Electronic Transactions on Numerical Analysis.
Volume 55, pp. 213–241, 2022.

ISSN 1068–9613.
DOI: 10.1553/etna_vol55s213

ETNA
Kent State University and
Johann Radon Institute (RICAM)

MONTE CARLO ESTIMATORS FOR THE SCHATTEN p-NORM OF SYMMETRIC POSITIVE SEMIDEFINITE MATRICES*

ETHAN DUDLEY‡, ARVIND K. SAIBABA†, AND ALEN ALEXANDERIAN†

Abstract. We present numerical methods for computing the Schatten p-norm of positive semi-definite matrices. Our motivation stems from uncertainty quantification and optimal experimental design for inverse problems, where the Schatten p-norm defines a measure of uncertainty. Computing the Schatten p-norm of high-dimensional matrices is computationally expensive. We propose a matrix-free method to estimate the Schatten p-norm using a Monte Carlo estimator and derive convergence results and error estimates for the estimator. To efficiently compute the Schatten p-norm for non-integer and large values of p, we use an estimator using Chebyshev polynomial approximations and extend our convergence and error analysis to this setting as well. We demonstrate the performance of our proposed estimators on several test matrices and in an application to optimal experimental design for a model inverse problem.

Key words. Schatten p-norm, Monte Carlo estimator, optimal experimental design, Chebyshev polynomials.

AMS subject classifications. 65F35, 65F50, 65C05,

1. Introduction.

The Schatten p-norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$|||\mathbf{A}|||_p = \left(\sum_{j=1}^{\min\{m,n\}} \sigma_j^p \right)^{1/p},$$

where $p \geq 1$ and σ_j is the jth singular value of $\mathbf{A}$ for $1 \leq j \leq \min\{m,n\}$. If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive semi-definite (SPSD) matrix, then the singular values of $\mathbf{A}$ are its eigenvalues, and the Schatten p-norm takes the form

$$|||\mathbf{A}|||_p = \left(\sum_{j=1}^{n} \lambda_j^p \right)^{1/p} = \left(\operatorname{tr}(\mathbf{A}^p)\right)^{1/p}, \tag{1.1}$$

where the λ_j's are the eigenvalues of $\mathbf{A}$. There are several notable special cases of the Schatten p-norm, including the nuclear norm ($p = 1$), the Frobenius norm ($p = 2$) and the spectral norm ($p \to \infty$). Since it encapsulates many well-known norms as special cases, the Schatten p-norm is frequently used in linear algebra and analysis [4].

Our motivation for computing the Schatten p-norm arises from uncertainty quantification and optimal experimental design (OED) for Bayesian inverse problems. An inverse problem seeks to estimate parameters of interest using experimental measurements. The goal of OED is to identify an optimal set of experiments by optimizing certain design criteria that measure the uncertainty in the estimated parameters, subject to budgetary or physical constraints. A well-known design criterion, known as the P-optimal design criterion, can be expressed in terms of the Schatten p-norm. Since optimization algorithms for OED require repeated evaluations of the design criterion for large matrices, efficient algorithms for estimating the Schatten p-norm are desirable.

In this article, we focus on computing the Schatten p-norm for large SPSD matrices. Computing the Schatten p-norm for such matrices is computationally challenging, because it

*Received May 20, 2020. Accepted November 3, 2021. Published online on December 20, 2021. Recommended by Y. Saad.

†Department of Mathematics, North Carolina State University, North Carolina, USA ({asaibab,aalexan3}@ncsu.edu).

‡Department of Mathematics, North Carolina State University, North Carolina, USA. Now at Department of Mathematics, University of Maryland, College Park, USA (edudley1@umd.edu).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

requires computing either the matrix pth power or all of its eigenvalues. However, if the matrix is large and its entries are not available explicitly, then the Schatten p-norm cannot be easily computed from its definition (1.1) and specialized numerical methods are necessary. Therefore, we consider computing the Schatten p-norm using matrix-free Monte Carlo methods. In a matrix-free method for computing $|||\mathbf{A}|||_p$ we only require matrix-vector products involving $\mathbf{A}$.

Related work. Hutchinson [12] developed a matrix-free Monte Carlo estimator using samples from the Rademacher distribution for computing $\mathrm{tr}(\mathbf{A})$, i.e., the Schatten 1-norm. Avron and Toledo [3] extended this idea to random variables from other distributions such as Gaussian and uniformly selected vectors from an orthogonal matrix. They devised several metrics for comparing the various trace estimators, including a single sample variance metric and a Chernoff-style lower bound on the minimum number of samples required to meet a given error tolerance with a given confidence level. This is made precise in the following definition:

DEFINITION 1.1. *Given $\varepsilon > 0$, $\delta \in (0,1)$, and an appropriate distribution for random samples $\mathbf{w}_j \in \mathbb{R}^n$, $j = 1, \ldots, M$, we say*

$$Z_M = \frac{1}{M} \sum_{j=1}^{M} \mathbf{w}_j^T \mathbf{A} \mathbf{w}_j$$

is an (ε, δ) estimator for $\mathrm{tr}(\mathbf{A})$ if

$$\mathbb{P}\left(\left| Z_M - \mathrm{tr}(\mathbf{A}) \right| \leq \varepsilon |\mathrm{tr}(\mathbf{A})| \right) \geq 1 - \delta.$$

This definition alternatively says that Z_M is an (ε, δ) estimator if it has a relative error at most ε with probability at least $1 - \delta$. Avron and Toledo [3] provided a lower bound on the number of samples so that Z_M is (ε, δ) estimator for $\mathrm{tr}(\mathbf{A})$ when $\mathbf{w}_j$ are drawn from the Gaussian, Rademacher and Uniform distributions. Roosta-Khorasani and Ascher [25] further reduced the lower bound on the number of samples needed for an (ε, δ) estimator for $\mathrm{tr}(\mathbf{A})$ when the estimators use random vectors from the Rademacher and Gaussian distributions. This Monte Carlo estimator has been extended to Schatten p-norm using Chebyshev polynomials [9] and Lanczos approach [29].

A recent survey paper by Martinsson and Tropp [19] reviews estimators for the Schatten p-norms, which avoid working with $\mathbf{A}^p$ directly. Let $\mathbf{X} = \mathbf{\Omega}^T \mathbf{A} \mathbf{\Omega}$, where the entries of $\mathbf{\Omega} \in \mathbb{R}^{n \times M}$ have zero mean and unit variance. The estimator V_p in Kong and Valiant [16] is

$$V_p = \binom{M}{p}^{-1} \mathrm{tr}(\mathcal{T}(\mathbf{X})^{p-1} \mathbf{X}),$$

where $\mathcal{T}(\mathbf{X})$ is a matrix that contains the strictly upper triangular part of $\mathbf{X}$ and zeroes out the rest of the entries. Note that V_p is an unbiased estimator for $|||\mathbf{A}|||_p$. A related estimator is

$$W_p = \frac{(M-p)!}{M!} \sum_{1 \leq i_1, \ldots, i_p \leq M} \mathbf{X}_{i_1, i_2} \mathbf{X}_{i_2, i_3} \cdots \mathbf{X}_{i_p, i_1},$$

where the summation is only over distinct indices. Similar to V_p, W_p is an unbiased estimator for $|||\mathbf{A}|||_p$. For both estimators, the recommended number of samples is $M \gtrsim n^{1-2/p}$. This lower bound was established by [17]. Stronger results have been developed in [18], where the authors studied the computation of the Schatten p-norm in the streaming setting. They showed the upper bound required was $\mathcal{O}(n^{2-4/p})$, where $p \geq 4$ is an even integer; the corresponding lower bound is $\Omega(n^{2-4/p})$. Both of these estimators are expensive for large p; however, the algorithm only requires M matrix-vector products involving $\mathbf{A}$. Theoretical analysis suggests

ETNA
Kent State University and
Johann Radon Institute (RICAM)

that the variance of these estimators is large, which makes their use for large-scale applications impractical [19].

The issue of estimating Schatten p-norms has also received considerable attention in the Theoretical Computer Science community; see [20] and references therein. In that paper, the authors estimate the Schatten p-norm and other spectral sums by estimating the histogram of the spectrum of the matrix. This is obtained by splitting the spectrum of $\mathbf{A}$ into small slices and using trace estimators to determine the number of singular values in each slice of the spectrum. In contrast, our approach builds a single global polynomial approximation of $\mathbf{A}^p$ over the spectrum of $\mathbf{A}$. Another interesting approach considered in [14] involves estimating the Schatten p-norm using only a limited number of entries of the matrix.

Our approach and contributions. We focus on the analysis and the development of efficient computational methods for the following estimator of $|||\mathbf{A}|||_p$

$$|||\mathbf{A}|||_p \approx \left(\frac{1}{M} \sum_{j=1}^{M} \mathbf{w}_j^T \mathbf{A}^p \mathbf{w}_j \right)^{1/p},$$

where $\mathbf{w}_j$ are random vectors from an appropriate distribution. To our knowledge, an analysis of the convergence of this (biased) estimator has not been performed in the literature. Computing the Monte Carlo estimator involves repeated applications of $\mathbf{A}^p$ to a vector, which is computationally expensive for large or non-integer values of p. To reduce this cost, two different approaches were proposed based on Chebyshev polynomial approximation [9] and on a Lanczos approach [29]. In this article, building on the work [9], we consider approximate Monte Carlo estimators based on Chebyshev polynomials.

The following are the main contributions of this article.

1. In our analysis of the new estimator we derive bounds on the expectation, bias, and variance (Section 3.2), and we show that the estimator converges almost surely both in L^1 and L^2 (Sections 3.1 and 3.3). In Section 3.4, we analyze the impact of p on the number of samples required to form an (ε, δ) estimator.
2. In Section 4, we consider a variation of the Chebyshev-Monte Carlo method proposed in [9]. This approach is applicable to non-integer values as well as large values of p. We extend our results from the standard Monte Carlo approach to the Chebyshev-Monte Carlo approach. Using results from polynomial approximation theory, we show that the degree of the polynomial approximation we use is optimal in an asymptotic sense.
3. We provide extensive numerical tests on synthetic matrices, matrices arising from real-world problems, and a model problem from OED, which illustrate the theoretical results. We also provide numerical evidence that a small degree Chebyshev approximation $\psi_N(\mathbf{A})$ to $\mathbf{A}^{p/2}$ is sufficient for an accurate estimator.

2. Background.

In this section, we review known results for the two largest contributing ideas in this article: Monte Carlo Trace Estimators (Section 2.1) and Chebyshev Polynomials (Section 2.2).

2.1. Monte Carlo trace estimators.

In what follows we let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

DEFINITION 2.1. *Let $\mathbf{w} : \Omega \to \mathbb{R}^n$ be a random n-vector with mean 0 and identity covariance matrix, and let $\mathbf{B}$ be a symmetric matrix. Then, the Monte Carlo trace estimator of*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$\mathbf{B}$ *is given by*

$$Z_M = \frac{1}{M}\sum_{j=1}^{M} \mathbf{w}_j^T \mathbf{B}\mathbf{w}_j, \tag{2.1}$$

where the $\mathbf{w}_j$, $j = 1, \ldots, M$, *are independent and distributed according to the law of* $\mathbf{w}$.

We call Z_M a trace estimator of $\mathbf{B}$ because $\mathbb{E}(\mathbf{w}^T\mathbf{B}\mathbf{w}) = \operatorname{tr}(\mathbf{B})$ and therefore by the linearity of expectation $\mathbb{E}(Z_M) = \operatorname{tr}(\mathbf{B})$ [12, 3]. Furthermore, since $\mathbf{w}_j^T\mathbf{B}\mathbf{w}_j \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, by the strong law of large numbers [13], we have

$$\mathbb{P}\left(\lim_{M\to\infty} Z_M = \operatorname{tr}(\mathbf{B})\right) = 1.$$

That is, Z_M converges to $\operatorname{tr}(\mathbf{B})$ almost surely (a.s.). Lastly, we can formulate a Chernoff-style lower bound on M to guarantee that Z_M is an (ε, δ) estimator. This means that M is the least number of samples to guarantee Z_M is an (ε, δ) estimator for $\operatorname{tr}(\mathbf{B})$, i.e., Z_M satisfies Definition 1.1. Note that the (ε, δ) bound on M depends on the distribution from which the $\mathbf{w}_j$ are chosen. This is summarized in Table 2.1.

TABLE 2.1
Variance and number of samples required for an (ε, δ) *bound for* $\operatorname{tr}(\mathbf{B})$. *Here, the* $\mathbf{w}_j$ *vectors are chosen from the Gaussian and Rademacher distributions [3, 25].*

	$\operatorname{Var}(Z_M)$	(ε, δ) bound
Gaussian	$\frac{2\|\mathbf{B}\|_F^2}{M}$	$M \geq 8\varepsilon^{-2}\ln\left(\frac{2}{\delta}\right)$
Rademacher	$\frac{2(\|\mathbf{B}\|_F^2 - \sum_{i=1}^n \mathbf{B}_{ii}^2)}{M}$	$M \geq 6\varepsilon^{-2}\ln\left(\frac{2}{\delta}\right)$

2.2. Chebyshev polynomials. Throughout this article, we will use Chebyshev polynomials of the first kind, which are defined as

$$T_j(x) = \cos(j \arccos(x)), \qquad x \in [-1, 1], \quad j = 0, 1, 2, \ldots$$

As is well-known, these polynomials are orthogonal with respect to the inner product $\langle u, v\rangle_w = \int_{-1}^{1} u(x)v(x)\, w(x)dx$, with the weight function $w(x) = 1/\sqrt{1-x^2}$. In particular,

$$\langle T_i, T_j\rangle_w = \begin{cases} \pi, & i = j = 0, \\ \frac{\pi}{2}, & i = j \neq 0, \\ 0, & i \neq j. \end{cases}$$

Moreover, any continuous function g on the interval $[-1, 1]$ can be expressed as [5]

$$g(x) = c_0 T_0(x) + \sum_{j=1}^{\infty} c_j T_j(x),$$

where the series converges uniformly. The coefficients can be computed by

$$c_j = \frac{2}{\pi}\int_{-1}^{1} \frac{g(x)T_j(x)}{\sqrt{1-x^2}}\, dx = \frac{2}{\pi}\int_0^{\pi} g(\cos(\theta))\cos(j\theta)\, d\theta, \tag{2.2}$$

and c_0 carries an additional factor of a half. Note that the Chebyshev polynomial approximation to a function is equivalent to the Fourier cosine series approximation of g [5]. Therefore the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

coefficients c_j can be computed using the real part of the Fast Fourier Transform (FFT) of g; see [28] for details, as well as computer code for doing so.

