbf{W}_k^T \mathbf{Q} \right)^\dagger \right\|_{2,F}.$$

As above, $\| \tan \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_2 = \tan \theta_k$, so the two-norm tangent is determined by the largest principal angle. The following lemma will be used in subsequent derivations, so we include its simple proof.

LEMMA 4.3 (Theorem 3.1 in [45]). *Let* $\mathbf{Q} \in \mathbb{R}^{n \times s}$ *have orthonormal columns, and let* $\mathbf{W} \equiv \begin{pmatrix} \mathbf{W}_k & \mathbf{W}_{k,\perp} \end{pmatrix} \in \mathbb{R}^{n \times n}$ *be an orthogonal matrix where* $\mathbf{W}_k \in \mathbb{R}^{n \times k}$ *with* $k \leq s$. *If* rank($\mathbf{W}_k^T \mathbf{Q}$) $= k$, *then*

$$\| \tan \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_{2,F} = \|(\mathbf{W}_{k,\perp}^T \mathbf{Q})(\mathbf{W}_k^T \mathbf{Q})^\dagger\|_{2,F}.$$

Proof. From (4.5) and $\mathbf{I} = \mathbf{W}_k \mathbf{W}_k^T + \mathbf{W}_{k,\perp} \mathbf{W}_{k,\perp}^T$ follows

$$\| \tan \boldsymbol{\Theta}(\mathbf{Q}, \mathbf{W}_k) \|_{2,F} = \left\| (\mathbf{I} - \mathbf{W}_k \mathbf{W}_k^T)\mathbf{Q} \left(\mathbf{W}_k^T \mathbf{Q} \right)^\dagger \right\|_{2,F}$$

$$= \left\| \mathbf{W}_{k,\perp} \mathbf{W}_{k,\perp}^T \mathbf{Q} \left(\mathbf{W}_k^T \mathbf{Q} \right)^\dagger \right\|_{2,F}$$

$$= \left\| \mathbf{W}_{k,\perp}^T \mathbf{Q} \left(\mathbf{W}_k^T \mathbf{Q} \right)^\dagger \right\|_{2,F}.$$

The last equality follows from the unitary invariance of the two- and the Frobenius norms. □

4.4. Gap-amplifying and Chebyshev polynomials. We generalize the Chebyshev-based *gap-amplifying polynomials* in [34, section 4.4], [43, section 2.2]. Given an integer $q \geq 1$, define the polynomial

$$\psi_{q'}(x) = \psi_{2q+1}(x) = \sum_{j=0}^{q} a_{2j+1} x^{2j+1}$$

of degree $q' = 2q + 1$ with only odd powers of x. The polynomial ψ is *gap-amplifying* if it satisfies three properties:

1. Small input values remain small,

$$\psi_{q'}(1) = 1, \qquad \text{and} \qquad |\psi_{q'}(x)| \leq 1 \quad \text{for } x \in [0, 1].$$

2. Large input values are amplified,

$$\psi_{q'}(x) \geq x\, c\, 2^{q'\, r(x)} \quad \text{for } x \geq 1,$$

where the constant c and the function $r(x)$ are parameters of ψ.

3. Large input values exhibit superlinear growth,

$$\frac{\psi_{q'}(x)}{x} \geq \frac{\psi_{q'}(y)}{y} \quad \text{for } x \geq y \geq 1.$$

The simplest gap-amplifying polynomial is a Chebyshev polynomial.

LEMMA 4.4 (Lemma 5 in [34]). *The Chebyshev polynomial $T_{q'}(x)$ of the first kind contains only odd powers of x and is gap-amplifying with $c = 1/4$ and $r(x) = \min\{\sqrt{x-1}, 1\}$.*

Proof. We give a quick sketch of the proof of [34, Lemma 5]. Clearly, the Chebyshev polynomial $T_{q'}(x)$ satisfies property 1. To prove property 3, it suffices to show

$$(4.6) \qquad T'_{q'}(z) \geq \frac{T_{q'}(y)}{y} \quad \text{for some } z \geq y \geq 1,$$

because the mean value theorem implies that there exists a $z \in [y, x]$ with

$$T_{q'}(x) = T_{q'}(y + x - y) = T_{q'}(y) + T'_{q'}(z)(x - y)$$
$$\geq T_{q'}(y) + \frac{T_{q'}(y)}{y}(x - y) = x\frac{T_{q'}(y)}{y}.$$

Our proof of property 2 corrects a small typo in a similar proof in [34]. Although a bound equivalent to property 2 is claimed in [34, Lemma 5], it is only proved that $T_{q'}(x) \geq c\, 2^{q'\, r(x)}$. However, only a slight modification is required for the stronger result. The proof of [34, Lemma 5] shows that

$$T_{q'}(x) \geq \tfrac{1}{2}\, 2^{q'\, \sqrt{x-1}} \quad \text{for } 1 \leq x \leq 2.$$

To derive a lower bound on $T_{q'}(x)/x$ for $x \geq 1$, first consider $1 \leq x \leq 2$, where

$$\frac{T_{q'}(x)}{x} \geq \frac{T_{q'}(x)}{2} \geq \tfrac{1}{4}\, 2^{q'\, \sqrt{x-1}}.$$

For the remaining case $x > 2$, property 3 implies

$$\frac{T_{q'}(x)}{x} \geq \frac{T_{q'}(2)}{2} \geq \tfrac{1}{4}\, 2^{q'}.$$

Hence $T_{q'}(x) \geq \frac{x}{4}\, 2^{q'\, \min\{\sqrt{x-1}, 1\}}$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

For our analysis, we use a rescaled version of the gap-amplifying polynomial $T_{q'}(x)$, which has properties similar to those of the original.

