

# A MULTILEVEL APPROACH TO STOCHASTIC TRACE ESTIMATION\*

ERIC HALLMAN<sup>†</sup> AND DEVON TROESTER

**Abstract.** This article presents a randomized matrix-free method for approximating the trace of  $f(\mathbf{A})$ , where  $\mathbf{A}$  is a large symmetric matrix and  $f$  is a function analytic in a closed interval containing the eigenvalues of  $\mathbf{A}$ . Our method uses a combination of stochastic trace estimation (i.e., Hutchinson’s method), Chebyshev approximation, and multilevel Monte Carlo techniques. We establish general bounds on the approximation error of this method by extending an existing error bound for Hutchinson’s method to multilevel trace estimators. Numerical experiments are conducted for common applications such as estimating the log-determinant, nuclear norm, and Estrada index, and triangle counting in graphs. We find that using multilevel techniques can substantially reduce the variance of existing single-level estimators.

**Key words.** Spectral function, trace estimation, Chebyshev approximation, Hutchinson’s trace estimator, multilevel Monte Carlo

**AMS subject classifications.** 68W25, 65C05, 65F60, 65F30

**1. Introduction.** Given a symmetric matrix  $\mathbf{A} \in \mathbb{R}^{d \times d}$  and a function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , we consider the problem of estimating

$$(1.1) \quad \text{tr}(f(\mathbf{A})) = \sum_{i=1}^d f(\lambda_i),$$

where  $\lambda_1, \dots, \lambda_d$  are the eigenvalues of  $\mathbf{A}$ . This could in theory be done by computing the eigenvalues of  $\mathbf{A}$ , but when  $\mathbf{A}$  is large this option is impractical. A cheaper option is to use stochastic trace estimation, which estimates  $\text{tr}(f(\mathbf{A}))$  by computing quantities of the form  $\mathbf{z}^T f(\mathbf{A}) \mathbf{z}$ , where  $\mathbf{z}$  is a random vector.

Four functions of particular interest are  $f(x) = \log(x)$ ,  $f(x) = 1/x$ ,  $f(x) = \exp(x)$ , and  $f(x) = x^{p/2}$ , which correspond respectively to the log-determinant of a matrix, the trace of the inverse, the Estrada index, and the Schatten  $p$ -norm<sup>1</sup>. For these functions it is not practical to compute  $\mathbf{z}^T f(\mathbf{A}) \mathbf{z}$  to machine precision, but neither is it necessary for the purpose of estimating the quantity in (1.1). Instead, it suffices to estimate  $\mathbf{z}^T f(\mathbf{A}) \mathbf{z}$  by constructing a polynomial or rational approximation to  $f$ , or by using Lanczos quadrature [1, 2]. The accuracy, and therefore the cost, of these approximations is governed by the accuracy to which one wishes to estimate  $\text{tr}(f(\mathbf{A}))$ . A typical analysis of one of these methods might provide a theorem along the following lines:

In order to estimate  $\text{tr}(f(\mathbf{A}))$  to tolerance  $\varepsilon$  with failure probability at most  $\delta$ , sample  $\mathbf{z}^T f(\mathbf{A}) \mathbf{z}$  at least  $m$  times with a level- $n$  approximation of  $f(\mathbf{A})$ .

In the above, the term “level- $n$ ” may refer to a degree- $n$  polynomial approximation or an  $n$ -point quadrature rule—either way, larger values of  $n$  correspond to more accurate and expensive approximations.

The aim of this article is to provide a general mechanism by which such methods might be improved.

\* This research was supported in part by the National Science Foundation through grant DMS-1745654.

<sup>†</sup>North Carolina State University ([erhallma@ncsu.edu](mailto:erhallma@ncsu.edu), <https://erhallma.math.ncsu.edu/>).

<sup>1</sup>In the latter case, we use  $\|\mathbf{X}\|_p^p = \text{tr}(\mathbf{X}^T \mathbf{X})^{p/2} = \text{tr} f(\mathbf{A})$ , where  $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ .

**1.1. Our approach.** We propose a method for reducing the cost of any stochastic trace estimation technique that approximates quantities of the form  $\mathbf{z}^T f(\mathbf{A})\mathbf{z}$  to variable accuracy. We focus specifically on Chebyshev approximation, but the method may be adapted to use Taylor series, rational approximations, or Lanczos quadrature. It may also be used in conjunction with other variance reduction methods such as those in [3].

The key idea is that by taking many samples with a crude approximation to  $f(\mathbf{A})$  and a few samples with an accurate approximation to  $f(\mathbf{A})$ , we can obtain a better estimate than we would have gotten simply by taking a moderate number of samples with an accurate approximation. This technique is known as *multilevel Monte Carlo* [4], which was originally developed for path simulation problems and has since found a wide variety of applications including chemical reaction networks [5], aerospace engineering [6], and rare event estimation [7]. Our application of multilevel techniques to trace estimation is outlined in Section 3.

In applying multilevel techniques to trace estimation problems, the user must choose how to set the levels: how crude should a “crude” approximation to  $f(\mathbf{A})$  be, and how many different approximations should be used? Under a certain framework it turns out that these questions have an optimal answer, summarized by Theorem 3.1. Based on this theorem, we propose a method for selecting the levels automatically based on a pilot sample.

We also show that existing error bounds for trace estimation using Hutchinson’s method may be extended to multilevel methods. This result is presented in Theorem 4.2, which offers a general framework for deriving  $(\delta, \epsilon)$ -type error guarantees for multilevel estimators.

Numerical experiments show that the multilevel estimator can have a significantly smaller variance than the single-level estimator, particularly on nuclear norm estimation problems. We also consider the problem of triangle counting in graphs, and show that using a certain set of control variates can modestly reduce the variance of existing trace estimates at minimal additional cost.

**1.2. Summary of contributions.** The key contributions of this article are as follows.

- Equations (3.1) and (3.2) show how multilevel Monte Carlo techniques may be applied to stochastic trace estimation problems.
- Using Theorem 3.1, we propose a method for selecting levels automatically and without the need for additional user input.
- Theorem 4.2 extends existing error bounds for single-level trace estimators to the multilevel framework.
- In Section 5.4 we propose a related variance reduction technique for estimating the number of triangles in a graph.
- Numerical experiments in Section 5 demonstrate the practical benefit of our methods.