Let $\psi_N(x)$ be the Nth degree Chebyshev approximation to g. The error in $\psi_N(x)$ is bounded tightly by [5]

$$|g(x) - \psi_N(x)| \leq \sum_{j=N+1}^{\infty} |c_j|.$$

Trefethen [28] presented a method for approximating this error without computing the remaining coefficients for both analytic functions and functions with singularities in the complex plane. Here we present the analytic version:

$$|g(x) - \psi_N(x)| \leq \frac{4U}{\rho^N(\rho - 1)},$$

where g is analytic in the interior of an ellipse E in the complex plane with foci at ± 1, $U = \sup_{z \in E} g(z)$, and ρ is the sum of the major and minor semi-axes of E with $\rho > 1$.

Finally, we recall that Chebyshev polynomials satisfy a three term recurrence relation [5]:

$$T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x), \quad x \in [-1, 1],$$

with $T_0(x) = 1$ and $T_1(x) = x$. This ensures that matrix-vector products using the Chebyshev matrix polynomials can be computed in a matrix-free manner, which is useful in constructing a Monte Carlo approximation to $|||A|||_p$.

3. Monte Carlo Estimators and their analysis.

In this section, we construct a Monte Carlo estimator for the Schatten p-norm (Section 3.1) and present a detailed analysis of convergence of the estimator (Section 3.3).

3.1. Building a Schatten p-norm estimator.

Recall that if $\mathbf{A}$ is SPSD and $\mathbf{w} : \Omega \to \mathbb{R}^n$ is an n-vector with mean 0 and identity covariance matrix, then

$$|||\mathbf{A}|||_p^p = \operatorname{tr}(\mathbf{A}^p) = \mathbb{E}(\mathbf{w}^T \mathbf{A}^p \mathbf{w}).$$

Therefore, consider the following Monte Carlo estimator for $|||\mathbf{A}|||_p$.

DEFINITION 3.1. *Let $\mathbf{A}$ be an SPSD matrix. We define the Monte Carlo estimator for $|||\mathbf{A}|||_p$ as*

$$X_M = \left(\frac{1}{M} \sum_{j=1}^{M} \mathbf{w}_j^T \mathbf{A}^p \mathbf{w}_j \right)^{1/p}, \quad M \geq 1,$$

where the $\mathbf{w}_j$ are realizations of a random variable $\mathbf{w} : \Omega \to \mathbb{R}^n$ with $\mathbb{E}(\mathbf{w}) = 0$ and $\mathbb{E}(\mathbf{w}\mathbf{w}^T) = \mathbf{I}$.

Note that X_M^p is an unbiased estimator for $|||\mathbf{A}|||_p^p$. Furthermore, if $p = 1$, then X_M is just the Monte Carlo trace estimator (2.1).

In Algorithm 1, we provide a pseudo-code for efficiently computing X_M for positive integer values of p. First, note that by using the symmetry of $\mathbf{A}$, computing X_M using Algorithm 1 requires $\lceil \frac{p}{2} \rceil M$ matrix-vector products with $\mathbf{A}$. Let the cost of a matrix-vector product with $\mathbf{A}$ be denoted by $T_{\mathbf{A}}$. If $\mathbf{A}$ is dense, then $T_{\mathbf{A}} = 2n^2$; on the other hand, if $\mathbf{A}$ is sparse, then $T_{\mathbf{A}} = \operatorname{nnz}(\mathbf{A})$, where nnz is the number of nonzeros of $\mathbf{A}$. Therefore the computational cost of Algorithm 1 is $T_{\mathbf{A}} M \lceil \frac{p}{2} \rceil + \mathcal{O}(Mp)$. Second, this algorithm can scale

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 1: Constructing the Monte Carlo Estimator X_M.

Input: a SPSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, positive integers M (number of samples) and p (Schatten p degree)
Initialize: $X_M \leftarrow 0$
$K \leftarrow \left\lfloor \frac{p}{2} \right\rfloor$
for $j = 1$ to M **do**
  $\mathbf{w}_j \leftarrow$ random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}$
  $\mathbf{y} \leftarrow \mathbf{A}^K \mathbf{w}_j$
  if p is odd **then**
    $X_M \leftarrow X_M + \mathbf{y}^T \mathbf{A} \mathbf{y}/M$
  else
    $X_M \leftarrow X_M + \mathbf{y}^T \mathbf{y}/M$
  end if
end for
$X_M \leftarrow (X_M)^{1/p}$

easily to large matrices, as one does not need to form or store $\mathbf{A}$ explicitly to compute the matrix-vector products. If we compute $\mathbf{A}^p$ explicitly, the cost is $\mathcal{O}(\lfloor \log_2 p \rfloor n^3)$; see [11, Section 4.1]. On the other hand, if we compute the eigenvalues of $\mathbf{A}$ explicitly, the cost is also $\mathcal{O}(n^3)$. Lastly, the algorithm is general in the sense that any distribution for the random vectors $\mathbf{w}_j$ can be used as long as the $\mathbf{w}_j$ are independent and drawn from a distribution that has mean zero and the identity matrix as its covariance. However, in our analysis we assume that the entries of the $\mathbf{w}_j$ are independent standard normal random variables. If a different distribution is used, then the number of samples required for an (ε, δ) estimator for $|||\mathbf{A}|||_p$ will have to be changed appropriately.

We first collect a series of results for the estimator X_M^p in Proposition 3.2. Then, in the rest of this section, we appropriately adapt these results to the estimator X_M.

PROPOSITION 3.2. *The estimator X_M satisfies the following properties:*

1. *(Expectation):* $\mathbb{E}(X_M^p) = |||\mathbf{A}|||_p^p$.
2. *(Variance):* $\mathrm{Var}(X_M^p) = \dfrac{2\|\mathbf{A}^p\|_F^2}{M}$.
3. *(Almost Sure Convergence):* $\lim_{M\to\infty} X_M^p = |||\mathbf{A}|||_p^p$ *a.s.*
4. *((ε, δ) Estimator): If* $M \geq 8 \frac{\|\mathbf{A}^p\|_2}{\mathrm{tr}(\mathbf{A}^p)} \varepsilon^{-2} \ln\left(\frac{2}{\delta}\right)$, *then X_M^p is an (ε, δ) estimator for* $|||\mathbf{A}|||_p^p$.
5. *(Non-negative):* $X_M^p \geq 0$ *for all* M.

Proof. The proof collects well-known results from the literature. The expressions for the expectation and variance of X_M^p follow from [3, Lemma 5]. To see the third statement, note that since $X_M^p \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, by the strong law of large numbers [13] we have

$$\lim_{M\to\infty} X_M^p = \mathbb{E}(X_M^p) = |||\mathbf{A}|||_p^p \qquad \text{a.s.}$$

Regarding the fourth statement, Roosta-Khorasani and Ascher [25, Theorem 3] showed that if $\varepsilon > 0$, $\delta \in (0, 1)$, and the number of samples M satisfies the bound given in the statement of the theorem, then X_M^p is an (ε, δ) estimator of $|||\mathbf{A}|||_p^p$. Finally, since $\mathbf{A}$ is SPSD, then so is $\mathbf{A}^p$. Thus, $\mathbf{w}_j^T \mathbf{A}^p \mathbf{w}_j \geq 0$ for all j. Hence, $X_M^p \geq 0$ for all M. ∎

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Since the quantity of interest is $|||\mathbf{A}|||_p$, we have to analyze its estimator X_M. While Proposition 3.2 states several properties of X_M^p, a natural question is to what extent these properties apply to X_M. We investigate this in the rest of this section.

3.2. Expectation and variance of X_M. In this section we will provide a bound for the first two moments of X_M. Specifically we show that X_M is biased for all finite values of M, and we provide an upper bound for the variance of X_M.

PROPOSITION 3.3 (Expectation). *For all $M \geq 1$, $\mathbb{E}(X_M) \leq |||\mathbf{A}|||_p$. Moreover, if $p > 1$ then the inequality is strict.*

Proof. Let M be any natural number. Since $X_M = (X_M^p)^{1/p}$ and $f(x) = x^{1/p}$ is concave, by Jensen's inequality [13], we have

$$\mathbb{E}(X_M) = \mathbb{E}\left((X_M^p)^{1/p}\right) \leq (\mathbb{E}(X_M^p))^{1/p} = |||\mathbf{A}|||_p.$$

Furthermore, recall that Jensen's inequality yields a strict inequality as long as X_M^p is a non-degenerate (i.e., non-constant) random variable and $f(x)$ is nonlinear. Note that X_M^p will be non-degenerate as the $\mathbf{w}_j$ are drawn from a non-degenerate distribution and $f(x)$ is nonlinear for all $p > 1$. ∎

EXAMPLE 3.4. To illustrate that Jensen's inequality becomes strict for $p > 1$, consider the following example: let $a > 0$ and define

$$\mathbf{A} = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}.$$

Notice that, for all $p \geq 1$, $|||\mathbf{A}|||_p = a$. Let $\mathbf{w}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \end{bmatrix}$, $j = 1, \ldots, M$ be independent standard normal random vectors. Then,

$$\mathbb{E}(X_M) = \mathbb{E}\left(\left(\frac{1}{M}\sum_{j=1}^{M} a^p w_{1,j}^2\right)^{1/p}\right) = \frac{a}{M^{1/p}}\mathbb{E}\left(\left(\sum_{j=1}^{M} w_{1,j}^2\right)^{1/p}\right) = \frac{a}{M^{1/p}}\mathbb{E}(z^{1/p}),$$

where $z = \sum_{j=1}^{M} w_{1,j}^2$, which is a Chi-squared random variable with M degrees of freedom. The expected value of X_M can be computed analytically and is given by

$$\mathbb{E}(X_M) = \frac{2^{1/p}\Gamma\left(\frac{M}{2} + \frac{1}{p}\right)}{M^{1/p}\Gamma\left(\frac{M}{2}\right)}a. \tag{3.1}$$

Note that, when $p = 1$, $\mathbb{E}(X_M) = a$. Now consider $p > 1$. As Γ is a strict logarithmically convex function,

$$\begin{aligned}\Gamma\left(\frac{M}{2} + \frac{1}{p}\right) &= \Gamma\left(\frac{1}{p}\frac{M+2}{2} + \left(1 - \frac{1}{p}\right)\frac{M}{2}\right) < \Gamma\left(\frac{M+2}{2}\right)^{1/p}\Gamma\left(\frac{M}{2}\right)^{1-1/p} \\ &= \left(\frac{M}{2}\Gamma\left(\frac{M}{2}\right)\right)^{1/p}\Gamma\left(\frac{M}{2}\right)^{1-1/p} = \frac{M^{1/p}\Gamma\left(\frac{M}{2}\right)}{2^{1/p}}.\end{aligned}$$

Substituting this inequality into (3.1), we get $\mathbb{E}(X_M) < a = |||\mathbf{A}|||_p$, for all $p > 1$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Similarly to the first moment, we will derive an upper bound for the variance of X_M.

PROPOSITION 3.5 (Variance). *If* $\mathbf{A}$ *is nonzero, then the variance in* X_M *is finite and satisfies*

$$\mathrm{Var}\,(X_M) \leq \frac{2\|\mathbf{A}^p\|_F^2}{M|||\mathbf{A}|||_p^{2p-2}}.$$

Proof. Without loss of generality, assume that $p > 1$; otherwise the variance bound holds trivially. By [23, Theorem 1, Corollary 1], if Y is a non-negative random variable with positive mean and finite variance, then for $\alpha \in [0, 1]$

$$\mathrm{Var}(Y^\alpha) \leq \mathbb{E}|Y^\alpha - (\mathbb{E}Y)^\alpha|^2 \leq \frac{\mathrm{Var}(Y)}{(\mathbb{E}Y)^{2-2\alpha}}. \tag{3.2}$$

We let $Y = X_M^p$ and $\alpha = 1/p$. Note that Y is non-negative, $\mathbb{E}(Y) = |||\mathbf{A}|||_p^p > 0$ since $\mathbf{A}$ is nonzero, and from Table 2.1 $\mathrm{Var}(Y) = 2\|\mathbf{A}^p\|_F^2/M < \infty$. Therefore, (3.2) applies and

$$\mathrm{Var}(X_M) = \mathrm{Var}\left((X_M^p)^{1/p}\right) \leq \frac{2\|\mathbf{A}^p\|_F^2}{M|||\mathbf{A}|||_p^{p(2-2/p)}} = \frac{2\|\mathbf{A}^p\|_F^2}{M|||\mathbf{A}|||_p^{2p-2}}. \qquad \square$$

3.3. Convergence of estimators. In this section, we show that X_M converges almost surely in L^1 and in L^2 as $M \to \infty$.

PROPOSITION 3.6 (Almost sure convergence). *We have* $\lim_{M\to\infty} X_M = |||\mathbf{A}|||_p$ *a. s.*

Proof. Let $f(x) = x^{1/p}$. Note that f is continuous for $x \geq 0$. Since $\mathbf{A}$ is SPSD, from Proposition 3.2, X_M^p is non-negative and converges almost surely to $|||\mathbf{A}|||_p^p$. Thus, we can apply the Continuous Mapping Theorem [13, Theorem 17.5] to obtain

$$\lim_{M\to\infty} X_M = \lim_{M\to\infty} (X_M^p)^{1/p} = \left(|||\mathbf{A}|||_p^p\right)^{1/p} = |||\mathbf{A}|||_p \text{ a.s.} \qquad \square$$

Recall that, by Proposition 3.3, $\mathbb{E}(X_M) \leq |||\mathbf{A}|||_p$. Now, we form a bound on the bias in X_M. This will also be useful for establishing convergence in L^1 and in L^2.

