# Optimal Query Complexity for Estimating the Trace of a Matrix

Karl Wimmer[1,⋆], Yi Wu[2,⋆⋆], and Peng Zhang[2]

[1] Duquesne University
[2] Purdue University

**Abstract.** Given an implicit $n \times n$ matrix $A$ with oracle access $x^T Ax$ for any $x \in \mathbb{R}^n$, we study the query complexity of randomized algorithms for estimating the *trace* of the matrix. This problem has many applications in quantum physics, machine learning, and pattern matching. Two metrics are commonly used for evaluating the estimators: i) variance; ii) a high probability multiplicative-approximation guarantee. Almost all the known estimators are of the form $\frac{1}{k}\sum_{i=1}^{k} x_i^T Ax_i$ for $x_i \in \mathbb{R}^n$ being i.i.d. for some special distribution.

Our main results are summarized as follows:

1. We give an *exact* characterization of the minimum variance unbiased estimator in the broad class of *linear nonadaptive* estimators (which subsumes all the existing known estimators).
2. We also consider the query complexity lower bounds for *any* (possibly nonlinear and adaptive) estimators:
   (a) We show that *any* estimator requires $\Omega(1/\epsilon)$ queries to have a guarantee of variance at most $\epsilon$.
   (b) We show that *any* estimator requires $\Omega(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ to achieve a $(1 \pm \epsilon)$-multiplicative approximation guarantee with probability at least $1 - \delta$.

   Both above lower bounds are asymptotically tight.

As a corollary, we also resolve a conjecture in the seminal work of Avron and Toledo (Journal of the ACM 2011) regarding the sample complexity of the Gaussian Estimator.

## 1 Introduction

Given an $n \times n$ matrix $A = \{A_{ij}\}_{1\le i\le n, 1\le j\le n}$, we study the problem of estimating its trace

$$trace(A) = \sum_{i=1}^{n} A_{ii}$$

⋆ Most of the work is done when the author is visiting the Simons Institute for the Theory of Computing, University of California-Berkeley. Supported in part by NSF grant CCF-1117079.

⋆⋆ Most of the work is done when the author is visiting the Simons Institute for the Theory of Computing, University of California-Berkeley.

J. Esparza et al. (Eds.): ICALP 2014, Part I, LNCS 8572, pp. 1051–1062, 2014.

with a randomized algorithm that can query $f_A(x) = x^T Ax$ for any $x \in \mathbb{R}^n$. The goal is to minimize the number of queries used to achieve certain type of accuracy guarantee, such as the variance of the estimate or a multiplicative approximation (which holds with high probability). Finding an estimator that achieves such an accuracy guarantee with few queries has several applications. For example, this problem is well studied in the subject of lattice quantum chromodynamics, since such queries are physically feasible and can be used to efficiently estimate the trace of a function of a large matrix $f(A)$. Such an estimator can also be used as a building block for many other applications including solving least-squares problems [Hut89], computing the number of triangles in a graph [Avr10, Tso08], and string pattern matching [ACD01, AGW13].

This problem has been well studied in the literature. All of the previously analyzed estimators are of the form $\frac{1}{k}\sum_{i=1}^{k} x_i^T Ax_i$ for $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$; nearly all take $x_1, x_2, \ldots, x_k$ to be independent and identically distributed (i.i.d.) from some well designed distribution. For example, in [Hut89], the author just takes each query to be a random vector whose entries are i.i.d. Rademacher random variables (i.e., each coordinate is a uniformly random sample from $\{-1, 1\}$); we call this the Rademacher estimator. There are also several other alternative distributions on $x_1, x_2, \ldots, x_k$, such as drawing each query from a multivariate normal distribution [SR97]; we call this the Gaussian estimator. Here, the coordinates of each vectors are i.i.d. Gaussian random variables. The work of [IE04] considers the case where only one query is allowed, but that query can be a unit vector in $\mathbb{C}^n$. Other estimators occur in [DS93, Wan94]. Recent work by [AT11], the authors propose several new estimators such as the unit vector estimator, normalized Rayleigh-quotient trace estimator, and the mixed unit vector estimator. One estimator that does not use i.i.d. queries is due to [RKA13]; in that work, the authors propose querying random standard basis vectors without replacement.

