SIAM J. MATRIX ANAL. APPL.
Vol. 39, No. 3, pp. 1208–1220

A PROBABILISTIC SUBSPACE BOUND WITH APPLICATION TO ACTIVE SUBSPACES*

JOHN T. HOLODNAK†, ILSE C. F. IPSEN‡, AND RALPH C. SMITH§

Abstract. Given a real symmetric positive semidefinite matrix E, and an approximation S that is a sum of n independent matrix-valued random variables, we present bounds on the relative error in S due to randomization. The bounds do not depend on the matrix dimensions but only on the numerical rank (intrinsic dimension) of E. Our approach resembles the low-rank approximation of kernel matrices from random features, but our accuracy measures are more stringent. In the context of parameter selection based on active subspaces, where S is computed via Monte Carlo sampling, we present a bound on the number of samples so that with high probability the angle between the dominant subspaces of E and S is less than a user-specified tolerance. This is a substantial improvement over existing work, as it is a nonasymptotic and fully explicit bound on the sampling amount n, and it allows the user to tune the success probability. It also suggests that Monte Carlo sampling can be efficient in the presence of many parameters, as long as the underlying function f is sufficiently smooth.

Key words. positive semidefinite matrices, principal angles, eigenvalue decomposition, eigenvalue gaps, matrix concentration inequality, intrinsic dimension, Monte Carlo sampling, active subspaces

AMS subject classifications. 15A18, 15A23, 15A60, 15B10, 35J25, 60G60, 65N30, 65C06, 65C30, 65F15, 65D05

DOI. 10.1137/17M1141503

1. Introduction.

We analyze the accuracy of approximating a symmetric positive semidefinite matrix $E \in \mathbb{R}^{m\times m}$ by a sum $\widehat{E} \equiv \frac{1}{n}\sum_{j=1}^{n} z_j z_j^T$ of n independently sampled outer products $z_j z_j^T$, each of which is an unbiased estimator of the mean $\mathbb{E}[z_j z_j^T] = E$, $1 \le j \le n$, thus producing an overall unbiased estimator $\widehat{E}$. We derive probabilistic bounds on the relative error due to randomization in $\widehat{E}$, and on the angle between equidimensional dominant subspaces of E and $\widehat{E}$. The bounds do not depend on the matrix dimension, but only on the numerical rank of E.

To avoid explicit dependence on the matrix dimensions, we use an intrinsic-dimension matrix Bernstein concentration inequality. This type of analysis can be found in low-rank approximations of kernel matrices via random features [2, 17]. However, our accuracy measures are more stringent, and the angle bounds necessitate assumptions that are strong enough to guarantee that the dominant subspaces are

*Received by the editors August 1, 2017; accepted for publication (in revised form) by D. B. Szyld May 23, 2018; published electronically August 2, 2018.
http://www.siam.org/journals/simax/39-3/M114150.html
Funding: The second author was supported in part by NSF grant CCF-1145383. The second author also acknowledges support from the XDATA Program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. The third author was supported in part by Air Force Office of Scientific Research (AFOSR) grant FA9550-15-1-0299 and in part by the Consortium for Advanced Simulation of Light Water Reactors (http://www.casl.gov), an Energy Innovation Hub (http://www.energy.gov/hubs) for Modeling and Simulation of Nuclear Reactors under U.S. Department of Energy contract DE-AC05-00OR22725.

†Massachusetts Institute of Technology, Cambridge, MA 02139 (jtholodn@ncsu.edu).
‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (ipsen@ncsu.edu, http://www4.ncsu.edu/∼ipsen/).
§Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (rsmith@ncsu.edu, http://www4.ncsu.edu/∼rsmith/).

well-defined. In contrast to existing probabilistic bounds for Krylov or subspace iterations (see [12, 14] and the references therein), our bounds are perturbation bounds and make no reference to actual methods for computing subspaces.

The motivation for this paper came from applications involving parameter selection, such as occurs in the solution of random ODEs and PDEs [7, 10, 27] and reduced-order nonlinear models [3], as well as various applications arising in engineering, math biology, and the sciences [4–6, 11, 20].

Given a function $f : \mathbb{R}^m \to \mathbb{R}$, which depends on m parameters and may be expensive to evaluate, one wants to select subspaces associated with a few influential parameters. This is a form of dimension reduction [23, Chapter 6], and one particular approach is to identify a low-dimensional *active subspace* in $\mathbb{R}^m$ along which f is, on average, most sensitive to change [7, 22]. This is done by replacing a "sensitivity" matrix $E \in \mathbb{R}^{m\times m}$ by a sum of n independent Monte Carlo samples $\widehat{E}$, followed by computing a dominant subspace of $\widehat{E}$.

Specifically, [23] illustrates that ODE models for HIV can easily have $m = 20$ parameters, whereas Boltzmann PDE models, quantifying neutron transport in a light water reactor, can have as many as $m = 10^6$. In both cases, it is critical to isolate active subspaces of parameters—which are identifiable in the sense that they are uniquely determined by observed responses—prior to frequentist or Bayesian inference. The matrices E arise when computing local sensitivities $\nabla f(x)$ for determining these subspaces.

In the following, we present a probabilistic bound (Theorem 2.5) that is tighter than existing work [9], represents a nonasymptotic and fully explicit bound on the sampling amount n, and allows the user to tune the success probability. The absence of the matrix dimension m in the bound suggests that Monte Carlo sampling can be efficient in the presence of many parameters, as long as the function f is sufficiently smooth.

Outline. Bounds are presented in section 2, with the proofs relegated to section 3.