**1.3. Outline.** Section 2 provides background on stochastic trace estimation, Chebyshev interpolation, and multilevel Monte Carlo methods. Section 3 describes how multilevel methods may be applied to trace estimation and provides a procedure for selecting the parameters for the multilevel estimator. Section 4 generalizes an existing error bound for single-level estimators to the multilevel case. Section 5 contains the results of numerical experiments on real data, and Section 6 offers our concluding remarks.

**1.4. Notation.** Matrices, vectors, integers, and scalars will typically be denoted as  $\mathbf{A}$ ,  $\mathbf{a}$ ,  $a$ , and  $\alpha$ , respectively, with  $\mathbf{I}$  denoting the identity matrix. The expressions  $\mathbb{E}[X]$  and  $\mathbb{V}[X]$  respectively denote the expected value and variance of a random variable  $X$ . The trace of a matrix  $\mathbf{A}$  is  $\text{tr}(\mathbf{A})$  and  $\|\mathbf{A}\|_F$  and  $\|\mathbf{A}\|_2$  are its Frobenius and operator norms, respectively. If  $\mathbf{A} \in \mathbb{R}^{d \times d}$  is a symmetric matrix with spectral decomposition  $\sum_{i=1}^d \lambda_i \mathbf{q}_i \mathbf{q}_i^T$ , then for a real-valued function  $f$  we define  $f(\mathbf{A}) = \sum_{i=1}^d f(\lambda_i) \mathbf{q}_i \mathbf{q}_i^T$ .

**2. Background.** Here we review Hutchinson's method, Chebyshev approximation, and multilevel Monte Carlo methods. For more background on these topics, see e.g. [8] and [9].

**2.1. Hutchinson's method.** A common method for trace estimation relies on the following theorem [10]:

**THEOREM 2.1.** *Let  $\mathbf{A} \in \mathbb{R}^{d \times d}$  be a symmetric matrix, and let  $\mathbf{z} \in \mathbb{R}^d$  be a random variable such that  $\mathbb{E}[\mathbf{z}\mathbf{z}^T] = \mathbf{I}$ . Then*

$$\mathbb{E}[\mathbf{z}^T f(\mathbf{A}) \mathbf{z}] = \text{tr}(f(\mathbf{A})).$$

If we generate random samples  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$  from a Rademacher distribution (entries  $\pm 1$  with equal probability), the *Hutchinson estimator* is then given by

$$(2.1) \quad \Gamma_m = \frac{1}{m} \sum_{i=1}^m \mathbf{z}^{(i)T} f(\mathbf{A}) \mathbf{z}^{(i)}.$$

Ideally, we will be able to compute or estimate each term  $\mathbf{z}^{(i)T} f(\mathbf{A}) \mathbf{z}^{(i)}$  with only a small number of matrix-vector products with  $\mathbf{A}$ . Thus if  $\mathbf{A}$  is sparse or otherwise permits fast matrix-vector multiplication, the estimator in (2.1) will be cheap to compute.

**EXAMPLE 2.2.** *If  $\mathbf{A}$  is the  $\{0, 1\}$ -valued adjacency matrix for a graph, the number of triangles in the graph is equal to  $\frac{1}{6} \text{tr}(\mathbf{A}^3)$ . Each term of the form  $\mathbf{z}^{(i)T} \mathbf{A}^3 \mathbf{z}^{(i)}$  may be evaluated using only three matrix-vector products<sup>2</sup>, and so stochastic trace estimation allows us to estimate the number of triangles in a graph without having to compute  $\mathbf{A}^3$  explicitly.*

In the case where  $f(\mathbf{A})$  is symmetric positive semi-definite (SPSD), the following error guarantee for the Hutchinson estimator is derived in [11]:

**THEOREM 2.3 (Roosta-Khorasani/Ascher).** *Let  $\mathbf{A}$  be SPSPD. For a given pair  $(\varepsilon, \delta)$  of positive numbers, the bound*

$$|\Gamma_m - \text{tr}(\mathbf{A})| \leq \varepsilon \text{tr}(\mathbf{A})$$

*holds with failure probability at most  $\delta$  if  $m \geq 6\varepsilon^{-2} \ln(2/\delta)$ .*

**2.2. Chebyshev interpolation.** We consider Chebyshev polynomials of the first kind, which follow the recurrence relation

$$(2.2) \quad T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x), \quad j = 1, 2, \dots$$

<sup>2</sup>Or two, if the symmetry of the quadratic form is exploited.

with  $T_0(x) = 1$  and  $T_1(x) = x$ . A given function  $f : [-1, 1] \rightarrow \mathbb{R}$  can then be approximated by the degree- $n$  interpolating polynomial

$$(2.3) \quad f(x) \approx p_n(x) = \sum_{j=0}^n c_j T_j(x).$$

The interpolating nodes for  $p_n$  are given<sup>3</sup> by

$$(2.4) \quad x_j = \cos \frac{j\pi}{n}, \quad 0 \leq j \leq n,$$

and the coefficients  $c_j$  can be elegantly computed using a fast Fourier transform [12].

Since  $p_n$  interpolates  $f$  on the interval  $[-1, 1]$ , it follows that  $p_n(\mathbf{A})$  will be a good approximation to  $f(\mathbf{A})$  if the spectrum of  $\mathbf{A}$  lies in the interval  $[-1, 1]$ . For a matrix whose spectrum lies in  $[a, b]$ , we can find an affine function  $g$  that maps  $[a, b]$  to  $[-1, 1]$ , define

$$\tilde{f} = f \circ g^{-1}, \quad \tilde{\mathbf{A}} = g(\mathbf{A}),$$

and approximate  $\text{tr}(\tilde{f}(\tilde{\mathbf{A}}))$  using a Chebyshev interpolation of  $\tilde{f}$ . We can therefore assume without loss of generality that the spectrum of  $\mathbf{A}$  is contained in  $[-1, 1]$ , although doing so requires at least a rough estimate of the maximum and minimum eigenvalues of  $\mathbf{A}$ .

Expressions of the form  $\mathbf{z}^T p_n(\mathbf{A}) \mathbf{z}$  can then be evaluated by computing  $\mathbf{z}_n = p_n(\mathbf{A}) \mathbf{z}$  using the recurrence in (2.2), then returning  $\mathbf{z}^T \mathbf{z}_n$ . Details can be found in [8], and the process requires  $n$  matrix-vector products (matvecs) with  $\mathbf{A}$ . It is observed in [13] that by exploiting the symmetry of the quadratic form the number of matvecs can be reduced to  $\lceil n/2 \rceil$ .