LEMMA 4.5. *Let*

$$(4.7) \qquad \phi(x) = \frac{(1+\gamma)\,\alpha}{\psi_{q'}\,(1+\gamma)}\psi_{q'}(x/\alpha)$$

be the rescaled gap-amplifying polynomial. Then

$$|\phi(x)| \leq \frac{4\alpha}{2^{q' \, \min\{\sqrt{\gamma},1\}}} \qquad \textit{for } 0 \leq x \leq \alpha.$$

Proof. The proof is immediate since $|\psi_{q'}(x/\alpha)| \leq 1$ for $0 \leq x \leq \alpha$, and property 2 implies

$$\psi_{q'}\,(1+\gamma) \geq \frac{(1+\gamma)}{4}\,2^{q' \, \min\{\sqrt{\gamma},1\}}. \qquad \qquad \Box$$

LEMMA 4.6. *The rescaled gap-amplifying polynomial $\phi(x)$ in (4.7) satisfies*

$$\phi(x) \geq x \qquad \textit{for } x \geq (1+\gamma)\alpha.$$

Proof. Property 3 implies

$$\frac{\psi_{q'}(x/\alpha)}{x/\alpha} \geq \frac{\psi_{q'}(1+\gamma)}{1+\gamma} \qquad \text{for } x \geq (1+\gamma)\alpha.$$

Now rearrange terms and apply the definition of $\phi(x)$ in (4.7). $\qquad \Box$

4.5. Proof of Lemma 2.4. Let $\phi(x)$ be the rescaled gap-amplifying polynomial in (4.7) with $\alpha = \sigma_{k+1}$ and γ in (2.6). The inequalities for $\phi(\sigma_i)$ follow from $q' = 2q+1$, Lemma 4.5, and Lemma 4.6.

From $\phi(\sigma_i) > 0$ for $1 \leq i \leq k$ and $\phi(\mathbf{\Sigma}_k)$ being a diagonal matrix follows

$$\|\phi(\mathbf{\Sigma}_k)^{-1}\|_2 = \max_{1 \leq i \leq k} \phi(\sigma_i)^{-1} \leq \max_{1 \leq i \leq k} \sigma_i^{-1} = \sigma_k^{-1}.$$

Furthermore

$$\|\phi(\mathbf{\Sigma}_{k,\perp})\|_2 = \max_{i \geq k+1} |\phi(\sigma_i)| \leq \frac{4\sigma_{k+1}}{2^{(2q+1) \, \min\{\sqrt{\gamma},1\}}}.$$

5. Proof of Theorem 2.1, for general and orthonormal X. We focus on the case where $\mathbf{X}$ is a general matrix or has orthonormal columns, and postpone the technicalities required for the full column-rank case to Appendix A.

The first and most critical step of the proof makes a connection between principal angles and least squares residuals.

Viewing the sine as a least squares residual. Let $\mathcal{P}_q$ be the orthogonal projector onto the Krylov space $\mathcal{K}_q$. For $\mathbf{\Phi}$ in (4.2), let $\mathbf{\Phi}\mathbf{\Phi}^\dagger$ be the orthogonal projector onto range($\mathbf{\Phi}$), with range($\mathbf{\Phi}\mathbf{\Phi}^\dagger$) $\subset$ range($\mathcal{P}_q$) due to (4.3). Hence, (4.4) implies

$$(5.1) \qquad \| \sin \mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} = \|(\mathbf{I} - \mathcal{P}_q)\,\mathbf{U}_k\|_{2,F} \leq \|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F}.$$

Lemma 4.2 implies that $\|(\mathbf{I}-\mathbf{\Phi}\mathbf{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F}$ is the residual of the least squares problem

$$\|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F} = \min_{\mathbf{\Psi}} \|\mathbf{U}_k - \mathbf{\Phi}\mathbf{\Psi}\|_{2,F} = \|\mathbf{U}_k - \mathbf{\Phi}\mathbf{\Psi}_{opt}\|_{2,F},$$

where $\mathbf{\Psi}_{opt} = \mathbf{\Phi}^\dagger\mathbf{U}_k$ is a least squares solution.

Focusing on the target space. Decompose $\mathbf{\Phi}$ into the target component $\mathrm{range}(\mathbf{U}_k)$ and the complementary subspace, $\mathbf{\Phi} = \mathbf{\Phi}_k + \mathbf{\Phi}_{k,\perp}$, where

$$(5.2) \qquad \mathbf{\Phi}_k \equiv \mathbf{U}_k \phi(\mathbf{\Sigma}_k) \mathbf{V}_k^T \mathbf{X}, \qquad \mathbf{\Phi}_{k,\perp} \equiv \mathbf{U}_{k,\perp} \phi(\mathbf{\Sigma}_{k,\perp}) \mathbf{V}_{k,\perp}^T \mathbf{X}.$$

From $\mathrm{rank}(\mathbf{V}_k^T \mathbf{X}) = k$ follows that $(\mathbf{V}_k^T \mathbf{X})^\dagger$ is a right inverse, $(\mathbf{V}_k^T \mathbf{X})(\mathbf{V}_k^T \mathbf{X})^\dagger = \mathbf{I}_k$. With (4.1) this gives

$$(5.3) \qquad \mathbf{\Phi}_k^\dagger = (\mathbf{V}_k^T \mathbf{X})^\dagger \phi(\mathbf{\Sigma}_k)^{-1} \mathbf{U}_k^T \quad \text{and} \quad \mathbf{\Phi}_k \mathbf{\Phi}_k^\dagger = \mathbf{U}_k \mathbf{U}_k^T,$$

meaning $\mathbf{\Phi}_k \mathbf{\Phi}_k^\dagger$ is the orthogonal projector onto the target space $\mathrm{range}(\mathbf{U}_k)$. The minimality of the least squares residual implies

$$\|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F} = \|\mathbf{U}_k - \mathbf{\Phi}\,(\mathbf{\Phi}^\dagger \mathbf{U}_k)\|_{2,F}$$
$$\leq \|\mathbf{U}_k - \mathbf{\Phi}\,(\mathbf{\Phi}_k^\dagger \mathbf{U}_k)\|_{2,F} = \|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}_k^\dagger)\,\mathbf{U}_k\|_{2,F}.$$