PROPOSITION 3.7 (Bias). *The bias in* X_M *is bounded as*

$$|\mathbb{E}(X_M) - |||\mathbf{A}|||_p| \leq |||\mathbf{A}|||_p \left(\frac{2}{M}\right)^{1/2}.$$

Proof. Without any loss of generality, assume that $\mathbf{A}$ is nonzero so that $|||\mathbf{A}|||_p \neq 0$. Assume that $p > 1$, otherwise the bias is zero and the bound holds trivially. As $\mathbf{A}$ is SPSD, by Proposition 3.2, $X_M \geq 0$ and $\mathbb{E}(X_M)$ is the L^1 norm of X_M. Then, by the reverse triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\mathbb{E}(|X_M|) - |||\mathbf{A}|||_p\right| &\leq \mathbb{E}\left(\left|X_M - |||\mathbf{A}|||_p\right|\right) \leq \left(\mathbb{E}\left|X_M - |||\mathbf{A}|||_p\right|^2\right)^{1/2} \\ &\leq \left(\frac{2\|\mathbf{A}^p\|_F^2}{M|||\mathbf{A}|||_p^{2p-2}}\right)^{1/2}.\end{aligned} \tag{3.3}$$

The last inequality follows from (3.2) with $Y = X_M^p$ and $\alpha = 1/p$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Now, using the fact that $\mathbf{A}$ is SPSD

$$\|\|\mathbf{A}\|\|_p^{2p} = \Big(\sum_{j=1}^{n} \lambda_j^p\Big)^2 \geq \sum_{j=1}^{n} \lambda_j^{2p} = \|\mathbf{A}^p\|_F^2.$$

Therefore, we have

$$\frac{\|\mathbf{A}^p\|_F^2}{\|\|\mathbf{A}\|\|_p^{2p-2}} = \|\|\mathbf{A}\|\|_p^2 \frac{\|\mathbf{A}^p\|_F^2}{\|\|\mathbf{A}\|\|_p^{2p}} \leq \|\|\mathbf{A}\|\|_p^2.$$

Substitute this into (3.3) and simplify to obtain the desired inequality. ∎

Proposition 3.7 can be readily used to establish L^1 convergence. For a fixed p, X_M converges in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ to $\|\|\mathbf{A}\|\|_p$ since

$$\lim_{M\to\infty} \Big|\mathbb{E}(X_M) - \|\|\mathbf{A}\|\|_p\Big| \leq \lim_{M\to\infty} \|\|\mathbf{A}\|\|_p \left(\frac{2}{M}\right)^{1/2} = 0.$$

Similarly, convergence in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ follows from the proof of Proposition 3.7.

3.4. Number of samples for an (ε, δ)-estimator of $\|\|\mathbf{A}\|\|_p$. In this section we determine the minimum number of samples required to form an (ε, δ) estimator for $\|\|\mathbf{A}\|\|_p$.

THEOREM 3.8 ((ε, δ) estimator). *For all $\varepsilon > 0$ and $\delta \in (0, 1)$, the number of samples required for X_M to be an (ε, δ) estimator for $\|\|\mathbf{A}\|\|_p$ satisfies*

$$M \geq \frac{8\|\mathbf{A}^p\|_2}{\operatorname{tr}(\mathbf{A}^p)} \varepsilon^{-2} \ln\left(\frac{2}{\delta}\right). \tag{3.4}$$

Proof. Consider the measurable sets

$$\mathcal{D} = \Big\{\omega \in \Omega : \Big|X_M(\omega) - \|\|\mathbf{A}\|\|_p\Big| \leq \varepsilon \|\|\mathbf{A}\|\|_p\Big\} \qquad \text{and}$$
$$\mathcal{E} = \Big\{\omega \in \Omega : \Big|X_M^p(\omega) - \|\|\mathbf{A}\|\|_p^p\Big| \leq \varepsilon \|\|\mathbf{A}\|\|_p^p\Big\}.$$

Note that if $\mathbf{A}$ is the zero matrix, then both of these events are equivalent and have probability 1. Now consider the case when $\mathbf{A}$ is a non-zero SPSD matrix. Roosta-Khorasani and Ascher [25, Theorem 3] showed that for ε, δ as in the statement of the theorem, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ if (3.4) holds. Thus, it is sufficient to show that $\mathcal{E} \subset \mathcal{D}$. Therefore, let $\omega \in \mathcal{E}$. One can show, using the difference of powers formula, that $f(x) = x^{1/p}$ satisfies

$$\Big|(x+h)^{1/p} - x^{1/p}\Big| \leq \frac{|h|}{x^{1-1/p}}$$

for all $x > 0$ and $h \geq -x$. Since $\mathbf{A}$ is nonzero we let $x = \|\|\mathbf{A}\|\|_p^p$ and $h = X_M^p(\omega) - \|\|\mathbf{A}\|\|_p^p$. Then,

$$\Big|X_M(\omega) - \|\|\mathbf{A}\|\|_p\Big| \leq \frac{|X_M^p(\omega) - \|\|\mathbf{A}\|\|_p^p|}{\|\|\mathbf{A}\|\|_p^{p-1}} \leq \frac{\varepsilon\|\|\mathbf{A}\|\|_p^p}{\|\|\mathbf{A}\|\|_p^{p-1}} = \varepsilon\|\|\mathbf{A}\|\|_p.$$

Thus, $\omega \in \mathcal{D}$ and $\mathcal{E} \subset \mathcal{D}$. Therefore,

$$1 - \delta \leq \mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{D}). \qquad \square$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

To understand the dependence of p and the spectrum of $\mathbf{A}$ on the number of samples for an (ε, δ) estimator, we consider the ratio $\operatorname{tr}(\mathbf{A}^p)/\|\mathbf{A}^p\|_2$, which is known as the intrinsic dimension of $\mathbf{A}^p$ and denoted by $\operatorname{intdim}(\mathbf{A}^p)$. Since $1 \leq \operatorname{intdim}(\mathbf{A}^p) \leq \operatorname{rank}(\mathbf{A})$, the minimum number of samples are $8\varepsilon^{-2}\ln(2/\delta)\,/\operatorname{rank}(\mathbf{A})$. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\mathbf{A}$. If $\lambda_1 > \lambda_2$, then it is easy to see that $\operatorname{intdim}(\mathbf{A}^p) = 1 + \mathcal{O}(|\lambda_2/\lambda_1|^p)$. If the largest eigenvalue has multiplicity $k < n$, then $\operatorname{intdim}(\mathbf{A}^p) = k + \mathcal{O}(|\lambda_{k+1}/\lambda_1|^p)$. These results suggest that the number of samples for an (ε, δ) estimator would increase as p increases. Additionally, a large number of samples will be required if $\lambda_2/\lambda_1 \ll 1$ (or $\lambda_{k+1}/\lambda_1 \ll 1$). On the other hand, if we want the number of samples to be independent of p and the spectrum of $\mathbf{A}$, we can take the number of samples to be $M \geq 8\varepsilon^{-2}\ln(2/\delta)$. In contrast, in Section 5 we show that the variance of the estimator decreases when p increases. Note that these two observations regarding the impact of increasing p are not contradictory, because we are analyzing different statistical properties of the estimator.

4. The Chebyshev Monte Carlo estimator and its analysis. Recall that Algorithm 1 requires $\mathcal{O}\left(\left\lceil \frac{p}{2} \right\rceil M\right)$ matrix-vector products and can be computationally expensive for large p. Also, the Algorithm is not applicable to non-integer values of p. To address these issues, we use a Chebyshev polynomial approximation to approximate $\mathbf{A}^p$ by a lower degree Chebyshev polynomial $\psi_N(\mathbf{A})$. A similar approach was used in [9] in the context of estimating the trace of matrix functions of which the computation of the Schatten p-norms was a special case. In this section, we propose a new estimator for the Schatten p-norm and extend our analysis of convergence for the standard Monte Carlo estimator to the estimator using Chebyshev polynomial approximation. In contrast to the previous section, where it was sufficient for $\mathbf{A}$ to be SPSD, in this section, we require $\mathbf{A}$ to be symmetric positive definite (SPD).

4.1. The Chebyshev polynomial approximation method. Recall that the Nth degree Chebyshev polynomial approximation of a continuous function $g(x)$ with $x \in [\lambda_{\min}, \lambda_{\max}]$, with $0 < a \leq \lambda_{\min} \leq \lambda_{\max} \leq b$, is given by

$$g(x) \approx \psi_N(x) = c_0 + \sum_{j=1}^{N} c_j T_j\left(\frac{2}{\lambda_{\max} - \lambda_{\min}}x + \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right),$$

where $T_j(x) = \cos(j \arccos(x))$ is the jth Chebyshev polynomial, and the coefficient c_j is defined in (2.2). In this article, since we are computing the Schatten p-norm, the function of interest is $g(x) = x^{p/2} \approx \psi_N(x)$. Based on this polynomial approximation, we can construct the Chebyshev polynomial approximation to $\mathbf{A}^p \approx [\psi_N(\mathbf{A})]^2$. This ensures that the Chebyshev polynomial approximation to $\mathbf{A}^p$ is symmetric positive semidefinite. This is an important point since approximating $\mathbf{A}^p \approx \psi_{N'}(\mathbf{A})$ using Chebyshev polynomials does not automatically guarantee semidefiniteness, unless N' is taken to be sufficiently large; see [9, Lemma 2.4].

In Algorithm 2 we present an efficient algorithm for approximating $\|\mathbf{A}\|_p$ using the Chebyshev-Monte Carlo method, based on the discussion in [9, 28]. The method combines the Chebyshev polynomial approximation for $x^{p/2}$ in $[\lambda_{\min}, \lambda_{\max}]$ along with the three-term recurrence formula of the Chebyshev polynomials. For Algorithm 2 to be cost effective compared to Algorithm 1, the degree of the Chebyshev approximation should satisfy $N < \frac{p}{2}$. Furthermore, observe that Algorithm 1 requires at least a crude estimate $[a, b]$ of the range for the spectrum of $\mathbf{A}$. This can be accomplished using matrix free methods such as Krylov subspace methods [26]. In our implementation, we use the MATLAB command `eigs`.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 2: Constructing the Monte Carlo Estimator $Y_{M,N}$.

Input: a SPD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with eigenvalues in $[a, b]$, sample number M, Chebyshev polynomial degree N, and Schatten p-norm degree p
Initialize: $Y_{M,N} \leftarrow 0$
$c \leftarrow N+1$ vector of Chebyshev Coefficients for $x^{p/2}$ (see (2.2))
for $j = 1$ **to** M **do**
  $\mathbf{w}_j \leftarrow$ random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}$
  $\mathbf{y}_0^{(j)} \leftarrow \mathbf{w}_j$ and $\mathbf{y}_1^{(j)} \leftarrow \frac{2}{b-a}\mathbf{A}\mathbf{w}_j - \frac{b+a}{b-a}\mathbf{w}_j$
  $\mathbf{z} \leftarrow c_0\mathbf{y}_0^{(j)} + c_1\mathbf{y}_1^{(j)}$
  for $k = 2$ **to** N **do**
    $\mathbf{y}_2^{(j)} \leftarrow \frac{4}{b-a}\mathbf{A}\mathbf{y}_1^{(j)} - \frac{2(b+a)}{b-a}\mathbf{y}_1^{(j)} - \mathbf{y}_0^{(j)}$
    $\mathbf{z} \leftarrow \mathbf{z} + c_k\mathbf{y}_2^{(j)}$
    $\mathbf{y}_0^{(j)} \leftarrow \mathbf{y}_1^{(j)}$ and $\mathbf{y}_1^{(j)} \leftarrow \mathbf{y}_2^{(j)}$
  end for
  $Y_{M,N} \leftarrow Y_{M,N} + \mathbf{z}^T\mathbf{z}/M$
end for
$Y_{M,N} \leftarrow (Y_{M,N})^{1/p}$

With the notation introduced in Section 3, the computational cost of Algorithm 2 is $T_{\mathbf{A}}MN + \mathcal{O}(MNn)$ flops. To obtain the eigenvalue estimates, we use a few iterations (say K) of a Krylov subspace method which costs $KT_{\mathbf{A}} + \mathcal{O}(nK)$ flops.

4.2. Error analysis. Given a Chebyshev polynomial approximation $\psi_N(x)$ of $x^{p/2}$ over the spectrum of $\mathbf{A}$, we define the following estimator

$$Y_{M,N} = \left(\frac{1}{M}\sum_{j=1}^{M}\mathbf{w}_j^T\phi_N(\mathbf{A})\mathbf{w}_j\right)^{1/p}, \tag{4.1}$$

where $\phi_N(\mathbf{A}) = [\psi_N(\mathbf{A})]^2$. Note that, by construction, $\phi_N(\mathbf{A})$ is SPSD matrix. We now extend the analysis in Section 3.