To characterize the performance of an estimator, perhaps the most natural metric is the variance of the estimator. It is known that the Gaussian estimator has variance $2\|A\|_F^2$ and the random Rademacher vector estimator has variance $2(\|A\|_F^2 - \sum_{i=1}^{n} A_{ii}^2)$, where $\|A\|_F = \sqrt{trace(A^T A)}$ is the Frobenius norm. In recent work by Avron and Toledo [AT11], it is suggested that the notion of a multiplicative approximation guarantee might be a better success metric of an estimator than the variance. Formally, we say an estimator is an $(\epsilon, \delta)$-estimator if it outputs an estimate in the interval $((1-\epsilon)trace(A), (1+\epsilon)trace(A))$ with probability at least $1-\delta$. It should be noted that some assumptions on the matrices need to be made to have a valid $(\epsilon, \delta)$-estimator, as it is impossible to achieve any multiplicative approximation when the matrix could have a trace of 0. A natural choice is to assume that $A$ comes from the class of symmetric positive semidefinite (SPD) matrices. For a SPD matrix, the authors in [AT11] prove that the Gaussian estimator with $k = O(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ queries to the oracle is an $(\epsilon, \delta)$-estimator. It was recently shown in [RKA13] that the random Rademacher vector estimator is also an $(\epsilon, \delta)$-estimator with the same sample complexity.

An open problem asked in [AT11] is the following: does the Gaussian estimator require $\Omega(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ in order to be an $(\epsilon,\delta)$-estimator? The authors showed that this number of queries suffices and conjectured that their analysis of the Gaussian estimator is tight with supporting evidence from empirical experiments. The paper gives some intuition on how to show an $\Omega(\frac{1}{\epsilon^2})$ lower bound. The authors suggested that the difficulty of turning this argument into a formal proof is that "current bounds [on the $\chi^2$ cumulative distribution function] are too complex to provide a useful lower bound". Regarding lower bounds for trace estimators, we note the related work of [LNW14], which considers the problem of sketching the nuclear norms of $A$ using bilinear sketches (which can be viewed as nonadaptive queries of the form $x^T Ay$). The problem is similar to estimating trace when the underlying matrix is positive semidefinite.

All of the above mentioned estimators (with one exception in [RKA13]) use independent identically distributed queries from some special distributions, and the output is a linear combination of the query results. On the other hand, we view an estimator as a randomized algorithm, so we can choose any distribution over the queries, and the output can be any (possibly randomized) function of the results of the queries. Given the success of the previously mentioned estimators, it is natural to ask whether these extensions are helpful. For example, can we get a significantly better estimator with a non i.i.d. distribution? Can we do better with adaptive queries? Can we do better with a nonlinear combination of the query results?

In this paper, we make progress on answering above questions and understanding the optimal query complexity for randomized trace estimators. Below is an informal summary of our results.

1. Among all the linear nonadaptive trace estimators (which subsumes all the existing trace estimators), we prove that the "random $k$ orthogonal vector" estimator is the minimum variance estimator. The distribution on the queries is *not* i.i.d., and we are unable to find an occurrence of this estimator in the literature regarding trace estimators.
2. We also prove two asymptotically optimal lower bounds for any (possibly adaptive and possibly nonlinear) estimator.
   (a) We show that *every* trace estimator requires $\Omega(1/\epsilon)$ queries to have a guarantee that the variance of the estimator is at most $\epsilon$.
   (b) We show that *every* $(\epsilon,\delta)$-estimator requires $\Omega(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ queries.

As a simple corollary, our result also confirms the above mentioned conjecture in [AT11] (as well as the tightness of the analysis of the Rademacher estimator in [RKA13]). Notice our result is a much stronger statement: the original conjectured lower bound is only for an estimator that returns a linear combination of i.i.d. Gaussian queries; we prove the lower bound holds for any estimator. Our lower bound also suggests that adaptiveness as well as nonlinearity will not help asymptotically as all these lower bounds are matched by the nonadaptive Gaussian estimator. On the other hand, our upper bound suggests that the exact minimum variance estimator might not use i.i.d. queries.