Now replace the other instance of $\mathbf{\Phi}$ by (5.2), and use (5.3) to simplify

$$\|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F} \leq \|(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}_k^\dagger)\,\mathbf{U}_k\|_{2,F} = \|(\mathbf{I} - \mathbf{\Phi}_k \mathbf{\Phi}_k^\dagger)\mathbf{U}_k - \mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{U}_k\|_{2,F}$$
$$= \|(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^T)\mathbf{U}_k - \mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{U}_k\|_{2,F}$$
$$(5.4) \qquad = \|\mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{U}_k\|_{2,F}.$$

Summary so far. Combining (5.1) with (5.4) gives

$$\|\sin\mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} \leq \|\mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{U}_k\|_{2,F}.$$

Extracting the polynomials. The expressions for $\mathbf{\Phi}_{k,\perp}$ in (5.2) and $\mathbf{\Phi}_k^\dagger$ in (5.3), and submultiplicativity (section 4.1) yield

$$\|\mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{U}_k\|_{2,F} = \|\mathbf{U}_{k,\perp}\,\phi(\mathbf{\Sigma}_{k,\perp})\,\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\,\phi(\mathbf{\Sigma}_k)^{-1}\mathbf{U}_k^T \mathbf{U}_k\|_{2,F}$$
$$= \|\phi(\mathbf{\Sigma}_{k,\perp})\,\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\,\phi(\mathbf{\Sigma}_k)^{-1}\|_{2,F}$$
$$\leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2\,\|\phi(\mathbf{\Sigma}_k)^{-1}\|_2\,\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_{2,F}.$$

Combining the previous two sets of inequalities gives

$$\|\sin\mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} \leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2\,\|\phi(\mathbf{\Sigma}_k)^{-1}\|_2\,\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_{2,F}.$$

This concludes the proof for general $\mathbf{X}$. The proof for the special case where $\mathbf{X}$ has linearly independent columns follows from Lemma 4.3.

6. Proof of Theorem 2.2. The proof imitates that of Theorem 2.1 and simply substitutes the vectors $\mathbf{u}_i$ for the matrix $\mathbf{U}_k$. Note that $(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^T)\mathbf{u}_i = 0$ for $1 \leq i \leq k$, which implies

$$|\sin\mathbf{\Theta}(\mathcal{K}_q, \mathbf{u}_i)| \leq \|\mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{u}_i\|_2, \qquad 1 \leq i \leq k.$$

The expressions for $\mathbf{\Phi}_{k,\perp}$ in (5.2) and $\mathbf{\Phi}_k^\dagger$ in (5.3), and submultiplicativity yield

$$\|\mathbf{\Phi}_{k,\perp}\mathbf{\Phi}_k^\dagger \mathbf{u}_k\|_2 = \|\mathbf{U}_{k,\perp}\,\phi(\mathbf{\Sigma}_{k,\perp})\,\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\,\phi(\mathbf{\Sigma}_k)^{-1}\mathbf{U}_k^T \mathbf{u}_i\|_2$$
$$= \|\phi(\mathbf{\Sigma}_{k,\perp})\,\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\,\phi(\sigma_i)^{-1}\|_2$$
$$\leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2\,|\phi(\sigma_i)^{-1}|\,\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_2.$$

Combining the previous two sets of inequalities gives

$$|\sin\mathbf{\Theta}(\mathcal{K}_q, \mathbf{u}_i)| \leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2\,|\phi(\sigma_i)^{-1}|\,\|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_2.$$

This concludes the proof of the case for general $\mathbf{X}$. The proof for the special case where $\mathbf{X}$ has orthonormal columns follows from Lemma 4.3.

7. Proof of Theorem 2.3. This proof is more involved than the previous ones and requires two auxiliary results, an alternative expression for the error (section 7.1), and a bound on its Frobenius norm (section 7.2).

7.1. An alternative expression for the error. Algorithm 2.1 approximates the dominant left singular vectors of $\mathbf{A}$ by the orthonormal matrix $\hat{\mathbf{U}}_k \in \mathbb{R}^{m \times k}$. Since bounding $\|\mathbf{A} - \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^T \mathbf{A}\|_F$ seems hard, we present an alternative expression that is easier to analyze.

LEMMA 7.1 (Lemma 8 in [11]). *Let* $\mathbf{U}_K$ *be an orthonormal basis for* $\mathcal{K}_q$ *and let* $\hat{\mathbf{U}}_i$ *be as defined in* (2.2), *containing the top* i *columns of the output of Algorithm* 2.1. *Then*

$$(7.1) \qquad \mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A} = \mathbf{A} - \mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_i, \qquad 1 \le i \le k.$$

In addition, $\mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_i$ *is a best rank-i approximation to* $\mathbf{A}$ *from* $\mathcal{K}_q$ *in the Frobenius norm,*

$$(7.2) \qquad \left\| \mathbf{A} - \mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_i \right\|_F^2 = \min_{\mathrm{rank}(\mathbf{Y}) \le i} \|\mathbf{A} - \mathbf{U}_K \mathbf{Y}\|_F^2, \qquad 1 \le i \le k.$$

Proof. Since the transition to best rank-i approximations is a key component, we illustrate how it comes about by proving the first assertion for the case $i = k$.

