RC25581 (WAT1512-022) December 13, 2015
Computer Science

IBM Research Report

How Accurately Should I Solve Linear Systems When Applying the Hutchinson Trace Estimator?

Jie Chen
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

HOW ACCURATELY SHOULD I SOLVE LINEAR SYSTEMS WHEN APPLYING THE HUTCHINSON TRACE ESTIMATOR?

JIE CHEN*

Abstract. The Hutchinson estimator defines an estimate of the trace of a matrix M, based on a bilinear form with independent vectors y of zero-mean unit-variance uncorrelated entries. This technique is particularly useful when M is only implicitly given but the matrix-vector product My can be efficiently computed without M being explicitly formed. Well-known examples in practice are $M = A^{-1}$, and more generally, $M = f(A)$. We study in this paper the conditions under which the numerical error incurred in computing My is comparable with the statistical uncertainty caused by the randomness of y. For the purpose of obtaining easily computable conditions, we focus on the use of random vectors consisting of normal variables, a precursor technique attributed to Girard by Hutchinson. As demonstrated in many practical scenarios, normal variables are as effective as symmetric Bernoulli variables (a more common definition under the name of Hutchinson), but are advantageous in that they enjoy a simultaneous estimation of the estimator variance.

Key words. Matrix trace, Hutchinson estimator, matrix inverse, matrix function

AMS subject classifications. 65C05, 65F10, 65F60

1. Introduction.

The trace of a large, implicit matrix M finds many applications in scientific computing. In estimation theory, if A is the Fisher information matrix of an unbiased estimator, then the trace of $M = A^{-1}$ gives a lower bound of the total variance of the estimator (see the Cramér–Rao inequality; e.g., [8]). Similarly, the log-determinant of a covariance matrix A, which is equivalent to the trace of $M = \log A$, appears naturally in the maximization of Gaussian log-likelihoods [1, 27]; across disciplines, this term serves as a barrier in interior point methods for solving semidefinite programs when A is the semidefinite constraint [28, 7]. In electronic structures, the trace of the Fermi–Dirac function

$$f_{\mathrm{FD}}(A) = \left[I + \exp\left(\frac{A - \mu I}{kT} \right) \right]^{-1} \tag{1.1}$$

gives the average number of electrons of a quantum system at chemical potential μ and temperature T, where A is the discretized Hamiltonian and k is the Boltzmann constant [25, 6]. Additionally, applications in lattice quantum chromodynamics [3, 26], density of states [5, 22, 21], and uncertainty quantification [4, 19] comprise a limited, yet informative, list that illustrates the importance of trace computation.

When the $n \times n$ matrix M is implicitly defined through a given matrix A, it may not be computationally economic, or even viable, to first form M before extracting the trace. If, on the other hand, matrix-vector products with M are relatively inexpensive to compute, then n such products suffice the recovery of the trace: $\mathrm{tr}(M) = \sum_{i=1}^{n} e_i^T M e_i$, where e_i is the ith column of the identity matrix. If, however, n is so large that even n matrix-vector products are too expensive to form, most of the existing work approximates the trace based on the following stochastic approach.

THEOREM 1.1 (Hutchinson [18]). *Let $M \in \mathbb{R}^{n\times n}$ be symmetric and $Y \in \mathbb{R}^n$ be a multivariate random variable of zero mean and unit covariance. Then, for a sample y of Y,*

$$\mathrm{E}[y^T M y] = \mathrm{tr}(M) \quad \textit{and} \quad \mathrm{Var}(y^T M y) = 2\,\mathrm{tr}(M^2) + \sum_{i=1}^{n} \big(\mathrm{E}[Y_i^4] - 3 \big) M_{ii}^2.$$

*IBM T. J. Watson Research Center, Yorktown Heights, NY 10598. Email: chenjie@us.ibm.com

Remark. The theorem straightforwardly extends to more general cases of M. For example, if M is unsymmetric, one may symmetrize the matrix to obtain the same trace: $\operatorname{tr}(M) = \operatorname{tr}(M + M^T)/2$. As another example, if M is Hermitian but not real, then $\operatorname{tr}(M) = \operatorname{tr}(\Re(M))$.

Practical uses of Theorem 1.1 form a sample average with N iid samples y_i, $i = 1, \ldots, N$, such that the variance is reduced by a factor of N:

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} y_i^T M y_i\right] = \mathrm{E}[y^T M y] \quad \text{with} \quad \operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^T M y_i\right) = \frac{1}{N}\operatorname{Var}(y^T M y). \tag{1.2}$$

This technique, while extremely useful, poses an often neglected issue on the numerical accuracy of the evaluation of the My_i's. Consider, for example, $My_i = A^{-1}y_i$. To one extreme, if the stochastic approximation (1.2) is highly accurate, meaning that the variance is sufficiently small, and if A is so ill conditioned that even a stable direct method for solving $Ax_i = y_i$ leaves a comparably large backward error, then the overall departure of the estimate from the truth $\operatorname{tr}(M)$ is dominated by numerical bias. Whereas such a scenario rarely occurs in practice, the opposite scenario is certainly not uncommon: the estimate yields a moderate variance, in the sense that it agrees with the truth on a small number of digits. Then, it is of little use to solve linear systems $Ax_i = y_i$ highly accurately (if possible). Instead, it suffices for one to use an iterative solver (possibly enhanced by using block iterations for improving convergence [23, 24, 13]) that terminates at a moderate residual.

Hence, the subject of this paper is to study the balance of stochastic uncertainty of the Hutchinson estimator and the numerical error incurred in the evaluation of M-vector products. We derive practical conditions that make the two sources of errors comparable. To this end, we first need to establish the concrete meaning of "comparable" on a statistical basis.

Let $h(y)$ be an unbiased estimator of some quantity μ, with variance σ as in

$$\mathrm{E}_y[h(y)] = \mu \quad \text{and} \quad \operatorname{Var}_y(h(y)) = \sigma^2.$$

Moreover, let y_i, $i = 1, \ldots, N$ be N independent samples from the same distribution and define

$$h_0 = \frac{1}{N}\sum_{i=1}^{N} h(y_i).$$

Clearly, h_0 as an estimator is also unbiased. The central limit theorem states that $\sqrt{N}(h_0 - \mu)$ converges to the normal distribution $\mathcal{N}(0, \sigma^2)$. Therefore, for large N, h_0 approximately follows $\mathcal{N}(\mu, \sigma^2/N)$. Because the variance of h_0 is nothing but σ^2/N, by applying the three-sigma rule we obtain

$$\Pr\left(|h_0 - \mu| \le 3\sqrt{\operatorname{Var}(h_0)}\right) \approx 99.7\%. \tag{1.3}$$

Although $\operatorname{Var}(h_0)$ in (1.3) could be replaced by the sample variance of $h(y_i)$ scaled by N, in this paper we consider the case when the $h(y_i)$'s are not computed accurately. Then, instead, we suppose that an unbiased estimator h_1 of $\operatorname{Var}(h_0)$ is available; that is,

$$\mathrm{E}[h_1] = \operatorname{Var}(h_0).$$

Let $\tilde{h}_0$ be the computed result of h_0 with error Δh_0 (i.e., $h_0 = \tilde{h}_0 + \Delta h_0$). If the error can be controlled to within an α portion of the estimate of the standard deviation of h_0:

$$|\Delta h_0| \leq \alpha\sqrt{h_1}, \qquad \alpha > 0, \tag{1.4}$$

then, we can maintain a confidence interval for $\tilde{h}_0$:

$$\Pr\left(|\tilde{h}_0 - \mu| \leq 3\sqrt{\mathrm{Var}(h_0)} + \alpha\sqrt{h_1}\right) \geq \Pr\left(|h_0 - \mu| \leq 3\sqrt{\mathrm{Var}(h_0)}\right) \approx 99.7\%. \tag{1.5}$$

The numerical interpretation of the quantity $s = 3\sqrt{\mathrm{Var}(h_0)} + \alpha\sqrt{h_1}$ is that with probability (approximately) at least 99.7%, the relative error of the actually obtainable numerical result $\tilde{h}_0$ is bounded by

$$\frac{|\tilde{h}_0 - \mu|}{|\mu|} \leq \frac{s}{|\mu|}.$$

In a (possibly crude) approximation, the right-hand side of the above inequality

$$\frac{s}{|\mu|} \approx \frac{\sigma}{|\mu|} \cdot \frac{3+\alpha}{\sqrt{N}}.$$

Therefore, the asymptotics with respect to N is key to the accuracy of $\tilde{h}_0$, whereas the magnitude of α (which controls the part of numerical error) plays a less significant role. Thus, one may safely consider α as large as 1.0. On the other hand, in order for $\tilde{h}_0$ to be one more digit accurate, one needs to increase the number N of samples by a factor of 100.