### 1.1 Problem Definitions

**Definition 1 (Estimator for the Trace).** *A trace estimator is a possibly randomized algorithm that, given query access to an oracle $f_A(\cdot)$ for an unknown $n \times n$ matrix $A$, makes a sequence of $k$ queries $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ to the oracle and receives $f_A(x_1), f_A(x_2), \ldots, f_A(x_k)$. The output of the estimator is a real number $h(A)$ determined by the queries and the answers to the queries (and possibly uses randomness).*

**Definition 2 (Nonadaptive Linear Unbiased Trace Estimator).** *We say a trace estimator is* nonadaptive *if the distribution of $x_i$ is not dependent on $f_A(x_1), f_A(x_2), \ldots, f_A(x_{i-1})$. A trace estimator is* linear *if we sample from a distribution over $k$ queries as well as their weights: $(x_1, x_2, \ldots, x_k)$, and $(w_1, w_2, \ldots, w_k)$, and output $\sum w_i f_A(x_i)$. In addition, a linear trace estimator is* unbiased *if*

$$\mathop{\mathbf{E}}_{w_1,w_2,\ldots,w_n,x_1,x_2,\ldots,x_n}[\sum_{i=1}^{n} w_i f_A(x_i)] = trace(A)$$

Without loss of generality, we can assume that all the queries in a linear estimator are of unit length, where the actual lengths of the queries are absorbed by the weights.

The most natural measure of quality of an estimator is its variance. There is a large body of work on the existence of and finding a *minimum variance unbiased estimator*. Such an estimator has a strong guarantee; it is the estimator for which the variance is minimized for all possible values of the parameter to estimate. In general, finding such an estimator is quite difficult. It is easy to see that the variance depends on the scale of the matrix. To normalize, we assume that the Frobenius norm of the matrix is fixed.

**Definition 3.** *We define the variance of a trace estimator as the worst case of variance over all matrices with Frobenius norm 1. To be specific, given a matrix $A$ let us define $Var(A,h) = \mathbf{E}[(h(A) - trace(A))^2]$, then*

$$Var(h) = \sup_{\|A\|_F^2 = 1} Var(A,h).$$

*If the variance of an estimator $h$ is at most $\delta$, we say that $h$ is a $\delta$-variance estimator.*

Given an unbiased estimator class, the *minimum variance unbiased estimator* has the minimum variance among all the (unbiased) estimators in the class.

Another natural accuracy guarantee for a trace estimator is the notion of $(\epsilon, \delta)$-estimator that is introduced in [AT11].

**Definition 4 ($(\epsilon, \delta)$-estimator).** *A trace estimator $h$ is said to be an $(\epsilon, \delta)$-estimator of the trace if, for every matrix $A$, we have that $|trace(A) - h(A)| \leq \epsilon \cdot trace(A)$ with probability at least $1 - \delta$.*

We stress that both Definitions 3 and 4 involve worst case estimates over the choice of the matrix, and the randomness only comes from the internal randomness of the estimator.

**Definition 5 (Random Gaussian Matrix and Random Orthogonal Matrix).**

- *We call a vector $g \in \mathbb{R}^n$ a random Gaussian vector if each coordinate is sampled independently from $N(0,1)$.*
- *We call an $n \times n$ matrix $G$ a random Gaussian matrix if its entries are sampled independently from $N(0,1)$.*
- *We call an $n \times n$ matrix $U$ a random orthogonal matrix if it is drawn from the distribution whose probability measure is the Haar measure on the group of orthogonal matrices; specifically, it is the unique probability measure that is invariant under orthogonal transformations.*
- *We call $k$ vectors $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ $k$ random orthogonal unit vectors if they are chosen as $k$ row vectors of a random orthogonal matrix.*

We note that one way to generate a random orthogonal matrix is to generate a random Gaussian matrix and perform Gram-Schmidt orthonormalization on its rows.

## 1.2 Main Results

Our main results are the following:

**Theorem 1.** *Among all linear nonadaptive unbiased trace estimators, the minimum variance unbiased estimator that makes $k$ queries is achieved by sampling $k$ random orthogonal unit vectors (see Definition 5) $x_1, x_2, \ldots, x_k$ and outputting $\frac{n}{k}\sum_i f_A(x_i)$.*

**Theorem 2.** *Any trace estimator with variance $\epsilon$ requires $\Omega(1/\epsilon)$ queries.*

**Theorem 3.** *Any $(\epsilon, \delta)$-estimator for the trace requires $\Omega(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ queries, even if the unknown matrix is known to be positive semidefinite.*

The bounds in Theorem 2 and 3 are tight: both bounds can be asymptotically matched by the Gaussian estimator and the uniform Rademacher vector estimator.