2. Our contributions. After stating the assumptions (section 2.1), we present an upper bound for the relative error due to randomization in $\widehat{E}$ (section 2.2), a lower bound on the sampling amount n (section 2.3), a bound on the angle between dominant subspaces of E and $\widehat{E}$ (section 2.4), and an application to active subspaces (section 2.5).

2.1. Assumptions. Let the nonzero matrix $E \in \mathbb{R}^{m\times m}$ be symmetric positive semidefinite, and let $\widehat{E} \equiv \frac{1}{n}\sum_{j=1}^{n} z_j z_j^T$ be an approximation, where $z_j \in \mathbb{R}^m$ are n independent random vectors with $\mathbb{E}[z_j z_j^T] = E$, $1 \le j \le n$. All quantities are uniformly bounded in the two norm; that is, there exists $L > 0$ with

$$\max_{1\le j\le n} \|z_j\|_2 \le L \qquad \text{and} \qquad \|E\|_2 \le L^2,$$

as well as

$$\|\widehat{E}\|_2 \le \tfrac{1}{n}\sum_{j=1}^{n} \|z_j z_j^T\|_2 \le L^2.$$

Our bounds do not depend on the matrix dimension m, but only on the numerical rank of E, which is quantified by

$$\operatorname{intdim}(E) \equiv \operatorname{trace}(E)/\|E\|_2 \ge 1$$

and equals the stable rank of $E^{1/2}$; see section 3.1.1.

2.2. Error due to randomization. The first bound is an expression for the relative error of $\widehat{E}$ in the two norm.

THEOREM 2.1. *Given the assumptions in section* 2.1*, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\frac{\|\widehat{E} - E\|_2}{\|E\|_2} \leq \gamma + \sqrt{\gamma(\gamma+6)}, \quad \textit{where} \quad \gamma \equiv \frac{1}{3n}\,\frac{L^2}{\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\mathrm{intdim}\,(E)\right).$$

Proof. See section 3.1 ∎

Theorem 2.1 implies that the relative error in $\widehat{E}$ is small if:

1. Many samples are used to compute $\widehat{E}$, that is, $n \gg 1$.
2. E has low numerical rank, that is, $\mathrm{intdim}(E) \ll m$.
3. The upper bound for E is tight, that is, $L^2/\|E\|_2 \approx 1$.
   In the context of active subspaces in Theorem 2.5, this is interpreted as the smoothness of an underlying function.

Section 3.1.3 contains a brief discussion on alternative, expectation-based bounds.

2.3. Required amount of sampling. We express Theorem 2.1 as a lower bound on the number of samples n required for a user-specified relative error.

COROLLARY 2.2. *Given the assumptions in section* 2.1*, let* $0 < \delta < 1$ *and* $0 < \epsilon < 1$*. If*

$$n \geq \frac{8}{3\epsilon^2}\,\frac{L^2}{\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\mathrm{intdim}\,(E)\right),$$

then with probability at least $1 - \delta$*,*

$$\|\widehat{E} - E\|_2/\|E\|_2 \leq \epsilon.$$

Proof. See section 3.2. ∎

Corollary 2.2 implies that only a few samples are required to compute an approximation $\widehat{E}$ that is highly likely to have specified accuracy ϵ if:

1. The requested accuracy for $\widehat{E}$ is low, that is, $\epsilon \approx 1$.
2. E has low numerical rank, that is, $\mathrm{intdim}(E) \ll m$.
3. The upper bound for E is tight, that is, $L^2/\|E\|_2 \approx 1$.

Whereas there are certain similarities to covariance estimation in the spirit of [29], our analysis in section 3 differs in two aspects. First, our matrices are not strictly covariance matrices since they are not centered. More importantly, we do not require that our matrices have linearly independent rows constructed as random vectors from an underlying distribution. Rather, our bounds show that good approximations require low-rank matrices, i.e., many linearly dependent rows.

2.4. Subspace bound. We bound the largest principal angle between equidimensional dominant subspaces of E and $\widehat{E}$. To identify the subspaces, consider the eigenvalue decompositions

$$E = V\Lambda V^T, \qquad \Lambda = \mathrm{diag}\begin{pmatrix}\lambda_1 & \cdots & \lambda_m\end{pmatrix}, \qquad V = \begin{pmatrix}v_1 & \ldots & v_m\end{pmatrix}, \tag{2.1}$$

where $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ and $V \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, and

$$\widehat{E} = \widehat{V}\widehat{\Lambda}\widehat{V}^T, \qquad \widehat{\Lambda} = \begin{pmatrix}\widehat{\lambda}_1 & \cdots & \widehat{\lambda}_m\end{pmatrix}, \qquad \widehat{V} = \begin{pmatrix}\widehat{v}_1 & \cdots & \widehat{v}_m\end{pmatrix}, \tag{2.2}$$

where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_m \geq 0$, and $\widehat{V} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix.

The bound below on the largest principal angle $\angle(\widehat{\mathcal{S}}, \mathcal{S})$ between dominant subspaces $\mathcal{S}$ and $\widehat{\mathcal{S}}$ requires the perturbation to be sufficiently small compared to the eigenvalue gap.

