A Tight Analysis of Hutchinson's Diagonal Estimator

Prathamesh Dharangutte* Christopher Musco†

Abstract

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a matrix with diagonal $\text{diag}(\mathbf{A}) \in \mathbb{R}^n$. We show that the simple and practically popular Hutchinson's estimator, run for m trials, returns a diagonal estimate $\tilde{d} \in \mathbb{R}^n$ such that with probability $1 - \delta$,

$$\|\tilde{d} - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{\frac{\log(2/\delta)}{m}}\|\bar{\mathbf{A}}\|_F.$$

Above c is a fixed constant and $\bar{\mathbf{A}}$ equals $\mathbf{A}$ with its diagonal set to zero. This result improves on recent work in [4] by a $\log(n)$ factor, yielding a bound that is independent of the matrix dimension, n. We show a similar bound for variants of Hutchinson's estimator that use non-Rademacher random vectors.

1 Introduction

We give a short and tight analysis of the popular Hutchinson's estimator for approximating the diagonal of a square matrix, $\mathbf{A}$, given only *implicit* matrix-vector multiplication access to the matrix [13, 5].

Definition 1.1. (Hutchinson's Diagonal Estimator) *Let $\mathbf{g}^1, \ldots, \mathbf{g}^m \in \{-1, +1\}^n$ be independent random vectors, each with i.i.d. Rademacher (random ± 1) entries. Hutchinson's diagonal estimator $\mathbf{r}^m(\mathbf{A}) \in \mathbb{R}^n$ is:*

$$\mathbf{r}^m(\mathbf{A}) = \frac{1}{m}\sum_{z=1}^{m} \mathbf{g}^z \odot \mathbf{A}\mathbf{g}^z,$$

where $\mathbf{a} \odot \mathbf{b} \in \mathbb{R}^n$ denotes the Hadamard product (entrywise product) between vectors $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$.

Computing $\mathbf{r}^m(\mathbf{A})$ requires m matrix-vector multiplications with $\mathbf{A}$, and it is not hard to check that it is an unbiased estimator for the diagonal, i.e., $\mathbb{E}[\mathbf{r}^m(\mathbf{A})] = \text{diag}(\mathbf{A})$, where $\text{diag}(\mathbf{A})$ is a vector containing $\mathbf{A}$'s diagonal elements. Hutchinson's diagonal estimator is simple to implement and is widely applied across applications in computational science [2, 15], machine learning [17, 8], and optimization [24, 7]. In these applications, it is used to estimate the diagonals of large Hessian matrices, matrix inverses, and other matrices that are expensive to construct explicitly, but for which matrix-vector multiplications can be implemented quickly (e.g. using backpropagation or an iterative linear system solver).

However, despite its popularity, there has been a lack of theoretical work on Hutchinson's diagonal estimator, and in particular on the question of how large m should be so that $\mathbf{r}^m(\mathbf{A})$ concentrates around its expectation. This is in contrast to the closely related Hutchinson's *trace* estimator, which has been heavily studied and for which a tight analysis is known [18, 6, 16, 23].

Two recent papers do provide bounds for the diagonal estimation problem [12, 4]. The second proves that if $m = O(\log(n/\delta)/\epsilon^2)$, then with probability $1-\delta$, $\|\mathbf{r}^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq \epsilon\|\bar{\mathbf{A}}\|_F$, where $\|\bar{\mathbf{A}}\|_F^2 = \|\mathbf{A}\|_F^2 - \|\text{diag}(\mathbf{A})\|_2^2$ denotes the squared Frobenius norm of $\mathbf{A}$ with its diagonal entries set to 0. Our goal is to tighten the analysis of [4] by removing the $\log(n)$, i.e., to prove that to achieve error $\epsilon\|\bar{\mathbf{A}}\|_F$, just $m = O\left(\log(1/\delta)/\epsilon^2\right)$ matrix-vector products are necessary. Formally, we prove:

Theorem 1.1. (Main Theorm) *For any $\delta \in (0, 1]$ and $m \geq 1$, with probability $1 - \delta$:*

$$\|\mathbf{r}^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{\frac{\log(2/\delta)}{m}}\|\bar{\mathbf{A}}\|_F,$$

where c is an absolute constant independent of $\mathbf{A}$ and all other problem parameters.

*Rutgers University, prathamesh.d@rutgers.edu
†New York University, cmusco@nyu.edu

The dependence on $\log(n)$ in the analysis of [4] arises through the use of a union bound argument: they show that Hutchinson's estimator separately obtains an accurate estimate for each entry of $\mathbf{A}$'s diagonal, and thus $\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2$ can be bounded[1]. A similar $\log(n)$ appeared in early analysis for the trace estimation problem [3] and was later removed [18]. We obtain a comparable improvement through a refined analysis of the stochastic diagonal estimator that relies on a symmetrization argument and techniques for proving vector-valued Bernstein inequalities [25].

We do note that it is possible to obtain a *low probability* result for Hutchinson's estimator which almost matches the bound of Theorem 1.1, but with a costly linear dependence on $1/\delta$. We discuss this result in Section 2.1. Additionally, using this low probability result, the same asymptotic complexity as Theorem 1.1 (with no n dependence, and just a $\log(1/\delta)$ dependence) can be obtained by combining Hutchinson's estimator with a multi-dimensional variant of the "median trick". We discuss this approach in Section 4.1. However, we are mostly interested in analyzing Hutchinson's estimator itself as the method is 1) simpler to implement 2) essentially parameter free (only requires specifying m) and 3) the most widely used diagonal estimator in practice.