PROPOSITION 4.1. *Let $Y_{M,N}$ be defined as in (4.1). For a fixed N, we have*

1. *(Non-negative): $Y_{M,N} \geq 0$ for all M;*
2. *(Almost Sure Convergence:) $\lim_{M\to\infty} Y_{M,N} = \left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p}$ a.s.;*
3. *(Expectation): $\mathbb{E}(Y_{M,N}) \leq \left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p}$, with a strict inequality for all $p > 1$;*
4. *(Variance): $\operatorname{Var}(Y_{M,N}) \leq \dfrac{2\|\phi_N(\mathbf{A})\|_F^2}{M\left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{2-2/p}}$.*

Proof. Note that, since $\phi_N(\mathbf{A})$ is SPSD, the non-negativity of $Y_{M,N}$ is immediate. Furthermore, applying the results of Propositions 3.6, 3.3, and 3.5 to $\phi_N(\mathbf{A})$, one can derive properties (2), (3), and (4) respectively. We omit the details. ∎

In the next result, we derive a bound for the smallest degree of the Chebyshev polynomial to ensure a user-defined relative error in the Schatten p-norm estimator.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

PROPOSITION 4.2. *Let* $0 < \varepsilon \leq 1$, $p \geq 1$, $q = p/2$, *and* $\kappa = \sqrt{\frac{b}{a}}$. *If the degree of the Chebyshev polynomial,* N, *satisfies*

$$N \geq \frac{\log\left(\frac{4}{\varepsilon}(\kappa^2+1)^q(\kappa-1)\left(\kappa^p + \sqrt{\frac{\varepsilon}{2}+\kappa^{2p}}\right)\right)}{\log\left(\frac{\kappa+1}{\kappa-1}\right)}, \tag{4.2}$$

then $|\mathrm{tr}(\phi_N(\mathbf{A})) - |||\mathbf{A}|||_p^p| \leq \frac{\varepsilon}{2}|||\mathbf{A}|||_p^p$.

Proof. The proof follows a similar strategy to [9, Theorem 3.1] and has several steps.

Error in terms of Chebyshev polynomials. The absolute error in $|||\mathbf{A}|||_p$ can be bounded using the approximation properties of the Chebyshev polynomials.

$$\begin{aligned}\left||||\mathbf{A}|||_p^p - \mathrm{tr}(\phi_N(\mathbf{A}))\right| &= \left|\sum_{j=1}^n \lambda_j^p - \sum_{j=1}^n \phi_N(\lambda_j)\right| \leq \sum_{j=1}^n \left|\lambda_j^p - \phi_N(\lambda_j)\right| \\ &\leq \max_{1\leq j\leq n} n\left|\lambda_j^p - \phi_N(\lambda_j)\right| \leq n \max_{x\in[a,b]} \left|x^p - \phi_N(x)\right|.\end{aligned}$$

Since $\phi_N(x) = \psi_N^2(x)$, by repeated use of the triangle inequality,

$$\begin{aligned}\left|x^p - \psi_N^2(x)\right| &= \left|x^{2q} + x^q\psi_N(x) - x^q\psi_N(x) + \psi_N^2(x)\right| \\ &\leq |x^q|\,|x^q - \psi_N(x)| + |\psi_N(x)|\,|x^q - \psi_N(x)| \\ &\leq 2\,|x^q|\,|x^q - \psi_N(x)| + |x^q - \psi_N(x)|^2.\end{aligned}$$

In the second step, we wrote $|\psi_N(x)| = |\psi_N(x) - x^q + x^q|$ and applied the triangle inequality.

Chebyshev polynomial approximation. Let E be the ellipse in the complex plane with foci at ± 1 and passing through the point $\left(\frac{b+a}{b-a}, 0\right)$. The sum of major and minor semi-axes, denoted by $\rho > 1$, can be computed as

$$\rho = \frac{b+a}{b-a} + \sqrt{\left(\frac{b+a}{b-a}\right)^2 - 1} = \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}} = \frac{\kappa+1}{\kappa-1},$$

where $\kappa = \sqrt{\frac{b}{a}}$ was defined in the statement of the proposition.

From [9, Corollary 2.2], since $g(x) = x^q$ is analytic in the interior of the ellipse E, we have

$$\max_{x\in[a,b]} |x^q - \psi_N(x)| \leq \frac{4U}{(\rho-1)\rho^N},$$

where the scalar U satisfies

$$U = \max_{z\in E}\left|g\left(\frac{b-a}{2}z + \frac{b+a}{2}\right)\right| = (b+a)^q.$$

Converting absolute error into relative error. Therefore, by the first two steps,

$$\max_{x\in[a,b]} |x^p - \phi_N(x)| \leq \left(2b^q + \frac{4U}{(\rho-1)\rho^N}\right)\frac{4U}{(\rho-1)\rho^N}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We want to find N such that $\max_{x\in[a,b]} |x^p - \phi_N(x)| \le \varepsilon a^p/2$. If such an N can be found, then

$$|\mathrm{tr}(\phi_N(\mathbf{A})) - |||\mathbf{A}|||_p^p| \le n \max_{x\in[a,b]} |x^p - \phi_N(x)| \le \frac{n\varepsilon a^p}{2} \le \frac{\varepsilon}{2}|||\mathbf{A}|||_p^p,$$

as desired. We now show that such an N can be found.

Solving for N. To this end, consider

$$\begin{aligned}\frac{\varepsilon a^p}{2} &\ge \left(2b^q + \frac{4U}{(\rho-1)\rho^N}\right)\frac{4U}{(\rho-1)\rho^N}\\ &= \left(\frac{4U}{(\rho-1)\rho^N}\right)^2 + 2b^q\frac{4U}{(\rho-1)\rho^N} + b^{2q} - b^{2q}\\ &= \left(\frac{4U}{(\rho-1)\rho^N} + b^q\right)^2 - b^{2q}.\end{aligned}$$

Simplifying this expression, we get

$$\rho^N \ge \frac{4U}{(\rho-1)\left(\sqrt{\frac{\varepsilon a^p}{2} + b^p} - b^q\right)}.$$

We have the elementary identity

$$\frac{1}{\sqrt{x+d}-\sqrt{x}}\cdot\frac{\sqrt{x+d}+\sqrt{x}}{\sqrt{x+d}+\sqrt{x}} = \frac{\sqrt{x+d}+\sqrt{x}}{d},$$

for all $x, d \ge 0$. Applying this inequality with $x = b^q$ and $d = \varepsilon a^p/2$, we obtain

$$\rho^N \ge \frac{4U}{(\rho-1)\left(\sqrt{\frac{\varepsilon a^p}{2} + b^p} - b^q\right)} = \frac{8U\left(b^q + \sqrt{\frac{\varepsilon a^p}{2} + b^p}\right)}{(\rho-1)\left(\varepsilon a^p\right)}.$$

Since $\rho > 1$, N is bounded from below as

$$N \ge \frac{1}{\log(\rho)}\log\left(\frac{8U\left(b^q + \sqrt{\frac{\varepsilon a^p}{2} + b^p}\right)}{(\rho-1)(\varepsilon a^p)}\right). \tag{4.3}$$

Substitute the expressions for U and ρ into (4.3) and simplify to get (4.2). ∎

In Figure 4.1, the bound in (4.2) is plotted for various values of p, $\kappa \in [1, 2]$, and $\varepsilon = 0.1$. Here, a dot is placed when the value of N is larger than $q = p/2$, suggesting that the bound is pessimistic for condition numbers larger than 2.

From Proposition 4.2, we can see that the degree of the polynomial N should be $\mathcal{O}(p)$. A natural question to ask is if there is a better polynomial approximation. Newman and Rivlin [22] showed that we can approximate a monomial ξ^p (with an integer value of p) by the best polynomial approximation $r_N^*(x)$ so that

$$\max_{\xi\in[0,1]} |\xi^p - r_N^*(\xi)| \le E_{N,p} = \frac{1}{2^{p-1}} \sum_{j>(p+N)/2} \binom{p}{j}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

As suggested in [22], the term $E_{N,p}$ has a probabilistic interpretation in terms of an experiment in which we toss p fair coins. Let B_i be independent Bernoulli$(1/2)$ random variables that model the coin tosses. The sum $B = \sum_{i=1}^{p} B_i$ has a binomial distribution, Binomial$(p, 1/2)$. Therefore, we can write $E_{N,p} = 2\mathbb{P}((N+p)/2 < B \leq p)$. Since $\mathbb{E}[B] = p/2$, using Hoeffding's inequality [30, Theorem 2.2.6], we can bound

$$\mathbb{P}((N+p)/2 < B \leq p) \leq \mathbb{P}(B \geq (N+p)/2) \leq \exp\left(-\frac{N^2}{2p}\right),$$

so that $E_{N,p} \leq 2\exp\left(-N^2/2p\right)$.

To estimate the trace of $\mathbf{A}^p$, we need to approximate x^p over $[a, b]$. We can do this by using the change of variables $x = \xi b$, so that the polynomial approximation takes the form $b^p r_N^*(x/b)$. The error in the resulting approximation is given by

$$\max_{x\in[a,b]} |x^p - b^p r_N^*(x/b)| \leq \max_{x\in[0,b]} |x^p - b^p r_N^*(x/b)| = \max_{\xi\in[0,1]} b^p |\xi^p - r_N^*(\xi)| \leq b^p E_{N,p}.$$

Since we require a relative error of $\varepsilon a^p/2$, we need $2b^p \exp(-N^2/2p) \leq \varepsilon a^p/2$. Solving for N, we get the bound $N \geq \sqrt{2p \log\left(\frac{4\kappa^{2p}}{\epsilon}\right)}$. Although Newman and Rivlin's analysis predicts that a polynomial of degree $N = \mathcal{O}(\sqrt{p})$ is sufficient for a small absolute error over $[0, 1]$, the above analysis suggests that a polynomial of degree $N = \mathcal{O}(p)$ is necessary for a small relative error over $[a, b]$. Note that this is asymptotically the same degree as that obtained in Proposition 4.2. Beyond polynomial approximations, one can use a low-degree rational approximation to accurately approximate the monomial x^p; see [21] for additional details.

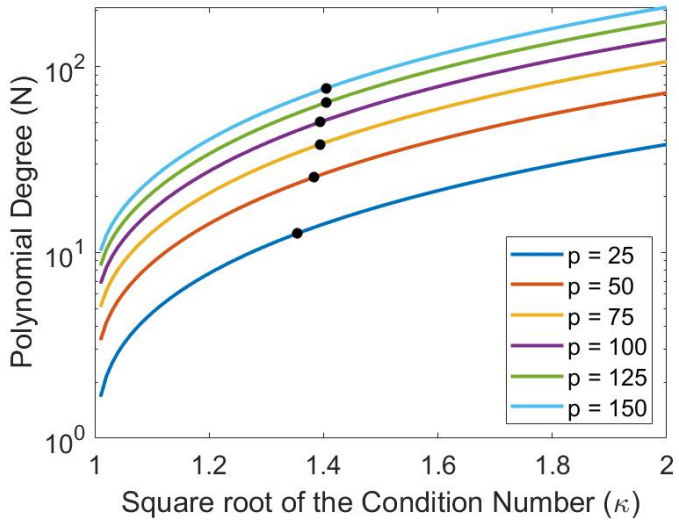

FIG. 4.1. *Plotting the values of (4.2) for $p = 25, 50, 75, 100, 125$, and 150, with $\kappa \in [1, 2]$ and $\varepsilon = 0.1$. This corresponds to approximating $\mathbf{A}^q$, where $p = 2q$ and $\mathbf{A}$ has a condition number between 1 and 4. A black dot has been placed to mark the first instance that $N > q$.*

Although Proposition 4.2 guarantees that $\mathrm{tr}(\phi_N(\mathbf{A}))$ has a small relative error, it is computationally challenging to implement, since it involves constructing $\phi_N(\mathbf{A})$ explicitly. Therefore, we approximate its trace using the Monte Carlo estimator $Y_{M,N}$. Using this

ETNA
Kent State University and
Johann Radon Institute (RICAM)

proposition, we derive bounds for the absolute error in the L^1 sense (i.e., the bias) and the number of samples required for an (ε, δ)-estimator for $|||\mathbf{A}|||_p$.

THEOREM 4.3. *Consider the same setup as in Proposition 4.2. Let $Y_{M,N}$ be defined as in (4.1) and let N satisfy (4.2). Then*

1. *(L^1 bound):* $\left|\mathbb{E}\left(Y_{M,N}\right) - |||\mathbf{A}|||_p\right| \leq (1+\frac{\varepsilon}{2})|||\mathbf{A}|||_p \left(\frac{2}{M}\right)^{1/2} + \frac{\varepsilon}{2}|||\mathbf{A}|||_p$;
2. *((ε,δ) estimator): if $M \geq 72\varepsilon^{-2}\ln\left(\frac{2}{\delta}\right)$, then $Y_{M,N}$ is an (ε,δ) estimator for $|||\mathbf{A}|||_p$.*

Proof. First consider the L^1 bound on $Y_{M,N}$. From Proposition 4.2, we have

$$\left(1-\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p^p \leq \operatorname{tr}(\phi_N(\mathbf{A})) \leq \left(1+\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p^p.$$

From the simple identity $(1-x)^p \leq (1-x) \leq (1+x) \leq (1+x)^p$, for $0 \leq x \leq 1$ and $p \geq 1$, we get

$$\left(1-\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p \leq (\operatorname{tr}(\phi_N(\mathbf{A}))^{1/p} \leq \left(1+\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p,$$

or $|(\operatorname{tr}(\phi_N(\mathbf{A}))^{1/p} - |||\mathbf{A}|||_p| \leq \frac{\varepsilon}{2}|||\mathbf{A}|||_p$.