## 1.3 Proof Techniques Overview

All of our results crucially use a powerful yet simple trick, which we call *symmetrization*. The heart of this trick lies in the fact that the trace of a matrix is unchanged under similarity transformations; $\mathit{trace}(A) = \mathit{trace}(U^T AU)$ for every $A$ and orthogonal $U$. If we have a nonadaptive estimator with query distribution $(x_1, x_2, \ldots, x_k) \sim P$ and an orthogonal matrix $U$, using the queries distributed as $(Ux_1, Ux_2, \ldots, Ux_k)$ should not be too different in terms of worst-case behavior. (We have to be more careful with adaptive estimators, which we discuss in Section 2.) Thus, applying symmetrization to a nonadaptive estimator yields a nonadaptive estimator where it draws queries as in the original estimator, but transforms the queries using a random orthogonal transformation. This "symmetrizes" the estimator. We prove that the performance of the estimator never decreases when symmetrization is applied, so we can exclusively consider symmetrized estimators.

In order to characterize the minimum variance linear nonadaptive unbiased estimator, we notice that after the symmetrization, the distribution over queries for any such estimator is defined by a distribution over the pairwise angles of the $k$ queries. We then show that the queries should be taken to be orthogonal with certainty in order to minimize variance.

As for the lower bounds for adaptive and nonlinear estimators, the symmetrization also plays an important role. Consider the problem of proving a query lower bound for $(\epsilon, \delta)$-approximation: the most common approach of proving such a lower bound is to use Yao's minimax principle. To apply this principle, we would need to construct two distributions of matrices such that the distributions cannot be distinguished after making a number of queries, even though the traces of the matrices are very different in the two distributions. There are several technical difficulties in applying the minimax principle directly here. First of all, the query space is $\mathbb{R}^n$, so it is unclear whether one can assert that there exists a sufficiently generalized minimax principle to handle this case. Second, even if one can apply a suitable version of minimax principle, we do not have general techniques of analyzing the distribution of $k$ adaptive queries, especially when the queries involve real numbers and thus the algorithm might have infinitely many branches.

We overcome the above two barriers and avoid using a minimax principle entirely by applying symmetrization. One nice property of the symmetrization process is that a symmetrized estimator outputs the same distribution of results on all matrices with the same diagonalization. In the proof we carefully construct two distributions of matrices with the same diagonalization in each distribution, while the traces are different for different distributions. Each distribution is simply the "orbit" of a single diagonal matrix $D$; the support consists of all matrices similar to $D$. Using the symmetrization, it suffices to show that we can not distinguish these two distributions of matrices by $k$ adaptive queries, as it is equivalent to distinguish two diagonal matrices for symmetrized trace estimators. The argument for achieving a lower bound for adaptive estimators is more subtle. Roughly speaking, we show that due to the structure of symmetrized estimators, we define a stronger query model such that adaptive estimators behave the same as the nonadaptive estimators while we achieve the same lower bound, even with the stronger query model.

### 1.4 Organization

In section 2, we introduce the idea of symmetrization. We prove Theorem 1 in section 3. Due to the space constraint, proofs of Theorem 2 and Theorem 3 are omitted, and they are included in our full version.

## 2 Symmetrization of an Estimator

In this section, we introduce the idea of symmetrization of an estimator which is a crucial element of all our remaining proofs. We first define the *rotation* of an estimator, which we will denote $h^U$ for an $n \times n$ orthogonal matrix $U$. Intuitively, the construction of $h^U$ is such that $h^U$ emulates the behavior of $h$ on a rotated version of the matrix $A$. More specifically, $h^U$ makes queries in the following way:

- Letting $q_1$ be a random variable whose distribution is the same as the first query of $h$, the distribution of the first query of $h^U$ is the same as the random variable $Uq_1$.
- Given queries $Uq_1, Uq_2, \ldots, Uq_{j-1}$ made by $h^U$ so far with responses $t_1, t_2, \ldots, t_{j-1}$, the distribution of the $j$th query of $h^U$ has the same distribution as $Uq_j$, where $q_j$ is distributed the same as the $j$th query that $h$ makes, given queries $q_1, q_2, \ldots, q_{j-1}$ with responses $t_1, t_2, \ldots, t_{j-1}$.