Algorithm 2.1 outputs $\hat{\mathbf{U}}_k = \mathbf{U}_K \mathbf{U}_{W,k}$, where $\mathbf{U}_{W,k}$ is the matrix of the dominant k left singular vectors of $\mathbf{W} = \mathbf{U}_K^T \mathbf{A}$. This means $\mathbf{U}_{W,k}$ spans the same range as $\mathbf{W}_k$, the best rank-k approximation to $\mathbf{W}$. Therefore

$$\mathbf{A} - \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^T \mathbf{A} = \mathbf{A} - \mathbf{U}_K \mathbf{U}_{W,k} \mathbf{U}_{W,k}^T \mathbf{U}_K^T \mathbf{A}$$
$$= \mathbf{A} - \mathbf{U}_K \mathbf{W}_k \mathbf{W}_k^\dagger \mathbf{W} = \mathbf{A} - \mathbf{U}_K \mathbf{W}_k = \mathbf{A} - \mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_k.$$

The penultimate equality follows from $\mathbf{W}_k \mathbf{W}_k^\dagger$ being the orthogonal projector onto range($\mathbf{W}_k$). □

Lemma 7.1 shows that (2.3) in Theorem 2.3 can be proved by bounding $\|\mathbf{A} - \mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_i \|_F$. Next we transition from the best rank-i approximation of the "projected" matrix $(\mathbf{U}_K^T \mathbf{A})_i$ to the best rank-i approximation $\mathbf{A}_i$ of the original matrix, by splitting, for $1 \le i \le k$,

$$(7.3) \quad \mathbf{A} = \mathbf{A}_i + \mathbf{A}_{i,\perp}, \qquad \text{where} \quad \mathbf{A}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T \text{ and } \mathbf{A}_{i,\perp} = \mathbf{U}_{i,\perp} \mathbf{\Sigma}_{i,\perp} \mathbf{V}_{i,\perp}^T.$$

LEMMA 7.2. *Let* $\mathbf{U}_K$ *be an orthonormal basis for* $\mathcal{K}_q$, *and let* $\hat{\mathbf{U}}_i$ *in* (2.2) *be the columns of the output of Algorithm* 2.1. *Then*

$$\|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_F^2 \le \|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F^2 + \|\mathbf{A}_{i,\perp}\|_F^2.$$

Proof. The optimality of (7.2) in Lemma 7.1 implies

$$\|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_F^2 = \|\mathbf{A} - \mathbf{U}_K \left(\mathbf{U}_K^T \mathbf{A} \right)_i \|_F^2$$
$$\le \|\mathbf{A} - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F^2$$
$$= \|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F^2 + \|\mathbf{A}_{i,\perp}\|_F^2.$$

The last equality follows from Lemma 4.1. □

7.2. Bounding the important part of the error. We bound the term in Lemma 7.2 over which we have control, namely $\|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F^2$.

As in section 5, let $\mathcal{P}_q$ be the orthogonal projector onto $\mathcal{K}_q$. For $\boldsymbol{\Phi}$ in (4.2) let $\boldsymbol{\Phi\Phi}^\dagger$ be the orthogonal projector onto range$(\boldsymbol{\Phi})$, with range$(\boldsymbol{\Phi\Phi}^\dagger) \subset$ range$(\mathcal{P}_q)$ due to (4.3). The leads to the obvious bound

$$(7.4) \qquad \|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F = \|\mathbf{A}_i - \mathcal{P}_q \mathbf{A}_i\|_F \leq \|\mathbf{A}_i - \boldsymbol{\Phi\Phi}^\dagger \mathbf{A}_i\|_F, \qquad 1 \leq i \leq k.$$

We don't stop here, though, but go further and pursue a bound in terms of polynomials.

LEMMA 7.3. *Let $\phi(x)$ be a polynomial of degree $2q + 1$ with odd powers only that satisfies $\phi(\sigma_j) \geq \sigma_j$ for $1 \leq j \leq k$. Then*

$$\|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F \leq \|\mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right) - \boldsymbol{\Phi\Phi}^\dagger \mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right)\|_F, \qquad 1 \leq i \leq k.$$

Proof. We use the abbreviation $\mathcal{P}_\phi^\perp \equiv \mathbf{I} - \boldsymbol{\Phi\Phi}^\dagger$ to denote the orthogonal projector onto range$(\boldsymbol{\Phi})^\perp$. From (7.3), (7.4), and the unitary invariance of the Frobenius norm follows

$$\|\mathbf{A}_i - \mathbf{U}_K \mathbf{U}_K^T \mathbf{A}_i\|_F \leq \|\mathcal{P}_\phi^\perp \mathbf{A}_i\|_F = \|\mathcal{P}_q^\perp \mathbf{U}_i \boldsymbol{\Sigma}_i\|_F, \qquad 1 \leq i \leq k.$$

Expressing the squared Frobenius norm as a sum of squared column norms and then applying the assumption $\sigma_j \leq \phi(\sigma_j)$ yields, for $1 \leq i \leq k$,

$$\|\mathcal{P}_\phi^\perp \mathbf{U}_i \boldsymbol{\Sigma}_i\|_F^2 = \sum_{j=1}^i \sigma_j^2 \|\mathcal{P}_\phi^\perp \mathbf{u}_j\|_2^2 \leq \sum_{j=1}^i \phi(\sigma_j)^2 \|\mathcal{P}_\phi^\perp \mathbf{u}_j\|_2^2 = \|\mathcal{P}_\phi^\perp \mathbf{U}_i \phi(\boldsymbol{\Sigma}_i)\|_F^2. \qquad \square$$

7.3. From projections to least squares residuals. Now we are ready to apply the approach from Theorem 2.1 and view the result of Lemma 7.3 as a least squares residual.

LEMMA 7.4. *Under the assumptions of Theorem 2.3,*

$$\|\mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right) - \boldsymbol{\Phi\Phi}^\dagger \mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right)\|_F \leq \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2 \ \|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_F.$$

Proof. Based on the orthogonality $(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^T)\mathbf{U}_i = \mathbf{0}$ for $1 \leq i \leq k$, we can deduce as in (5.4) that

$$\|(\mathbf{I} - \boldsymbol{\Phi\Phi}^\dagger)\mathbf{U}_i \phi(\boldsymbol{\Sigma}_i)\|_F \leq \|\boldsymbol{\Phi}_{k,\perp} \boldsymbol{\Phi}_k^\dagger \mathbf{U}_i \phi(\boldsymbol{\Sigma}_i)\|_F.$$