By the triangle inequality and by applying Proposition 3.7 to $\phi_N(\mathbf{A})$, we find

$$\begin{aligned}\left|\mathbb{E}\left(Y_{M,N}\right) - |||\mathbf{A}|||_p\right| &\leq \left|\mathbb{E}\left(Y_{M,N}\right) - \left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p}\right| + \left|\left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p} - |||\mathbf{A}|||_p\right| \\ &\leq \left(\frac{2}{M}\right)^{1/2}\left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p} + \left|\left(\operatorname{tr}\left(\phi_N\left(\mathbf{A}\right)\right)\right)^{1/p} - |||\mathbf{A}|||_p\right| \\ &\leq \left(1+\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p\left(\frac{2}{M}\right)^{1/2} + \frac{\varepsilon}{2}|||\mathbf{A}|||_p.\end{aligned}$$

If $M \geq 72\varepsilon^2 \ln\left(\frac{2}{\delta}\right)$ then, by [25, Theorem 3],

$$\Pr\left(\left|Y_{M,N}^p - \operatorname{tr}(\phi_N(\mathbf{A}))\right| \leq \frac{\varepsilon}{3}\operatorname{tr}(\phi_N(\mathbf{A}))\right) \geq 1-\delta.$$

Furthermore, as $\varepsilon \in (0,1)$, we have $\operatorname{tr}(\phi_N(\mathbf{A})) \leq \left(1+\frac{\varepsilon}{2}\right)|||\mathbf{A}|||_p^p \leq \frac{3}{2}|||\mathbf{A}|||_p^p$. Then, with probability at least $1-\delta$,

$$\left|Y_{M,N}^p - \operatorname{tr}(\phi_N(\mathbf{A}))\right| \leq \frac{\varepsilon}{2}|||\mathbf{A}|||_p^p.$$

Using the triangle inequality, with the same probability

$$\begin{aligned}\left|Y_{M,N}^p - |||\mathbf{A}|||_p^p\right| &\leq \left|Y_{M,N}^p - \operatorname{tr}(\phi_N(\mathbf{A}))\right| + |\operatorname{tr}(\phi_N(\mathbf{A})) - |||\mathbf{A}|||_p^p| \\ &\leq \frac{\varepsilon}{2}|||\mathbf{A}|||_p^p + \frac{\varepsilon}{2}|||\mathbf{A}|||_p^p = \varepsilon|||\mathbf{A}|||_p^p.\end{aligned}$$

Now consider the measurable sets

$$\begin{aligned}\mathcal{E} &= \left\{\omega \in \Omega \middle| |Y_{M,N}^p(\omega) - |||\mathbf{A}|||_p^p| \leq \varepsilon|||\mathbf{A}|||_p^p\right\}, \\ \mathcal{D} &= \left\{\omega \in \Omega \middle| |Y_{M,N}(\omega) - |||\mathbf{A}|||_p| \leq \varepsilon|||\mathbf{A}|||_p\right\}.\end{aligned}$$

Using a similar argument as in Proposition 3.8 we can show $1-\delta \leq \Pr(\mathcal{E}) \leq \Pr(\mathcal{D})$. Thus $Y_{M,N}$ is an (ε,δ) estimator for $|||A|||_p$. ∎

We point out that the number of samples required for an (ε,δ) estimator for $|||A|||_p$ is independent of the degree p, provided N is sufficiently large.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

5. Numerical experiments. In this section, we will present numerical experiments demonstrating the performance of the estimators and the convergence analysis for several test matrices. The first set of test matrices are synthetically generated, the second set is extracted from the SuiteSparse collection, and the final test matrix arises in an application to Optimal Experimental Design (OED). In our numerical experiments, we will be using matrices of relatively small size for which we can estimate errors by computing the Schatten p-norms exactly using the eigendecomposition of $\mathbf{A}$. However, the methods in this article can be scaled up to arbitrarily large matrices that do not need to be stored.

5.1. Choice of matrices.

5.1.1. Synthetic test matrices. Here we construct several 100×100 test matrices with prespecified sets of eigenvalues. The test matrices are of the form $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$, where $\mathbf{D}$ is a diagonal matrix with the eigenvalues on its diagonal and $\mathbf{Q}$ is an orthogonal matrix. The orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{100\times 100}$ is constructed by first generating a standard Gaussian random matrix and then computing its QR factorization.

1. **Linear Decay:** The first test matrix $\mathbf{A}_{\text{linear}} = \mathbf{Q}\mathbf{D}_{\text{lin}}\mathbf{Q}^T \in \mathbb{R}^{100\times 100}$ has eigenvalues
$$\mathbf{D}_{\text{lin}} = \text{diag}(6, 7, \dots, 105).$$
2. **Clustered:** The second test matrix takes the form $\mathbf{A}_{\text{clustered}} = \mathbf{Q}\mathbf{D}_{\text{clus}}\mathbf{Q}^T$, where
$$\mathbf{D}_{\text{clus}} = \text{diag}(\underbrace{100, \dots 100}_{20}, \underbrace{1, \dots, 1}_{80}).$$
3. **Quadratic Decay:** The test matrix takes the form $\mathbf{A}_{\text{quad}} = \mathbf{Q}\mathbf{D}_{\text{quad}}\mathbf{Q}^T$, with
$$\mathbf{D}_{\text{quad}} = \text{diag}(1, 2^{-2}, \dots, 100^{-2}).$$
4. **Exponential Decay:** The test matrix takes the form $\mathbf{A}_{\text{exp}} = \mathbf{Q}\mathbf{D}_{\text{exp}}\mathbf{Q}^T$, with
$$\mathbf{A}_{\text{exp}} = \text{diag}(0.9^1, \dots, 0.9^{100}).$$

The test matrices simulate different scenarios of eigenvalue distributions for an SPSD matrix. We have plotted the eigenvalue distributions in Figure 5.1 to illustrate these distributions.

5.1.2. Test matrices from the SuiteSparse collection. In addition to the test matrices described above, we also consider two relatively large matrices from the SuiteSparse matrix collection [15]. In particular, we consider

1. Trefethen_700 matrix, a 700×700 SPD matrix from an application in combinatorics with a condition number of approximately 4.71×10^3;
2. mhd4800b matrix, a 4800×4800 SPD matrix from an application in electrohydrodynamics with a condition number of approximately 8.16×10^{13}.

5.1.3. Application to optimal experimental design. For the last test matrix, we return to our motivating problem from Optimal Experimental Design (OED). Our goal is to compute the Schatten p-norm of the posterior covariance operator, arising from a Bayesian linear inverse problem.

We consider the inverse problem of estimating the initial state in the following 1D heat equation:

$$\begin{cases} u_t = k u_{xx}, & x \in [0,1],\ t \in (0, t_f], \\ u(x,0) = \phi(x), & x \in [0,1], \\ u(0,t) = u(1,t) = 0, & t \in (0, t_f]. \end{cases} \tag{5.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

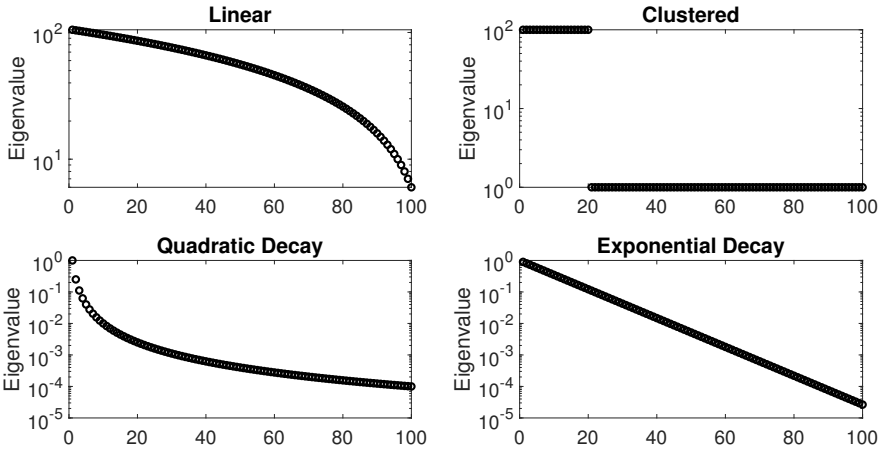

FIG. 5.1. *Semi-log plots of the eigenvalue distributions for the test matrices, each of size* 100×100.

Here $\phi(x)$ is an unknown initial state, which we seek to estimate using sensor measurements of the temperature at a few observation times. In (5.1), k is the diffusion coefficient, which we choose to be $k = 2 \times 10^{-4}$.

After discretization, the goal is to estimate the discretized parameter $\boldsymbol{\phi}$ from

$$\mathbf{F}\boldsymbol{\phi} + \boldsymbol{\eta} = \mathbf{d}. \tag{5.2}$$

Here $\mathbf{F}$ is the parameter-to-observable map, which maps the (discretized) initial state $\boldsymbol{\phi}$ to spatio-temporal observations, $\boldsymbol{\phi}$ is the discretized inversion parameter (the initial state), $\boldsymbol{\eta}$ is a random variable modeling measurement noise, and $\mathbf{d}$ is measurement data.

We discretize the problem (5.1) using finite-differences in space and implicit Euler in time. Thus, an application of $\mathbf{F}$ to a vector requires solving (5.1), and extracting solution values at the measurement points and at measurement times. Here we take measurements at 17 equally spaced sensors in the spatial domain $[0, 1]$ and at observation times $\{0.25, 0.5, 0.75, 1\}$. We assume that $\boldsymbol{\eta}$ in (5.2) is multivariate Gaussian distributed with mean zero and covariance given by $\sigma\mathbf{I}$ with $\sigma = 0.002$ (corresponding to 0.1% noise).

We consider a Bayesian formulation [27] of this inverse problem. Assuming a Gaussian prior, since the forward operator is linear, we also have a Gaussian posterior. One possible measure of the posterior uncertainty is given by the Schatten p-norm of the posterior covariance operator

$$\boldsymbol{\Gamma}_{\text{post}} = (\sigma^{-2}\mathbf{F}^T\mathbf{F} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})^{-1}.$$

Here $\boldsymbol{\Gamma}_{\text{prior}}$ is the covariance operator of the Gaussian prior. For this example, we choose $\boldsymbol{\Gamma}_{\text{prior}} = (\gamma\mathbf{K})^{-1}$ where $\gamma = 10^{-4}$ is a regularization parameter and $\mathbf{K}$ is the discretized Laplacian operator, with homogeneous Dirichlet boundary conditions.

As mentioned in the introduction, the Schatten p-norm of $\boldsymbol{\Gamma}_{\text{post}}$ is related to the P-optimal experimental design criterion and provides a measure of uncertainty. Alternative approaches involve computing the trace and determinant of the posterior covariance matrix, which are related to the A- and D-optimal design criteria, respectively. Since repeated evaluations of the P-optimal criterion are needed to compute a P-optimal design, we focus on addressing the computational cost of computing $|||\boldsymbol{\Gamma}_{\text{post}}|||_p$.

In our application, the matrix $\boldsymbol{\Gamma}_{\text{post}}$ need not be computed explicitly and is handled using matrix-free techniques. This is a common approach in large-scale Bayesian inverse problems

ETNA
Kent State University and
Johann Radon Institute (RICAM)

in general; see, e.g., [6, 2, 10]. To see this, first note that the matrix $\mathbf{F}$ is not available, and a matrix-vector product (denoted as matvec) using $\mathbf{F}$ involves one time-dependent PDE solve, with cost $T_{\mathbf{F}}$; similarly the action of $\mathbf{F}^T$ requires one adjoint time-dependent PDE solve [1, 7]. Applying $\mathbf{\Gamma}_{\text{prior}}^{-1}$ to a vector involves a sparse matvec, so is relatively cheap; denote this cost as T_{prior}. Therefore, forming $\mathbf{H} = \sigma^{-2}\mathbf{F}^T\mathbf{F} + \mathbf{\Gamma}_{\text{prior}}^{-1}$ involves computing $2n$ time-dependent PDE solves where n is the size of $\mathbf{H}$. Computing $\mathbf{\Gamma}_{\text{post}} = \mathbf{H}^{-1}$ and its Schatten p-norm (using the eigendecomposition of $\mathbf{\Gamma}_{\text{post}}$) both require $\mathcal{O}(n^3)$ flops. Therefore, the total cost of this approach is $n(2T_{\mathbf{F}} + T_{\text{prior}}) + \mathcal{O}(n^3)$. When the domain is discretized using fine-scale grids, the number of degrees of freedom n can be as high as 10^6. For these problem sizes, forming $\mathbf{\Gamma}_{\text{post}}$ is neither feasible nor advisable. It is important to note that even forming $\mathbf{H}$ is often infeasible in such large-scale problems, due to the associated computational expense—$2n$ PDE solves—and storage requirements.

Following [6], we use an iterative solver to compute the action of $\mathbf{\Gamma}_{\text{post}}$. Consider $\mathbf{\Gamma}_{\text{post}}\mathbf{w}$, which can be computed as the solution to $\mathbf{H}\mathbf{x} = \mathbf{w}$. Assuming $\mathbf{\Gamma}_{\text{prior}}$ is positive definite, $\mathbf{H}$ is positive definite; thus, we can use a preconditioned Conjugate Gradient (CG) solver [8]. Each application of $\mathbf{H}$ to a vector requires one forward and adjoint solve and a multiplication times a sparse matrix. If we use the Cholesky factor of $\mathbf{\Gamma}_{\text{prior}}^{-1}$ as the preconditioner, numerical evidence suggests that the number of required CG iterations, say m, is independent of the size of the problem. Therefore, the cost of applying $\mathbf{\Gamma}_{\text{post}}$ to a vector involves $m(2T_{\mathbf{F}} + T_{\text{prior}})$. Here we have assumed that the cost of forming and applying the preconditioner is negligible compared to the other costs. We can then use either Algorithm 1 or 2, which are both matrix-free, to compute the Schatten p-norm of $\mathbf{\Gamma}_{\text{post}}$.

In the numerical experiments in Sections 5.2 and 5.3, we examine the effectiveness of our proposed estimators for computing $\|\|\mathbf{\Gamma}_{\text{post}}\|\|_p$. We take the number of degrees of freedom $n = 254$; note that we choose a relatively small problem size to study the accuracy of the method, but the estimators that we propose are scalable to larger problem sizes.

5.2. Monte Carlo estimator results. For each test matrix described in the previous subsection, we apply Algorithm 1 and compute the error statistics as a function of the sample size M. For each fixed sample size M, we generated 500 different realizations of X_M using Algorithm 1 and then computed the average, the 97.5th quantile, and the 2.5th quantile of the relative errors. Note that the interval between the 2.5th and 97.5th quantiles is the same as the central 95th confidence interval for the error.