We also note that Theorem 1.1 is tight, and the bound cannot be further improved for Hutchinson's estimator. To see that this is the case, consider the matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. We can check that $r^m(\mathbf{A}) = \begin{bmatrix} S/m \\ 0 \end{bmatrix}$ where S is a sum of m independent ± 1 random variables. By the well-known tightness of the Chernoff bound (see e.g. [14]) we will only have that $S/m \leq \epsilon$ with probability $1 - \delta$ if $m = O(\log(1/\delta)/\epsilon^2)$, which matches the upper bound implied by Theorem 1.1. It is possible that a different estimator could improve on Theorem 1.1, either in general or for some classes of matrices. Proving a strong lower bound showing the result is optimal in e.g. the matrix-vector product model of computation is a nice open question [21, 16].

2 Preliminaries

Notation. For a vector $\mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{y}\|_2 = (\sum_{i=1}^n y_i^2)^{1/2}$ denotes the Euclidean norm. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\|\mathbf{A}\|_F = (\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2)^{1/2}$ denotes the Frobenius norm and $\|\mathbf{A}\|_2 = \max_{\mathbf{x} \in \mathbb{R}^m} \|\mathbf{A}\mathbf{x}\|_2 / \|\mathbf{x}\|_2$ denotes the spectral norm. When $\mathbf{A}$ is square, $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n A_{ii}$ denotes the trace. We use c, c', C, etc. to denote absolute constants that are independent of the problem input and all other parameters. The exact value of these constants changes depending on context.

Random Variables. When analyzing random variables, we will make use of the properties of sub-Gaussian and sub-exponential random variables, using the notation of [22]. Formally we define:

Definition 2.1. (Sub-Gaussian Random Variable) *A random variable X is sub-Gaussian with parameter K if we have $\mathbb{E}\left[e^{X^2/K^2}\right] \leq 2$.*

Definition 2.2. (Sub-exponential Random Variable) *A random variable X is sub-exponential with parameter K if we have $\mathbb{E}\left[e^{|X|/K}\right] \leq 2$.*

Trace Estimation. To prove Theorem 1.1 we will relate Hutchinson's diagonal estimator (Definition 1.1) to the well-known Hutchinson's trace estimator, which we define below:

Definition 2.3. (Hutchinson's Trace Estimator) *Let $\mathbf{g} \in \{-1, +1\}^n$ be a vectors with i.i.d. Rademacher entries. Hutchinson's trace estimator T for a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ is:*

$$T(\mathbf{B}) = \mathbf{g}^T \mathbf{B} \mathbf{g}.$$

It is not hard to show that $\mathbb{E}[T(\mathbf{B})] = \mathrm{tr}(\mathbf{B})$ and $\mathrm{Var}[T(\mathbf{B})] = 2\|\bar{\mathbf{B}}\|_F^2$, where $\bar{\mathbf{B}}$ denotes $\mathbf{B}$ with its diagonal entries set to 0. By averaging repeated copies of the estimator we can obtain a lower variance estimate. To prove high probability error bounds, a tight analysis can be obtained via the Hanson-Wright inequality, which implies that $T(\mathbf{B})$ exhibits exponential concentration [6]. We will use an intermediate result stated in Section 6.2 of [22] as a step towards proving Hanson-Wright[2]:

[1] It is possible to replace n with a natural "intrinsic dimension" parameter that is smaller for some problem instances [11].

[2] Note that when $\mathbf{g}$ contains Rademacher random variables, $Z(\mathbf{B}) = \mathrm{tr}(\mathbf{B}) - T(\mathbf{B})$ exactly equals $\sum_{i \neq j} g_i g_j B_{ij}$, which is precisely the "off-diagonal sum" random variable bounded in [22].

LEMMA 2.1. ([22]) *Let $Z(\mathbf{B}) = T(\mathbf{B}) - \mathrm{tr}(\mathbf{B})$ be the error of Hutchinson's trace estimator as in Definition 2.3. For absolute constants c, C, we have:*

$$\mathbb{E}\left[e^{\lambda Z(\mathbf{B})}\right] \leq e^{C\lambda^2 \|\mathbf{B}\|_F^2} \qquad \textit{for all} \qquad |\lambda| \leq c/\|\mathbf{B}\|_2.$$

2.1 Relation Between Diagonal Estimator and Trace Estimator Consider $\mathbf{r}^m(\mathbf{A})$ and as before let $\mathrm{diag}(\mathbf{A})$ denote the diagonal of $\mathbf{A}$. Let $\mathbf{g}^1, \ldots, \mathbf{g}^m$ be the m random ± 1 vectors used to obtain $\mathbf{r}^m(\mathbf{A})$. We can rewrite the mean zero random vector $\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})$ as:

$$\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A}) = \frac{1}{m}\sum_{z=1}^{m} \mathbf{e}_z \qquad \text{where for } z = 1, \ldots, m \qquad \mathbf{e}_z = \mathbf{g}^z \odot \mathbf{A}\mathbf{g}^z - \mathrm{diag}(\mathbf{A}).$$

Note that the i^{th} entry in $\mathbf{e}_z$ equals $\sum_{j \neq i} A_{ij} g_i^z g_j^z$. Using that $(g_i^z)^2 = 1$ for all i, z and recalling that $\bar{\mathbf{A}}$ denotes $\mathbf{A}$ with its diagonal set to zero, we have:

$$\|\mathbf{e}_z\|_2^2 = \sum_{i=1}^{d}\left(\sum_{j\neq i} A_{ij} g_i^z g_j^z\right)^2 = \sum_{i=1}^{d}\sum_{j\neq i}\sum_{k\neq i} A_{ij}A_{ik} g_i^z g_j^z g_j^z g_k^z = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \bar{A}_{ij}\bar{A}_{ik} g_j^z g_k^z = \sum_{j=1}^{d}\sum_{k=1}^{d} g_j^z g_k^z \sum_{i=1}^{d} \bar{A}_{ij}\bar{A}_{ik}.$$