The expressions for $\boldsymbol{\Phi}_{k,\perp}$ in (5.2) and $\boldsymbol{\Phi}_k^\dagger$ in (5.3), along with the strong submultiplicativity in section 4.1, yield

$$\|\boldsymbol{\Phi}_{k,\perp} \boldsymbol{\Phi}_k^\dagger \mathbf{U}_i \phi(\boldsymbol{\Sigma}_i)\|_F = \|\mathbf{U}_{k,\perp} \, \phi(\boldsymbol{\Sigma}_{k,\perp}) \, \mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger \, \phi(\boldsymbol{\Sigma}_k)^{-1} \mathbf{U}_k^T \mathbf{U}_i \phi(\boldsymbol{\Sigma}_i)\|_F$$

$$\leq \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2 \ \|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_F.$$

The above inequality is obtained by noting that for $i = k$ we have $\phi(\boldsymbol{\Sigma}_k)^{-1} \mathbf{U}_k^T \mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right) = \mathbf{I}_k$, while for $1 \leq i < k$,

$$\phi(\boldsymbol{\Sigma}_k)^{-1} \mathbf{U}_k^T \mathbf{U}_i \phi\left(\boldsymbol{\Sigma}_i\right) = \begin{pmatrix} \mathbf{I}_i \\ \mathbf{0}_{(k-i) \times i} \end{pmatrix}. \qquad \square$$

7.4. Concluding the proof of Theorem 2.3. We prove each of the three inequalities in turn. Recall that

$$\Delta \equiv \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2 \, \|\mathbf{V}_{k,\perp}^T \mathbf{X} \, (\mathbf{V}_k^T \mathbf{X})^\dagger\|_F.$$

Proof of (2.3). Combining Lemmas 7.2, 7.3, and 7.4 and recognizing the expression for Δ yields

$$\|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T\|_F^2 \leq \|\mathbf{A}_{i,\perp}\|_F^2 + \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2^2 \, \|\mathbf{V}_{k,\perp}^T \mathbf{X}(\mathbf{V}_k^T \mathbf{X})^\dagger\|_F^2,$$

(7.5)
$$= \|\mathbf{A}_{i,\perp}\|_F^2 + \Delta^2, \qquad 1 \leq i \leq k.$$

Inserting $\|\mathbf{A}_{i,\perp}\|_F = \|\mathbf{A} - \mathbf{A}_i\|_F$ gives

(7.6)
$$\|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_i\|_F^2 + \Delta^2, \qquad 1 \leq i \leq k.$$

Taking advantage of the inequality below for scalars $\alpha, \beta \geq 0$,

(7.7)
$$\sqrt{\alpha^2 + \beta^2} \leq \sqrt{\alpha^2 + \beta^2 + 2\alpha\beta} = \sqrt{(\alpha + \beta)^2} = \alpha + \beta,$$

gives the weaker, but square-free, bound (2.3).

Proof of (2.4). We use Lemma 7.5 below, which shows that an additive error bound for a low-rank approximation in the Frobenius norm implies the same in the two norm, and apply it to (7.6). This gives $\|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_i\|_2^2 + \Delta^2$. Taking square roots based on (7.7) produces the desired bound (2.4).

LEMMA 7.5 (Theorem 3.4 in [20]). *Given* $\mathbf{A}, \tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ *with* $\mathrm{rank}(\tilde{\mathbf{A}}) = k < \mathrm{rank}(\mathbf{A})$. *If* $\|\mathbf{A} - \tilde{\mathbf{A}}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \delta$, *then*

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2^2 + \delta.$$

Proof of (2.5). The upper bounds follow from the *minimax theorem for singular values* [19, Theorem 8.6.1].

This leaves the lower bounds. Recall the non-increasing ordering of the singular values $\sigma_1 \geq \cdots \geq \sigma_k$ and the fact that $\hat{\mathbf{U}}_i$ in (2.2) has orthonormal columns.

Case $i = 1$. Apply Lemma 4.1 to (7.5)

$$\|\mathbf{A}\|_F^2 - \|\hat{\mathbf{u}}_1^T \mathbf{A}\|_F^2 = \|\mathbf{A} - \hat{\mathbf{u}}_1 \hat{\mathbf{u}}_1^T \mathbf{A}\|_F^2 \leq \|\mathbf{A}_{1,\perp}\|_F^2 + \Delta^2.$$

From $\|\mathbf{A}\|_F^2 - \|\mathbf{A}_{1,\perp}\|_F^2 = \sigma_1^2$ follows $\sigma_1^2 \leq \|\mathbf{A} - \hat{\mathbf{u}}_1 \hat{\mathbf{u}}_1^T \mathbf{A}\|_F^2 + \Delta^2$. Taking square roots based on (7.7) proves (2.5) for $i = 1$.

Case $2 \leq i \leq k$. Among all matrices of rank $i - 1$, the matrix $\mathbf{A}_{i-1}$ is closest to $\mathbf{A}$ in the Frobenius norm. Hence

$$\|\mathbf{A}_{i-1,\perp}\|_F = \|\mathbf{A} - \mathbf{A}_{i-1}\|_F \leq \|\mathbf{A} - \hat{\mathbf{U}}_{i-1} \hat{\mathbf{U}}_{i-1}^T \mathbf{A}\|_F.$$

The above, together with the outer product representation $\hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T = \hat{\mathbf{U}}_{i-1} \hat{\mathbf{U}}_{i-1}^T + \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T$, Lemma 4.1, and (7.5), gives

$$\|\mathbf{A}_{i-1,\perp}\|_F^2 - \|\hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \hat{\mathbf{U}}_{i-1} \hat{\mathbf{U}}_{i-1}^T \mathbf{A}\|_F^2 - \|\hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \mathbf{A}\|_F^2$$
$$= \|\mathbf{A} - \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^T \mathbf{A}\|_F^2 \leq \|\mathbf{A}_{i,\perp}\|_F^2 + \Delta^2.$$

At last, applying $\|\mathbf{A}_{i-1,\perp}\|_F^2 - \|\mathbf{A}_{i,\perp}\|_F^2 = \sigma_i^2$ and taking square roots based on (7.7) proves (2.5) for $2 \leq i \leq k$.

This concludes the proof for general $\mathbf{X}$. The proof for the special case where $\mathbf{X}$ has orthonormal columns follows from Lemma 4.3.