THEOREM 2.3. *In addition to the assumptions in section* 2.1*, let E have an eigenvalue gap $\lambda_k - \lambda_{k+1} > 0$ for some $1 \leq k < m$, so that $\mathcal{S} \equiv \operatorname{range}\begin{pmatrix} v_1 & \cdots & v_k \end{pmatrix}$ is well-defined. Also, let $0 < \epsilon < \frac{\lambda_k - \lambda_{k+1}}{4\,\|E\|_2}$ and $0 < \delta < 1$.*

If the sampling amount is sufficiently large,

$$n \geq \frac{8}{3\,\epsilon^2}\,\frac{L^2}{\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\operatorname{intdim}(E)\right),$$

then with probability at least $1 - \delta$, the dominant subspace $\widehat{\mathcal{S}} = \operatorname{range}\begin{pmatrix} \widehat{v}_1 & \cdots & \widehat{v}_k \end{pmatrix}$ is well-defined, and the largest principal angle $\angle(\widehat{\mathcal{S}}, \mathcal{S})$ is bounded by

$$\sin\angle(\widehat{\mathcal{S}}, \mathcal{S}) \leq \frac{4\,\|E\|_2}{\lambda_k - \lambda_{k+1}}\,\epsilon.$$

Proof. See section 3.3. ∎

Theorem 2.3 implies that $\widehat{\mathcal{S}}$ is likely to be ϵ-close to $\mathcal{S}$ if:

1. E has a large eigenvalue gap $(\lambda_k - \lambda_{k+1})/\lambda_1$. This is a relative gap, but it is weak because the denominator contains λ_1 rather than λ_k. The inverse of the gap is a measure of sensitivity for the subspace $\mathcal{S}$.
2. The matrix E has low numerical rank, that is, $\operatorname{intdim}(E) \ll m$.
3. The upper bound for E is tight, that is, $L^2/\|E\|_2 \approx 1$.

2.5. Application to active subspaces. After setting the context (section 2.5.1) we improve an existing bound on the number of Monte Carlo samples required to approximate an active subspace to a user-specified error subject to a user-specified success probability (section 2.5.2).

2.5.1. Problem setting. Assume the nonconstant function $f(x) : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, with gradient vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_m}(x) \end{pmatrix}^T \in \mathbb{R}^m$$

and Lipschitz constant $L > 0$ so that $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^m$. Assume also that f is square integrable on $\mathbb{R}^m$ with respect to a positive and bounded probability density function $\rho(x)$, and that all products of partial derivatives of f are also integrable with respect to $\rho(x)$.

Let $\mathbf{X} \in \mathbb{R}^m$ be a random vector with the associated probability density function $\rho(x)$, and denote by

$$\mathbb{E}\left[h(\mathbf{X})\right] \equiv \int_{\mathbb{R}^m} h(x)\,\rho(x)\,dx$$

the expected value of a function $h(\mathbf{X}) : \mathbb{R}^m \to \mathbb{R}$ with regard to $\mathbf{X}$. Then the sensitivity of f along a unit-norm direction v can be estimated from the expected value of the squared directional derivative of f along v,

$$\mathbb{E}\left[\left(v^T\,\nabla f(\mathbf{X})\right)^2\right] = \int_{\mathbb{R}^m} \left(v^T\,\nabla f(x)\right)^2 \rho(x)\,dx.$$

Directional derivatives [1] can measure sensitivity in any direction, while mean squared derivatives [24] are limited to coordinate directions. Informative directional derivatives can be obtained from the $m \times m$ matrix [7, Lemma 2.1]

$$E \equiv \mathbb{E}\left[\nabla f(\mathbf{X})\,(\nabla f(\mathbf{X}))^T\right] = \int_{\mathbb{R}^m} \nabla f(x)\,(\nabla f(x))^T\,\rho(x)\,dx \tag{2.3}$$

and its eigenvalue decomposition (2.1). For an eigenpair (λ_j, v_j), with $Ev_j = \lambda_j v_j$, we have $\lambda_j = \mathbb{E}\left[(v_j^T\,\nabla f(\mathbf{X}))^2\right]$. This means that eigenvector v_j indicates a direction of sensitivity for f, while λ_j represents the average amount of sensitivity of f along v_j. In particular, eigenvectors associated with the dominant eigenvalues represent directions along which f, on average, is the most sensitive.

This leads to the concept of active subspace [7–9]. We assume that for some $1 \leq k < m$, the matrix E has an eigenvalue gap $\lambda_k > \lambda_{k+1}$. Then the dominant subspace $\mathcal{S} = \text{range}\begin{pmatrix} v_1 & \cdots & v_k \end{pmatrix}$ is well-defined and called the *active subspace of dimension k of f*.

Since explicit computation of E in (2.3) is often not feasible because the elements are high-dimensional integrals, one can use a Monte Carlo method [7, (2.16)] to independently sample n vectors $x_j \in \mathbb{R}^m$ according to $\rho(x)$, and approximate E by

$$\widehat{E} = \frac{1}{n}\sum_{j=1}^{n} \nabla f(x_j)\,(\nabla f(x_j))^T.$$

If $\widehat{E}$ happens to have an eigenvalue gap at the same location as E, so that $\widehat{\lambda}_k > \widehat{\lambda}_{k+1}$, then the perturbed dominant subspace $\widehat{\mathcal{S}} = \text{range}\begin{pmatrix} \widehat{v}_1 & \cdots & \widehat{v}_k \end{pmatrix}$ is also well-defined and called *approximate active subspace* of dimension k for f.

2.5.2. Accuracy of approximate active subspace. The bounds below are conceptual as they depend on unknown quantities like the eigenvalues of E and a global bound on the gradient norm. Nevertheless, a sufficiently tight bound is informative because it suggests that Monte Carlo sampling can be efficient in the presence of many parameters, as long as the function f is sufficiently smooth with most eigenvalues of E being small.

Below we tried to collect all the required assumptions for [9, Corollary 3.7].