The central contribution of this work is the conditions that ensure (1.4). To materialize the estimators h_0 and h_1, we first define the random vector Y in Theorem 1.1. As oppose to the common use of independent ± 1's (symmetric Bernoulli variables) as the elements of Y, in Section 2, we justify that normal variables are often as effective. In fact, the invention of the use of normal variables was attributed to Girard [15] by Hutchinson [18], before the symmetric Bernoulli variables became popular. One advantage of symmetric Bernoulli variables is that they minimize the variance in (1.2); however, we show two examples, motivated by electronic structure calculations with matrices of structural decay [6], that demonstrate that symmetric Bernoulli variables often cannot improve the estimation accuracy over normal variables by even one digit. Normal variables, on the other hand, allow an estimate of the variance with almost negligible cost, which is challenging for symmetric Bernoulli variables to achieve, unless the variance is replaced by sample variance and the evaluation of the samples (i.e., the My_i's) is accurate to machine precision.

In Section 3, we establish the condition ensuring (1.4) for the case $M = A^{-1}$. The condition is with respect to the *absolute* residual in the solution of linear equations $Ax_i = y_i$. This condition can be straightforwardly used as the *absolute* residual tolerance in an iterative solver. Additionally, an example with symmetric tridiagonal matrices is shown. Similar to the example in the section that follows, these are "toy" matrices because many properties (e.g., $\mathrm{tr}(A^{-1})$ and the estimator variance) can be analytically derived. The purpose of the toy examples, however, is to show the asymptotics with respect to the matrix size n and the sample size N and to give a flavor of the numerical results under randomness. A numerical example with matrices from applications is given two sections later.

In Section 4, we establish the condition ensuring (1.4) in a probabilistic for the case $M = f(A)$. A general approach for computing matrix-vector products of the form $f(A)y$ is to replace f by an approximate function p (e.g., a polynomial or a rational function) such that the evaluation of $p(A)y$ renders to matrix-vector multiplications with A. This approach should bare no surprise since for the case of linear systems, a Krylov solver can be interpreted as building a polynomial that interpolates $f(x) = x^{-1}$ at the approximate eigenvalues of A. Different from the condition for $M = A^{-1}$, however, the condition here is with respect to the *relative* error of the approximant p in the uniform norm. This condition is straightforwardly applicable when p is a polynomial, such as in the approach proposed by Chen et al. [12], because the approximation error can be monitored without the knowledge of y and $p(A)y$ can be evaluated accurately to machine precision. For rational or other approximations (see, e.g., [17]), one must take into account the numerical error in evaluating $p(A)y$ in addition to the approximation error of p. As before, we show a numerical example with Toeplitz matrices with structural decay to illustrate the use of the condition. These matrices are model matrices for electronic structures and the attainable relative error scales as $\Theta(n^{-\frac{1}{2}}N^{-\frac{1}{2}})$ with high probability.

We show further computational experiences in Section 5, by using the PARSEC collection[1] of matrices arising from density functional theory [10, 9]. The function f therein is defined based on the Fermi–Dirac function (1.1). In this case, the number N of samples is chosen to be $1,000$ and the trace estimate is generally two to four digits accurate (with sufficiently high probability). Interestingly, the general trend of results suggests that the relative accuracy improves as the matrix size increases, even though the same N is used throughout. This observation agrees with that of the model matrices with structural decay in Section 4.

Related work, discussions, and concluding remarks are given in Section 6.

2. Hutchinson estimator with normal variables. In this section, we justify the use of normal variables in the Hutchinson estimator and study the properties of the estimator, the variance of the estimator, and the estimator of the variance.

2.1. Normal v.s. symmetric Bernoulli. A symmetric Bernoulli variable is a discrete random variable that takes the values ± 1 with equal probabilities. The following result is straightforward.

COROLLARY 2.1 (Hutchinson [18]). *Under the condition of Theorem 1.1,*

1. *If the entries of Y are independent symmetric Bernoulli variables, then*

$$\mathrm{Var}(y^T My) = 2\,\mathrm{tr}(M^2) - 2\sum_{i=1}^{n} M_{ii}^2.$$

This variance is the minimum among all possible distributions of Y.

2. *If $Y \sim \mathcal{N}(0, I)$, then*

$$\mathrm{Var}(y^T My) = 2\,\mathrm{tr}(M^2).$$

Proof. The result immediately follows from the fact that the fourth moment of a symmetric Bernoulli variable is 1 whereas that of a normal variable is 3. Moreover, for any random variable X of zero-mean and unit-variance, the fourth moment is lower

[1] Available from the University of Florida Sparse Matrix Collection `https://www.cise.ufl.edu/research/sparse/matrices/`

bounded by 1 because $0 \le \mathrm{Var}(X^2) = \mathrm{E}[X^2 - 1]^2 = \mathrm{E}[X^4] - 1$. Thus, symmetric Bernoulli variables yield the minimum variance. □

A direct consequence of the corollary is that the variance may vanish (when M is diagonal) for symmetric Bernoulli variables; but for normal variables, the standard-deviation-to-mean ratio admits a lower bound $\Omega(n^{-\frac{1}{2}})$.

PROPOSITION 2.2. *Let $M \in \mathbb{R}^{n\times n}$ be symmetric and $Y \sim \mathcal{N}(0, I_n)$. Then, for a sample y of Y,*

$$\frac{\sqrt{\mathrm{Var}(y^T My)}}{|\,\mathrm{E}[y^T My]|} \ge \sqrt{\frac{2}{n}}.$$

Proof. Let λ be the vector of eigenvalues of M. One obtains the inequality by noting that

$$\mathrm{Var}(y^T My) = 2\,\mathrm{tr}(M^2) = 2\|\lambda\|_2^2, \qquad |\,\mathrm{E}[y^T My]| = |\,\mathrm{tr}(M)| \le \|\lambda\|_1,$$

and that $\|\lambda\|_1 \le \sqrt{n}\|\lambda\|_2$. □

As such, it may appear that normal variables are inferior to symmetric Bernoulli variables, because if the energy of the matrix (in the sense of Frobenius norm $\|M\|_F^2 = \mathrm{tr}(M^2)$) is concentrated on the diagonal, then the latter will yield highly accurate estimates. In many scenarios, however, this is an impractical assumption. In the following, we show two examples, both of which entail a decaying structure, and demonstrate that the standard-deviation-to-mean ratio attains the rate $\Theta(n^{-\frac{1}{2}})$ in both estimators. To maintain clarity, we use subscript "N" to mean normal and "B" to mean symmetric Bernoulli.

Example: Toeplitz matrix with exponential decay. Consider $M_{ij} = \theta^{|i-j|}$ where $0 < \theta < 1$. Then, $\mathrm{tr}(M) = n$ and

$$\mathrm{tr}(M^2) = n\frac{1+\theta^2}{1-\theta^2} - 2\frac{\theta^2(1-\theta^{2n})}{(1-\theta^2)^2}, \qquad \sum_{i=1}^{n} M_{ii}^2 = n.$$

Thus,

$$\frac{\sqrt{\mathrm{Var_N}(y^T My)}}{|\,\mathrm{E}[y^T My]|} = \sqrt{\frac{2}{n}\frac{1+\theta^2}{1-\theta^2}} + O\left(\frac{1}{n}\right),$$

and

$$\frac{\sqrt{\mathrm{Var_B}(y^T My)}}{|\,\mathrm{E}[y^T My]|} = \sqrt{\frac{2}{n}\frac{2\theta^2}{1-\theta^2}} + O\left(\frac{1}{n}\right).$$

Asymptotically, θ needs to be $\le 1/\sqrt{199} \approx 0.07$ in order that $\sqrt{\mathrm{Var_B}}$ is a factor of 10 smaller than $\sqrt{\mathrm{Var_N}}$ (i.e., one more digit accurate under the same probability).