Synthetic Test Matrices. In Figures 5.2 and 5.3, we display the mean and central 95th confidence interval for each of the four synthetic test matrices, using a value of $p = 5$ and $p = 120$, respectively. We call the shaded region within the 95th confidence interval the error envelope for X_M. First, we observe that the error statistics for the Monte Carlo estimator X_M does not depend significantly on the eigenvalue distributions. Second, we observe that for $p = 120$ the average relative error is lower, compared to the average relative error for $p = 5$, for all four eigenvalue distributions; see Figure 5.4. Furthermore, the error envelopes appear tighter, suggesting smaller empirical variance with increasing p. A partial explanation of these observations is as follows:

1. Bias: from Proposition 3.7, we see for the bias

$$|\mathbb{E}(X_M) - \|\|\mathbf{A}\|\|_p| \leq \|\|\mathbf{A}\|\|_p (2/M)^{1/2}.$$

For a fixed sample size, as $p \to \infty$ the bias is decreasing.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

2. Variance: combining the statement of Proposition 3.5 with the proof of Proposition 3.7, we have

$$\mathrm{Var}(X_M) \leq \frac{2\|\mathbf{A}^p\|_F^2}{M|||\mathbf{A}|||_p^{2p-2}} \leq \frac{2|||\mathbf{A}|||_p^2}{M}.$$

As $p \to \infty$ the upper bound $\frac{2|||\mathbf{A}|||_p^2}{M}$ decreases, suggesting that the estimators are more "concentrated" about their mean.

A more precise statement can be derived by using Chebyshev's inequality, from which we obtain for $\zeta > 0$

$$\mathbb{P}\left\{|Z_M - \mathbb{E}[Z_M]| \geq \zeta\sqrt{\frac{2}{M}}|||\mathbf{A}|||_p\right\} \leq \frac{1}{\zeta^2}.$$

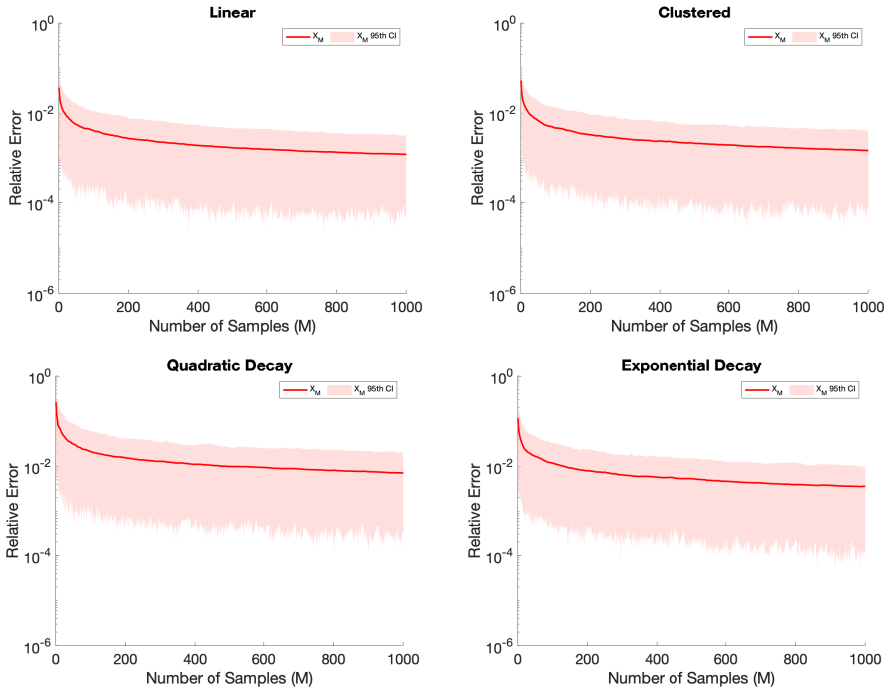

FIG. 5.2. *Relative error in the Monte Carlo Estimator X_M for the 100 × 100 synthetic test matrices with $p = 5$. The error statistics were generated based on 500 realizations, each for a fixed sample size M.*

SuiteSparse Matrices. We consider the two test matrices from the SuiteSparse matrix collection. In Figure 5.5 we display the error envelope when $p = 5$ and in Figure 5.6 we plot the error envelope when $p = 80$. Once again, as p increases the mean relative error decreases (see Figure 5.7) and the error envelope appears to tighten. This further provides evidence that the relative error does not show strong dependency on the eigenvalue distribution.

Posterior Covariance Matrix. In Figure 5.8 and 5.9, we display the relative error for the posterior covariance matrix generated using the setup in Section 5.1.3. The main conclusions from this plot are essentially the same as from the other two sets of test matrices.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

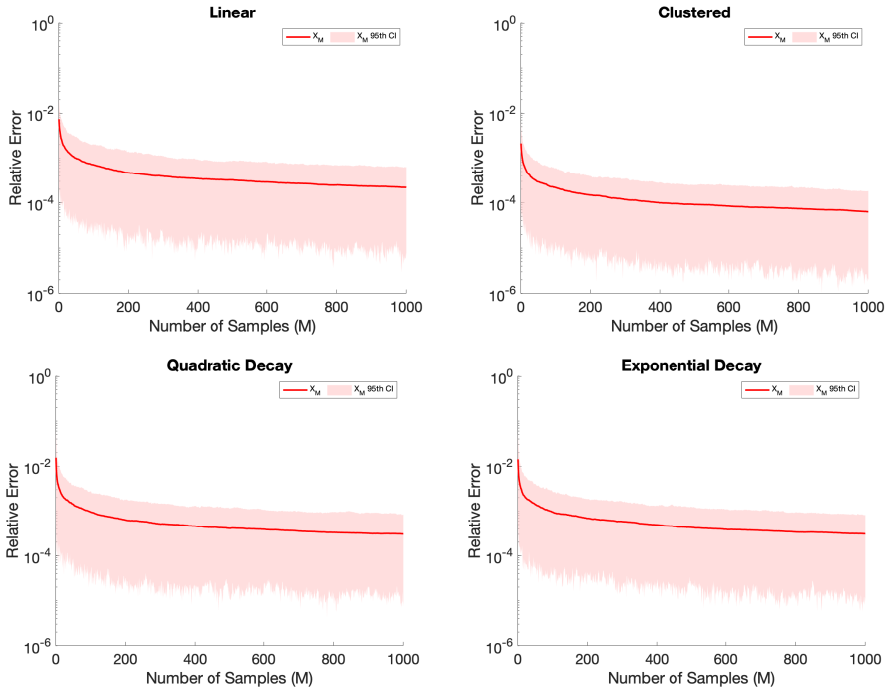

FIG. 5.3. *Relative error in X_M for the 100×100 test matrices with $p = 120$. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

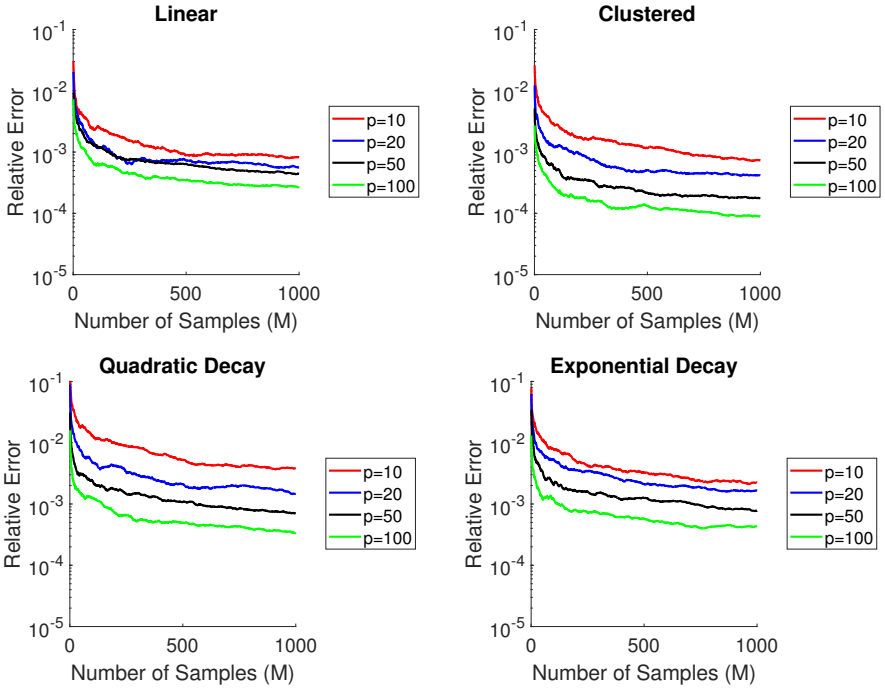

FIG. 5.4. *Comparing the mean relative error in X_M for the 100×100 test matrices for $p = 10, 20, 50, 100$. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

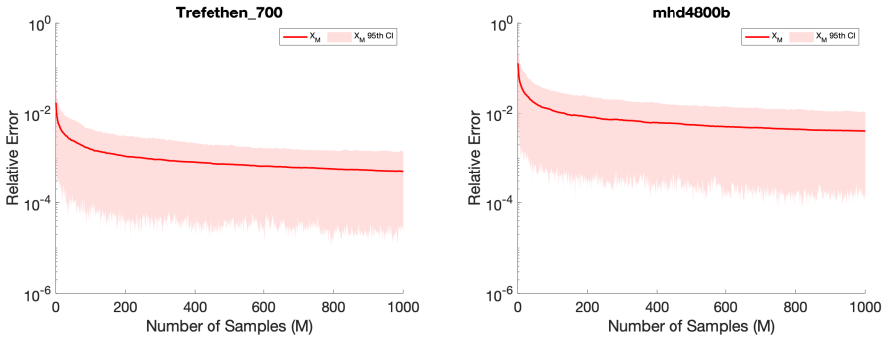

FIG. 5.5. *Relative error of X_M, with $p = 5$, for each of the matrices from the SuiteSparse matrix collection. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

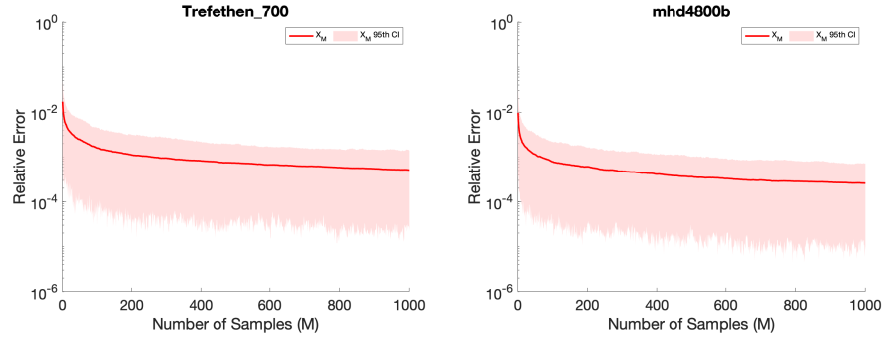

FIG. 5.6. *Relative error of X_M, with $p = 80$, for each of the matrices from the SuiteSparse matrix collection. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

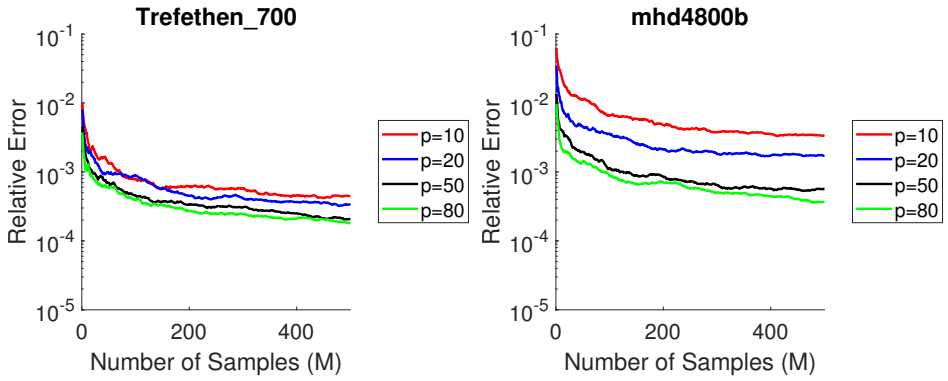

FIG. 5.7. *Comparing the mean relative error in X_M for the SuiteSparse matrices for $p = 10, 20, 50, 80$. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

5.3. Chebyshev Monte Carlo estimator. We recall that the bound on N derived in Proposition 4.2 was pessimistic. In this section, we present numerical evidence that a relatively small N is sufficient for accurately estimating $|||\mathbf{A}|||_p$. For the synthetic test matrices and the posterior covariance matrix, we choose $p = 120$ and $N = 5, 10, 20, 60$, whereas for the test matrices from the SuiteSparse collection, we use $p = 80$ and $N = 5, 10, 20, 40$. For all the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

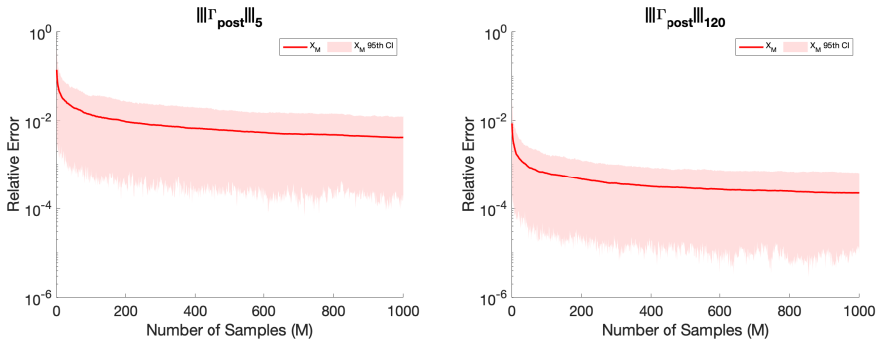

FIG. 5.8. *Relative error of X_M for a 254×254 matrix Γ_{post} with $p = 5$ and $p = 120$. Similar to the test matrices, we ran 500 different simulations for a fixed sample size M and we computed the mean error and the 97.5th quantile and 2.5th quantile in the errors.*

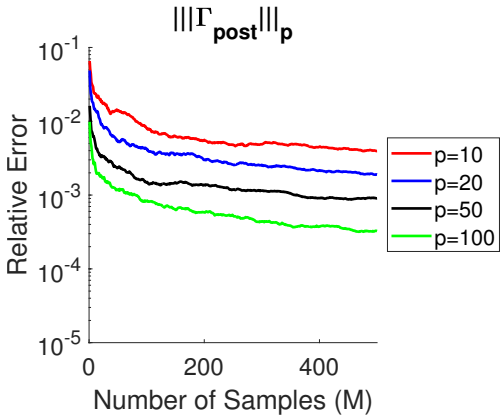

FIG. 5.9. *Comparing the mean relative error in X_M for the OED posterior covariance matrix when $p = 10, 20, 50, 80$. The error statistics were generated based on 500 realizations each for a fixed sample size M.*

test matrices, we compute the error using Algorithm 2. Similarly to the "standard" Monte Carlo method in Section 5.2, we compute the average error by using 500 realizations for a fixed sample size and value of N.