Let $\mathbf{B} = \bar{\mathbf{A}}^T\bar{\mathbf{A}}$. We have that $B_{jk} = \sum_{i=1}^{d} \bar{A}_{ij}\bar{A}_{ik}$, so we can rewrite the above as:

$$\|\mathbf{e}_z\|_2^2 = \sum_{j=1}^{d}\sum_{k=1}^{d} g_j^z g_k^z B_{jk} = \mathbf{g}^{z\,T}\mathbf{B}\mathbf{g}^z. \tag{2.1}$$

In other words, $\|\mathbf{e}_z\|_2^2$ is identically distributed to Hutchinson's *trace estimator* applied to the positive semi-definite matrix $\mathbf{B}$. An immediate consequence of Eq. (2.1) is that $\mathbb{E}\,\|\mathbf{e}_z\|_2^2 = \mathrm{tr}(\mathbf{B}) = \|\bar{\mathbf{A}}\|_F^2$. This in turn yields the following:

LEMMA 2.2. (EXPECTED SQUARED ERROR OF HUTCHINSON'S DIAGONAL ESTIMATOR) *Let $\mathbf{r}^m(\mathbf{A})$ as in Definition 1.1. We have:*

$$\mathbb{E}\left[\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2^2\right] = \frac{1}{m}\|\bar{\mathbf{A}}\|_F^2.$$

Proof.

$$\mathbb{E}\left[\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2^2\right] = \mathbb{E}\left[\left\|\frac{1}{m}\sum_{z=1}^{m}\mathbf{e}_z\right\|_2^2\right] = \frac{1}{m^2}\,\mathbb{E}\left[\sum_{z=1}^{m}\|\mathbf{e}_z\|_2^2 + \sum_{z=1}^{m}\sum_{w\neq z}\mathbf{e}_z{}^T\mathbf{e}_w\right] = \frac{1}{m^2}\left[\sum_{z=1}^{m}\|\bar{\mathbf{A}}\|_F^2 + 0\right]$$

In the last inequality we used that $\mathbf{e}_z{}^T\mathbf{e}_w = 0$ because the random vectors are mean zero and independent. ∎

Applying Markov's inequality, an immediate consequence of Lemma 2.2 is that $\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2^2 \leq \frac{1}{m\delta}\|\bar{\mathbf{A}}\|_F^2$ with probability $1-\delta$. Setting $m = \frac{1}{\epsilon^2\delta}$ we thus have that with probability $1-\delta$, $\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2 \leq \epsilon\|\bar{\mathbf{A}}\|_F$. Notably this simple bound already avoids the $\log(n)$ dependence from [4], but it incurs a suboptimal $1/\delta$ dependence in comparison to that result and Theorem 1.1, which depend on $\log(1/\delta)$.

3 Proof of Main Theorem

In this section we prove Theorem 1.1, which requires bounding the norm of $\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})$. This random vector can be written as the average of m mean-zero random vectors $\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A}) = \frac{1}{m}\sum_{z=1}^{m}\mathbf{e}_z$. Via the connection to Hutchinson's trace estimator, we know the expected norm of each $\mathbf{e}_z$ is equal to $\|\bar{\mathbf{A}}\|_F$. Moreover, each norm should not be much larger than $\|\bar{\mathbf{A}}\|_F$ with high probability due to the concentration of Hutchinson's trace estimator. Thus a natural approach might be to apply a "vector valued Bernstein" inequality for sums of norm-bounded random vectors [25]. However a direct application of prior work yields a suboptimal polynomial dependence on $\log(1/\delta)$.

Alternatively, since $\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2^2$ is a low-degree (degree 4) polynomial in Rademacher random variables, we might hope to prove concentration by applying techniques based on hypercontractivity to bound the random variable's higher moments, as done e.g. in [20] for Hutchinson's trace estimator. However, doing so would require establishing a bound on the second moment $\mathbb{E}[\|\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2^4]$, which is already challenging. Relying directly on hypercontractivity also seems to be limited to yielding a suboptimal dependence on $\log(1/\delta)$.

We take another approach, providing an analysis that loosely follows the same approach used for Lemma 2 in Yurinskii's proof of the vector-valued Bernstein inequality and does not require $\mathbf{e}_1, \ldots, \mathbf{e}_m$ to be strictly bounded.

3.1 Symmetrization and Scalar Comparison Let $\mathbf{e} = \sum_{z=1}^m \mathbf{e}_z$ and note that $\mathbf{e} = m \cdot (\mathbf{r}^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A}))$. Our goal will be to upper bound the moments of $\mathbf{e}$'s squared norm by the moments of an easier to analyze scalar random variable. To do so, we start with a symmetrization argument. First, consider the alternative random vector $\tilde{\mathbf{e}} = \mathbf{e}_1 - \mathbf{e}_2$ where $\mathbf{e}_1$ and $\mathbf{e}_2$ are i.i.d. copies of $\mathbf{e}$. Using that $f(\mathbf{x}) = \|\mathbf{x}\|_2^{2k}$ is a convex function and that $\mathbb{E}[\mathbf{e}_1] = \mathbb{E}[\mathbf{e}_2] = 0$, we can apply Jensen's inequality to show that:

$$\mathbb{E}\left[\|\mathbf{e}\|_2^{2k}\right] \leq \mathbb{E}\left[\|\mathbf{e}_1 - \mathbf{e}_2\|_2^{2k}\right] = \mathbb{E}\left[\|\tilde{\mathbf{e}}\|_2^{2k}\right]. \tag{3.2}$$