Example: Toeplitz matrix with algebraic decay. Consider $M_{ij} = |i - j + 1|^{-1}$. Then, $\mathrm{tr}(M) = n$ and

$$\mathrm{tr}(M^2) = -n + 2(n+1)\sum_{i=1}^{n}\frac{1}{i^2} - 2\sum_{i=1}^{n}\frac{1}{i}, \qquad \sum_{i=1}^{n} M_{ii}^2 = n.$$

By applying the inequalities

$$\frac{\pi^2}{6} - \frac{1}{n} < \sum_{i=1}^{n} \frac{1}{i^2} < \frac{\pi^2}{6} - \frac{1}{n+1} \quad \text{and} \quad \ln(n+1) < \sum_{i=1}^{n} \frac{1}{i} \leq \ln n + 1,$$

we obtain

$$\left(\frac{\pi^2}{3} - 1\right) n - \frac{2}{n} - 2\ln n + \frac{\pi^2}{3} - 4 < \operatorname{tr}(M^2) < \left(\frac{\pi^2}{3} - 1\right) n - 2\ln(n+1) + \frac{\pi^2}{3} - 2.$$

Therefore,

$$\frac{\sqrt{\operatorname{Var}_{\mathrm{N}}(y^T M y)}}{|\operatorname{E}[y^T M y]|} = \sqrt{\frac{2}{n}\left(\frac{\pi^2}{3} - 1\right)} + O\left(\frac{\ln n}{n}\right),$$

and

$$\frac{\sqrt{\operatorname{Var}_{\mathrm{B}}(y^T M y)}}{|\operatorname{E}[y^T M y]|} = \sqrt{\frac{2}{n}\left(\frac{\pi^2}{3} - 2\right)} + O\left(\frac{\ln n}{n}\right).$$

Asymptotically, the ratio between $\sqrt{\operatorname{Var}_{\mathrm{B}}}$ and $\sqrt{\operatorname{Var}_{\mathrm{N}}}$ is approximately 0.75, which means that the relative error resulting from the use of symmetric Bernoulli variables is only slightly better than that of normal variables. Improvement on the number of accurate digits is impossible.

2.2. Estimator, variance, and estimator of variance. Structural decay is an important property in electronic structures [6]. As demonstrated above, for model matrices with such a property, normal variables are generally as effective as symmetric Bernoulli variables. An advantage of the former is that it allows for a simultaneous estimation of the variance with negligible costs. Thus, in this subsection, we define the variance estimator, which itself has a variance that in turn admits an estimator. The recurrence of estimator and variance interestingly repeats endlessly. From now on, we use the sample average in place of a single sample in the estimator.

PROPOSITION 2.3. *Let $M \in \mathbb{R}^{n\times n}$ be symmetric and y_i, $i = 1, \ldots, N$, be random iid vectors from $\mathcal{N}(0, I_n)$. Then, for all $j = 0, 1, \ldots$*

$$\operatorname{Var}\left(\frac{2^{2^j-1}}{N^{2^j}} \sum_{i=1}^{N} y_i^T M^{2^j} y_i\right) = \operatorname{E}\left[\frac{2^{2^{j+1}-1}}{N^{2^{j+1}}} \sum_{i=1}^{N} y_i^T M^{2^{j+1}} y_i\right].$$

Proof. With basic properties of the variance,

$$\begin{aligned}\operatorname{Var}\left(\frac{2^{2^j-1}}{N^{2^j}} \sum_{i=1}^{N} y_i^T M^{2^j} y_i\right) &= \left(\frac{2^{2^j-1}}{N^{2^j-1}}\right)^2 \operatorname{Var}\left(\frac{1}{N} \sum_{i=1}^{N} y_i^T M^{2^j} y_i\right) \\ &= \frac{1}{N}\left(\frac{2^{2^j-1}}{N^{2^j-1}}\right)^2 \operatorname{Var}\left(y_1^T M^{2^j} y_1\right).\end{aligned}$$

Invoking Corollary 2.1 followed by Theorem 1.1, we obtain

$$\operatorname{Var}\left(y_1^T M^{2^j} y_1\right) = 2\operatorname{tr}\left(M^{2^{j+1}}\right) = 2\operatorname{E}\left[y_1^T M^{2^{j+1}} y_1\right].$$

Then, together with

$$\mathrm{E}\left[y_1^T M^{2^{j+1}} y_1\right] = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} y_i^T M^{2^{j+1}} y_i\right],$$

we conclude the proposition. ∎

Based on the proposition, we define

$$h_j := \frac{2^{2^j-1}}{N^{2^j}} \sum_{i=1}^{N} y_i^T M^{2^j} y_i, \qquad j = 0, 1, \ldots \tag{2.1}$$

This definition is consistent with the notation h_0 and h_1 introduced earlier in the introduction. Then, Equation (1.2) together with Proposition 2.3 state that h_0 is an estimator of $\mathrm{tr}(M)$, h_1 is an estimator of the variance of h_0, and generally, h_{j+1} is an estimator of the variance of h_j. Clearly, all estimators are unbiased. We have the following result.

THEOREM 2.4. *Denote by $\lambda_{|\min|}$ and $\lambda_{|\max|}$ the smallest the largest singular value of a symmetric nonzero matrix M, respectively. For all $j > 0$,*

$$\left(\frac{2\lambda_{|\min|}}{N}\right)^{2^j} \le \frac{h_{j+1}}{|h_j|} \le \left(\frac{2\lambda_{|\max|}}{N}\right)^{2^j}. \tag{2.2}$$

Additionally,

1. (2.2) *holds when $j = 0$ and M is definite.*
2. *The left half of* (2.2) *holds when $j = 0$ and M is indefinite.*

Proof. When $j > 0$, $h_j > 0$. Write $z_i = M^{2^{j-1}} y_i$. Then,

$$h_j = \frac{2^{2^j-1}}{N^{2^j}} \sum_{i=1}^{N} \|z_i\|^2 \quad \text{and} \quad h_{j+1} = \frac{2^{2^{j+1}-1}}{N^{2^{j+1}}} \sum_{i=1}^{N} z_i^T M^{2^j} z_i.$$

Because for each i,

$$\lambda_{|\min|}^{2^j} \|z_i\|^2 \le z_i^T M^{2^j} z_i \le \lambda_{|\max|}^{2^j} \|z_i\|^2,$$

summing over i we conclude (2.2).

When $j = 0$ and M is positive semidefinite, the same argument can be used to establish (2.2), by noting that z_i is well defined. When $j = 0$ and M is negative semidefinite, $|h_0| = -h_0$. Then, replacing M by $-M$ will prove (2.2).

We now proceed to the remaining case: $j = 0$ and M is indefinite. We concatenate all the vectors y_i to form a vector z, and duplicate M diagonally to form a matrix $\widetilde{M}$ (i.e., $\widetilde{M}$ is block diagonal and the diagonal blocks are M). Then,

$$|h_0| = \frac{1}{N}|z^T \widetilde{M} z| \quad \text{and} \quad h_1 = \frac{2}{N^2} z^T \widetilde{M}^2 z.$$

Because $\widetilde{M}$ can be diagonalized and the unitary factor can be absorbed to z, we treat $\widetilde{M}$ a diagonal matrix where the diagonal elements are the eigenvalues of M. Then,

$$\left(\frac{2\lambda_{|\min|}}{N}\right)|h_0| = \left(\frac{2\lambda_{|\min|}}{N^2}\right)\left|\sum_k \lambda_k z_k^2\right| \le \frac{2}{N^2}\sum_k \lambda_{|\min|}|\lambda_k z_k^2| \le \frac{2}{N^2}\sum_k \lambda_k^2 z_k^2 = h_1,$$

where z_k are the elements of z and λ_k are the eigenvalues of M. This shows the left half of (2.2). ∎

Remark. The right half of (2.2) may not hold when $j = 0$ and M is indefinite, because h_0 may attain 0.

An immediate consequence of Theorem 2.4 is that

$$\left(\frac{2\lambda_{|\min|}}{N}\right)^{2^j-2} h_1 \le h_j \le \left(\frac{2\lambda_{|\max|}}{N}\right)^{2^j-2} h_1 \quad \text{and} \quad \left(\frac{2\lambda_{|\min|}}{N}\right)|h_0| \le h_1.$$

The significance of this result is three fold. First, when N is considered fixed and sufficiently large, h_j decreases doubly exponentially with respect to j; such a decrease is faster than the exponential. Second, when N varies, h_j decreases as $\Theta(N^{-2^j+1})$ if $h_1 = \Theta(N^{-1})$. Such a decrease is a very-high-order algebraic decrease. Third, for all $j > 0$, the standard-deviation-to-mean ratio $\sqrt{h_{j+1}}/h_j = \Theta(N^{-\frac{1}{2}})$ if h_1 is $\Theta(N^{-1})$. Moreover, $\sqrt{h_1}/|h_0|$ is at least $\Omega(N^{-\frac{1}{2}})$. Since the ratio quantifies the relative error of using h_j as an estimator of $\mathrm{Var}(h_{j-1})$ when $j > 0$, and of $\mathrm{tr}(M)$ when $j = 0$, the j-independent decrease rate implies that the quality of estimators h_j is similar across j, in the relative term. Then, in the absolute term, h_j is more and more accurate as j becomes large.