8. Conclusions and open problems. Motivated by the emergence of randomized Krylov space methods for low-rank approximations [34, 43], we presented a "proof of concept," that is, structural results for the accuracy of approximate dominant subspaces.

Several open problems arise from our work:

1. Can we better understand and close the disconnect between low-rank approximations and dominant subspace computations?

   A singular value gap is a must for dominant subspace computations, if only to ensure well-posedness of the mathematical problem. In contrast, low-rank approximations can do without a gap for special starting guesses $\mathbf{X}$ [34]. This comes at the expense of accuracy, though. Bolstered by a gap, subspace accuracy exhibits the logarithmic dependence (2.7) on ϵ, while, without a gap, the accuracy of a low-rank approximation has only polynomial dependence on ϵ. To the best of our knowledge, gap-independent results are not known for arbitrary $\mathbf{X}$. As the analysis [34] only exploits the fact that $\mathbf{X}$ can give an approximation that is polynomially close to optimal in the Frobenius norm, it could potentially be extended to a variety of random starting guesses.

2. Is it possible to relax the full-rank assumption for $\mathbf{V}_k^T\mathbf{X}$?

   Our proofs require $\text{rank}(\mathbf{V}_k^T\mathbf{X}) = k$, which forces starting guesses to have at least $s \geq k$ columns. Thus, even in the presence of the requisite singular value gaps, our proofs collapse for starting guesses that consist of a single column.

3. Are our bounds tight enough to be informative, and how relevant are they for practical numerical implementations of block Krylov methods?

Appendix A. More general proof of Theorem 2.1. The proof below applies to starting guesses $\mathbf{X}$ with linearly independent columns and consists of several steps.

Preparing $\mathbf{X}$. Since subspace angles are defined by matrices with orthonormal columns, we perform a thin QR decomposition $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{n \times s}$ has orthonormal columns, and $\mathbf{R} \in \mathbb{R}^{s \times s}$ is nonsingular. Then $\text{range}(\mathbf{Q}) = \text{range}(\mathbf{X})$.

The expression for $\boldsymbol{\Phi}$ contains a basis transformation on $\mathbf{X}$ with the orthogonal matrix $\mathbf{V}$, resulting in an $n \times s$ matrix

$$\mathbf{V}^T\mathbf{Q} = \begin{pmatrix} \mathbf{V}_k^T\mathbf{Q} \\ \mathbf{V}_{k,\perp}^T\mathbf{Q} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_k \\ \mathbf{Q}_{k,\perp} \end{pmatrix}$$

with orthonormal columns. It remains to account for $\mathbf{R}$:

$$\mathbf{V}^T\mathbf{X} = (\mathbf{V}^T\mathbf{Q})\,\mathbf{R} = \begin{pmatrix} \mathbf{Q}_k^T\mathbf{R} \\ \mathbf{Q}_{k,\perp}^T\mathbf{R} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_k \\ \mathbf{X}_{k,\perp} \end{pmatrix}.$$

By assumption, $k = \text{rank}(\mathbf{V}_k^T\mathbf{Q}) = \text{rank}(\mathbf{X}_k) = \text{rank}(\mathbf{Q}_k)$, so that $\mathbf{X}_k \in \mathbb{R}^{k \times s}$ and $\mathbf{Q}_k \in \mathbb{R}^{k \times s}$ have full row rank. In particular,

(A.1) $$\mathbf{Q}_k\mathbf{Q}_k^\dagger = \mathbf{I}_k, \qquad \mathbf{Q}_k^\dagger = \mathbf{Q}_k^T(\mathbf{Q}_k\mathbf{Q}_k^T)^{-1}.$$

For $\mathbf{X}_k$, though, we forego the Moore–Penrose inverse and choose instead a $(1,2,3)$ inverse [12, Definition 6.2.4]. The matrix

$$\mathbf{X}_k^+ \equiv \mathbf{R}^{-1}\mathbf{Q}_k^\dagger$$

is a right inverse, $\mathbf{X}_k\mathbf{X}_k^+ = \mathbf{I}_k$, and satisfies three of the four Moore–Penrose conditions,

$$\mathbf{X}_k\,\mathbf{X}_k^+\,\mathbf{X}_k = \mathbf{X}_k, \qquad \mathbf{X}_k^+\,\mathbf{X}_k\,\mathbf{X}_k^+ = \mathbf{X}_k^+, \qquad (\mathbf{X}_k\,\mathbf{X}_k^+)^T = (\mathbf{X}_k\,\mathbf{X}_k^+).$$

Viewing the sine as a least squares residual. Let $\mathcal{P}_q$ be the orthogonal projector onto the Krylov space $\mathcal{K}_q$, and let $\boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger$ be the orthogonal projector onto range($\boldsymbol{\Phi}$). From range($\boldsymbol{\Phi}$) $\subset \mathcal{K}_q$ follows range($\boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger$) $\subset$ range($\mathcal{P}_q$); hence (4.4) implies

$$(A.2) \qquad \|\sin\boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} = \|(\mathbf{I} - \mathcal{P}_q)\,\mathbf{U}_k\|_{2,F}\|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F}$$
$$= \|\sin\boldsymbol{\Theta}(\boldsymbol{\Phi}, \mathbf{U}_k)\|_{2,F}.$$

Lemma 4.2 shows that $\|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F}$ is the residual of the least squares problem

$$\|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F} = \min_{\boldsymbol{\Psi}} \|\mathbf{U}_k - \boldsymbol{\Phi}\boldsymbol{\Psi}\|_{2,F} = \|\mathbf{U}_k - \boldsymbol{\Phi}\boldsymbol{\Psi}_{opt}\|_{2,F},$$

where $\boldsymbol{\Psi}_{opt} = \boldsymbol{\Phi}^\dagger\mathbf{U}_k$ is a least squares solution.