Synthetic Test Matrices. In Figure 5.10, we display the mean relative error in $Y_{M,N}$ for $N = 5, 10, 20, 30$, for each of the synthetic test matrices, when $p = 120$. Notice that with $N = 20$ and $N = 30$ the average relative error behaves similarly as in Figures 5.3. Next, we observe that the estimator $Y_{M,N}$ is accurate for all the test matrices here. However, if N is small, i.e., 5–10, then we see that increasing the number of samples does not decrease the average relative error due to the bias, i.e., the error due to Chebyshev polynomial approximation. On the other hand, if N is sufficiently large, we see that increasing the sample size M can reduce the average relative error.

SuiteSparse Matrices. In Figure 5.11 we display the mean relative error when using Chebyshev approximation to accelerate the computation of $|||\mathbf{A}|||_p$ for the the Trefethen_700 and the mhd4800b matrices when $p = 80$. Here we used $N = 5, 10, 20, 30$, and notice similar trends as in the numerical experiments using synthetic matrices. For example, once again we observe that if N is too small, then increasing the sample size does not reduce the error in the Chebyshev approximation. Also we find that, for both test matrices, $N = 20$ is sufficient for accurately approximating the Schatten p-norm.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

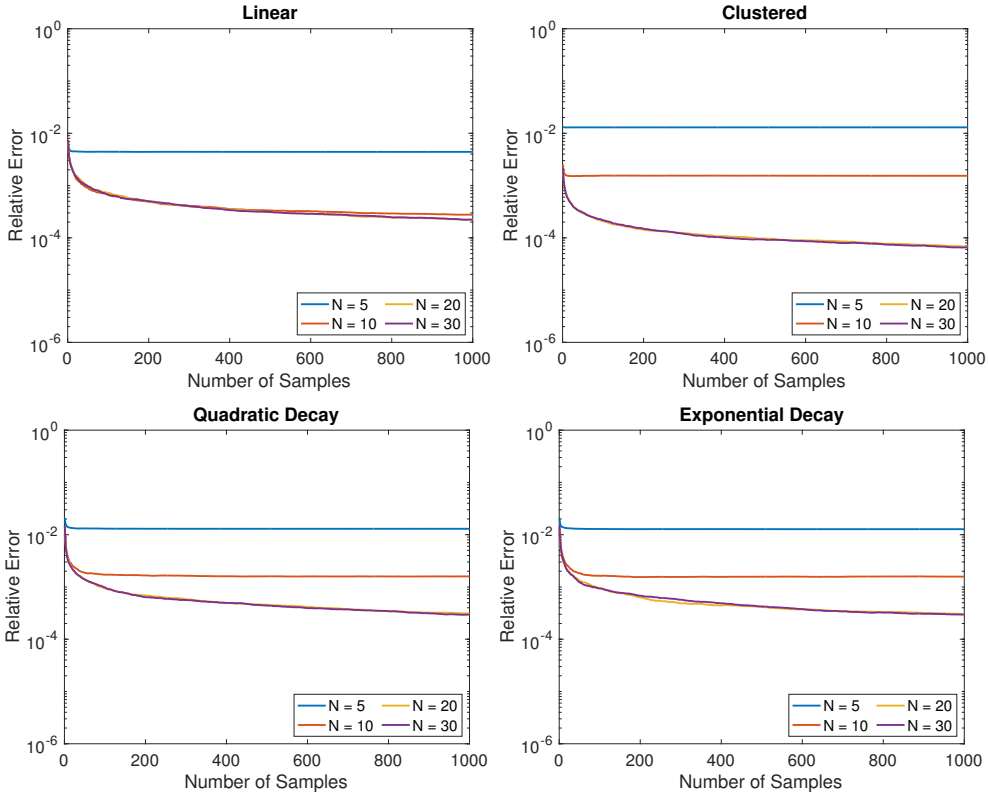

FIG. 5.10. *Average relative error of $Y_{M,N}$ for each of the 100 $\times$ 100 test matrices with $p = 120$. We used 500 different realizations for a fixed sample size M and degree N.*

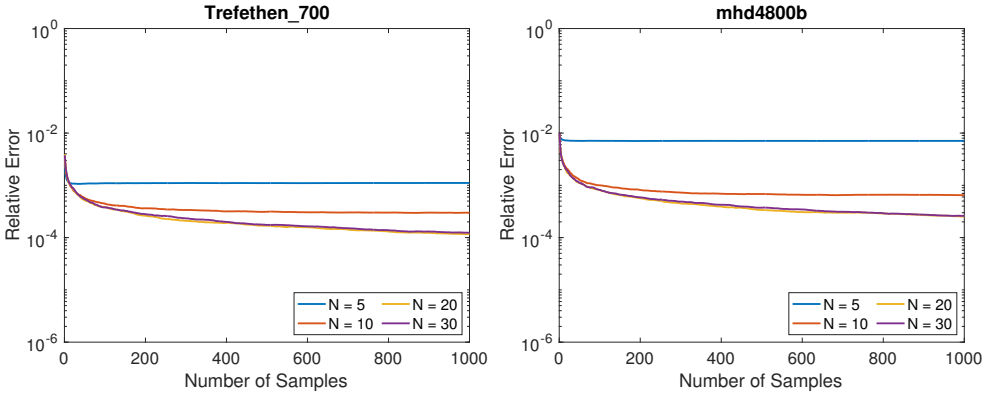

FIG. 5.11. *Average relative error of $Y_{M,N}$ for each of the SuiteSparse matrices with $p = 80$. We used 500 different realizations for a fixed sample size M and degree N .*

Posterior Covariance Matrix. In Figure 5.12 we display the mean relative error in $Y_{M,N}$ for the Posterior Covariance Matrix from our OED example problem with $p = 120$. Here we use $N = 5, 10, 20, 30$ and, similar to the Test matrices, we find that $N = 20$ is sufficient to approximate $\|\|\Gamma_{post}\|\|_p$, which is a speedup of a factor of 3 in terms on number of matrix-vector products.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

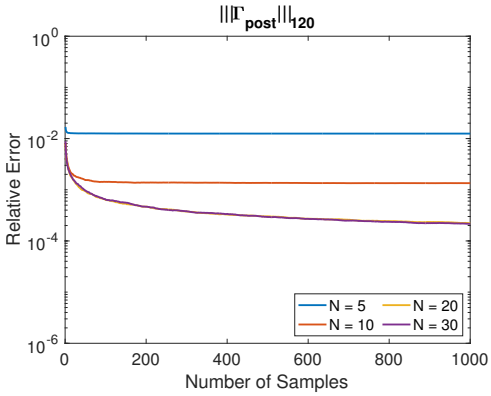

FIG. 5.12. *Average relative error of $Y_{M,N}$ for a 254 $\times$ 254 matrix Γ_{post} with $p = 120$. We used 500 different realizations for a fixed sample size M and degree N.*

We return to the question of the degree of Chebyshev polynomials. Numerical evidence suggested that $N = 20$ was sufficient for $p = 120$ and $N = 10$ is sufficient for $p = 80$ even with condition numbers as large as 8×10^{13}. This suggests that the bound in Proposition 4.2 is pessimistic and that there is potential room for improvement.

Another point worth mentioning here is the trade-off between the degree of the polynomial and the number of samples used. If the degree of the polynomial is small, then even with a large number of samples the error may be dominated by the bias in the Chebyshev polynomial approximation. On the other hand, if the degree of the polynomial is sufficiently high, then the error may be determined by the sample size. Suppose we are given a fixed computational budget for a certain number of matrix-vector products. For a given relative error, and a certain user defined probability, one can use Theorem 4.3 to give insight into apportioning the computational budget between the degree of the polynomial and the number of Monte Carlo samples.

5.4. Comparison with SLQ methods. An alternative approach for computing the Schatten p-norms is provided by the Stochastic Lanczos Quadrature (SLQ) method proposed in [29]. In contrast to our approach which uses Chebyshev polynomial approximations, the SLQ method also uses the Monte Carlo estimators, but approximates quadratic forms of the type $\mathbf{w}^T f(\mathbf{A})\mathbf{w}$ using the Lanczos process along with Gauss quadrature. The Lanczos process has the advantage of not requiring estimates of the extreme eigenvalues of the matrix, and theoretical results suggest that the SLQ method is more accurate than the Chebyshev approach [29]. However, while the Lanczos method is also based on a three-term recurrence, due to round-off errors the Lanczos basis vectors lose orthogonality. To fix this issue one can use full or partial reorthogonalization of the basis vectors [24]. In the next set of numerical experiments, we compare the accuracy of the Chebyshev approach with the SLQ approach. We use the implementation of SLQ provided by the authors of [29], in which full reorthogonalization is used. This makes the computational cost of computing the Schatten p-norm $T_{\mathbf{A}}MN + \mathcal{O}(MN^2n)$. If the degree of the polynomial N is large, the cost of a full reorthogonalization may dominate the computational cost. Note that the SLQ approach approximates $\mathrm{tr}(\mathbf{A}^p)$; however, since this approximation preserves the non-negativity of $\mathbf{A}^p$ when $\mathbf{A}$ is SPSD, we can take the p-root of this approximation to estimate $|||\mathbf{A}|||_p$.

Synthetic Test Matrices. To compare the mean relative error of the SLQ method to the mean relative error in the Chebyshev method for each of the synthetic test matrices, we conducted two different numerical tests. First, in Figure 5.13 we fix the degree of the approximation $N = 20$ and vary $p = 10, 20, 50, 100$. Note that for the Chebyshev approach

ETNA
Kent State University and
Johann Radon Institute (RICAM)

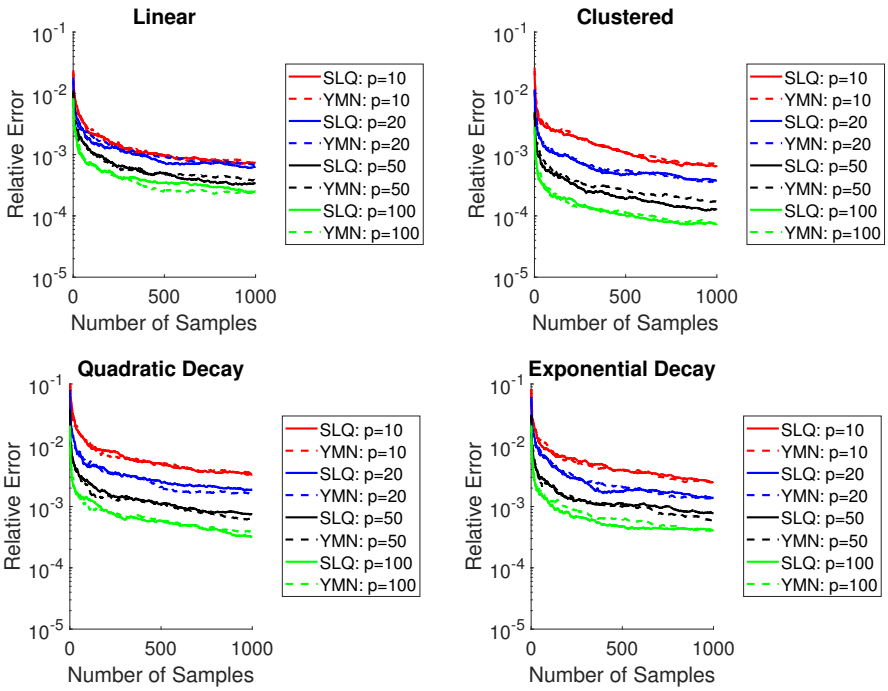

FIG. 5.13. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing the degree of approximation $N = 20$ and varying $p = 10, 20, 50, 100$.*

N is the maximum polynomial degree, whereas in SLQ it is the number of Lanczos steps. Next, in Figure 5.14 we fix $p = 100$ and vary the degree of approximation $N = 5, 10, 15, 20$. Notice that when fixing N and varying p the two methods perform similarly on all cases, whereas when p is fixed SLQ performs better for low order approximations, except in the linear distribution case. Despite this, the Chebyshev method provides good approximations for the $p = 100$ case with a polynomial of degree 20, while being easier to implement in the sense that no reorthogonalization cost has to be incurred.

SuiteSparse Matrices. Similar to the test matrices we have plotted the mean relative error for both the Chebyshev and SLQ methods for each of our SuiteSparse matrices. First, in Figure 5.15, we fixed $N = 20$ and found the the relative error when $p = 10, 20, 40, 80$, Then, in Figure 5.16, we fixed $p = 80$ and varied $N = 5, 10, 15, 20$. Once again we notice that when N is fixed, both the SLQ and Chebyshev methods perform roughly equally well when p varies for the mhd4800b matrix. Furthermore, the relative error appears to decrease as p increases. In the case of the Trefethen_700 matrix, the Chebyshev method works well for a fixed N. This is most likely due to the fact that the highest order eigenvalues decay roughly linearly.