**2.3. Multilevel Monte Carlo.** Given a sequence  $P_1, \dots, P_{L-1}$  of random variables approximating  $P_L$  with increasing accuracy, the quantity of interest  $\mathbb{E}[P_L]$  can be rewritten as the telescoping sum

$$(2.5) \quad \mathbb{E}[P_L] = P_1 + \sum_{k=2}^L \mathbb{E}[P_k - P_{k-1}].$$

We can then estimate  $\mathbb{E}[P_L]$  by estimating each term on the right hand side independently. The insight of multilevel Monte Carlo methods is that if the low-level terms in (2.5) are cheap to compute and the high-level terms have small variance, this strategy can be more efficient than sampling  $P_L$  alone [4, 9].

Let  $C_1$  and  $m_1$  denote the cost of computing a single sample of  $P_1$  and the number of times it was sampled, and let  $V_1 = \mathbb{V}[P_1]$ . Similarly, for  $2 \leq k \leq L$  let  $C_k$  and  $m_k$  respectively denote the cost of a single sample of  $P_k - P_{k-1}$  and the number of times it was sampled, and let  $V_k = \mathbb{V}[P_k - P_{k-1}]$ . For a fixed variance  $\varepsilon^2$ , the total cost  $C$  is minimized by setting

$$(2.6) \quad m_k = \mu \sqrt{V_k / C_k}, \quad \text{where}$$

$$(2.7) \quad \mu = \varepsilon^{-2} \sum_{k=1}^L \sqrt{V_k C_k}.$$

<sup>3</sup>Other variants exist; see [12] for details.

The total cost of the estimate is therefore given by

$$(2.8) \quad C = \sum_{k=1}^L m_k C_k = \varepsilon^{-2} \left( \sum_{k=0}^L \sqrt{V_k C_k} \right)^2.$$

If a computational budget  $C$  is prescribed rather than a target variance, then the formula for  $\mu$  will change but equations (2.6) and (2.8) will still hold. Note that the solution in (2.6) is only an approximation as it allows the terms  $m_k$  to take on non-integer values.

Giles [9] observes that if the terms  $V_k C_k$  are generally decreasing with  $k$  then the first term  $V_1 C_1$  will make the largest contribution to the overall cost. If this is the case, we can hope to attain a total cost along the lines of  $C \approx \varepsilon^{-2} V_1 C_1$ , as opposed to the standard cost  $C \approx \varepsilon^{-2} V_1 C_L$  that would come from estimating  $P_L$  alone.

**3. Multilevel trace estimation.** The form of the interpolating polynomial in (2.3) suggests a natural way to apply multilevel techniques to Chebyshev approximation. For a fixed degree  $n$  for the interpolant  $p_n$  and indices  $-1 \leq \ell' < \ell \leq n$ , we define the variables

$$(3.1) \quad Q_{\ell'\ell} = \sum_{j=\ell'+1}^{\ell} c_j \mathbf{z}^T T_j(\mathbf{A}) \mathbf{z}.$$

Given a sequence  $0 \leq \ell_1 < \ell_2 < \dots < \ell_L = n$  we obtain the decomposition

$$(3.2) \quad \mathbf{z}^T p_n(\mathbf{A}) \mathbf{z} = \sum_{k=1}^L Q_{\ell_{k-1}\ell_k},$$

where for convenience we will always take  $\ell_0$  to be equal to  $-1$ . Thus a choice of levels  $\{\ell_k\}_{k=1}^L$  corresponds to a partition of  $\mathbf{z}^T p_n(\mathbf{A}) \mathbf{z}$  into  $L$  parts, each of which is a sum of consecutive terms in the polynomial.

Altering the notation of the previous section somewhat, we define  $V_{\ell'\ell}$  and  $C_{\ell'\ell}$  to be the variance and cost of estimating  $Q_{\ell'\ell}$ . The basic framework for the multilevel method is then as follows: we choose a set of levels  $\{\ell_k\}_{k=1}^L$ , then take a pilot sample to estimate the variance at each level. Given a desired variance  $\varepsilon^2$  or computational budget  $C$ , we then use (2.6) and (2.7) to determine the optimal number of samples  $m_k$  for each level.

**3.1. Cost estimates.** For Chebyshev interpolation, the cost of a sample will be more or less proportional to the number of matvecs required. The cost of sampling  $Q_{\ell'\ell}$  can therefore be modeled as  $\ell$  if we use the methods of [8], or as  $\lceil \ell/2 \rceil$  if we exploit the symmetry of the quadratic form as in [13].

**3.2. Optimal level selection.** Considering the form of the cost in (2.8), it is critical to note that using a large number of levels may be counterproductive, particularly if the corresponding variances decay slowly. A judicious choice of levels is therefore necessary if we want our multilevel method to outperform the single-level estimator. Here we present a method for choosing the levels with the aim of minimizing the total cost as given in (2.8).

Recalling that the approximate cost of the multilevel method is given by (2.8), we define for  $0 \leq \ell \leq n$  the variables

$$(3.3) \quad \mathcal{C}_\ell := \min_L \min_{\{\ell_k\}} \sum_{k=1}^L \sqrt{V_{\ell_{k-1}\ell_k} C_{\ell_{k-1}\ell_k}},$$

where the minimization is taken over indices satisfying  $0 \leq \ell_1 < \dots < \ell_L = \ell$ . In particular,  $\mathcal{C}_n$  corresponds to the optimal multilevel cost of approximating  $\mathbf{z}^T p_n(\mathbf{A})\mathbf{z}$ . Our goal is to find the set of levels corresponding to this optimal cost.

With perfect information about the variances and costs, it turns out that we can efficiently find the set of levels corresponding to  $\mathcal{C}_n$  through dynamic programming. We summarize this finding in the form of the following theorem.