We illustrate an example for the last point. Suppose $\mathrm{tr}(M) =$ `316.22` and the estimate $h_0 =$ `314.70` is two-digit accurate. The true standard deviation $\sqrt{\mathrm{Var}(h_0)} =$ `1.99`. In practice, when we use h_1 to estimate $\mathrm{Var}(h_0)$, we may obtain $\sqrt{h_1} =$ `1.90`, again two-digit accurate. In such a case, we may safely treat h_1 the "same" as $\mathrm{Var}(h_0)$ when establishing the confidence interval for h_0, because the absolute difference `1.99-1.90` is very small compared with the true trace `316.22`.

On closing this section, we note that later we will frequently refer to the quantity

$$\frac{\sqrt{\mathrm{tr}(M^2)}}{\mathrm{tr}(M)} = \frac{\sqrt{h_1}}{h_0}\sqrt{\frac{N}{2}}, \tag{2.3}$$

which is $\sqrt{N/2}$ times the standard-deviation-to-mean ratio. If this quantity scales as $\Theta(n^{-\frac{1}{2}})$ (cf. the lower bound in Proposition 2.2), then the standard-deviation-to-mean ratio $\sqrt{h_1}/h_0$ scales as $\Theta(n^{-\frac{1}{2}}N^{-\frac{1}{2}})$. This means that the relative error of the estimate of the trace decreases not only with the sample size N but also with the matrix size n.

3. Estimating $\mathrm{tr}(A^{-1})$. In this section, we consider the case $M = A^{-1}$, where the My_i's are computed through solving linear systems $Ax_i = y_i$. The following is the main result.

THEOREM 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric nonsingular and y_i, $i = 1, 2, \ldots, N$, be independent vectors from $\mathcal{N}(0, I_n)$. For each i, denote by $r_i = y_i - Ax_i$ the residual of the linear system with matrix A and right-hand side y_i, where x_i is an approximate solution. Decompose the trace estimator $h_0 = \tilde{h}_0 + \Delta h_0$, where*

$$h_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T A^{-1} y_i \quad \textit{and} \quad \tilde{h}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T x_i,$$

and let h_1 be the estimator of the variance of h_0 defined in (2.1)*. For any $\alpha > 0$, if*

$$\|r_i\| \le \alpha\sqrt{\frac{2}{N}} \quad \textit{for all } i, \tag{3.1}$$

then $|\Delta h_0| \le \alpha\sqrt{h_1}$.

Proof. We express Δh_0 in terms of r_i:

$$\Delta h_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T A^{-1} r_i.$$

Let z be the column concatenation of the vectors $A^{-1}y_i$ and similarly let r be the concatenation of the r_i's. Then,

$$\Delta h_0 = \frac{1}{N} z^T r \quad \text{and} \quad h_1 = \frac{2}{N^2} z^T z.$$

By Cauchy–Schwarz, $|z^T r| \le \|z\|\|r\|$. Therefore, if all vectors r_i satisfy (3.1), then $\|r\| \le \alpha\sqrt{2}$. Thus,

$$|\Delta h_0| = \frac{1}{N}|z^T r| \le \frac{1}{N}\|z\|\|r\| \le \frac{\alpha\sqrt{2}}{N}\|z\| = \alpha\sqrt{h_1},$$

which concludes the theorem. ∎

We note that the condition (3.1) concerns the absolute residual, but in software implementations, the tolerance of a Krylov solver is typically or the relative residual. Hence, care is called for when one uses a software. Nevertheless, the following result indicates that the absolute residual is approximately $\sqrt{n}$ times the relative one.

PROPOSITION 3.2. *If* $y \sim \mathcal{N}(0, I_n)$, *then*

$$\mathrm{E}[\|y\|] = \frac{n\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{n}{2}+1\right)} \quad \textit{and} \quad \lim_{n\to\infty} \frac{\mathrm{E}[\|y\|]}{\sqrt{n}} = 1.$$

Proof. The first equality is a straightforward calculation:

$$\begin{aligned}
\mathrm{E}[\|y\|] &= \int_{\mathbb{R}^n} \frac{\|y\|}{(2\pi)^{n/2}} \exp\left(-\frac{\|y\|^2}{2}\right) dy && \text{by definition} \\
&= \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \int_0^\infty \frac{r}{(2\pi)^{n/2}} \exp\left(-\frac{r^2}{2}\right) r^{n-1}\, dr && \text{spherical coordinate } r = \|y\| \\
&= \frac{n\sqrt{2}}{\Gamma\left(\frac{n}{2}+1\right)} \int_0^\infty \exp\left(-r^2\right) r^n\, dr && \text{change of variable } r/\sqrt{2} \to r \\
&= \frac{n\sqrt{2}}{\Gamma\left(\frac{n}{2}+1\right)} \cdot \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right).
\end{aligned}$$

The second equality follows from

$$\lim_{m\to\infty} \frac{\Gamma\left(m-\frac{1}{2}\right)\sqrt{m}}{\Gamma(m)} = 1. \qquad \blacksquare$$

Remark. Clearly, because $\|y\|^2 \sim \chi^2_n$, we have $\mathrm{E}[\|y\|^2] = n$.

3.1. Example: Symmetric tridiagonal matrix. Consider the following tridiagonal matrix

$$A = \begin{bmatrix} a & -1 & & & \\ -1 & a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & a \end{bmatrix}. \tag{3.2}$$

An interesting fact of this example is that the standard-deviation-to-mean ratio behaves differently depending on the value of a. When $|a| = 2$, in a later proposition we show that the ratio (2.3) is $\Theta(1)$, and hence $\sqrt{h_1}/h_0$ is $\Theta(N^{-\frac{1}{2}})$. This means that the relative error decreases only with the sample size but not with the matrix size. For a moderate N, one expects that the approximation is, say, one-digit accurate. When $|a| < 2$, even though the ratio (2.3) admits a closed-form expression, its asymptotics with respect to n is unclear. If one arbitrarily pick a in this interval, one expects that the relative error is as bad as that in the case $|a| = 2$. However, when $|a| > 2$, the ratio (2.3) is $\Theta(n^{-\frac{1}{2}})$, and thus $\sqrt{h_1}/h_0 = \Theta(n^{-\frac{1}{2}}N^{-\frac{1}{2}})$. This encouraging result indicates that the relative error generally decreases when the matrix becomes larger, in addition to when N increases. Hence, one expects that the relative error is a few digits accurate. In all cases, we demonstrate that the computed result of the trace estimate, $\tilde{h}_0$, is close to the theoretical value h_0, if the linear systems are solved to an accuracy dictated by the condition (3.1).

We first present some analytical results. When A is invertible, one may show through a direct verification that the inverse takes the form

$$A^{-1} = \begin{bmatrix} & & \frac{u_i u_{n+1-j}}{u_{n+1}} \\ & \iddots & \\ \frac{u_{n+1-i} u_j}{u_{n+1}} & & \end{bmatrix},$$

where the u_i's satisfy the recurrence relation

$$u_1 = 1, \qquad u_2 = a, \qquad u_{i+1} = a u_i - u_{i-1}, \quad i = 2, 3, \ldots$$

For different cases of a, we give the expressions of u_i.

1. When $a = 2$, $u_i = i$ for all i. Hence,

$$(A^{-1})_{ij} = \min(i, j) - \frac{ij}{n+1}, \qquad i, j = 1, \ldots, n.$$

2. When $a = -2$, $u_i = (-1)^{i+1} i$ for all i. Hence,

$$(A^{-1})_{ij} = (-1)^{i-j+1} \left[\min(i, j) - \frac{ij}{n+1} \right], \qquad i, j = 1, \ldots, n.$$

3. When $|a| < 2$, let θ be the unique value in the interval $(0, \pi)$ such that $\tan\theta = \sqrt{4 - a^2}/a$. If $(n+1)\theta$ is a multiple of π, then A is singular; otherwise,

$$u_i = \frac{2\sin(i\theta)}{\sqrt{4 - a^2}} \quad \text{for all } i.$$

Hence, the elements of the lower triangular part of A^{-1} are

$$(A^{-1})_{ij} = \frac{2\sin[(n+1-i)\theta]\sin(j\theta)}{\sqrt{4 - a^2}\sin[(n+1)\theta]}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, i.$$

4. When $|a| > 2$,

$$u_i = \frac{1}{\sqrt{a^2 - 4}} \left[\left(\frac{a + \sqrt{a^2 - 4}}{2} \right)^i - \left(\frac{a - \sqrt{a^2 - 4}}{2} \right)^i \right] \quad \text{for all } i.$$

Based on these results, we obtain $\operatorname{tr}(A^{-1})$ and $\operatorname{tr}(A^{-2})$, which lead to the ratio (2.3).