If p is fixed and N is varied instead, we see that SLQ method does not perform significantly better with the Trefethen_700 matrix, as the relative errors are around 10^{-3}, but the Chebyshev method works well for low to medium values of N relative to p. This is again most likely due to the linear decay relationship between the highest order eigenvalues for the Trefethen_700 matrix. On the other hand, the mhd4800b matrix does not exhibit these issues. In fact, low order SLQ methods work quite well, as there is no significant decrease in the relative error by increasing the number of Lanczos steps. However, the Chebyshev method does not perform well if the degree of the approximation is not sufficiently high and the total relative error is dominated by the error in the approximation.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

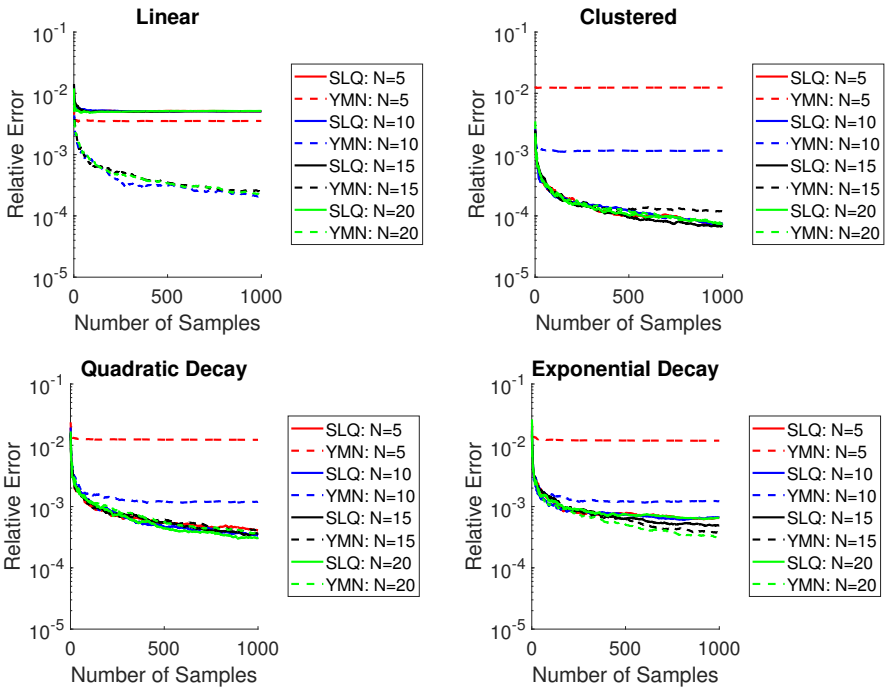

FIG. 5.14. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing* $p = 100$ *and varying the degree of approximation* $N = 5, 10, 15, 20$.

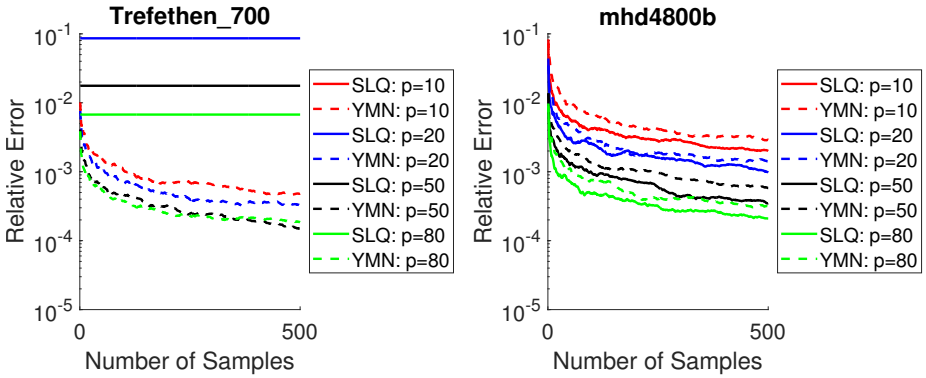

FIG. 5.15. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing the degree of approximation* $N = 20$ *and varying* $p = 10, 20, 50, 80$.

Posterior Covariance Matrix. Lastly, in Figures 5.17 and 5.18 we compare the SLQ and Chebyshev approaches for approximating the Schatten p-norm for our OED problem. In Figure 5.17, we fix $N = 20$ and vary the degree $p = 10, 20, 50, 100$, whereas in Figure 5.18 we fix $p = 100$ and vary $N = 5, 10, 15, 20$. As seen before, fixing N and varying p causes both methods to work roughly equivalently in terms of relative error. Moreover, as the OED matrix does not have a linear eigenvalue decay, we notice that the error in the SLQ method is minimal even at low order approximations (say $N = 5$), while the Chebyshev method accomplishes the same objective with a slightly higher order, $N = 15$ in our example.

In conclusion, we found that both the SLQ method and our Chebyshev polynomial approximation are competitive in terms of computing time and accuracy. SLQ has the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

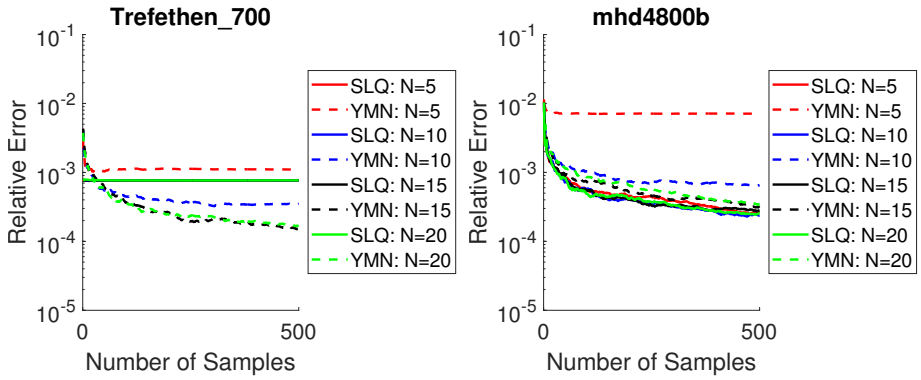

FIG. 5.16. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing* $p = 80$ *and varying the degree of approximation* $N = 5, 10, 15, 20$.

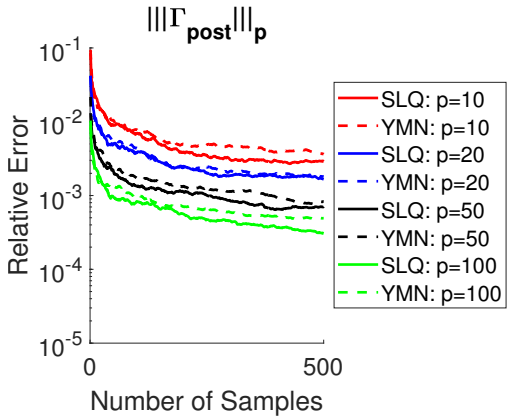

FIG. 5.17. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing the degree of approximation* $N = 20$ *and varying* $p = 10, 20, 50, 100$.

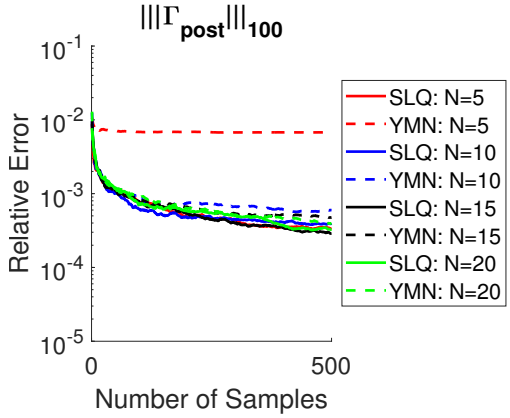

FIG. 5.18. *Mean relative error in SLQ method compared to the Chebyshev approach when fixing* $p = 100$ *and varying the degree of approximation* $N = 5, 10, 15, 20$.

advantage that it does not require estimates of the extreme eigenvalues, but it demands for a careful implementation that pays attention to reorthogonalization.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

6. Conclusion. Computation of the Schatten p-norm is frequently used in linear algebra and analysis. However, its computation by a straightforward application of the definition can be computationally difficult for large matrices. We propose two different estimators and present a probabilistic analysis of their convergence and accuracy. The numerical results show that our estimators are efficient and accurate. They also illustrate the main theoretical analysis developed in this paper, but show room for improvement. Specifically, we would like to show that the number of samples for an (ε, δ) estimator for $|||\mathbf{A}|||_p$ decreases with p. Similarly, we would like to show that a small degree N is sufficient for accurately estimating $|||\mathbf{A}|||_p$ using $Y_{M,N}$. Other possible future directions involve further exploring the stochastic Lanczos quadrature approach presented in [29], which has the advantage that it does not require estimates of the extreme points of the spectrum and promises to be more accurate compared to the Chebyshev polynomial approximation. Another possible approach is to use a rational approximation to x^p [29]; while a rational function of relatively small degree is sufficient, computing a rational matrix function can be computationally expensive.

Acknowledgements. We are grateful to Eric Hallman for his suggestion of using the symmetry of $\mathbf{A}$ to half the computational cost in Algorithm 1. We are also grateful to Shashanka Ubaru for sharing his implementation of the SLQ method with us. The authors would like to acknowledge support from the National Science Foundation through the grant "RTG: Randomized Numerical Analysis" DMS - 1745654.

REFERENCES

[1] V. AKÇELIK, G. BIROS, A. DRĂGĂNESCU, O. GHATTAS, J. HILL, AND B. VAN BLOEMAN WAANDERS, *Dynamic data-driven inversion for terascale simulations: Real-time identification of airborne contaminants*, in Proceedings of SC2005, IEEE Conference Proceedings, Los Alamitos, 2005, Art. 43, 15 pages.
[2] A. ALEXANDERIAN AND A. K. SAIBABA, *Efficient D-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems*, SIAM J. Sci. Comput., 40 (2018), pp. A2956–A2985.
[3] H. AVRON AND S. TOLEDO, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), Art. 8, 17 pages.
[4] R. BHATIA, *Matrix Analysis*, Springer, New York, 1997.
[5] J. P. BOYD, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, Mineola, 2001.
[6] T. BUI-THANH, O. GHATTAS, J. MARTIN, AND G. STADLER, *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.
[7] H. P. FLATH, L. C. WILCOX, V. AKÇELIK, J. HILL, B. VAN BLOEMEN WAANDERS, AND O. GHATTAS, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432.
[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, 2013.
[9] I. HAN, D. MALIOUTOV, H. AVRON, AND J. SHIN, *Approximating spectral sums of large-scale matrices using stochastic Chebyshev approximations*, SIAM J. Sci. Comput., 39 (2017), pp. A1558–A1585.
[10] E. HERMAN, A. ALEXANDERIAN, AND A. K. SAIBABA, *Randomization and reweighted ℓ_1-minimization for A-optimal design of linear inverse problems*, SIAM J. Sci. Comput., 42 (2020), pp. A1714–A1740.
[11] N. J. HIGHAM, *Functions of Matrices*, SIAM, Philadelphia, 2008.
[12] M. F. HUTCHINSON, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Commun. Stat.-Simul. Comput., 19 (1990), pp. 433–450.
[13] J. JACOD AND P. PROTTER, *Probability Essentials*, 2nd ed., Springer, Berlin, 2003.
[14] A. KHETAN AND S. OH, *Matrix norm estimation from a few entries*, in Advances in Neural Information Processing Systems 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., NIPS, La Jolla, 2017, pp. 6424–6433.
[15] S. KOLODZIEJ, M. AZNAVEH, M. BULLOCK, J. DAVID, T. DAVIS, M. HENDERSON, Y. HU, AND R. SANDSTROM, *The SuiteSparse matrix collection website interface*, J. Open Source Software, 4 (2019), Art. 1244, 4 pages.
[16] W. KONG AND G. VALIANT, *Spectrum estimation from samples*, Ann. Stat., 45 (2017), pp. 2218–2247.
[17] Y. LI, H. L. NGUYEN, AND D. P. WOODRUFF, *On sketching matrix norms and the top singular vector*, in Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

2014, pp. 1562–1581.
[18] ———, *On approximating matrix norms in data streams*, SIAM J. Comput., 48 (2019), pp. 1643–1697.
[19] P.-G. MARTINSSON AND J. TROPP, *Randomized numerical linear algebra: Foundations & algorithms*, ArXiv Preprint, 2020. https://arxiv.org/abs/2002.01387
[20] C. MUSCO, P. NETRAPALLI, A. SIDFORD, S. UBARU, AND D. P. WOODRUFF, *Spectrum approximation beyond fast matrix multiplication: algorithms and hardness*, ArXiv Preprint, 2017. https://arxiv.org/abs/1704.04163
[21] Y. NAKATSUKASA AND L. N. TREFETHEN, *Rational approximation of* x^n, Proc. Amer. Math. Soc., 146 (2018), pp. 5219–5224.
[22] D. J. NEWMAN AND T. J. RIVLIN, *Approximation of monomials by lower degree polynomials*, Aequationes Math., 14 (1976), pp. 451–455.
[23] V. NOLLAU, *Inequalities for variances of some functions of random variables*, Stat. Pap., 36 (1995), pp. 163–174.
[24] A. QUARTERONI, R. SACCO, AND F. SALERI, *Numerical Mathematics*, Springer, Berlin, 2007.
[25] F. ROOSTA-KHORASANI AND U. ASCHER, *Improved bounds on sample size for implicit matrix trace estimators*, Found. Comput. Math., 15 (2015), pp. 1187–1212.
[26] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, SIAM, Philadelphia, 2011.
[27] A. TARANTOLA, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2005.
[28] L. N. TREFETHEN, *Is Gauss quadrature better than Clenshaw-Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.
[29] S. UBARU, J. CHEN, AND Y. SAAD, *Fast estimation of tr(f(A)) via stochastic Lanczos quadrature*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1075–1099.
[30] R. VERSHYNIN, *High-Dimensional Probability*, Cambridge University Press, Cambridge, 2018.