**THEOREM 3.1.** *For  $0 \leq \ell \leq n$ , let  $\mathcal{C}_\ell$  be defined as in (3.3). Then  $\mathcal{C}_n$  can be computed by the recurrence*

$$(3.4) \quad \mathcal{C}_\ell = \begin{cases} 0 & \ell = 0, \\ \min_{0 \leq \ell' < \ell} \mathcal{C}_{\ell'} + \sqrt{V_{\ell'\ell} \mathcal{C}_{\ell'}} & 1 \leq \ell \leq n. \end{cases}$$

Assuming we already know the variances  $V_{\ell'\ell}$ , Theorem 3.1 implies that we can compute  $\mathcal{C}_n$  in  $\mathcal{O}(n^2)$  time. The optimal levels associated with  $\mathcal{C}_n$  can be obtained at minimal extra cost. Since  $n$  is small compared to the size of  $\mathbf{A}$ , determining the optimal levels will be inexpensive compared to the overall cost of trace estimation.

**3.2.1. Application to Chebyshev interpolation.** In applying the level selection method of Theorem 3.1 to Chebyshev interpolation, we face two complications. The first is that we do not have prior knowledge of the variances and so must estimate them. The second is that equations (2.6) and (2.7) assume that the sample sizes  $\{m_k\}_{k=1}^L$  may take on non-integer values. As a result, our method as described runs the risk of selecting too many levels and recommending  $0 < m_k \ll 1$  for the more expensive levels.

We propose to estimate the variances by taking a pilot sample. For  $1 \leq i \leq m_{\text{pilot}}$  we compute the terms  $\{c_j \mathbf{z}^{(i)T} T_j(\mathbf{A}) \mathbf{z}^{(i)}\}_{j=0}^n$ , storing them in a matrix of size  $m_{\text{pilot}} \times (n+1)$ . The variances can then be estimated from this information in  $\mathcal{O}(n^2 m_{\text{pilot}})$  time, which will generally be small compared to the overall cost of trace estimation. The pilot samples may subsequently be reused for the trace estimate.

**REMARK 1.** *An alternate method might be to use the Chebyshev coefficients to bound the variances at each level. We tried but ultimately rejected this approach, as the resulting bounds were too pessimistic. Performance improved when we made the assumption that  $\mathbf{z}^T T_j(\mathbf{A})\mathbf{z}$  and  $\mathbf{z}^T T_{j'}(\mathbf{A})\mathbf{z}$  were uncorrelated whenever  $j \neq j'$ , but it is not clear whether this assumption is realistic enough to be reliable.*

To resolve the issue of the sample sizes taking on non-integer values, we make the following modification: for  $0 \leq \ell \leq n-1$  we compute  $\mathcal{C}_\ell$  using the recursion in (3.4), but when computing  $\mathcal{C}_n$  we add the additional constraint that the number of samples recommended for the highest level should be at least  $m_{\text{pilot}}$ . To obtain a “recommended” number of samples, we require either a target variance  $\varepsilon^2$  or a computational budget  $C$ .

**4. Error bounds for multilevel methods.** In this section, we derive error guarantees for multilevel methods. Given a set of levels  $\{\ell_k\}_{k=1}^L$  and sample sizes  $\mathbf{m} = \{m_k\}_{k=1}^L$ , we define the estimator

$$(4.1) \quad \Gamma_{\mathbf{m}} = \sum_{k=1}^L \sum_{i=1}^{m_k} \frac{1}{m_k} Q_{\ell_{k-1}\ell_k}^{(i,k)},$$

where the  $(i, k)$  superscripts denote independent samples. Alternately, we may define for  $1 \leq k \leq L$  the matrices

$$(4.2) \quad \mathbf{A}_k = \sum_{j=\ell_{k-1}+1}^{\ell_k} c_j T_j(\mathbf{A}).$$

Then  $p_n(\mathbf{A}) = \sum_{k=1}^L \mathbf{A}_k$ , and we can express the multilevel estimator in the form

$$(4.3) \quad \Gamma_{\mathbf{m}} = \sum_{k=1}^L \sum_{i=1}^{m_k} \frac{1}{m_k} \mathbf{z}^{(i,k)T} \mathbf{A}_k \mathbf{z}^{(i,k)},$$

where the  $\mathbf{z}^{(i,k)}$  are independently drawn Rademacher vectors. We can then obtain bounds on the accuracy of the estimator  $\Gamma_{\mathbf{m}}$  by using the following theorem, due to [14]:

**THEOREM 4.1** (Cortinovis/Kressner). *Let  $\mathbf{z} \in \mathbb{R}^d$  be a Rademacher vector and let  $\mathbf{A} \in \mathbb{R}^{d \times d}$  be a nonzero symmetric matrix with all-zero diagonal entries. Then for all  $\varepsilon > 0$ ,*

$$(4.4) \quad \Pr(|\mathbf{z}^T \mathbf{A} \mathbf{z}| \geq \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{8\|\mathbf{A}\|_F^2 + 8\varepsilon\|\mathbf{A}\|_2}\right).$$

Cortinovis and Kressner subsequently use Theorem 4.1 to derive error bounds for single-level estimates. We use the same proof technique to extend their bounds to multilevel methods.

**THEOREM 4.2.** *Let  $\widehat{\mathbf{A}} \in \mathbb{R}^{d \times d}$  be a nonzero symmetric matrix. Let  $\{\mathbf{A}_k\}_{k=1}^L$  be symmetric matrices such that  $\widehat{\mathbf{A}} = \sum_{k=1}^L \mathbf{A}_k$ , and for  $1 \leq k \leq L$  let  $\mathbf{B}_k$  equal  $\mathbf{A}_k$  but with the diagonal entries set to zero. For sample sizes  $\mathbf{m} = \{m_k\}_{k=1}^L$ , let  $\Gamma_{\mathbf{m}}$  be defined as in (4.3). Then for all  $\varepsilon > 0$ ,*

$$(4.5) \quad \Pr\left(|\Gamma_{\mathbf{m}} - \text{tr}(\widehat{\mathbf{A}})| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-\varepsilon^2/8}{\sum_{k=1}^L \|\mathbf{B}_k\|_F^2/m_k + \varepsilon \max_{1 \leq k \leq L} \|\mathbf{B}_k\|_2/m_k}\right).$$

Furthermore, for  $1 \leq k \leq L$  let  $V_k = \|\mathbf{B}_k\|_F^2 + \varepsilon\|\mathbf{B}_k\|_2$  and let  $C_k$  represent the cost of sampling from  $\mathbf{B}_k$ . Then if  $m_k \geq \mu\sqrt{V_k/C_k}$  where

$$(4.6) \quad \mu = 8\varepsilon^{-2} \log(2/\delta) \sum_{k=1}^L \sqrt{V_k C_k},$$

it follows that  $\Pr\left(|\Gamma_{\mathbf{m}} - \text{tr}(\widehat{\mathbf{A}})| \geq \varepsilon\right) \leq \delta$ .