PROPOSITION 3.3. *For A defined in* (3.2),

1. *When $a = \pm 2$, respectively,*

$$\operatorname{tr}(A^{-1}) = \pm\frac{n^2+2n}{6} \quad \textit{and} \quad \operatorname{tr}(A^{-2}) = \frac{2n^4+8n^3+17n^2+18n}{180}.$$

2. *When $|a| < 2$, let θ be the unique value in the interval $(0,\pi)$ such that $\tan\theta = \sqrt{4-a^2}/a$. If $(n+1)\theta$ is not a multiple of π,*

$$\operatorname{tr}(A^{-1}) = \frac{-(n+1)\cot[(n+1)\theta] + \cot\theta}{\sqrt{4-a^2}}$$

and

$$\operatorname{tr}(A^{-2}) = \frac{(n+1)^2\csc^2[(n+1)\theta] + (n+1)\cot[(n+1)\theta]\cot\theta + 1 - 2\csc^2\theta}{4-a^2}.$$

3. *When $|a| > 2$, let*

$$\zeta = \frac{|a|+\sqrt{a^2-4}}{2} \quad \textit{and} \quad \delta = \frac{|a|-\sqrt{a^2-4}}{|a|+\sqrt{a^2-4}}.$$

Then,

$$\frac{(1-\delta)^2 n}{\sqrt{a^2-4}} \le |\operatorname{tr}(A^{-1})| \le \frac{n}{\sqrt{a^2-4}}$$

and

$$\frac{(1-\delta)^4}{a^2-4}\left[n\frac{\zeta^2+1}{\zeta^2-1} - 2\frac{\zeta^2(\zeta^{2n}-1)}{(\zeta^2-1)^2\zeta^{2n}}\right] \le \operatorname{tr}(A^{-2}) \le \frac{1}{a^2-4}\left[n\frac{\zeta^2+1}{\zeta^2-1} - 2\frac{\zeta^2(\zeta^{2n}-1)}{(\zeta^2-1)^2\zeta^{2n}}\right].$$

Proof. The first two parts of the proposition can be straightforwardly verified through tedious algebraic calculations. For the third part, we prove only the case $a > 2$. The proof of the other case $a < -2$ is analogous.

Based on the definition of ζ and δ, we note that $\zeta > 1$, $0 < \delta < 1$ and $u_i = (1-\delta^i)\zeta^i/\sqrt{a^2-4}$. Then, when $i \ge j$,

$$\frac{u_{n+1-i}u_j\sqrt{a^2-4}}{u_{n+1}} = \frac{(1-\delta^{n+1-i})(1-\delta^j)}{1-\delta^{n+1}}\zeta^{j-i},$$

which is bounded on both sides by $(1-\delta)^2\zeta^{j-i}$ and ζ^{j-i}. Therefore,

$$(1-\delta)^2\zeta^{-|i-j|} \le (A^{-1})_{ij}\sqrt{a^2-4} \le \zeta^{-|i-j|}.$$

Summing over $i = j$, we obtain the inequality for $\operatorname{tr}(A^{-1})$; and summing over i and j, we obtain the inequality for $\operatorname{tr}(A^{-2}) = \|A^{-1}\|_F^2$. ∎

Remark. In the inequalities of the third part, the right half is generally tight but not the left half. This is because when $\delta < 1$ is away from 1,

$$\frac{(1-\delta^{n+1-i})(1-\delta^j)}{1-\delta^{n+1}}$$

is close to 1, unless when $i \approx n$ and $j \approx 1$, which makes it close to $(1-\delta)^2$. However, the number of cases when $i \approx n$ and $j \approx 1$ is limited and thus when summing over i (and j), these cases contribute little to the overall sum. Therefore, the trace terms leaned toward the upper bounds. As a result,

$$\frac{\sqrt{\mathrm{tr}(A^{-2})}}{\mathrm{tr}(A^{-1})} \approx \sqrt{\frac{1}{n}\frac{\zeta^2+1}{\zeta^2-1}}.$$

With the analytic understanding of $\mathrm{tr}(A^{-1})$ and $\mathrm{tr}(A^{-2})$, we now show numerical results for the following six quantities:

(a) $\mathrm{tr}(A^{-1})$ (b) h_0 (c) h_0, iterative solve (`tol`)

(d) $\sqrt{\frac{2}{N}\mathrm{tr}(A^{-2})}$ (e) $\sqrt{h_1}$ (f) $\sqrt{h_1}$, iterative solve (`tol`).

See Table 3.1.

TABLE 3.1

Computational results for the tridiagonal matrix A defined in (3.2). *Parameters:* $n = 1000$, $N = 100$, $\alpha = 1$. *Residual tolerance computed from* (3.1)*:* `tol = 1.41e-01`, E[`rtol`] `= 4.47e-03`.

Case $a = 2$, average residual = `1.23e-01`,
average relative residual = `3.90e-03`

	Truth	Estim. (full solve)	Estim. (`tol`)
$\mathrm{tr}(A^{-1})$	167000	175960	175708
`stddev(estim)`	14937	15507	15470

Case $a = 1.7$, average residual = `9.75e-02`,
average relative residual = `3.08e-03`

	Truth	Estim. (full solve)	Estim. (`tol`)
$\mathrm{tr}(A^{-1})$	1138.61	1030.96	1029.46
`stddev(estim)`	209.47	201.93	201.72

Case $a = 2.6$, average residual = `9.19e-02`,
average relative residual = `2.91e-03`

	Truth	Estim. (full solve)	Estim. (`tol`)
$\mathrm{tr}(A^{-1})$	601.589	600.135	600.088
`stddev(estim)`	3.365	3.358	3.358

Let us recall the meaning of these quantities. Term (a) is the truth. Term (b) is the estimate, with variance in Term (d). In the context of this paper, we approximate Term (b) by using Term (c), which is computed approximately based on the condition (3.1) in Theorem 3.1; and Term (e) is the estimator of Term (d). We approximate Term (e) by using Term (f), a byproduct of the calculation of Term (c).

We pick three values of a for demonstrating the numerical results, each corresponding to one part of Proposition 3.3. Because the case $a = 2$ corresponds to the standard 1D Laplacian, whose eigenvalues lie in the interval $(0, 4)$, we easily see that A is positive definite when $a = 2$ and 2.6, but is indefinite when $a = 1.7$. In light of the indefiniteness, we use GMRES as the linear solver and block Jacobi as the preconditioner.

When the sample size N is 100, the condition (3.1) indicates that we need only an absolute residual tolerance `1.41e-01`, which corresponds to an average relative residual tolerance `4.47e-03` for a matrix of size $n = 1,000$. In such a setting, the results in Table 3.1 indicate that the estimates/approximations are one- (close to two-) digit accurate for $a = 2$ and $a = 1.7$, but are two- (close to three-) digit accurate for $a = 2.6$. The absolute errors are all within the standard deviation.

4. Estimating $\mathrm{tr}(f(A))$**.** In this section, we consider the case $M = f(A)$, where the function f is approximated by p. The following is the main result.

THEOREM 4.1. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, f and p be functions well defined on the spectrum of A, and y_i, $i = 1, 2, \ldots, N$, be independent vectors from $\mathcal{N}(0, I_n)$. Decompose the trace estimator $h_0 = \tilde{h}_0 + \Delta h_0$, where*

$$h_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T f(A) y_i \quad \textit{and} \quad \tilde{h}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T p(A) y_i,$$

and let h_1 be the estimator of the variance of h_0 defined in (2.1)*. For any $\alpha > 0$ and $\delta \in (0, 1/2)$, if*

$$|1 - p(\lambda)/f(\lambda)| \le \alpha\sqrt{\frac{2}{(1+\delta)nN}} \tag{4.1}$$

for all eigenvalues λ of A, then $|\Delta h_0| \le \alpha\sqrt{h_1}$ with probability at least $1 - e^{-\delta^2 nN/6}$.