THEOREM 2.4 (Theorem 3.5, Corollaries 3.6 and 3.7 in [9]). *In addition to the assumptions in section* 2.5.1*, also assume that* $0 < \epsilon \leq \min\left\{1, \frac{\lambda_k - \lambda_{k+1}}{5\,\lambda_1}\right\}$ *and*

$$\nu \equiv \left\| \int \left(\nabla f(x)(\nabla f(x))^T - E\right)^2 \rho(x)\,dx \right\| > 0.$$

If the number of samples is

$$n = \Omega\left(\max\left\{\frac{L^2}{\lambda_1\,\epsilon},\, \frac{\nu}{\lambda_1^2\,\epsilon^2}\right\} \ln(2\,m)\right),$$

then with high probability

$$\sin\angle(\widehat{\mathcal{S}}, \mathcal{S}) \leq \frac{4\,\lambda_1\,\epsilon}{\lambda_k - \lambda_{k+1}}.$$

We improve on Theorem 2.4 by presenting a bound that is tighter and more informative. More specifically, Theorem 2.5 below (i) specifies a nonasymptotic, fully

explicit, computable bound for the sampling amount n; (ii) specifies an explicit expression for the failure probability δ and its impact on the sampling amount n, thus allowing tuning by the user; (iii) depends on the numerical rank of E, which can be much smaller than the total number m of parameters; and (iv) guarantees that the approximate active subspace $\widehat{\mathcal{S}}$ is well-defined.

THEOREM 2.5. *With the assumptions in section* 2.5.1*, let* $0 < \epsilon < \frac{\lambda_k - \lambda_{k+1}}{4\,\|E\|_2}$ *and* $0 < \delta < 1$*. If the number of Monte Carlo samples is at least*

$$n \geq \frac{8}{3\,\epsilon^2}\,\frac{L^2}{\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\mathrm{intdim}\,(E)\right),$$

then with probability at least $1-\delta$ *the approximate active subspace* $\widehat{\mathcal{S}}$ *is well-defined, and the largest principal angle* $\angle(\widehat{\mathcal{S}},\mathcal{S})$ *is bounded by*

$$\sin\angle(\widehat{\mathcal{S}},\mathcal{S}) \leq \frac{4\,\|E\|_2}{\lambda_k - \lambda_{k+1}}\,\epsilon.$$

Proof. See section 3.4. ∎

Theorem 2.5 implies that $\widehat{\mathcal{S}}$ is likely to be an ϵ-accurate approximation to the active subspace $\mathcal{S}$ if:

1. E has a large relative eigenvalue gap $(\lambda_k - \lambda_{k+1})/\lambda_1$.
2. The matrix E has low numerical rank, that is, $\mathrm{intdim}(E) \ll m$.
3. The function f is smooth, in the sense that $L^2/\|E\|_2 \approx 1$; see section 3.4.

Remark 2.6. Monte Carlo sampling of x_j, according to $\rho(x)$, does not necessarily produce gradients $\nabla f(x_j)\,(\nabla f(x_j))^T$ that concentrate tightly around the mean E. This could be remedied with some form of importance sampling.

For instance, in order to speed up the sampling of Fourier features for kernel ridge regression, [2, section 4] proposes sampling according to the leverage function of the kernel. However, it is not clear how this can be implemented efficiently in practice; plus the required sampling amount appears to exhibit a much stronger dependence on the problem dimension than is acceptable in our context.

3. Proofs. We present all the materials required for the proofs of Theorem 2.1 (section 3.1), Corollary 2.2 (section 3.2), Theorem 2.3 (section 3.3), and Theorem 2.5 (section 3.4).

3.1. Everything for the proof of Theorem 2.1. The idea is to view $\widehat{E}$ as a sum of random variables. To this end we review a matrix Bernstein concentration inequality and the definition of intrinsic dimension (section 3.1.1), and then apply the concentration inequality to prove Theorem 2.1 (section 3.1.2), followed by a brief discussion of expectation-based bounds (section 3.1.3).

3.1.1. Matrix Bernstein concentration inequality, and intrinsic dimension. Concentration inequalities bound the deviation of a sum of random variables from the mean.

Here, the random variables are matrix-valued and bounded, and have zero mean and bounded "variance" in the sense of the Löwner partial order.[1] We use a matrix Bernstein concentration inequality with intrinsic dimension [28, section 7.2] in the

[1] If P_1 and P_2 are real symmetric matrices, then $P_1 \preceq P_2$ means that $P_2 - P_1$ is positive semidefinite [16, Definition 7.7.1].

context of a random sampling model [28, p. 83]. The *intrinsic dimension* quantifies the numerical rank of a symmetric positive semidefinite matrix and is instrumental in avoiding an explicit dependence on the matrix dimension.

DEFINITION 3.1 (section 2.1 in [19], Definition 7.1.1 in [28]). *The* intrinsic dimension *or* effective rank *of a nonzero, symmetric positive semidefinite matrix* $P \in \mathbb{R}^{m\times m}$ *is* $\operatorname{intdim}(P) \equiv \operatorname{trace}(P)/\|P\|_2$.

The symmetric positive semidefiniteness of P implies that its *intrinsic dimension* is bounded by the rank

$$1 \leq \operatorname{intdim}(P) \leq \operatorname{rank}(P) \leq m$$

and is equal to the *stable rank* of a square root [28, section 6.5.4]

$$\begin{aligned}\operatorname{intdim}(P) &= \frac{\lambda_1(P) + \cdots + \lambda_m(P)}{\lambda_1(P)} = \frac{\sigma_1(P^{1/2})^2 + \cdots + \sigma_m(P^{1/2})^2}{\sigma_1(P^{1/2})} \\ &= \left(\frac{\|P^{1/2}\|_F}{\|P^{1/2}\|_2}\right)^2 = \operatorname{sr}(P^{1/2}).\end{aligned}$$