Focusing on the target space. Decompose $\boldsymbol{\Phi}$ into the target component range($\mathbf{U}_k$) and the complementary subspace, $\boldsymbol{\Phi} = \boldsymbol{\Phi}_k + \boldsymbol{\Phi}_{k,\perp}$, where

$$(A.3) \qquad \boldsymbol{\Phi}_k \equiv \mathbf{U}_k\phi(\boldsymbol{\Sigma}_k)\mathbf{X}_k, \qquad \boldsymbol{\Phi}_{k,\perp} \equiv \mathbf{U}_{k,\perp}\phi(\boldsymbol{\Sigma}_{k,\perp})\mathbf{X}_{k,\perp}.$$

It is easy to verify that

$$(A.4) \qquad \boldsymbol{\Phi}_k^+ \equiv \mathbf{X}_k^+\phi(\boldsymbol{\Sigma}_k)^{-1}\mathbf{U}_k^T$$

satisfies the conditions of a $(1,2,3)$ inverse. The minimality of the least squares residual implies

$$\|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger)\,\mathbf{U}_k\|_{2,F} = \|\mathbf{U}_k - \boldsymbol{\Phi}\,(\boldsymbol{\Phi}^\dagger\mathbf{U}_k)\|_{2,F}$$
$$\leq \|\mathbf{U}_k - \boldsymbol{\Phi}\,(\boldsymbol{\Phi}_k^+\mathbf{U}_k)\|_{2,F} = \|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}_k^+)\,\mathbf{U}_k\|_{2,F}.$$

Now replace the other instance of $\boldsymbol{\Phi}$ by (A.3),

$$\|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^+)\,\mathbf{U}_k\|_{2,F} \leq \|(\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}_k^+)\,\mathbf{U}_k\|_{2,F}$$
$$(A.5) \qquad = \|(\mathbf{I} - \boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^+)\mathbf{U}_k - \boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F}.$$

Since a $(1,2,3)$ inverse is an orthogonal projector [8, section 2.2.1, (2.2.13)], it follows that $\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^+ = \mathbf{U}_k\mathbf{U}_k^T$ is the orthogonal projector onto the target space range($\mathbf{U}_k$). This observation simplifies (A.5),

$$\|(\mathbf{I} - \boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^+)\mathbf{U}_k - \boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F} = \|(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T)\mathbf{U}_k - \boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F}$$
$$= \|\boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F}.$$

Summary so far. Combining the above with (A.2) and (A.5) gives

$$\|\sin\boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} \leq \|\sin\boldsymbol{\Theta}(\boldsymbol{\Phi}, \mathbf{U}_k)\|_{2,F} \leq \|\boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F}.$$

Extracting the polynomials. The expressions for $\boldsymbol{\Phi}_{k,\perp}$ in (A.3) and $\boldsymbol{\Phi}_k^+$ in (A.4), and submultiplicativity (section 4.1) yield

$$\|\boldsymbol{\Phi}_{k,\perp}\boldsymbol{\Phi}_k^+\mathbf{U}_k\|_{2,F} = \|\phi(\boldsymbol{\Sigma}_{k,\perp})\,\mathbf{X}_{k,\perp}\mathbf{X}_k^+\,\phi(\boldsymbol{\Sigma}_k)^{-1}\|_{2,F}$$
$$\leq \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2\,\|\phi(\boldsymbol{\Sigma}_k)^{-1}\|_2\,\|\mathbf{X}_{k,\perp}\mathbf{X}_k^+\|_{2,F}.$$

Combining the previous two sets of inequalities gives

$$\|\sin\boldsymbol{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} \leq \|\phi(\boldsymbol{\Sigma}_{k,\perp})\|_2\,\|\phi(\boldsymbol{\Sigma}_k)^{-1}\|_2\,\|\mathbf{X}_{k,\perp}\mathbf{X}_k^+\|_{2,F}.$$

We chose the $(1, 2, 3)$ inverse so that $\mathbf{R}$ cancels out, $\|\mathbf{X}_{k,\perp}\mathbf{X}_k^+\|_{2,F} = \|\mathbf{Q}_{k,\perp}\mathbf{Q}_k^\dagger\|_{2,F}$, and

$$\|\sin\mathbf{\Theta}(\mathcal{K}_q, \mathbf{U}_k)\|_{2,F} \leq \|\phi(\mathbf{\Sigma}_{k,\perp})\|_2 \, \|\phi(\mathbf{\Sigma}_k)^{-1}\|_2 \, \|\mathbf{Q}_{k,\perp}\mathbf{Q}_k^\dagger\|_{2,F}.$$

At last, Lemma 4.3 and $\mathrm{range}(\mathbf{Q}) = \mathrm{range}(\mathbf{X})$ imply

$$\|\mathbf{Q}_{k,\perp}\mathbf{Q}_k^\dagger\|_{2,F} = \|\tan\mathbf{\Theta}(\mathbf{Q}, \mathbf{V}_k)\|_{2,F} = \|\tan\mathbf{\Theta}(\mathbf{X}, \mathbf{V}_k)\|_{2,F}.$$

Acknowledgment. We thank Mark Embree for many useful discussions.

REFERENCES

[1] D. ACHLIOPTAS, *Database-friendly random projections*, in Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '01), ACM Press, New York, 2001, pp. 274–281.

[2] N. AILON AND B. CHAZELLE, *The fast Johnson–Lindenstrauss transform and approximate nearest neighbors*, SIAM J. Comput., 39 (2009), pp. 302–322, https://doi.org/10.1137/060673096.

[3] F. ARRIGO, M. BENZI, AND C. FENU, *Computation of generalized matrix functions*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 836–860, https://doi.org/10.1137/15M1049634.

[4] J. BAGLAMA AND L. REICHEL, *Augmented implicitly restarted Lanczos bidiagonalization methods*, SIAM J. Sci. Comput., 27 (2005), pp. 19–42, https://doi.org/10.1137/04060593X.

[5] J. BAGLAMA AND L. REICHEL, *Restarted block Lanczos bidiagonalization methods*, Numer. Algorithms, 43 (2006), pp. 251–272.

[6] C. BEATTIE, M. EMBREE, AND J. ROSSI, *Convergence of restarted Krylov subspaces to invariant subspaces*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1074–1109, https://doi.org/10.1137/S0895479801398608.