The proof of the theorem relies on the following lemma, whose result appears in various slightly different forms; see, e.g., [14, 20].

LEMMA 4.2. *Let $z \sim \mathcal{N}(0, I_d)$. Then, for any $\delta \in (0, 1/2)$,*

$$\Pr\Big(\|z\|^2 \le (1+\delta)d\Big) \ge 1 - e^{-\delta^2 d/6}.$$

Proof. For any $\lambda > 0$, we apply the monotone transformation followed by the Markov inequality:

$$\Pr\Big(\|z\|^2 > (1+\delta)d\Big) = \Pr\Big(\exp(\lambda\|z\|^2) > \exp(\lambda(1+\delta)d)\Big) \le \frac{\mathrm{E}[\exp(\lambda\|z\|^2)]}{\exp(\lambda(1+\delta)d)}.$$

One easily calculates that the above expectation evaluates to $\mathrm{E}[\exp(\lambda\|z\|^2)] = (1 - 2\lambda)^{-d/2}$ when $\lambda < 1/2$. Therefore,

$$\Pr\Big(\|z\|^2 > (1+\delta)d\Big) \le \frac{(1-2\lambda)^{-d/2}}{\exp(\lambda(1+\delta)d)}$$

when $0 < \lambda < 1/2$. In this interval, the right-hand side of the above inequality attains minimum when $1 - 2\lambda = (1+\delta)^{-1}$. Hence,

$$\Pr\Big(\|z\|^2 > (1+\delta)d\Big) \le \frac{(1+\delta)^{d/2}}{\exp(\delta d/2)}.$$

By using the inequality $\ln(1+\delta) \le \delta - \delta^2/3$ for $\delta \in (0, 1/2)$, we conclude the lemma. ∎

Proof of Theorem 4.1. When $f(A)$ is nonsingular, write

$$\Delta h_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^T f(A)(I - f(A)^{-1}p(A))y_i = \frac{1}{N}\sum_{i=1}^{N} y_i^T f(A) r_i,$$

where $r_i = (I - f(A)^{-1}p(A))y_i$. Let z be the column concatenation of the vectors $f(A)y_i$ and similarly let r be the concatenation of r_i. Then,

$$\Delta h_0 = \frac{1}{N} z^T r \quad \text{and} \quad h_1 = \frac{2}{N^2} z^T z.$$

If (4.1) is satisfied, we have

$$\|r\| \le \alpha \sqrt{\frac{2}{(1+\delta)nN}} \|w\|,$$

where w is the column concatenation of the vectors y_i. By Lemma 4.2, with probability at least $1 - e^{-\delta^2 nN/6}$, we have $\|w\| \le \sqrt{(1+\delta)nN}$. Therefore, with at least this probability, $\|r\| \le \alpha\sqrt{2}$. Then, by Cauchy–Schwarz $|z^T r| \le \|z\|\|r\|$, we immediately conclude the theorem. □

Remark. The Theorem suffers no loss of generality when $f(\lambda) = 0$ for some eigenvalue λ. In such a case, because f is bounded within the spectrum interval, one may define $\tilde{f} = f + c$ for some constant c so that $\tilde{f}(\lambda) \neq 0$ for all eigenvalues. Then, the theorem applies to the new function $\tilde{f}$.

Different from the condition in the preceding section, the condition (4.1) here ensures $|\Delta h_0| \le \alpha\sqrt{h_1}$ only in the probabilistic sense. Hence, the establishment of the confidence interval (1.5) for $\tilde{h}_0$ needs a modification:

$$\begin{aligned}&\Pr\left(|\tilde{h}_0 - \mu| \le 3\sqrt{\mathrm{Var}(h_0)} + \alpha\sqrt{h_1}\right)\\ &\quad \ge \Pr\left(|h_0 - \mu| \le 3\sqrt{\mathrm{Var}(h_0)}\right)\Pr\left(|\Delta h_0| \le \alpha\sqrt{h_1}\right) \approx (1 - e^{-\delta^2 nN/6}) \times 99.7\%.\end{aligned}$$

Such a modification, however, is minor because with three-digit precision,

$$(1 - e^{-\delta^2 nN/6}) \times 99.7\% \approx 99.7\%$$

as long as $\delta^2 nN/6 > 10$.

4.1. Example: Toeplitz matrices with decay. To demonstrate the use of Theorem 4.1, here we consider Toeplitz matrices with structural decay (exponential or algebraic). Similar to the example in the preceding section, decaying Toeplitz matrices enjoy interesting properties. As we will show later, the standard-deviation-to-mean ratio decreases as $\Theta(n^{-\frac{1}{2}}N^{-\frac{1}{2}})$ for any continuous function f. Hence, one expects that the estimate can be more accurate when the matrix becomes larger. As model matrices for electronic structures, they hint on the effective use of the trace estimator in this application.

Let $A_{ij} = t_{i-j}$, where for symmetry $t_k = t_{-k}$. Assume that the infinite sequence $\ldots, t_{-2}, t_{-1}, t_0, t_1, t_2, \ldots$ consists of the coefficients of the Fourier series of a 2π-periodic function $q(\omega)$ in that

$$q(\omega) = \sum_{k=-\infty}^{\infty} t_k e^{\mathbf{i}\,k\omega} \quad \text{with} \quad t_k = \frac{1}{2\pi}\int_0^{2\pi} q(\omega) e^{-\mathbf{i}\,k\omega}\, d\omega. \tag{4.2}$$

Then, for any length-n vector x,

$$x^T A x = \frac{1}{2\pi} \int_0^{2\pi} q(\omega) \left| \sum_{j=1}^{n} x_j e^{-\mathbf{i}\, j\omega} \right|^2 d\omega.$$

Because

$$\frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{j=1}^{n} x_j e^{-\mathbf{i}\, j\omega} \right|^2 d\omega = \|x\|_2^2,$$

if we choose an x with a unit norm, then

$$\inf_{\omega\in[0,2\pi]} q(\omega) \le x^T A x \le \sup_{\omega\in[0,2\pi]} q(\omega),$$

which means that the eigenvalues of A are bounded within the range of q.

We will add a subscript n to A when asymptotics is in concern. A useful result for Toeplitz matrices is that the eigenvalues of A_n, denoted as $\lambda_j^{(n)}$, $j = 0, \ldots, n-1$, are close to equally-spaced samples of q. For this, we need the definition of *equal distribution*.

DEFINITION 4.3. *Two sets of real numbers* $\{a_j^{(n)}\}_{j=0,\ldots,n-1}$ *and* $\{b_j^{(n)}\}_{j=0,\ldots,n-1}$ *are* equally distributed *in the interval* $[M_1, M_2]$ *if for any continuous function* $F : [M_1, M_2] \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} [F(a_j^{(n)}) - F(b_j^{(n)})] = 0.$$

It is well known that the eigenvalues $\{\lambda_j^{(n)}\}$ of A_n and the set $\{q(2\pi j/n)\}_{j=0,\ldots,n-1}$ are equally distributed (see, e.g, [16, 11]). An immediate consequence is that for a matrix function f, if it is continuous on the range of q, then the trace of $f(A_n)$ and that of $f^2(A_n)$ can be characterized by

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{tr}(f(A_n)) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(\lambda_j^{(n)}) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(q(2\pi j/n)) = \frac{1}{2\pi} \int_0^{2\pi} f(q(\omega))\, d\omega$$

and

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{tr}(f^2(A_n)) = \frac{1}{2\pi} \int_0^{2\pi} f^2(q(\omega))\, d\omega.$$

Therefore, we have the following result.

THEOREM 4.4. *Given an infinite sequence* $\ldots, t_{-2}, t_{-1}, t_0, t_1, t_2, \ldots$ *that satisfies the symmetry* $t_k = t_{-k}$ *and the assumption* (4.2) *for some* 2π*-periodic function* $q(\omega)$*, define a sequence of matrices* A_n, $n = 1, 2, \ldots$, *where* $(A_n)_{ij} = t_{i-j}$*. Then, for any* f *continuous on the range of* q,

$$\lim_{n\to\infty} n^{\frac{1}{2}} \frac{\sqrt{\operatorname{tr}(f^2(A_n))}}{\operatorname{tr}(f(A_n))} = \frac{\left(\int_0^{2\pi} 2\pi f^2(q(\omega))\, d\omega \right)^{\frac{1}{2}}}{\int_0^{2\pi} f(q(\omega))\, d\omega}.$$

We now consider two examples.