See Lemma 6.1.2 in [22] for a detailed derivation of the above inequality. Next, we can turn our attention to bounding $\mathbb{E}\left[\|\tilde{\mathbf{e}}\|_2^{2k}\right]$. Letting $\mathbf{e}_{z,1}$ and $\mathbf{e}_{z,2}$ be i.i.d. copies of $\mathbf{e}_z$ (i.e. error vectors of Hutchinson's diagonal estimator applied with a single random vector), we have that $\tilde{\mathbf{e}} = \sum_{z=1}^m \mathbf{e}_{z,1} - \mathbf{e}_{z,2}$. Let $\mathbf{w}_z$ denote $\mathbf{w}_z = \mathbf{e}_{z,1} - \mathbf{e}_{z,2}$ and note that $\tilde{\mathbf{e}} = \sum_{z=1}^m \mathbf{w}_z$. Let W_z be a scalar random variable equal to $r_z \cdot \|\mathbf{w}_z\|_2$, where r_z is a ± 1 Rademacher random variable. For all $k = 1, 2, \ldots$ we have that:

$$\mathbb{E}[W_z] = 0 \qquad \text{and} \qquad \mathbb{E}[W_z^{2k}] = \mathbb{E}[\|\mathbf{w}_z\|_2^{2k}].$$

Let $\tilde{E} = \sum_{z=1}^m W_z$. We will bound the moments of $\|\tilde{\mathbf{e}}\|_2^2$ by comparing to $\tilde{E}$. In particular, we will show that for all k,

$$\mathbb{E}\left[\|\tilde{\mathbf{e}}\|_2^{2k}\right] \leq \mathbb{E}\left[\tilde{E}^{2k}\right]. \tag{3.3}$$

To do so, we compare the expansions:

$$\mathbb{E}\left[\|\tilde{\mathbf{e}}\|_2^{2k}\right] = \mathbb{E}\left[\left(\|\mathbf{w}_1\|_2^2 + \ldots + \|\mathbf{w}_m\|_2^2 + 2\mathbf{w}_1^T\mathbf{w}_2 + \ldots + 2\mathbf{w}_{m-1}^T\mathbf{w}_m\right)^k\right]$$
$$\mathbb{E}\left[\tilde{E}^{2k}\right] = \mathbb{E}\left[\left(W_1^2 + \ldots + W_m^2 + 2W_1W_2 + \ldots + 2W_{m-1}W_m\right)^k\right]$$

Consider each term obtained when expanding out the k^{th} powers above and apply linearity of expectation. Because each $\mathbf{w}_z$ is a symmetric random variable – i.e. $\Pr(\mathbf{w}_z = X) = \Pr(\mathbf{w}_z = -X)$ – we can verify that the expectation of any term where $\mathbf{w}_z^T\mathbf{w}_j$ appears an odd number of times (for *any* fixed j) is equal to zero. Similarly, the corresponding term in the second sum has expectation zero because some W_z must appear an odd number of times. For all other terms, we can use that $\mathbf{w}_z^T\mathbf{w}_j \leq \|\mathbf{w}_z\|_2\|\mathbf{w}_j\|_2$ (Cauchy–Schwarz) and that $\mathbb{E}[\|\mathbf{w}_z\|_2^{2k}] = \mathbb{E}[W_z^{2k}]$ to see that each term in the bottom expansion upper bounds the corresponding term in the top. We conclude Eq. (3.3).

A Taylor expansion of e^x combined with Eq. (3.2) and Eq. (3.3) implies a bound on the moment generating function (MGF) of $\|\mathbf{e}\|_2^2$, which we will use to obtain a final concentration result. Specifically, we have that for any $\lambda \geq 0$:

$$\mathbb{E}\left[e^{\lambda\|\mathbf{e}\|_2^2}\right] \leq \mathbb{E}\left[e^{\lambda\|\tilde{\mathbf{e}}\|_2^2}\right] \leq \mathbb{E}\left[e^{\lambda\tilde{E}^2}\right]. \tag{3.4}$$

3.2 Moment Bound With Eq. (3.4) in place, we prove our main result by bounding the exponential $\mathbb{E}\left[e^{\lambda\tilde{E}^2}\right]$ for our scalar random variable $\tilde{E}$. Specifically, we will show that $\tilde{E}$ is a sub-exponential random variable (Definition 2.2), and thus $\|\mathbf{e}\|_2^2$ is as well by Eq. (3.4). We can then apply a standard tail bound for sub-exponential random variables.

Proof. [Proof of Theorem 1.1] Recall that $\tilde{E} = \sum_{z=1}^m W_z$ is the sum of i.i.d. random variables. Recall that $W_z = r_z \cdot \|\mathbf{w}_z\|_2$, where r_z is a random ± 1 and $\mathbf{w}_z = \mathbf{e}_{z,1} - \mathbf{e}_{z,2}$. We have that $\|\mathbf{w}_z\|_2^2 \leq 2\|\mathbf{e}_{z,1}\|_2^2 + 2\|\mathbf{e}_{z,2}\|_2^2$ and

since $\mathbf{e}_{z,1}$ and $\mathbf{e}_{z,2}$ are just i.i.d. copies of $\mathbf{e}_z$, we have that for all $\lambda \geq 0$:

$$\mathbb{E}\left[e^{\lambda\|\mathbf{w}_z\|_2^2}\right] \leq \mathbb{E}\left[e^{4\lambda\|\mathbf{e}_z\|_2^2}\right]. \tag{3.5}$$