THEOREM 3.2 ([18, 19] and Theorem 7.3.1 in [28]). *If*

1. *(independence)* X_j *are* n *independent real symmetric random matrices;*
2. *(boundedness)* $\max_{1\leq j\leq n} \|X_j\|_2 \leq \beta$ *for some* $\beta > 0$*;*
3. *(zero mean)* $\mathbb{E}[X_j] = 0$, $1 \leq j \leq n$*;*
4. *(bounded matrix variance)* $\sum_{j=1}^n \mathbb{E}[X_j^2] \preceq P$ *for some* P*;*
5. *(sufficient tolerance)* $\epsilon \geq \|P\|_2^{1/2} + \beta/3$*,*

then

$$\mathbb{P}\left[\left\|\sum_{j=1}^n X_j\right\|_2 \geq \epsilon\right] \leq 4\ \operatorname{intdim}(P)\ \exp\left(\frac{-\epsilon^2/2}{\|P\|_2 + \beta\epsilon/3}\right).$$

Since $\mathbb{E}\left[\sum_{j=1}^n X_j\right] = 0$, Theorem 3.2 is a bound for the deviation of the sum from its mean, and implies that a large deviation is unlikely if the matrix variance has low rank. Most importantly, the bound does not depend on the dimension of the matrices X_j.

3.1.2. Proof of Theorem 2.1. The proof is inspired by [15, Theorem 7.8] and similar in part to the proof of [17, Theorem 3]. Set

$$X_j \equiv z_j z_j^T \quad \text{and} \quad Y_j \equiv \tfrac{1}{n}(X_j - E) \ \text{ for } 1 \leq j \leq n, \qquad S \equiv \tfrac{1}{n}\sum_{j=1}^n X_j,$$

so that

$$\sum_{j=1}^n Y_j = \tfrac{1}{n}\sum_{j=1}^n (X_j - E) = S - E. \tag{3.1}$$

Before applying Theorem 3.2 to the sum of the Y_j, we need to verify that they satisfy the assumptions.

Independence. By assumption, the X_j are independent, and so are the Y_j.

Zero mean. Since $\mathbb{E}[X_j] = E$, the linearity of the expected value implies

$$\mathbb{E}\,[Y_j] = \tfrac{1}{n}\,(\mathbb{E}[X_j] - E) = \tfrac{1}{n}(E - E) = 0.$$

Boundedness. The positive semidefiniteness and boundedness of X_j and E from section 2.1 imply

$$\|Y_j\|_2 \leq \frac{1}{n}\,\max\{\|X_j\|_2, \|E\|_2\} \leq \frac{1}{n}\max\{L^2, \|E\|_2\} \leq \beta \equiv \frac{L^2}{n}. \tag{3.2}$$

Matrix variance. Multiply out, and apply the definition of E:

$$\begin{aligned}\mathbb{E}\,[Y_j^2] &= \frac{1}{n^2}\,\mathbb{E}\left[(X_j - E)^2\right] = \frac{1}{n^2}\left(\mathbb{E}[X_j^2] - E\;\mathbb{E}[X_j] - \mathbb{E}[X_j]\,E + E^2\right)\\ &= \frac{1}{n^2}\left(\mathbb{E}[X_j^2] - E^2\right).\end{aligned}$$

Since E is positive semidefinite, we can drop it without decreasing semidefiniteness:

$$\mathbb{E}\,[Y_j^2] \preceq \frac{1}{n^2}\,\mathbb{E}\,[X_j^2]. \tag{3.3}$$

Bounded matrix variance. Since X_j is an outer product,

$$X_j^2 = z_j^T z_j\, X_j = \|z_j\|_2^2\, X_j \preceq L^2 X_j,$$

thus $\mathbb{E}[X_j^2] \preceq L^2\,\mathbb{E}[X_j] = L^2 E$. This, together with (3.3), gives $\mathbb{E}\left[Y_j^2\right] \preceq \frac{L^2}{n^2}\,E$. The linearity of the expected value across the n identically distributed summands implies

$$\sum_{j=1}^{n}\mathbb{E}\,[Y_j^2] \preceq \sum_{j=1}^{n}\frac{L^2}{n^2}E = P \equiv \frac{L^2}{n}\,E, \tag{3.4}$$

where P is symmetric positive semidefinite since E is.

From $\|P\|_2 = \frac{L^2}{n}\,\|E\|_2$ and $\operatorname{trace}(P) = \frac{L^2}{n}\operatorname{trace}(E)$ follows

$$\operatorname{intdim}(P) = \operatorname{trace}(E)/\|E\|_2.$$

Application of Theorem 3.2. Substituting (3.1), (3.2), and (3.4) into Theorem 3.2 gives the probability for the absolute error:

$$\mathbb{P}\left[\|S - E\|_2 \geq \hat{\epsilon}\right] \leq 4\,\operatorname{intdim}(E)\,\exp\left(-\frac{n}{L^2}\,\frac{\hat{\epsilon}^2/2}{\|E\|_2 + \hat{\epsilon}/3}\right).$$

Setting $\hat{\epsilon} = \|E\|_2\,\epsilon$ gives the probability for the relative error:

$$\mathbb{P}\left[\frac{\|S - E\|_2}{\|E\|_2} \geq \epsilon\right] \leq 4\,\operatorname{intdim}(E)\,\exp\left(-n\,\frac{\|E\|_2}{L^2}\,\frac{\epsilon^2/2}{1 + \epsilon/3}\right).$$

Setting the above right-hand side equal to δ and solving for ϵ gives

$$\epsilon = \gamma + \sqrt{\gamma\,(\gamma + 6)}, \qquad \text{where} \quad \gamma = \frac{1}{3n}\,\frac{L^2}{\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\operatorname{intdim}(E)\right).$$