In the case that $h$ is a nonadaptive estimator, the queries of $h^U$ are just $Ux_1, Ux_2, \ldots, Ux_k$, where $x_1, x_2, \ldots, x_k$ is a set of queries from the distribution of queries that $h$ makes.

**Lemma 1.** *For any estimator $h$ and orthogonal matrix $U$,*

- *$Var(h^U) = Var(h)$.*
- *$h$ is an $(\epsilon, \delta)$-approximation estimator if and only if $h^U$ is also an $(\epsilon, \delta)$-estimator.*

*Proof.* We know that given a matrix $A$, the behavior of $h^U$ is the same as $h$ on estimating $U^T AU$. On the other hand, we know that $trace(U^T AU) = trace(A)$ and $\|A\|_F = \|U^T AU\|_F$. Therefore, the variance of $h^U$ on $A$ is the same as the variance of $h$ on $U^T AU$. Now suppose $h$ is an $(\epsilon, \delta)$-estimator. We know that the approximation guarantee of $h^U$ on $A$ is the same as $h$ on $U^T AU$. Therefore, we know that with probability at least $(1 - \delta)$, the estimator $h^U$'s output is within

$$\left((1-\epsilon)trace(U^T AU), (1+\epsilon)trace(U^T AU)\right) = ((1-\epsilon)trace(A), (1+\epsilon)trace(A)) .$$

**Definition 6 (Averaging Estimators over a Distribution).** *Suppose we have a collection of estimators $H$, for any probability distribution $P$ on $H$, we define $h_{H,P}$ as the following estimator:*

1. *Randomly sample an estimator $h \sim P$.*
2. *Output according to the estimation of $h$.*

**Lemma 2.** *Averaging a collection of estimators cannot increase variance or weaken an $(\epsilon, \delta)$-guarantee. Specifically:*

- *If all the estimators $H$ are unbiased and have variance at most $c$, then $h_{H,P}$'s variance is also at most $c$.*
- *If all the estimators in $H$ are $(\epsilon, \delta)$-estimators, then $h_{H,P}$ is also an $(\epsilon, \delta)$-estimator.*

*Proof.* For the first, we apply the law of total variance conditioned on the draw of $h \sim P$:

$$Var[h_{H,P}] = \mathop{\mathbf{E}}_{h \sim P}[Var[h]] + \mathop{Var}_{h \sim P}[\mathbf{E}[h]]$$

The second term above is 0, since all estimators in $H$ are unbiased. Since $Var[h] \leq c$ for every $h \in H$, $\mathbf{E}_{h \sim P}[Var[h]] \leq c$ as well.

For the second claim, assuming that

$$\mathbf{Pr}[h(A) \in ((1-\epsilon)trace(A), (1+\epsilon)trace(A))] \geq 1 - \delta$$

for each $h \in H$, we have

$$\mathbf{Pr}[h_{H,P}(A) \in ((1-\epsilon)trace(A), (1+\epsilon)trace(A))] \geq$$
$$\inf_{h \in H} \mathbf{Pr}[h(A) \in ((1-\epsilon)trace(A), (1+\epsilon)trace(A))] \geq 1 - \delta.$$

**Definition 7 (Symmetrization of a Trace Estimator).** *Given an estimator $h$, we define the symmetrization $h^{sym}$ of $h$ to be the estimator where we*

1. *sample a random orthogonal matrix $U$ (see definition 5), and*
2. *use $h^U$ to estimate the trace.*

*We say an estimator is* symmetric *if it is equivalent to the symmetrization of some estimator.*

By Lemma 1 and Lemma 2, we know that $h^{sym}$'s variance as well as its $(\epsilon, \delta)$-approximation is always no worse than $h$. Therefore, without loss of generality, we can always assume that the optimal estimator is symmetric.

One nice property of the symmetric estimator is that it has the same performance on all matrices with the same diagonalization.

**Lemma 3.** *Given a symmetrized estimator $h^{sym}$, its variance and approximation guarantee is the same for any matrix $A$ and $U^T A U$ for any orthogonal matrix $U$.*

*Proof.* Given any matrix $A$, we know that the variance of $h^{sym}$ is $\mathbf{E}_{U_1}[Var(h, U_1^T A U_1)]$ and the variance of $h^{sym}$ on the matrix $U^T A U$ is $\mathbf{E}_{U_1}[Var(h, (U_1 U)^T A U_1 U)]$. We know that $U_1 U$ is also is a "uniformly" random orthogonal matrix. Therefore, $h^{sym}$ has the same estimation variance on $A$ and $U^T A U$.