As discussed before, $\|\mathbf{e}_z\|_2^2$ is exactly equal to Hutchinson's estimator applied to the matrix $\mathbf{B} = \bar{\mathbf{A}}^T\bar{\mathbf{A}}$. Under the notation of Lemma 2.1, $\|\mathbf{e}_z\|_2^2 = T(\mathbf{B}) = Z(\mathbf{B}) + \operatorname{tr}(\mathbf{B})$. We can thus apply Lemma 2.1 to obtain that for $0 \leq 4\lambda \leq c/\|\mathbf{B}\|_2$,

$$\mathbb{E}\left[e^{4\lambda\|\mathbf{e}_z\|_2^2}\right] = e^{4\lambda\operatorname{tr}(\mathbf{B})}\mathbb{E}\left[e^{4\lambda Z(\mathbf{B})}\right] \leq e^{4\lambda\operatorname{tr}(\mathbf{B})}e^{16C\lambda^2\|\mathbf{B}\|_F^2}.$$

Since $B = \bar{\mathbf{A}}^T\bar{\mathbf{A}}$ is a positive semidefinite matrix, $\|\mathbf{B}\|_F^2/\|\mathbf{B}\|_2 \leq \operatorname{tr}(\mathbf{B})$ and thus $4\lambda\|\mathbf{B}\|_F^2 \leq c\operatorname{tr}(\mathbf{B})$. Continuing we have:

$$\mathbb{E}\left[e^{4\lambda\|\mathbf{e}_z\|_2^2}\right] \leq e^{4\lambda\operatorname{tr}(\mathbf{B})+4Cc\lambda\operatorname{tr}(\mathbf{B})} \tag{3.6}$$

We conclude from Eq. (3.5) and Eq. (3.6) that for all $\lambda \leq \frac{c}{4\|\mathbf{B}\|_2} \leq \frac{c}{4\operatorname{tr}(\mathbf{B})}$,

$$\mathbb{E}\left[e^{\lambda\|\mathbf{w}_z\|_2^2}\right] \leq e^{c'\lambda\operatorname{tr}(\mathbf{B})}, \tag{3.7}$$

where $c' = 4 + 4Cc$ is a constant.

3.3 Completing the Proof Applying Definition 2.2, we can check that Equation (3.7) implies that $\|\mathbf{w}_z\|_2^2$ is a sub-exponential random variable with parameter $C' \cdot \operatorname{tr}(\mathbf{B})$, where $C' = \max(2c', 4/c)$. Equivalently, $\|\mathbf{w}_z\|_2$ is sub-Gaussian with parameter $K = \sqrt{C'}\sqrt{\operatorname{tr}(\mathbf{B})}$. Since it has the same moments as $\|\mathbf{w}_z\|_2$, $W_z = r_z \cdot \|\mathbf{w}_z\|_2$ is also sub-Gaussian with the same parameter.

Proposition 2.6.1 from [22] states that the sum of m mean 0, independent, sub-Gaussian random variables, each with parameter K, is itself sub-Gaussian with parameter $C \cdot \sqrt{m}K$ for a fixed constant C. We conclude that $\tilde{E} = \sum_{i=1}^m W_z$ is sub-Gaussian with parameter $C\sqrt{C'} \cdot \sqrt{m}\sqrt{\operatorname{tr}(\mathbf{B})}$. Finally, it follows that $\tilde{E}^2$ is sub-exponential with parameter $c'' \cdot m\operatorname{tr}(\mathbf{B})$, where $c'' = C^2C'$. From Eq. (3.4), we know that $\|\mathbf{e}\|_2^2$ is sub-exponential with the same parameter. Finally, from Proposition 2.7.1 in [22] we have that $\Pr\left[\|\mathbf{e}\|_2^2 \geq t\right] \leq 2e^{-\frac{t}{c''\cdot m\operatorname{tr}(\mathbf{B})}}$ and thus:

$$\Pr\left[\frac{1}{m^2}\|\mathbf{e}\|_2^2 \geq \operatorname{tr}(\mathbf{B}) \cdot \frac{c''\log(2/\delta)}{m}\right] \leq \delta.$$

Recalling that $[\|\mathbf{r}^m(\mathbf{A}) - \operatorname{diag}(\mathbf{A})\|_2^2 = \frac{1}{m^2}\|\mathbf{e}\|_2^2$ and $\operatorname{tr}(\mathbf{B}) = \|\bar{\mathbf{A}}\|_F^2$, Theorem 1.1 follows. ∎

4 General Stochastic Diagonal Estimators

In addition to the standard Hutchinson's estimator, prior work on stochastic diagonal and trace estimation also considers estimators involving Gaussian random vectors, or more generally, vectors filled with arbitrary mean 0, variance 1 random variables [9, 4].

Definition 4.1. (Generalized Diagonal Estimator[3]) *Let $\mathbf{g}^1, \ldots, \mathbf{g}^m \in \mathbb{R}^n$ be independent random vectors, each with i.i.d. entries that have mean 0 and variance 1. The generalized stochastic diagonal estimator $\mathbf{d}^m(\mathbf{A})$ has the form:*

$$\mathbf{d}^m(\mathbf{A}) = \frac{1}{m}\sum_{z=1}^m \mathbf{g}^z \odot \mathbf{A}\mathbf{g}^z.$$

When each g_i^z has bounded 4th moment, we can prove a statement comparable to Lemma 2.2.