Sufficient tolerance. We still need to check the lower bound for $\hat{\epsilon}$ and verify that $\hat{\epsilon} \geq \beta/3 + \|P\|_2^{1/2}$, which is equivalent to verifying that

$$\epsilon \geq \frac{\beta}{3\|E\|_2} + \frac{\|P\|_2^{1/2}}{\|E\|_2}.$$

From $0 < \delta < 1$ and $\operatorname{intdim}(E) \geq 1$ follows $e < \frac{4}{\delta} \leq \frac{4}{\delta} \operatorname{intdim}(E)$, which implies $\ln(\frac{4}{\delta} \operatorname{intdim}(E)) \geq 1$. Together with (3.2) this gives

$$\frac{\beta}{3\|E\|_2} = \frac{1}{3n} \frac{L^2}{\|E\|_2} \leq \frac{1}{3n} \frac{L^2}{\|E\|_2} \ln\left(\frac{4}{\delta} \operatorname{intdim}(E)\right) = \gamma,$$

and with (3.4)

$$\frac{\|P\|_2^{1/2}}{\|E\|_2} = \frac{1}{\|E\|_2} \sqrt{\frac{L^2}{n}\|E\|_2} = \sqrt{\frac{1}{n} \frac{L^2}{\|E\|_2}} \leq \sqrt{6\gamma} \leq \sqrt{\gamma(\gamma+6)}.$$

Adding the two previous inequalities gives the required lower bound:

$$\epsilon \geq \gamma + \sqrt{\gamma(\gamma+6)} \geq \frac{\beta}{3\|E\|_2} + \frac{\|P\|_2^{1/2}}{\|E\|_2}. \tag{3.5}$$

3.1.3. Expectation-based bounds. Following the observation [28, section 4.1] that matrix concentration bounds may not always give satisfactory information about the tail, we consider an alternative to the exponential Bernstein concentration inequality [28, Theorem 7.3.1], represented here by Theorem 3.2, and combine an expectation bound with the scalar Markov inequality.

The intrinsic dimension expectation bound [28, Corollary 7.3.2], together with the assumptions in Theorem 3.2 and the bounds in [28, section 7.4.4], implies for matrices P with $\operatorname{intdim}(P) \geq 2$ that

$$\mathbb{E}\left[\left\|\sum_{j=1}^{n} X_j\right\|_2\right] \leq \frac{10}{3}\left(\sqrt{\|P\|_2\,\theta} + \beta\,\theta\right), \qquad \theta \equiv \ln(1 + \operatorname{intdim}(P)).$$

Combined with Markov's inequality [28, (2.2.1)] this gives

$$\delta \equiv \mathbb{P}\left[\|\widehat{E} - E\|_2 \geq \epsilon\right] \leq \mathbb{E}\left[\left\|\sum_{j=1}^{n} X_j\right\|_2\right] / \epsilon.$$

Thus, the error bound is inversely proportional to the failure probability,

$$\epsilon = \mathcal{O}(1/\delta).$$

In contrast, Theorem 2.1 implies the much weaker dependence

$$\epsilon = \mathcal{O}(\ln(1/\delta)).$$

Since we are interested in extremely small failure probabilities, on the order of machine epsilon, $\delta = 10^{-15}$, exponential concentration inequalities are preferable to expectation-based bounds in our context.

3.2. Proof of Corollary 2.2. A lower bound on ϵ is required by the last assumption of Theorem 3.2 and was established in (3.5), which is equivalent to $2\,\gamma\,(3+\epsilon) \leq \epsilon^2$. This bound, and therefore (3.5), definitely holds if

$$\gamma \leq \epsilon^2/8 \tag{3.6}$$

because $\epsilon < 1$. Setting $\gamma = \alpha/n$ and $\alpha \equiv \frac{L^2}{3\,\|E\|_2} \ln\left(\frac{4}{\delta}\,\mathrm{intdim}\,(E)\right)$ shows that (3.6) is equivalent to the desired lower bound for n.

3.3. Everything for the proof of Theorem 2.3. We verify the conditions and apply a deterministic bound for the subspace angle (section 3.3.1), and then present the proof of Theorem 2.3 (section 3.3.2).

3.3.1. Deterministic subspace angle bound. We bound $\sin\angle(\widehat{\mathcal{S}}, \mathcal{S})$ in terms of the absolute error $\|\widehat{E} - E\|_2$.

The keep the notation simple, partition the eigenvectors in (2.1),

$$V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}, \qquad \text{where} \qquad V_1 \equiv \begin{pmatrix} v_1 & \cdots & v_k \end{pmatrix} \in \mathbb{R}^{m\times k}, \tag{3.7}$$

and conformally partition the eigenvectors in (2.2),

$$\widehat{V} = \begin{pmatrix} \widehat{V}_1 & \widehat{V}_2 \end{pmatrix}, \qquad \text{where} \qquad \widehat{V}_1 \equiv \begin{pmatrix} \widehat{v}_1 & \cdots & \widehat{v}_k \end{pmatrix} \in \mathbb{R}^{m\times k}.$$

Next is a straightforward specialization of [25, Theorems 2.7 and 4.11], [26, Theorem V.2.7], and [13, Corollary 8.1.11] to real symmetric matrices and the two norm.

LEMMA 3.3. *Partition as in* (3.7),

$$F = \begin{pmatrix} F_{11} & F_{12} \\ F_{12}^T & F_{22} \end{pmatrix} \equiv \begin{pmatrix} V_1 & V_2 \end{pmatrix}^T (\widehat{E} - E) \begin{pmatrix} V_1 & V_2 \end{pmatrix}.$$

If $\mathrm{gap} \equiv \lambda_k - \lambda_{k+1} > 0$,

$$\eta \equiv \mathrm{gap} - \|F_{11}\|_2 - \|F_{22}\|_2 > 0, \qquad \textit{and} \qquad \frac{\|F_{12}\|_2}{\mathrm{gap}} < \tfrac{1}{2}, \tag{3.8}$$

then

$$\|V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T\|_2 \leq 2\,\|F_{12}\|_2/\eta.$$

Now comes the deterministic basis for Theorem 2.3, and it requires the perturbation $\|\widehat{E} - E\|_2$ to be sufficiently small compared to the eigenvalue gap. The conclusions are spelled out in more detail than usual to ensure a correct interface with the matrix concentration bounds for section 3.3.2.