*Proof.* Let  $m = \sum_{k=1}^L m_k$ , and let  $\mathbf{B}$  be a block diagonal matrix in  $\mathbb{R}^{md \times md}$  with  $m_k$  copies of  $\mathbf{B}_k/m_k$  as its diagonal blocks. The matrix  $\mathbf{B}$  has zero diagonal and satisfies

$$\begin{aligned} \|\mathbf{B}\|_F^2 &= \sum_{k=1}^L \|\mathbf{B}_k\|_F^2/m_k, \\ \|\mathbf{B}\|_2 &= \max_{1 \leq k \leq L} \|\mathbf{B}_k\|_2/m_k. \end{aligned}$$

The first result follows by applying Theorem 4.1 to  $\mathbf{B}$ . The second follows by using the relaxation  $\max_{1 \leq k \leq L} \|\mathbf{B}_k\|_2/m_k \leq \sum_{k=1}^L \|\mathbf{B}_k\|_2/m_k$  and setting the failure probability in (4.5) to  $\delta$ .  $\square$

When using the sampling strategy proposed in Theorem 4.2, the total cost of the multilevel estimator for a given pair  $(\varepsilon, \delta)$  can be approximated as

$$(4.7) \quad C = 8\varepsilon^{-2} \log(2/\delta) \left( \sum_{k=1}^L \sqrt{V_k C_k} \right)^2.$$

This expression closely resembles the one in (2.8), with the caveat that  $V_k$  and  $\varepsilon$  refer to different quantities in these two equations. The single-level estimator, by comparison, guarantees an error of  $\varepsilon$  with failure probability  $\delta$  at a cost of  $8\varepsilon^{-2} V_{\text{tot}} C_{\text{tot}}$ , where  $V_{\text{tot}} = \|\mathbf{B}\|_F^2 + \varepsilon \|\mathbf{B}\|_2$  and  $C_{\text{tot}}$  is the cost of sampling from  $\mathbf{B}$ .

In short, multilevel estimators can be expected not only to have smaller variances than their single-level counterparts, but better  $(\varepsilon, \delta)$ -type error bounds as well. One limitation of Theorem 4.2 is that since we do not know  $\|\mathbf{B}_k\|_F$  or  $\|\mathbf{B}_k\|_2$  in advance, it does not directly give practical advice on how to choose the number of samples. More work must be done to derive error guarantees for individual functions of interest, as is done in [8] or [1], but we leave this matter for a future study.

**5. Numerical experiments.** In this section we conduct several experiments to examine the behavior of our multilevel estimator, particularly in comparison to single-level methods. All experiments were conducted using MATLAB 2020b on an Intel Core i7 3.5GHz machine, and the code used to produce all figures and tables is available at <https://github.com/erhallma/multilevel-trace-estimation/>.

Table 5.1 contains a list of the matrices used in our experiments. Most come from the SuiteSparse database [15]. Matrices ca-GrQc and wiki-Vote are the exceptions, which were obtained from the Stanford Large Network Dataset Collection<sup>4</sup>. We examine four functions in particular:  $f(x) = \sqrt{x}$  (for estimating the nuclear norm),  $f(x) = \log(x)$  (log determinant),  $f(x) = \exp(x)$  (Estrada index), and  $f(x) = x^3$  (triangle counting). For information on practical applications, we refer the reader to [16] or [8] and the references therein.

**5.1. Automated level selection.** In section 3.2 we propose a method for choosing the levels on the basis of a pilot sample and without the need for further user input. Here we illustrate how this method behaves in practice.

We use our multilevel method to estimate the nuclear norm of the matrix FA (see Table 5.1), representing a directed unweighted graph with 10617 nodes and 72176 edges. We estimate  $\text{tr}((\mathbf{A}^T \mathbf{A})^{1/2})$ , approximating the function  $f(x) = x^{1/2}$  with a degree 300 polynomial and using a budget of 15,000 matvecs, the equivalent of 50 samples for a single-level method. In order to estimate the variances at each level for the purpose of level selection, we take 10 samples using the degree 300 approximation.

Figure 5.1 shows the behavior of the multilevel method over 100 trials. In general, we make the following observations:

- The number of levels chosen is highly variable. The median trial uses twenty levels, but the number of levels ranges from as few as seven to as many as fifty-nine.
- The selected levels tend to appear in a smaller number of clusters. Table 5.2 shows one fairly typical case using eighteen levels. Aside from the consecutive

<sup>4</sup>See <https://snap.stanford.edu/data/>.



Matrix	Application	Size	nnz
thermal2	Thermal	1228045	8580313
thermomechTC	Thermal	102158	711558
boneS01	Model reduction	127224	5516602
ecology2	2D/3D	999999	4995991
ukerbe1	2D/3D	5981	15704
dictionary28	undirected graph	52652	178076
Erdos02	undirected graph	6927	16944
fe_4elt2	undirected graph	11143	65636
California	Web search	9664	16150
deter3	Linear program	$7647 \times 21777$	44547
FA	Pajek network	10617	72176
Roget	Pajek network	1022	7297
ca-GrQc	undirected graph	5242	28968
wiki-Vote	undirected graph	7115	201524

TABLE 5.1

Matrices used in our numerical experiments. All matrices with the exception of *deter3* are square.

levels 1–11, the method also selects the smaller clusters (45,46,47) and (63,65) in this example.

- Despite the variability in the *number* of levels selected, the budget allocation by level is fairly consistent between trials. For example, a typical trial spends around 60-70 percent of its computational budget on polynomials of degree 50 or less.
- The median degree of the second most expensive level is 86 over the 100 trials, and the maximum degree is 137. Thus although a high-degree polynomial may be needed to obtain a certain approximation accuracy, the multilevel method devotes most of its effort to estimating terms of significantly lower degree.