Lemma 4.1. (Expected Squared Error of Generalized Diagonal Estimator) *Let $\mathbf{d}^m(\mathbf{A})$ be as in Definition 1.1 and suppose each g_i^z has 4th moment bounded by some constant c_4. I.e. $\mathbb{E}[(g_i^z)^4] \leq c_4$. Then we have:*

$$\mathbb{E}\left[\|\mathbf{d}^m(\mathbf{A}) - \operatorname{diag}(\mathbf{A})\|_2^2\right] = \frac{1}{m}\left(\|\bar{\mathbf{A}}\|_F^2 + (1 + c_4 - 2)\sum_{i=1}^d A_{ii}^2\right)$$

Proof. As before, fix z and let $\mathbf{e}_z = \mathbf{g}^z \odot \mathbf{A}\mathbf{g}^z - \operatorname{diag}(\mathbf{A})$. Let $h_i = A_{ii} - A_{ii}(g_i^z)^2$ and note that, since $\mathbb{E}(g_i^z)^2 = 1$, we have that $\mathbb{E}\, h_i = 0$. More over, we have:

$$\mathbb{E}\left[h_i^2\right] = \mathbb{E}\left[A_{ii}^2 + A_{ii}^2(g_i^z)^4 - 2A_{ii}^2(g_i^z)^2\right] = (c_4 + 1 - 2)A_{ii}^2.$$

We then have that:

$$\begin{aligned}\|\mathbf{e}_z\|_2^2 = \sum_{i=1}^d \left(h_i + \sum_{j\neq i} A_{ij} g_i^z g_j^z\right)^2 &= \sum_{i=1}^d h_i^2 + \sum_{i=1}^d h_i \sum_{j\neq i} A_{ij} g_i^z g_j^z + \sum_{i=1}^d \left(\sum_{j\neq i} A_{ij} g_i^z g_j^z\right)^2 \\ &= \sum_{i=1}^d h_i^2 + \sum_{i=1}^d h_i \sum_{j\neq i} A_{ij} g_i^z g_j^z + \sum_{i=1}^d \sum_{j\neq i}\sum_{k\neq i} A_{ij}A_{ik} g_i^z g_i^z g_j^z g_k^z\end{aligned}$$

Considering each term separately, we can bound the expectation of $\|\mathbf{e}_z\|_2^2$. Noting that $\mathbb{E}[g_i^z g_i^z g_j^z g_k^z] = 0$ if $j \neq k$ and 1 otherwise since $j \neq i$, we have:

$$\mathbb{E}\,\|\mathbf{e}_z\|_2^2 = \sum_{i=1}^d (1 + c_4 - 2)A_{ii}^2 + 0 + \sum_{i=1}^d \sum_{j\neq i} A_{ij}^2 = (1 + c_4 - 2)\|\operatorname{diag}(\mathbf{A})\|_2^2 + \|\bar{\mathbf{A}}\|_F^2.$$

□

Combining Lemma 4.1 with Markov's inequality yields a simple dimension independent bound:

COROLLARY 4.1. *Let $\mathbf{d}^m$ be implemented with any mean 0 variance 1 random variable with 4th moment upper bounded by c_4 and let $E = \sqrt{(1 + c_4 - 2)\|\operatorname{diag}(\mathbf{A})\|_2^2 + \|\bar{\mathbf{A}}\|_F^2}$. For any $\delta \in (0, 1)$ and $m \geq 1$, with probability $1 - \delta$:*

$$\|\mathbf{d}^m(\mathbf{A}) - \operatorname{diag}(\mathbf{A})\|_2 \leq \sqrt{\frac{1}{m\delta}} \cdot E.$$

When $\mathbf{d}^m$ is implemented with Gaussian random vectors, we have fourth moment $c_4 = 3$, so obtain the upper bound:

$$\|\mathbf{d}^m(\mathbf{A}) - \operatorname{diag}(\mathbf{A})\|_2 \leq \sqrt{\frac{2}{m\delta}} \cdot \|\mathbf{A}\|_F.$$

4.1 High Probability Bounds To obtain an error bound of $\epsilon \cdot E$ with probability $1 - \delta$, Corollary 4.1 requires $m = O\left(1/\epsilon^2\delta\right)$ matrix-vector products with $\mathbf{A}$. The linear dependence on $1/\delta$ is worse than the logarithmic dependence in Theorem 1.1, which requires $m = O\left(\log(1/\delta)/\epsilon^2\right)$ matrix-vector products for a comparable guarantee. It is possible to improve the dependence on δ using a high-dimensional analog of the standard "median trick". Specifically, we have:

COROLLARY 4.2. *Consider the following estimation procedure that computes multiple independent generalized stochastic diagonal estimators (Definition 4.1), all implemented with mean 0 variance 1 random variables with 4th moment $\leq c_4$.*

- *Compute $r = \lceil 10 \log(1/\delta) \rceil$ independent generalized diagonal estimators $\mathbf{d}_1^m(\mathbf{A}), \ldots, \mathbf{d}_q^m(\mathbf{A})$.*
- *For all $i \in 1, \ldots, r$, compute the distance $\|\mathbf{d}_i^m(\mathbf{A}) - \mathbf{d}_j^m(\mathbf{A})\|_2$ for all $j \neq i$. Let B_i be the $\lfloor \frac{r}{2} \rfloor$ smallest distance.*
- *Return $\mathbf{d}_{i^*}^m(\mathbf{A})$, where $i^* = \arg\min_{i \in 1,\ldots,r} B_i$.*

There is an absolute constant c so that, for any $\delta \in (0, 1)$ and $m \geq 1$, with probability $1 - \delta$:

$$\|\mathbf{d}_{i^*}^m(\mathbf{A}) - \operatorname{diag}(\mathbf{A})\|_2 \leq \sqrt{\frac{c}{m}} \cdot E,$$

where $E = \sqrt{(1 + c_4 - 2)\|\operatorname{diag}(\mathbf{A})\|_2^2 + \|\bar{\mathbf{A}}\|_F^2}$, as before.

As desired, Corollary 4.2 implies that $m = O\left(\log(1/\delta)/\epsilon^2\right)$ matrix-vector multiplies are required to obtain error $\epsilon \cdot E$ with probability $(1-\delta)$.