THEOREM 3.4. *If for some* $1 \leq k < m$ *the matrix* E *has an eigenvalue gap* $\lambda_k - \lambda_{k+1} > 0$ *and* $\|\widehat{E} - E\|_2 < (\lambda_k - \lambda_{k+1})/4$, *then*

1. $\widehat{E}$ *has an eigenvalue gap at the same location as* E, *that is,* $\widehat{\lambda}_k - \widehat{\lambda}_{k+1} > 0$.
2. *The dominant subspaces* $\mathcal{S} = \mathrm{range}\,(V_1)$ *and* $\widehat{\mathcal{S}} = \mathrm{range}\,(\widehat{V}_1)$ *are well-defined.*
3. *The largest principal angle* $\angle(\widehat{\mathcal{S}}, \mathcal{S})$ *is bounded by*

$$\sin\angle(\widehat{\mathcal{S}}, \mathcal{S}) \leq 4\,\frac{\|\widehat{E} - E\|_2}{\lambda_k - \lambda_{k+1}}.$$

Proof. With the abbreviations $\tau \equiv \|\widehat{E} - E\|_2$ and $\text{gap} \equiv \lambda_k - \lambda_{k+1} > 0$, the all-important assumption takes the form

$$\|F\|_2 = \tau < \text{gap}/4. \tag{3.9}$$

The three statements will now be proved in the order listed.

1. To show that $\widehat{\lambda}_k - \widehat{\lambda}_{k+1} > 0$, invoke the Cauchy interlace theorem [21, section 10-1], $\max_{1\le j\le m} |\lambda_j - \widehat{\lambda}_j| \le \tau$, which implies in particular $\widehat{\lambda}_k \ge \lambda_k - \tau$ and $\widehat{\lambda}_{k+1} \le \lambda_{k+1} + \tau$. Together with (3.9), this gives
$$\widehat{\lambda}_k - \widehat{\lambda}_{k+1} \ge (\lambda_k - \tau) - (\lambda_{k+1} + \tau) = \text{gap} - 2\tau > 4\tau - 2\tau = 2\tau \ge 0.$$
2. The existence of the eigenvalue gaps $\lambda_k - \lambda_{k+1} > 0$ and $\widehat{\lambda}_k - \widehat{\lambda}_{k+1} > 0$ implies that $\mathcal{S}$ and $\widehat{\mathcal{S}}$ are simple invariant subspaces [26, section V.1].
Since the columns of $V_1, \widehat{V}_1 \in \mathbb{R}^{m\times k}$ are orthonormal bases for $\mathcal{S} = \text{range}(V_1)$ and $\widehat{\mathcal{S}} = \text{range}(\widehat{V}_1)$, respectively, the matrices $V_1 V_1^T$ and $\widehat{V}_1 \widehat{V}_1^T$ are orthogonal projectors onto $\mathcal{S}$ and $\widehat{\mathcal{S}}$, respectively.
3. The two norm difference between orthogonal projectors onto equidimensional spaces is the sine of the largest principal angle [13, sections 2.5.3, 6.4.3], [25, Corollary 2.6],
$$\sin \angle(\widehat{\mathcal{S}}, \mathcal{S}) = \|V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T\|_2. \tag{3.10}$$
To bound this difference in terms of τ, we apply Lemma 3.3, but need to verify first that its conditions (3.8) hold.
The first condition in (3.8) follows from (3.9) and
$$\eta = \text{gap} - \|F_{11}\|_2 - \|F_{22}\|_2 \ge \text{gap} - 2\|F\|_2 \ge \text{gap} - \tfrac{1}{2}\text{gap} = \tfrac{1}{2}\text{gap} > 0.$$
The second condition in (3.8) also follows from (3.9) and
$$\frac{\|F_{12}\|_2}{\text{gap}} \le \frac{\|F\|_2}{\text{gap}} < \tfrac{1}{4} < \tfrac{1}{2}.$$
The desired bound follows from combining (3.10), Lemma 3.3, and $\eta \ge \frac{1}{2}\text{gap}$. ∎

Theorem 3.4 implies that the subspace $\mathcal{S}$ is well-conditioned if the eigenvalue gap $\lambda_k - \lambda_{k+1}$ is large compared to the matrix perturbation $\|\widehat{E} - E\|_2$.

3.3.2. Proof of Theorem 2.3. We combine the probabilistic bound in Corollary 2.2 and the deterministic bound in Theorem 3.4.

Corollary 2.2 implies: If
$$n \ge \frac{8}{3\,\epsilon^2} \frac{L^2}{\|E\|_2} \ln\left(\frac{4}{\delta}\ \text{intdim}\,(E)\right), \tag{3.11}$$
then with probability at least $1 - \delta$ we have $\|\widehat{E} - E\|_2 \le \|E\|_2\,\epsilon$. Theorem 2.3 guarantees, by assumption, that $\|E\|_2\,\epsilon < (\lambda_k - \lambda_{k+1})/4$.

Combining the two gives: If (3.11) holds, then with probability at least $1 - \delta$ we have $\|\widehat{E} - E\|_2 \le (\lambda_k - \lambda_{k+1})/4$. This in turn means: If (3.11) holds, then with probability at least $1 - \delta$ the assumptions for Theorem 3.4 are satisfied, and its conclusions hold.