Similarly, for the approximation guarantee, suppose $h^{sym}$ is an $(\epsilon, \delta)$-estimator, which means that

$$\mathop{\mathbf{E}}_{U_1}\left[\mathbf{Pr}\left(h(U_1^T A U_1) \in ((1-\epsilon)trace(A), (1+\epsilon)trace(A))\right)\right] \geq 1-\delta.$$

We know that for the matrix $A' = UAU$, $U_1^T A' U_1 = U_1^T U^T A U U_1$ has the same distribution as $U_1^T A U_1$. Therefore,

$$\begin{aligned}&\mathop{\mathbf{E}}_{U_1}\left[\mathbf{Pr}\left(h(U_1^T A U_1) \in ((1-\epsilon)trace(A), (1+\epsilon)trace(A))\right)\right]\\ &\qquad = \mathop{\mathbf{E}}_{U_1}\left[\mathbf{Pr}\left(h(U_1^T A' U_1) \in ((1-\epsilon)trace(A'), (1+\epsilon)trace(A'))\right)\right]\end{aligned}$$

## 3 Optimal Linear Nonadaptive Estimator

Without loss of generality, we can assume that the optimal estimator is symmetric. For a symmetric nonadaptive estimator, we can think of $x_1, x_2, \ldots, x_k$ as generated by the following process.

- Sample a configuration $\theta = \{\theta_{ij}\}_{1\leq i<j\leq k}$ from some distribution $P_\Theta$ . For each configuration $\theta$, there is a corresponding weight vector $w^\theta = (w_1^\theta, w_2^\theta, \ldots, w_k^\theta)$.
- Generate $x_1, x_2, \ldots, x_k$ by drawing $k$ random unit vectors conditioned on the angle between $x_i, x_j$ being $\theta_{ij}$ for all $i < j$. (This can be done efficiently.)
- Output $\sum_{i=1}^k w_i^\theta f_A(x_i)$.

The proof of Theorem 1 consists of two steps. First we will show that we can set all of the angles (deterministically) to be $\frac{\pi}{2}$ without increasing the variance, so we can assume that the queries are orthogonal. In the second step, we will then show that the optimal way of assigning weight is to (deterministically) set each weight to be $\frac{n}{k}$.

We first prove that we can replace the queries $x_1, x_2, \ldots, x_k$ by $k$ random orthogonal unit vectors without increasing the variance.

**Lemma 4.** *Let $y_1, y_2, \ldots, y_k$ be $k$ randomly orthogonal unit vectors. We have that*

$$Var\left(\sum_{i=1}^k w_i^\theta f_A(y_i)\right) \leq Var\left(\sum_{i=1}^k w_i^\theta f_A(x_i)\right) \tag{1}$$

*Proof.* It is easy to see that the marginal distribution on each $x_i$ is the same as the marginal distribution on $y_i$. Therefore, we have that

$$\mathop{\mathbf{E}}_{\theta, y_1, \ldots, y_k}\left[\sum_{i=1}^k w_i^\theta f_A(y_i)\right] = \mathop{\mathbf{E}}_{x_1, \ldots, x_k, \theta}\left[\sum_{i=1}^k w_i^\theta f_A(x_i)\right] = trace(A)$$

This implies that $\sum_{i=1}^k w_i^\theta f_A(y_i)$ is also an unbiased estimator.

Since both estimators have the same expectation, in order to show (1), it suffices to prove that

$$\mathop{\mathbf{E}}_{\theta,x_1,\ldots,x_n}\left[\left(\sum_{i=1}^{k} w_i^\theta f_A(x_i)\right)^2\right] \geq \mathop{\mathbf{E}}_{\theta,y_1,\ldots,y_n}\left[\left(\sum_{i=1}^{k} w_i^\theta f_A(y_i)\right)^2\right] \tag{2}$$

By the process of generating $x_1, x_2, \ldots, x_k$, we know that the marginal distribution of $x_i$ is independent of $\theta$ and equal to the marginal distribution of $y_i$. Therefore, in order to prove (2), it suffices to prove that for any $i$ and $j$, we have

$$\mathop{\mathbf{E}}_{\theta,x_i,x_j}[w_i^\theta w_j^\theta f_A(x_i) f_A(x_j)] \geq \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta] \mathop{\mathbf{E}}_{y_i,y_j}[f_A(y_i) f_A(y_j)] \tag{3}$$