Proof. By Corollary 4.1, for each $i \in 1, \ldots, r$ and a constant c, we have that $\|\mathbf{d}_i^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{1/m} \cdot E$ with probability 19/20. By a standard Chernoff bound argument, it follows that, with probability greater than $1 - e^{-r/10} = 1 - \delta$, $\|\mathbf{d}_i^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{1/m} \cdot E$ for at least half of all values of i (see e.g. Proposition 2.4 in [1]). Accordingly, by triangle inequality, we have that:

$$\text{there is at least one } i \in 1, \ldots, r \text{ for which } B_i \leq 2c\sqrt{1/m} \cdot E. \tag{4.8}$$

Also by pigeonhole principal, there must be at least one value of j which is both one of the $\lfloor r/2 \rfloor$ closest points to $\mathbf{d}_{i^*}^m$ and for which $\|\mathbf{d}_j^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{1/m} \cdot E$. I.e., there is some j such that $\|\mathbf{d}_j^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{1/m} \cdot E$ and $\|\mathbf{d}_j^m(\mathbf{A}) - \mathbf{d}_{i^*}^m(\mathbf{A}) \leq B_{i^*}$. By triangle inequality we thus have:

$$\|\mathbf{d}_{i^*}^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq c\sqrt{1/m} \cdot E + B_{i^*} \leq 3c\sqrt{1/m} \cdot E.$$

The last inequality follows from Eq. (4.8) since $B_{i^*} \leq B_i$ for all $i \in 1, \ldots, r$. This completes the proof. ∎

If instead of just assuming that $\mathbf{g}^1, \ldots, \mathbf{g}^m$ contain entries with bounded 4^{th} moment, if we make the stronger assumption that they contain i.i.d. sub-Gaussian entries, then we can obtain a bound for the generalized diagonal estimator that is more comparable to Theorem 1.1 and does not require the median trick to obtain a dependence the ideal dependence on $\log(1/\delta)$. Specifically, in Appendix A, we prove the following result:

THEOREM 4.1. *Let $\mathbf{d}^m$ be a generalized stochastic diagonal estimator (Definition 4.1) for $\mathbf{A} \in \mathbb{R}^{n \times n}$ implemented with any symmetric, mean 0, and variance 1 random variable that is sub-Gaussian with parameter K. Then with probability $1 - \delta$:*

$$\|\mathbf{d}^m(\mathbf{A}) - \text{diag}(\mathbf{A})\|_2 \leq cK^2 \cdot \sqrt{\frac{\log(2/\delta)}{m} + \frac{\log^4(2/\delta)}{m^2}} \|\mathbf{A}\|_F.$$

Typically K^2 is a small constant (e.g. for the previously studied case of Gaussian random variables [4]), so Theorem 4.1 nearly matches Theorem 1.1, except in two ways. First, as in Corollary 4.1, it has a dependence on $\|\mathbf{A}\|_F^2$ instead of $\|\bar{\mathbf{A}}\|_F^2$. In general, $\|\bar{\mathbf{A}}\|_F^2$ is always smaller. This is inherent: as shown in Lemma 4.1, the expected error of $\mathbf{d}^m(\mathbf{A})$ has a dependence on $\|\mathbf{A}\|_F^2$ unless the fourth moment equals 1, but this is only the case for ± 1 Rademacher random variables. All other random variables with variance 1 have higher 4^{th} moment.

Second, Theorem 4.1 has an extra dependence on $\frac{\log^4(2/\delta)}{m^2}$ that does not appear in Theorem 1.1. While this is a lower order term for large m – specifically, the bound matches Theorem 1.1 when $m \geq \log^{1.5}(1/\delta)$ – we believe it can likely be improved or removed entirely, possibly by following a different proof technique.

Acknowledgements

We would like to thank Yuji Nakatsukasa, Eric Hallman, and Cameron Musco for helpful discussions. Christopher Musco was supported by NSF CAREER award No. 2045590. Prathamesh Dharangutte was supported by NSF award No. CCF-2118953.

References

[1] Dana Angluin and Leslie G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, 1979.

[2] Richard C. Aster, Brian Borchers, and Clifford H. Thurber. *Parameter estimation and inverse problems*. Elsevier, 3rd edition, 2019.

[3] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM*, 58(2), 2011.

[4] Robert A. Baston and Yuji Nakatsukasa. Stochastic diagonal estimation: probabilistic bounds and an improved algorithm. *arXiv:2201.10684*, 2022.

[5] C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Appl. Numer. Math.*, 57(11–12):1214–1229, 2007.
[6] Alice Cortinovis and Daniel Kressner. On randomized trace estimates for indefinite matrices with an application to determinants. *Foundations of Computational Mathematics*, 22(3):875–903, 2022.
[7] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, volume 28, 2015.
[8] David Eriksson, Kun Dong, Eric Lee, David Bindel, and Andrew G Wilson. Scaling Gaussian process regression with derivatives. *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 31, 2018.
[9] Didier Girard. Un algorithme simple et rapide pour la validation croisee géénéralisée sur des problémes de grande taille. Technical report, 1987.
[10] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26:1 – 22, 2021.
[11] Eric Hallman. Personal Communication, 2022.
[12] Eric Hallman, Ilse C.F. Ipsen, and Arvind Saibaba. Monte Carlo methods for estimating the diagonal of a real symmetric matrix. *arXiv:2202.02887*, 2022.
[13] Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.
[14] Philip Klein and Neal E. Young. On the number of iterations for Dantzig–Wolfe optimization and packing-covering approximation algorithms. *SIAM Journal on Computing*, 44(4):1154–1172, 2015.
[15] L. Métivier, F. Bretaudeau, R. Brossier, S. Operto, and J. Virieux. Full waveform inversion and the truncated newton method: quantitative imaging of complex subsurface structures. *Geophysical Prospecting*, 62(6):1353–1375, 2014.
[16] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David Woodruff. Hutch++: optimal stochastic trace estimation. *Proceedings of the 4th Symposium on Simplicity in Algorithms (SOSA)*, 2021.
[17] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
[18] Farbod Roosta-Khorasani and Uri M. Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.
[19] Holger Sambale. Some notes on concentration for α-subexponential random variables. *arXiv:2002.10761*, 2020.
[20] Maciej Skórski. Modern analysis of hutchinson's trace estimator. In *55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.
[21] Xiaoming Sun, David P. Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. In *Proceedings of the 46th International Colloquium on Automata, Languages and Programming (ICALP)*, volume 132, pages 94:1–94:16, 2019.
[22] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.
[23] David Woodruff, Fred Zhang, and Richard Zhang. Optimal query complexities for dynamic trace estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.
[24] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. AdaHessian: An adaptive second order optimizer for machine learning. In *Proceedings of the AAAI Conference on Artificial (AAAI)*, volume 35, pages 10665–10673, 2021.
[25] V. V. Yurinskii. On an infinite-dimensional version of S. N. Bernstein's inequalities. *Theory of Probability & Its Applications*, 15(1):108–109, 1970.