[7] C. BEATTIE, M. EMBREE, AND D. C. SORENSEN, *Convergence of polynomial restart Krylov methods for eigenvalue computations*, SIAM Rev., 47 (2005), pp. 492–515, https://doi.org/10.1137/S0036144503433077.

[8] Å. BJÖRCK, *Numerical Methods in Matrix Computations*, Springer, Heidelberg, 2015.

[9] N. BOSNER AND Z. DRMAČ, *Subspace gap residuals for Rayleigh–Ritz approximations*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 54–67, https://doi.org/10.1137/070689425.

[10] C. BOUTSIDIS, P. DRINEAS, AND M. MAGDON-ISMAIL, *Near-optimal column-based matrix reconstruction*, in Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011, IEEE Computer Society, Los Alamitos, CA, 2011, pp. 305–314.

[11] C. BOUTSIDIS, P. DRINEAS, AND M. MAGDON-ISMAIL, *Near-optimal column-based matrix reconstruction*, SIAM J. Comput., 43 (2014), pp. 687–717, https://doi.org/10.1137/12086755X.

[12] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Dover, New York, 1991.

[13] K. L. CLARKSON AND D. P. WOODRUFF, *Low rank approximation and regression in input sparsity time*, in Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC), ACM Press, New York, 2013, pp. 81–90.

[14] J. D. DIXON, *Estimating extreme eigenvalues and condition numbers of matrices*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814, https://doi.org/10.1137/0720053.

[15] P. DRINEAS AND M. W. MAHONEY, *RandNLA: Randomized Numerical Linear Algebra*, Commun. ACM, 59 (2016), pp. 80–90.

[16] P. DRINEAS, M. W. MAHONEY, AND S. MUTHUKRISHNAN, *Relative error CUR matrix decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 844–881, https://doi.org/10.1137/07070471X.

[17] P. DRINEAS, M. W. MAHONEY, S. MUTHUKRISHNAN, AND T. SARLÓS, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 219–249.

[18] Z. DRMAČ, *On relative residual bounds for the eigenvalues of a Hermitian matrix*, Linear Algebra Appl., 244 (1996), pp. 155–163.

[19] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, 2013.

[20] M. GU, *Subspace iteration randomization and singular value problems*, SIAM J. Sci. Comput., 37 (2015), pp. A1139–A1173, https://doi.org/10.1137/130938700.

[21] N. HALKO, P.-G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.

[22] J. B. HAWKINS AND A. BEN-ISRAEL, *On generalized matrix functions*, Linear Multilinear Algebra, 1 (1973), pp. 163–171.

[23] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[24] E. R. JESSUP AND I. C. F. IPSEN, *Improving the accuracy of inverse iteration*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–572, https://doi.org/10.1137/0913031.

[25] Z. JIA, *The Regularization Theory of the Krylov Iterative Solvers LSQR and CGLS for Linear Discrete Ill-Posed Problems, Part* I: *The Simple Singular Value Case*, preprint, https://arxiv.org/abs/1701.05708v1, 2017.

[26] A. V. KNYAZEV, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396. Translated from the Russian.

[27] J. KUCZYŃSKI AND H. WOŹNIAKOWSKI, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122, https://doi.org/10.1137/0613066.

[28] J. KUCZYŃSKI AND H. WOŹNIAKOWSKI, *Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 672–691, https://doi.org/10.1137/S0895479892230456.

[29] R.-C. LI AND L.-H. ZHANG, *Convergence of the block Lanczos method for eigenvalue clusters*, Numer. Math., 131 (2015), pp. 83–113.

[30] A. MAGEN AND A. ZOUZIAS, *Low rank matrix-valued Chernoff bounds and approximate matrix multiplication*, in Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, Philadelphia, ACM, New York, 2011, pp. 1422–1436, https://doi.org/10.1137/1.9781611973082.109.

[31] P. J. MAHER, *Some norm inequalities concerning generalized inverses*, Linear Algebra Appl., 174 (1992), pp. 99–110.

[32] M. MAHONEY AND P. DRINEAS, *Structural properties underlying high-quality randomized numerical linear algebra algorithms*, in Handbook of Big Data, CRC Press, Boca Raton, FL, 2016, pp. 137–154.

[33] X. MENG AND M. W. MAHONEY, *Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression*, in Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing (STOC '13), 2013, pp. 91–100.

[34] C. MUSCO AND C. MUSCO, *Randomized block Krylov methods for stronger and faster approximate singular value decomposition*, in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, 2015, pp. 1396–1404.

[35] J. NELSON AND H. L. NGUYEN, *OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings*, in Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013, IEEE Computer Society, Los Alamitos, CA, 2013.

[36] N. H. NGUYEN, T. T. DO, AND T. D. TRAN, *A fast and efficient algorithm for low-rank approximation of a matrix*, in Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing (STOC '09), ACM Press, New York, 2009, pp. 215–224.

[37] Y. SAAD, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706, https://doi.org/10.1137/0717059.

[38] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems: Revised Edition*, Classics Appl. Math. 66, SIAM, Philadelphia, 2011, https://doi.org/10.1137/1.9781611970739.

[39] T. SARLÓS, *Improved approximation algorithms for large matrices via random projections*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS '06), 2006, pp. 143–152.

[40] H. D. SIMON AND H. ZHA, *Low-rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput., 21 (2000), pp. 2257–2274, https://doi.org/10.1137/S1064827597327309.

[41] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.

[42] J. A. TROPP, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal., 3 (2011), pp. 115–126.

[43] S. WANG, Z. ZHANG, AND T. ZHANG, *Improved Analyses of the Randomized Power Method and Block Lanczos Method*, preprint, https://arxiv.org/abs/1508.06429v2, 2015.

[44] D. P. WOODRUFF, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. 1–157.

[45] P. ZHU AND A. V. KNYAZEV, *Angles between subspaces and their tangents*, J. Numer. Math., 21 (2013), pp. 325–340.