Remark 3.5. We do not see how to transfer the eigenvalue gap $\lambda_k - \lambda_{k+1}$ from the angle bound into the number of samples.

This is because the deterministic bound in Theorem 3.4 and the probabilistic bound in Corollary 2.2 make competing demands. The former requires an upper bound on the matrix perturbation $\|\widehat{E} - E\|_2$, while the latter requires a lower bound.

3.4. Proof of Theorem 2.5. Once the required bound for E has been established below, the proof of Theorem 2.5 is a direct consequence of Theorem 2.3.

LEMMA 3.6. *The matrix E in Theorem* 2.5 *satisfies $\|E\|_2 \leq L^2$.*

Proof. This follows from the assumptions in section 2.5 and the fact that ρ is a probability density function,

$$\begin{aligned}\|E\|_2 &= \left\| \int \nabla f(x)(\nabla f(x))^T \rho(x)\, dx \right\|_2 \leq \int \|\nabla f(x)(\nabla f(x))^T\|_2\, \rho(x)\, dx \\ &\leq \max_x \{\|\nabla f(x)\|_2^2\} \int \rho(x)\, dx \leq L^2. \end{aligned}$$ ∎

Thus we can interpret

$$L^2/\|E\|_2 \geq 1 \tag{3.12}$$

as a measure for the smoothness of f.

Acknowledgments. We thank Haim Avron, David Bindel, Serkan Gugercin, Mert Gurbuzbalaban, Tim Kelley, and Jim Nagy for helpful discussions.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

[2] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh, *Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees*, in Proceedings of the 34th International Conference on Machine Learning (ICML), Proc. Mach. Learn. Res. 70, 2017.

[3] Y. Bang, H. S. Abdel-Khalik, and J. M. Hite, *Hybrid reduced order modeling applied to nonlinear models*, Internat. J. Numer. Methods Engrg., 91 (2012), pp. 929–949.

[4] M. Bauerheim, A. Ndiaye, P. Constantine, S. Moreau, and F. Nicoud, *Symmetry breaking of azimuthal thermoacoustic modes: The UQ perspective*, J. Fluid Mech., 789 (2016), pp. 534–566.

[5] H. Chen, Q. Wang, R. Hu, and P. G. Constantine, *Conditional sampling and experiment design for quantifying manufacturing error of transonic airfoil*, in 49th AIAA Aerospace Sciences Meeting, 2011.

[6] P. G. Constantine, A. Doostan, Q. Wang, and G. Iaccarino, *A surrogate accelerated Bayesian inverse analysis of the HyShot* II *flight data*, in Proceedings of the 52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Material Conference, 2011.

[7] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524, https://doi.org/10.1137/130916138.

[8] P. G. Constantine, E. Dow, and Q. Wang, *Erratum: Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A3030–A3031, https://doi.org/10.1137/140983598.

[9] P. G. Constantine and D. Gleich, *Computing Active Subspaces with Monte Carlo*, preprint, https://arxiv.org/abs/1408.0545, 2015.

[10] P. G. Constantine, Q. Wang, and G. Iaccarino, *A Method for Spatial Sensitivity Analysis*, Tech. report, Center for Turbulence Research, Stanford University, 2012.

[11] P. G. Constantine, B. Zaharatos, and M. Campanelli, *Discovering an active subspace in a single-diode solar cell model*, Stat. Anal. Data Min., 8 (2015), pp. 264–273.
[12] P. Drineas, I. C. F. Ipsen, E.-M. Kontopoulou, and M. Magdon-Ismail, *Structural convergence results for approximation of dominant subspaces from block Krylov spaces*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 567–586, https://doi.org/10.1137/16M1091745.
[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, 2013.
[14] 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.
[15] J. T. Holodnak and I. C. F. Ipsen, *Randomized approximation of the Gram matrix: Exact computation and probabilistic bounds*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 110–137, https://doi.org/10.1137/130940116.
[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, 2013.
[17] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, *Randomized nonlinear component analysis*, in Proceedings of the 31st International Conference on Machine Learning (ICML), Proc. Mach. Learn. Res. 32, 2014, pp. 1359–1367.
[18] S. Minsker, *On Some Extensions of Bernstein's Inequality for Self-Adjoint Operators*, preprint, version 1, https://arxiv.org/abs/1112.5448v1, 2011.
[19] S. Minsker, *On Some Extensions of Bernstein's Inequality for Self-Adjoint Operators*, preprint, version 3, https://arxiv.org/abs/1112.5448v3, 2017.
[20] N. Namura, K. Shimoyama, and S. Obayashi, *Kriging surrogate model enhanced by coordinate transformation of design space based on eigenvalue decomposition*, in Evolutionary Multi-Criterion Optimization, Lecture Notes in Comput. Sci. 9018, Springer, 2015, pp. 321–335.
[21] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980.
[22] T. M. Russi, *Uncertainty Quantification with Experimental Data and Complex System Models*, Ph.D. thesis, University of California, Berkeley, 2010.
[23] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, SIAM, 2014.
[24] I. M. Sobol' and S. Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Math. Comput. Simulation, 79 (2009), pp. 3009–3017.
[25] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764, https://doi.org/10.1137/1015095.
[26] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.
[27] M. Stoyanov and C. G. Webster, *A gradient-based sampling approach for dimension reduction of partial differential equations with stochastic coefficients*, Int. J. Uncertain. Quantif., 5 (2015), pp. 49–72.
[28] J. A. Tropp, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learning, 8 (2015), pp. 1–230.
[29] R. Vershynin, *Introduction to the Non-Asymptotic Analysis of Random Matrices*, preprint, https://arxiv.org/abs/1011.3027, 2011.