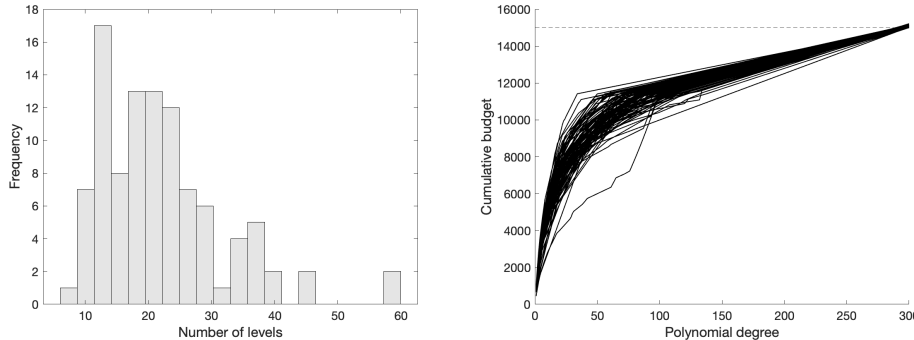


FIG. 5.1. Behavior of automated level selection over 100 trials. Left: number of levels chosen. Right: budget allocation.

We then compare the performance of the multilevel method with automated level selection against the single-level estimator, as well as the multilevel estimator with two different sets of prescribed levels. The first of these uses the three levels  $\{3, 30, 300\}$ , the sort of selection one might make with no other knowledge of the system. The second uses the seventeen levels  $\{1, 2, \dots, 15, 29, 300\}$ . This latter choice was

Level	1	2	3	4	5	6	7	8	9
Samples	776	330	180	134	90	68	52	44	33
Level	10	11	45	46	47	63	65	119	300
Samples	21	22	94	2	2	11	3	11	12

TABLE 5.2

The levels and sampling numbers from a fairly typical trial for a nuclear norm estimation problem. The multilevel method used 15,070 matvecs given a budget of 15,000.

informed by using automated selection on a pilot of 100 samples, so we expect it to be reasonably close to the optimal choice for this problem.

Results are shown in Figure 5.2, where 100 trials are run for each method. As expected, the 17-level method is the most accurate with a standard error of approximately 0.47. Automated level selection performs about as well as the 3-level method, with a standard error of approximately 0.54. The single-level approximation has a standard error of about 1.53, lagging significantly behind all of the multilevel variants.

These results suggest that choosing the levels on the basis of a pilot sample can work well in practice despite the variation in exactly which levels are chosen. It does not appear to be necessary to choose too many levels, as even the three-level method showed significant improvement over the single-level method. In practice, we recommend erring on the side of using too few levels rather than too many since taking a larger number of samples at each level will make the variance estimates more accurate.

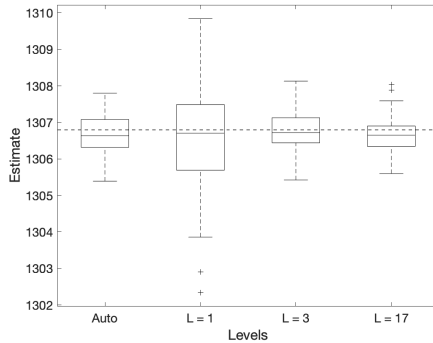


FIG. 5.2. Nuclear norm estimates over 100 trials, comparing automated level selection with using a fixed set of levels.

**5.2. Degree of approximating polynomial.** When using stochastic trace estimation, one faces the problem of deciding how to set the degree of the approximating polynomial  $p_n$ . Ideally, the degree  $n$  and number of samples  $m$  should be chosen so that the errors  $|\Gamma_m - \text{tr}(p_n(\mathbf{A}))|$  and  $|\text{tr}(p_n(\mathbf{A})) - \text{tr}(f(\mathbf{A}))|$  are similar in magnitude—a large difference between the two suggests a waste of computation, either from drawing too many samples or from using too accurate a polynomial approximation.

In this experiment, we explore the behavior of single-level and multilevel methods as the degree of the approximating polynomial changes. We again estimated the nuclear norm of the matrix  $\mathbf{FA}$ , this time allowing the degree  $n$  to range from 25 to 350. The single-level method used 50 samples for each trial, and the multilevel method used the equivalent computational budget (i.e.,  $50n$  matvecs for a degree  $n$

polynomial).

Results are shown in Figure 5.3, where we ran 100 trials for each method and each polynomial degree  $n$ . The plot on the left shows the approximate standard errors at each degree, as well as the error  $|\text{tr}(p_n(\mathbf{A})) - \text{tr}(f(\mathbf{A}))|$  due to the polynomial approximation for reference. For the single-level method this quantity is essentially constant, which is to be expected since we take the same number of samples at each degree. The multilevel method outperforms the single-level method even on the coarsest approximation, and continues improving as the approximation degree increases. The reason for this is that as the computational budget increases, the multilevel method devotes most of its effort to taking more samples at the lower levels. The single-level method, by contrast, takes the same number of samples as before but just at higher degrees.

The plot on the right shows the median relative approximation errors over 100 trials, along with the 25th and 75th percentile errors. For smaller degrees, the accuracy of both single-level and multilevel methods is constrained by the accuracy of the polynomial approximation rather than the number of samples. Somewhere between  $n = 150$  and  $n = 200$ , the single-level method becomes constrained by the number of samples and shows no further improvement as the degree increases. The multilevel method remains close to optimal until around  $n = 250$  and continues to improve afterwards.

One implication of these results is that the advantage of using multilevel methods will be greater when more accurate estimates are desired (and therefore, when higher-degree polynomial approximations are needed). A second implication is that multilevel methods are significantly less sensitive to the choice of degree than single-level methods, whose cost for a fixed number of samples grows proportionally to  $n$ . Various authors [17, 1, 8] derive error bounds for specific functions that make recommendations for the degree  $n$  and number of samples  $m$ . Although these theoretical bounds are not necessarily tight (particularly for  $m$ ), our results suggest that when using multilevel methods there is little downside to choosing  $n$  conservatively.

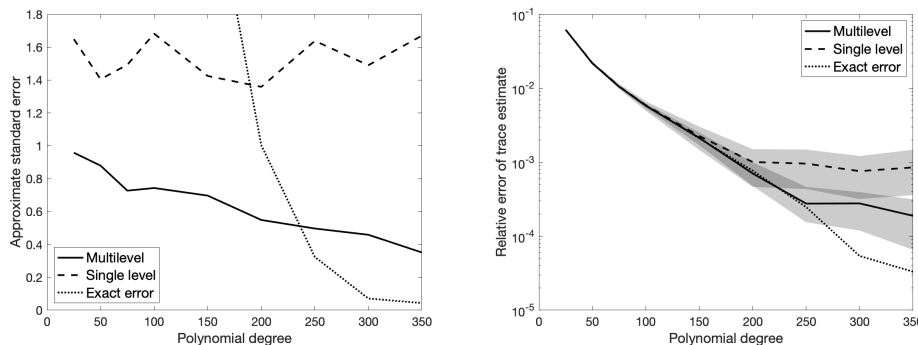


FIG. 5.3. Performance as the approximation degree  $n$  changes. Left: approximate standard errors. Right: median relative errors over 100 trials.