Exponential decay. Let

$$t_k = \theta^{|k|}, \qquad 0 < \theta < 1. \tag{4.3}$$

We have

$$\begin{aligned} q(\omega) &= \sum_{k=-\infty}^{\infty} t_k e^{\mathbf{i}\, k\omega} = -1 + 2\Re\left(\sum_{k=0}^{\infty} e^{k\ln\theta} e^{\mathbf{i}\, k\omega}\right) \\ &= -1 + 2\Re\left(\frac{1}{1 - e^{\ln\theta + \mathbf{i}\,\omega}}\right) = \frac{1-\theta^2}{1 - 2\theta\cos\omega + \theta^2}. \end{aligned}$$

Therefore,

$$q_{\max} = q(0) = \frac{1+\theta}{1-\theta} \quad \text{and} \quad q_{\min} = q(\pi) = \frac{1-\theta}{1+\theta}.$$

Algebraic decay. Let

$$t_k = (k^2+1)^{-1}. \tag{4.4}$$

Note that this decay has a different order from that in Section 2. We make use of the well known Fourier transform

$$\frac{2a}{a^2+\omega^2} = \int_{-\infty}^{+\infty} e^{-a|t|} e^{-\mathbf{i}\,\omega t}\, dt, \qquad a > 0$$

to write

$$\begin{aligned} q(\omega) &= \sum_{k=-\infty}^{+\infty} \frac{e^{\mathbf{i}\, k\omega}}{1+k^2} = \frac{1}{2} \sum_{k=-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-|t|} e^{-\mathbf{i}\, kt}\, dt\, e^{\mathbf{i}\, k\omega} \\ &= \frac{1}{2} \int_{-\infty}^{+\infty} e^{-|t|} \left[\sum_{k=-\infty}^{+\infty} e^{-\mathbf{i}\, kt} e^{\mathbf{i}\, k\omega}\right] dt \\ &= \pi \int_{-\infty}^{+\infty} e^{-|t|}\, \mathrm{III}(t-\omega)\, dt = \pi \sum_{j=-\infty}^{+\infty} e^{-|\omega + 2\pi j|}, \end{aligned}$$

where III denotes the Dirac comb. Then, in the interval $[0, 2\pi]$,

$$q(\omega) = \frac{\pi e^{2\pi}}{e^{2\pi}-1} e^{-\omega} + \frac{\pi}{e^{2\pi}-1} e^{\omega}.$$

Therefore,

$$q_{\max} = q(0) = \pi \frac{e^{2\pi}+1}{e^{2\pi}-1} \quad \text{and} \quad q_{\min} = q(\pi) = \frac{2\pi e^{\pi}}{e^{2\pi}-1}.$$

In both examples, q is positive and hence A is positive definite. We use the square root function $f(x) = \sqrt{x}$ to demonstrate numerical results of the following eight quantities:

(a) $\mathrm{tr}(f(A))$ (b) $\dfrac{n}{2\pi}\displaystyle\int f(q(\omega))$ (c) h_0, exact f (d) h_0, using p

(e) $\sqrt{\dfrac{2}{N}\mathrm{tr}(f^2(A))}$ (f) $\sqrt{\dfrac{n}{N\pi}\displaystyle\int f^2(q(\omega))}$ (g) $\sqrt{h_1}$, exact f (h) $\sqrt{h_1}$, using p.

TABLE 4.1

Computational results for Toeplitz matrices defined in (4.3) *and* (4.4). *Parameters:* $n = 1000$, $N = 100$, $\alpha = 1$, $\delta = 0.1$. *Tolerance on relative approximation error computed from* (4.1)*:* `rtol = 4.26e-03`.

Case: exponential decay, $\theta = 0.7$, interpolation interval $[0.176, 5.667]$, polynomial degree $k = 10$, relative approximation error = `3.40e-03`

	Truth	Approx. Truth	Estim. (exact f)	Estim. (`rtol`)
$\mathrm{tr}(A^{1/2})$	`839.299`	`839.122`	`837.889`	`837.944`
`stddev(estim)`	`4.472`	`4.472`	`4.464`	`4.464`

Case: algebraic decay, interpolation interval $[0.272, 3.154]$, polynomial degree $k = 6$, relative approximation error = `2.35e-03`

	Truth	Approx. Truth	Estim. (exact f)	Estim. (`rtol`)
$\mathrm{tr}(A^{1/2})$	`930.241`	`930.169`	`929.038`	`929.046`
`stddev(estim)`	`4.472`	`4.472`	`4.465`	`4.465`

See Table 4.1.

Let us recall the meaning of these quantities. Term (a) is the truth. Term (c) is the estimate, with variance in Term (e). Terms (b) and (f) are approximations of Terms (a) and (e), respectively, based on equal distributions at large n. In the context of this paper, we approximate Term (c) by using Term (d), where the approximation $p \approx f$ achieves a relative error dictated by (4.1) of Theorem 4.1; and Term (g) is the estimator of Term (e). Moreover, we approximate Term (g) by using Term (h), a byproduct of the calculation of Term (d).

Because the matrix A is well conditioned (see the values of $q_{\min}$ and $q_{\max}$ previously analyzed), the square root function f can be well approximated by using the simple Chebyshev interpolation. Let k be the degree of the interpolating polynomial p that interpolates f at the (shifted) Chebyshev nodes in the interval $[q_{\min}, q_{\max}]$. These nodes are the roots of the (shifted) Chebyshev polynomial T_{k+1} of degree $k+1$. Then, the relative approximation error $\max |1 - p/f|$, which is required in (4.1), can be well approximated by the maximum of $|1 - p/f|$ evaluated at the extrema of T_{k+1}, because these nodes interleave with the Chebyshev nodes.

When the sample size N is 100, the condition (4.1) indicates that we need only a relative error `4.26e-03`. In such a setting, the results in Table 4.1 indicate that the estimates/approximations yield a relative error between 10^{-2} and 10^{-3}. The absolute errors are all within the standard deviation.

5. Further numerical examples in electronic structures. In this section, we show further computational results by using the PARSEC collection of matrices arising from density functional theory. The function f is a simple scaling and shift of the Fermi–Dirac function f_{FD} in (1.1):

$$f(x) = 2f_{\mathrm{FD}}(x) - 1 = \frac{1 - \exp[\beta(x-\mu)]}{1 + \exp[\beta(x-\mu)]}, \qquad \beta = (kT)^{-1} > 0. \tag{5.1}$$

The reason of performing this transformation is that f_{FD} approaches 0 as $x \to \infty$. Then, the relative approximation error is hard to control were f_{FD} used as the function. According to the remark of Theorem 4.1, we perform the transformation to resolve this challenge. Clearly, the function f is nothing but the negative hyperbolic tangent: $f(x) = -\tanh[\beta(x-\mu)/2]$.

We set β to be a large number 100 so that f appears close to the negative sign function. We set the chemical potential μ to lie at 1/3 of the spectrum interval, and scale the shifted matrix $A - \mu I$ such that its spectral radius is 1. Hence, function approximation is carried out in the unit interval $[-1, 1]$. The extreme eigenvalues of A are computed by using the Lanczos method. Numerical results are shown in Table 5.1.

TABLE 5.1
Computational results for the PARSEC collection of matrices. The function f is defined in (5.1). *Parameters:* $N = 1000$, $\alpha = 1$, $\delta = 0.1$.

Matrix	n	rtol	Degree	h_0	$\sqrt{h_1}$
Si2	769	1.54e-03	241	-269.096	1.210
SiH4	5,041	6.01e-04	273	-2049.98	3.09
SiNa	5,743	5.63e-04	275	-2324.90	3.30
Na5	5,832	5.58e-04	275	-2006.35	3.32
benzene	8,219	4.70e-04	281	-2817.46	3.94
Si10H16	17,077	3.26e-04	291	-6694.64	5.70
Si5H12	19,896	3.02e-04	295	-7290.60	6.14
SiO	33,401	2.33e-04	303	-12601.3	7.9
Ga3As3H12	61,349	1.72e-04	313	61321.1	11.0
GaAsH6	61,349	1.72e-04	313	61334.8	11.0
H2O	67,024	1.65e-04	315	-22269.1	11.2
Si34H36	97,569	1.37e-04	321	-36723.1	13.6
Ge87H76	112,985	1.27e-04	323	-43704.8	14.6
Ge99H100	112,985	1.27e-04	323	-43371.0	14.6
Ga10As10H30	113,081	1.27e-04	323	113020.	15.
Ga19As19H42	133,123	1.17e-04	327	133034.	16.
SiO2	155,331	1.08e-04	329	-52984.2	17.1
Si41Ge41H72	185,639	9.90e-05	333	-67395.9	18.7
CO	221,119	9.07e-05	335	-81894.4	20.4
Si87H76	240,369	8.70e-05	337	-89993.4	21.3
Ga41As41H72	268,096	8.24e-05	339	267859.	23.