To compare $\mathbf{E}_{\theta,x_i,x_j}[w_i^\theta w_j^\theta f_A(x_i) f_A(x_j)]$ and $\mathbf{E}_\theta[w_i^\theta w_j^\theta]\,\mathbf{E}_{y_i,y_j}[f_A(y_i) f_A(y_j)]$, we note that the marginal distribution on the pair $(x_i, x_j)$ is equivalent to drawing $x_i, x_j$ from the following process:

1. Draw $\theta \sim P_\Theta$.
2. Set $x_i = y_i$ and $x_j = y_i \cos\theta_{ij} + y_j \sin\theta_{ij}$

It is easy to check that the joint distribution on $x_i$ and $x_j$ has the same distribution as two random unit vectors with angle $\theta_{ij}$.

Therefore,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\theta,x_i,x_j}[w_i^\theta w_j^\theta f_A(x_i) f_A(x_j)] \\
&= \mathop{\mathbf{E}}_{\theta,y_i,y_j}[w_i^\theta w_j^\theta y_i^T A y_i (\cos\theta_{ij}\cdot y_i + \sin\theta_{ij}\cdot y_j)^T A(\cos\theta_{ij}\cdot y_i + \sin\theta_{ij}\cdot y_j)] \\
&= \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta \cos^2\theta_{ij}] \mathop{\mathbf{E}}_{y_i}[y_i^T A y_i \cdot y_i^T A y_i] + \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta \sin^2\theta_{ij}] \mathop{\mathbf{E}}_{y_i,y_j}[y_i^T A y_i y_j^T A y_j] \\
&+ \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta \sin\theta_{ij}\cos\theta_{ij}] \mathop{\mathbf{E}}_{y_i,y_j}[y_i^T A y_i y_i^T A y_j + y_i^T A y_i y_j^T A y_i]
\end{aligned} \tag{4}$$

In order to simplify the above expression, we first claim that

$$\mathop{\mathbf{E}}_{y_i,y_j}[y_i^T A y_i y_i^T A y_j + y_i^T A y_i y_j^T A y_i] = 0.$$

To see this, note that $y_j$ is a random unit vector orthogonal to $y_i$. Although $y_i, y_j$ are dependent, conditioned on any fixed realization of $y_i$, the distribution on $y_j$ is symmetric about $\mathbf{0}$; $y_j$ has the same distribution as $-y_j$.

In addition, using Cauchy-Schwarz and the fact that $y_i$ and $y_j$ have the same distribution, we have that

$$\mathop{\mathbf{E}}_{y_i}[(y_i^T A y_i)^2] = \sqrt{\mathop{\mathbf{E}}_{y_i}[(y_i^T A y_i)^2] \mathop{\mathbf{E}}_{y_j}[(y_j^T A y_j)^2]} \geq \mathop{\mathbf{E}}_{y_i,y_j}[y_j^T A y_j \cdot y_i^T A y_i].$$

Therefore, we have that

$$\begin{aligned}
(4) &\geq \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta \cos^2\theta] \mathop{\mathbf{E}}_{y_i,y_j}[y_j^T A y_j \cdot y_i^T A y_i] + \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta \sin^2\theta] \mathop{\mathbf{E}}_{y_i,y_j}[y_i^T A y_i y_j^T A y_j] \\
&= \mathop{\mathbf{E}}_{\theta}[w_i^\theta w_j^\theta] \cdot \mathop{\mathbf{E}}_{y_i,y_j}[y_j^T A y_j \cdot y_i^T A y_i]
\end{aligned}$$

which proves (3), completing the proof of Lemma 4.

Now that we can assume that the queries are mutually orthogonal, we can view this as an estimator with randomized weights $w_i^\theta$ for $\theta \sim P_\Theta$. Below we will use the random variable $w_i$ to denote $w_i^\theta$ as $\theta$ is independent from $y_1, y_2, \ldots, y_k$.