**5.3. SuiteSparse test cases.** Here we show results for the multilevel method and single-level method on a variety of test cases drawn from the SuiteSparse matrix collection. We generally set the degree  $n$  large enough to allow for 3-4 digits of precision in the estimate.

The most promising results for the multilevel method are when estimating the

Matrix	Exact norm	$n$	Multilevel		Single Level	
			Estimate	std	Estimate	std
California	3803.74	100	3800.78	3.46	3802.04	10.82
FA	1306.80	300	1306.55	0.44	1305.26	1.48
Erdos02	3478.23	100	3481.65	5.03	3492.99	15.31
fe_4elt2	22677.4	70	22677.1	6.68	22726.3	30.05
deter3	16518.1	70	16514.5	3.19	16501.0	11.45
uberke1	7641.44	20	7637.79	4.69	7620.45	11.75

TABLE 5.3

Nuclear norm estimates with  $m = 50$  and  $m_{pilot} = 10$ .

Matrix	Exact logdet	$n$	Multilevel		Single Level	
			Estimate	std	Estimate	std
thermomechTC	-546787	75	-546784	9.36	-546805	30.9
boneS01	1.1039e6	150	1.1040e6	25.7	1.1039e6	77.4
ecology2	3.3943e6	60	3.3933e6	158	3.3935e6	229
thermal2	1.3869e6	100	1.3864e6	182	1.3870e6	266

TABLE 5.4

Log-determinant estimates with  $m = 30$  and  $m_{pilot} = 5$ .

nuclear norm, shown in Figure 5.3. The singular values of all of these test matrices are available in the SuiteSparse database, and the exact norms are computed using these values. For matrices with low numerical rank such as California, we follow the procedure recommended in [1] and compute the nuclear norm of  $\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ , where  $\lambda$  is a small regularization term. This procedure does not change the norm by much, but it does circumvent the problem of the square root function being nondifferentiable at  $x = 0$ . The single-level method takes 50 samples for each test case, and given an equivalent computational budget our multilevel method delivers estimates whose standard errors are smaller by a factor of 2.5-4.5. Since the accuracy of an estimate scales with the square root of the number of samples, these results suggest that a multilevel approach could deliver estimates of quality comparable to the single level method while lowering the cost by as much as an order of magnitude.

For the test cases estimating the log determinant (Figure 5.4), the exact values are taken from [18], in which the values are computed using a Cholesky factorization. Here, results are somewhat more modest—the single level method uses 100 samples for each test case, and for the same computational budget the multilevel method gives estimates whose standard errors are smaller by a factor of 1.5-3.5.

For the test cases estimating the Estrada index ( $f(x) = \exp(x)$ ), shown in Figure 5.5), the exact values are computed directly. Here, the multilevel method shows little to no improvement over the single-level method. At least part of the reason is that the spectra of these matrices are typically contained in a small interval, and so a small degree  $n$  suffices to approximate the exponential function to high accuracy. We observe that the multilevel method typically uses just two levels in this case, the smaller of which was generally around  $n/2$ . Since the ratio between the lowest and highest levels is small, the multilevel method had little chance to improve over the single-level method.

In the case of the Estrada index, we also note that the standard errors for our estimates are quite large. This is because the Estrada index of a matrix is dominated

Matrix	Exact index	$n$	Multilevel		Single Level	
			Estimate	std	Estimate	std
fe_4elt2	2.2737e5	15	2.272e5	5.88e2	2.261e5	8.89e2
ErDOS02	1.6705e11	20	2.206e11	2.31e10	2.303e11	2.50e10
Roget	2.3797e5	20	2.113e5	1.53e4	2.378e5	2.37e4

TABLE 5.5

Estrada index estimates with  $m = 100$  and  $m_{pilot} = 10$ .

by its largest eigenvalues, to a far greater extent than the nuclear norm or log determinant. As a result, it will be particularly helpful to apply the variance reduction methods of [3] when estimating the Estrada index.

**5.4. Graph triangle counting.** If  $\mathbf{A}$  is the adjacency matrix for an undirected graph, the number of triangles in the graph is known to be equal to  $\text{tr}(\mathbf{A}^3)/6$ . We could apply multilevel techniques to estimate this quantity, but it is simpler to just use a control variate instead. For any real numbers  $a_1$  and  $a_2$ , we have that

$$\begin{aligned} \text{tr}(\mathbf{A}^3) &= \text{tr}(\mathbf{A}^3 - a_2\mathbf{A}^2 - a_1\mathbf{A}) + a_1 \text{tr}(\mathbf{A}) + a_2 \text{tr}(\mathbf{A}^2) \\ &= \mathbb{E}[\mathbf{z}^T(\mathbf{A}^3 - a_2\mathbf{A}^2 - a_1\mathbf{A})\mathbf{z}] + a_2 \text{nnz}(\mathbf{A}). \end{aligned}$$

The quantities  $a_1$  and  $a_2$  can then be chosen to minimize the standard deviation of  $\mathbf{z}^T(\mathbf{A}^3 - a_2\mathbf{A}^2 - a_1\mathbf{A})\mathbf{z}$ . These quantities could be chosen *a priori* using the Chebyshev expansion of  $x^3$ , but for our experiments we compute and store the values  $\mathbf{z}^{(i)T}\mathbf{A}^j\mathbf{z}^{(i)}$  for  $1 \leq i \leq m$  and  $1 \leq j \leq 3$ , then find  $a_1$  and  $a_2$  through linear regression. The added cost is minimal—in particular, no extra matvecs with  $\mathbf{A}$  are required.

We test this variance reduction method on ca-GrQc and wiki-Vote, two standard test graphs. Results are shown in Figure 5.4, where we report the median relative error over 100 trials along with the 25th and 75th percentile errors. We find that the benefit of using control variates is fairly modest, typically reducing the relative error by around 30% in the first case and 20% in the second. Nonetheless, this method is both simple to implement and inexpensive, so there appears to be little drawback to using it.