Because of the steep slope of f at the origin, Chebyshev interpolation is no longer the best choice. We use the approach proposed by Chen et al. [12] to perform the spline/polynomial approximation. In this approach, f is first approximated by a cubic spline, where the knots are at geometrically progressing locations $\pm\theta^k$ together with the origin (for Table 5.1 we set $\theta = 9/10$ and $k = 0, 1, \ldots, 99$). Then, the spline is in turn approximated by a least squares polynomial, where the inner product is defined as the sum of the $(1-x^2)^{-\frac{1}{2}}$-weighted inner products in each subinterval. The relative approximation error is approximated by the maximum of $|1 - p/f|$ evaluated at the mid-points of the spline knots. Under such a scheme, the degree of the polynomial p which satisfies the relative error tolerance dictated by (4.1) is on the order of several hundreds, as listed in Table 5.1.

Unlike the examples in the preceding sections, where a number of quantities can be computed because of the known expressions and the small size of the matrix, here we compute only the (approximate) h_0 and h_1. We increase the sample size N to $1,000$ for a more accurate estimation. One observes from the table that overall, the estimates are two to four digits accurate and the accuracy improves when the matrix

size becomes larger.

6. Concluding remarks. Computing the trace of a large, implicit matrix M has diverse interesting applications in scientific computing. The Hutchinson trace estimator is a matrix-free approach that makes use of efficient M-vector multiplications to remedy the expensive cost of the explicit construction of M. We have studied the effect of the numerical error in the evaluation of M-vector products. In particular, we derive conditions (3.1) and (4.1) that ensure that the numerical error is comparable to the uncertainty of the stochastic estimation. These conditions are readily applicable as a computational guidance for the approximate evaluation of M-vector products. Several examples with special matrices and matrices from an application demonstrate the effective use of the conditions. Particularly, in the experiments with the PARSEC collection of matrices from density functional theory, we observe that the relative error of the estimation decreases as the matrix size n increases, a qualitative agreement with the theoretical order $\Theta(n^{-\frac{1}{2}}N^{-\frac{1}{2}})$ for model matrices with structural decay.

Many additional methods and variants exist for the trace computation. Avron and Toledo [2] proposed and analyzed several distributions where the random vectors are drawn. The effectiveness of the resulting estimators, as long as they are unbiased, may be measured by the estimator variance, because the central limit theorem ensures that confidence intervals can be established by treating the sample average approximately Gaussian for large N. In that vein, variance reduction is a valuable resort. For example, Stein et al. [27] proposed using dependent random vectors in groups so that the variance of any diagonal block of M corresponding to the group is eliminated. The grouping of the rows and columns of M, in this case, respects the closeness of spatial data in order that the eliminated variance majorizes the remained covariance in the off-diagonal blocks. Interestingly, an opposite approach for grouping, coined "probing," was proposed as well. In this approach, rows and columns far apart are grouped together. The rationale is simple. Consider, for the moment, that M is tridiagonal. It suffices to use three deterministic vectors

$$y_1 = \sum_{i=0}^{\lfloor (n-1)/3 \rfloor} e_{3i+1}, \quad y_2 = \sum_{i=0}^{\lfloor (n-2)/3 \rfloor} e_{3i+2}, \quad y_3 = \sum_{i=0}^{\lfloor (n-3)/3 \rfloor} e_{3i+3}$$

to exactly recover the trace:

$$\mathrm{tr}(M) = y_1^T M y_1 + y_2^T M y_2 + y_3^T M y_3,$$

because $e_i^T M e_j$ vanishes whenever $|i-j| \geq 3$. In other words, every other three columns (and rows) of M are grouped together. Then, the diagonal blocks of the permuted matrix is diagonal. Hence, each deterministic vector is used to compute the trace of one block. Such a technique can be generalized for a sparse matrix by coloring the graph representation of the matrix, so that no two same-colored nodes share neighbors [5]. In practice, however, the sparsity pattern of an implicit matrix M is unknown; but if M is the inverse of A whose sparsity is given, it is often a reasonable assumption that the magnitude of M_{ij} decreases as the distance between nodes i and j in the graph of A increases. Hence, the heuristic is to group graph nodes that are a certain distance apart [26]. In all these methods, if the variance of the resulting estimator is formulated, the technique for deriving computational conditions similar to those in this paper is possibly transferable.

REFERENCES

[1] M. Anitescu, J. Chen, and L. Wang, *A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem*, SIAM J. Sci. Comput., 34 (2012), pp. A240–A262.
[2] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. Assoc. Comput. Mach., 58 (2011).
[3] Z. Bai, G. Fahey, and G. Golub, *Some large-scale matrix computation problems*, J. Comput. Appl. Math., 74 (1996), pp. 71–89.
[4] C. Bekas, A. Curioni, and I. Fedulova, *Low cost high performance uncertainty quantification*, in Proceedings of the 2nd Workshop on High Performance Computational Finance, 2009.
[5] C. Bekas, E. Kokiopoulou, and Y. Saad, *An estimator for the diagonal of a matrix*, Appl. Numer. Math., 57 (2007), pp. 1214–1229.
[6] M. Benzi, P. Boito, and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Rev., 55 (2013), pp. 3–64.
[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[8] G. Casella and R. L. Berger, *Statistical Inference*, Duxbury Press, 2nd ed., 2001.
[9] J. Chelikowsky, *The pseudopotential-density functional method applied to nanostructures*, J. Phys. D: Appl. Phys., 33 (2000), pp. R33–R50.
[10] J. R. Chelikowsky, N. Troullier, and Y. Saad, *Finite-difference-pseudopotential method: Electronic structure calculations without a basis*, Phys. Rev. Lett., 72 (1994), pp. 1240–1243.
[11] J. Chen, *On the use of discrete Laplace operator for preconditioning kernel matrices*, SIAM J. Sci. Comput., 36 (2013), pp. A289–A309.
[12] J. Chen, M. Anitescu, and Y. Saad, *Computing $f(A)b$ via least squares polynomial approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 195–222.
[13] J. Chen, T. L. H. Li, and M. Anitescu, *A parallel linear solver for multilevel Toeplitz systems with possibly several right-hand sides*, Parallel Comput., 40 (2014), pp. 408–424.
[14] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures & Algorithms, 22 (2003), pp. 60–65.
[15] D. Girard, *Un algorithme simple et rapide pour la validation croisé généralisée sur des problèmes de grande taille*, Tech. Rep. RR 669-M, Inf. et Math. Appl. de Grenoble, Grenoble, France, 1987.
[16] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*, Now Publishers Inc, 2006.
[17] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, 2008.
[18] M. F. Hutchinson, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Communications in Statistics – Simulation and Computation, 19 (1990), pp. 433–450.
[19] V. Kalantzis, C. Bekas, A. Curioni, and E. Gallopoulos, *Accelerating data uncertainty quantification by solving linear systems with multiple right-hand sides*, Numerical Algorithms, 62 (2013), pp. 637–653.
[20] P. Li, T. J. Hastie, and K. W. Church, *Nonlinear estimators and tail bounds for dimension reduction in l_1 using Cauchy random projections*, in Learning Theory, vol. 4539 of Lecture Notes in Computer Science, 2007, Springer Berlin Heidelberg, pp. 514–529.
[21] L. Lin, *Randomized estimation of spectral densities of large matrices made accurate*. arXiv:1504.07690 [math.NA], 2015.
[22] L. Lin, Y. Saad, and C. Yang, *Approximating spectral densities of large matrices*, SIAM Rev., (in press).
[23] D. P. O'Leary, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.
[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 2nd ed., 2003.
[25] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, *Numerical methods for electronic structure calculations of materials*, SIAM Rev., 52 (2010), pp. 3–54.
[26] A. Stathopoulos, J. Laeuchli, and K. Orginos, *Hierarchical probing for estimating the trace of the matrix inverse on toroidal lattices*, SIAM J. Sci. Comput., 35 (2013), pp. S299–S322.
[27] M. L. Stein, J. Chen, and M. Anitescu, *Stochastic approximation of score functions for Gaussian processes*, Annals of Applied Statistics, 7 (2013), pp. 1162–1191.
[28] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.