**Lemma 5.** *Let $(y_1, \ldots, y_k)$ be $k$ random orthogonal unit vectors. Then the estimator $h = \sum_{i=1}^k w_i f_A(y_i)$ has minimum variance when $w_1 = w_2 = \cdots = w_k = n/k$.*

*Proof.* First we must have $\mathbf{E}[\sum_{i=1}^k w_i] = n$ to make the estimator unbiased, since $\mathbf{E}[f_A(y_i)] = \frac{1}{n} trace(A)$. Also,

$$\mathbf{E}\left[\left(\sum_{i=1}^k w_i\right)^2\right] = \mathbf{E}\left[\sum_{i=1}^k w_i^2\right] + 2 \cdot \mathbf{E}\left[\sum_{1\le i<j\le k} w_i w_j\right] \ge n^2$$

Minimizing the variance is equivalent to minimizing

$$\begin{aligned}
&\mathop{\mathbf{E}}_{w,y}\left[\left(\sum_{i=1}^k w_i f_A(y_i)\right)^2\right] \\
&= \sum_{i=1}^k \mathbf{E}[f_A(y_i)^2]\,\mathbf{E}[w_i^2] + 2 \cdot \sum_{1\le i<j\le k} \mathbf{E}[f_A(y_i) f_A(y_j)] \cdot \mathbf{E}[w_i w_j] \\
&= \mathbf{E}\left[\sum_{i=1}^k w_i^2\right] \mathbf{E}[f_A^2(y_1)] + \left(\mathbf{E}\left[\left(\sum_{i=1}^k w_i\right)^2\right] - \mathbf{E}\left[\sum_{i=1}^k w_i^2\right]\right) \mathbf{E}[f_A(y_1) f_A(y_2)] \\
&\ge \frac{n^2}{k}\,\mathbf{E}[f_A(y_1)^2] + \left(n^2 - \frac{n^2}{k}\right) \mathbf{E}[f_A(y_1) f_A(y_2)] \\
&= \mathbf{E}\left[\left(\sum_{i=1}^k \frac{n}{k} f_A(y_i)\right)^2\right]
\end{aligned}$$

Equality holds for $w_1 = \cdots = w_k = n/k$, completing the proof.

Combining Lemma 4 and Lemma 5, the minimum variance linear nonadaptive unbiased estimator making $k$ queries is $\sum_{i=1}^k \frac{n}{k} f_A(y_i)$, where $y_1, y_2, \ldots, y_k$ is a collection of random orthogonal unit vectors. This completes the proof of Theorem 1.

**Acknowledgement.** The second author is grateful for Yi Li and Siu-On Chan for helpful discussions.

## References


[ACD01] Atallah, M.J., Chyzak, F., Dumas, P.: A randomized algorithm for approximate string matching. Algorithmica 29(3), 468–486 (2001)

[AGW13] Atallah, M.J., Grigorescu, E., Wu, Y.: A lower-variance randomized algorithm for approximate string matching. Information Processing Letters 113(18), 690–692 (2013)

[AT11] Avron, H., Toledo, S.: Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. Journal of the ACM (JACM) 58(2), 8 (2011)

[Avr10] Avron, H.: Counting triangles in large graphs using randomized matrix trace estimation. In: Proceedings of KDD-LDMTA, vol. 10 (2010)

[DS93] Drabold, D.A., Sankey, O.F.: Maximum entropy approach for linear scaling in the electronic structure problem. Physical Review Letters 70(23), 3631 (1993)

[Hut89] Hutchinson, M.F.: A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Communications in Statistics-Simulation and Computation 18(3), 1059–1076 (1989)

[IE04] Iitaka, T., Ebisuzaki, T.: Random phase vector for calculating the trace of a large matrix. arXiv preprint cond-mat/0401202 (2004)

[LNW14] Li, Y., Nguyên, H.L., Woodruff, D.P.: On sketching matrix norms and the top singular vector. In: Proceedings of the Twenty Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (2014)

[RKA13] Roosta-Khorasani, F., Ascher, U.: Improved bounds on sample size for implicit matrix trace estimators (2013)

[SR97] Silver, R.N., Röder, H.: Calculation of densities of states and spectral functions by chebyshev recursion and maximum entropy. Physical Review E 56(4), 4822 (1997)

[Tso08] Tsourakakis, C.E.: Fast counting of triangles in large real networks without counting: Algorithms and laws. In: Eighth IEEE International Conference on Data Mining, ICDM 2008, pp. 608–617. IEEE (2008)

[Wan94] Wang, L.-W.: Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method. Physical Review B 49(15), 10154 (1994)