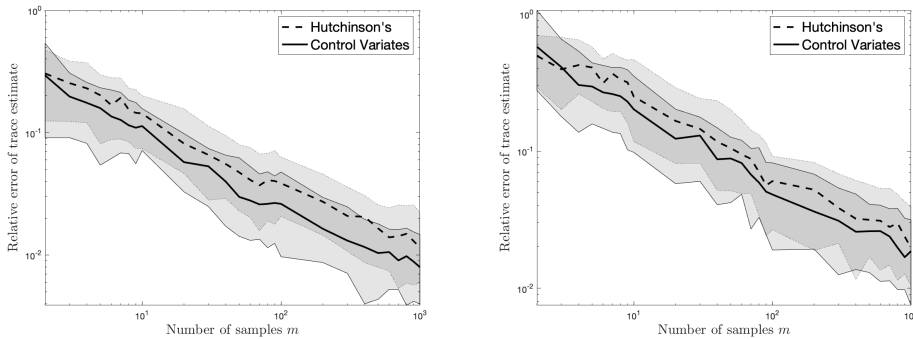


FIG. 5.4. Triangle counting with  $f(\mathbf{A}) = \frac{1}{6}\mathbf{A}^3$ . Left: ca-GrQc, an ArXiv.org collaboration network. Right: wiki-Vote, a Wikipedia administrator voting network.

In theory, we could use these same control variates to estimate the trace of polynomials  $p_n(\mathbf{A})$  of larger degree, such as when approximating the nuclear norm. It

is simple to compute  $\text{tr}(\mathbf{A}^j)$  or  $\text{tr}(T_j(\mathbf{A}))$  for  $0 \leq j \leq 2$ , so these low-degree terms may effectively be removed from our variables  $Q_{\ell\ell}$  in (3.1). When the degree of the matrix polynomial is large, however, the coefficients of  $p_n$  will decay more slowly and so the effect of using control variates will likely be fairly small.

**6. Conclusion.** In this paper, we have shown how multilevel techniques can be used to improve existing methods for stochastic trace estimation. We have derived general error bounds for our multilevel trace estimator, and through numerical experiments have demonstrated the efficacy of the multilevel estimator as compared with single-level methods.

One avenue for further study is in deriving multilevel error guarantees that are specific to the function  $f$ , such as those for single-level methods in [8, 1, 17]. Another possibility is to explore whether other variance reduction techniques for Monte Carlo methods might find applications in stochastic trace estimation problems: for example, tools for modeling rare events could potentially be used to determine whether a given matrix is positive definite, a problem which trace estimation is used to solve in [8]. Our hope is that this paper will encourage further exploration in these directions.

**Acknowledgements.** The authors would like to thank Michael Merritt, Alen Alexanderian, and Pierre Gremaud for their helpful remarks.

#### REFERENCES

- [1] S. Ubaru, J. Chen, and Y. Saad, “Fast estimation of  $\text{tr}(f(\mathbf{A}))$  via stochastic Lanczos quadrature,” *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1075–1099, 2017.
- [2] G. H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*. Princeton University Press, 2009, vol. 30.
- [3] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff, “Hutch++: Optimal stochastic trace estimation,” in *Symposium on Simplicity in Algorithms (SOSA)*. SIAM, 2021, pp. 142–155.
- [4] M. B. Giles, “Multilevel Monte Carlo path simulation,” *Operations research*, vol. 56, no. 3, pp. 607–617, 2008.
- [5] D. F. Anderson and D. J. Higham, “Multilevel monte carlo for continuous time markov chains, with applications in biochemical kinetics,” *Multiscale Modeling & Simulation*, vol. 10, no. 1, pp. 146–179, 2012.
- [6] G. Geraci, M. S. Eldred, and G. Iaccarino, “A multifidelity multilevel monte carlo method for uncertainty propagation in aerospace applications,” in *19th AIAA Non-Deterministic Approaches Conference*, 2017, p. 1951.
- [7] E. Ullmann and I. Papaioannou, “Multilevel estimation of rare events,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 3, no. 1, pp. 922–953, 2015.
- [8] I. Han, D. Malioutov, H. Avron, and J. Shin, “Approximating spectral sums of large-scale matrices using stochastic Chebyshev approximations,” *SIAM Journal on Scientific Computing*, vol. 39, no. 4, pp. A1558–A1585, 2017.
- [9] M. B. Giles, “Multilevel Monte Carlo methods,” *Acta Numerica*, vol. 24, p. 259, 2015.
- [10] M. F. Hutchinson, “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines,” *Communications in Statistics-Simulation and Computation*, vol. 18, no. 3, pp. 1059–1076, 1989.
- [11] F. Roosta-Khorasani and U. Ascher, “Improved bounds on sample size for implicit matrix trace estimators,” *Foundations of Computational Mathematics*, vol. 15, no. 5, pp. 1187–1212, 2015.
- [12] L. N. Trefethen, “Is Gauss quadrature better than Clenshaw–Curtis?” *SIAM review*, vol. 50, no. 1, pp. 67–87, 2008.
- [13] E. Hallman, “Faster stochastic trace estimation with a Chebyshev product identity,” *arXiv preprint arXiv:2101.00325*, 2021.
- [14] A. Cortinovis and D. Kressner, “On randomized trace estimates for indefinite matrices with an application to determinants,” *arXiv preprint arXiv:2005.10009*, 2020.
- [15] T. A. Davis and Y. Hu, “The University of Florida Sparse Matrix Collection,”

- ACM Trans. Math. Softw.*, vol. 38, no. 1, Dec. 2011. [Online]. Available: <https://doi.org/10.1145/2049662.2049663>
- [16] S. Ubaru and Y. Saad, “Applications of trace estimation techniques,” in *International Conference on High Performance Computing in Science and Engineering*. Springer, 2017, pp. 19–33.
  - [17] E. Dudley, A. K. Saibaba, and A. Alexanderian, “Monte Carlo estimators for the Schatten  $p$ -norm of symmetric positive semidefinite matrices,” *arXiv preprint arXiv:2005.10174*, 2020.
  - [18] C. Boutsidis, P. Drineas, P. Kambadur, E.-M. Kontopoulou, and A. Zouzias, “A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix,” *Linear Algebra and its Applications*, vol. 533, pp. 95–117, 2017.