A General Sub-Gaussian Analysis

In this section, we focus on proving Theorem 4.1 for general sub-Gaussian stochastic diagonal estimators. The proof follows a different approach and is more involved than our proof for Hutchinson's estimator in Section 3, which strongly relies on the fact that the estimator uses Rademacher random variables.

A.1 Initial Symmetrization We first show how Theorem 4.1 can be reduced to an equivalent statement involving a symmetric random vector:

LEMMA A.1. *Let* $\mathbf{d}_1^m, \mathbf{d}_2^m$ *be independent generalized stochastic diagonal estimators (Definition 4.1) for* $\mathbf{A}$ *implemented with any symmetric, mean* 0*, and variance* 1 *random variable that is sub-Gaussian with parameter* K*.*

For any $\delta \in (0,1)$ and $m \geq 1$, let $m = O(\log(1/\delta)/\epsilon^2)$. Then with probability $1-\delta$:

$$\|\mathbf{d}_1^m(\mathbf{A}) - \mathbf{d}_2^m(\mathbf{A})\|_2 \leq cK^2 \cdot \sqrt{\frac{\log(2/\delta)}{m} + \frac{\log^4(2/\delta)}{m^2}}\|\mathbf{A}\|_F.$$

Before proving Lemma A.1, we show how it can be used to prove Theorem 4.1.

Proof. [Proof of Theorem 4.1] Let $\ell = \log_2(1/\delta')$ for some $\delta' < \delta$ to be chosen later and consider independent diagonal estimators $\mathbf{d}_2^m, \ldots, \mathbf{d}_{\ell+1}^m$. We will not actually compute these estimators – they are hypothetical and introduced for the purpose of analysis. Any random variable with sub-Gaussian parameter K has fourth moment bounded by $O(K^4)$ (see [22], Proposition 2.5.2). Accordingly, by Corollary 4.1, for some constant c, we have that, with probability $1/2$, $\|\mathbf{d}_i^m(A) - \mathrm{diag}(\mathbf{A})\|_2 \leq cK^2\|\mathbf{A}\|_F$ for each $i \in 2, \ldots, \ell+1$. It follows that, with probability δ', $\|\mathbf{d}_j^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2 \leq \frac{cK^2}{\sqrt{m}}\|\mathbf{A}\|_F$ for at least one value of j. At the same time, combining Lemma A.1 with a union bound, we know that *for all* i simultaneously, with probability $1 - \delta'\log_2(1/\delta')$, $\|\mathbf{d}_1^m(\mathbf{A}) - \mathbf{d}_i^m(\mathbf{A})\|_2 \leq cK^2 \cdot \sqrt{\frac{\log(2/\delta')}{m} + \frac{\log^4(2/\delta')}{m^2}}\|\mathbf{A}\|_F$. It follows by triangle inequality and another union bound that with probability $1 - \delta'\log_2(1/\delta') - \delta'$,

$$\|\mathbf{d}_1^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2 \leq \|\mathbf{d}_1^m(\mathbf{A}) - \mathbf{d}_i^m(\mathbf{A})\|_2 + \|\mathbf{d}_i^m(\mathbf{A}) - \mathrm{diag}(\mathbf{A})\|_2 \leq 2cK^2 \cdot \sqrt{\frac{\log(2/\delta)}{m} + \frac{\log^4(2/\delta)}{m^2}}\|\mathbf{A}\|_F.$$

Setting $\delta' = c'\delta^2$ for sufficiently small constant c' yields Theorem 4.1. ∎

A.2 Single Sample Norm Bound In order to prove Lemma A.1, we first prove a tail bound on the norm of a single-sample sub-Gaussian stochastic diagonal estimator. This intermediate result is the crux of our analysis, and from it Lemma A.1 follows relatively directly.

LEMMA A.2. *Let $\mathbf{e} = \mathbf{g} \odot \mathbf{Ag} - \mathrm{diag}(\mathbf{A})$, where $\mathbf{g} \in \mathbb{R}^n$ contains i.i.d. symmetric, mean 0, and variance 1 sub-Gaussian random variables with parameter K. For any $\gamma \geq 0$ and a fixed constant c we have that*

$$\Pr\left[\|\mathbf{e}\|_2^2 \geq \gamma K^4\|\mathbf{A}\|_F^2\right] \leq 2e^{-c\gamma^{1/3}}.$$

Proof. In what follows, we will assume that $\gamma \geq C$ for some sufficiently large constant C. If we can prove the result with some constant c' in the exponent under this assumption, than we immediately have that $\Pr\left[\|\mathbf{e}\|_2^2 \geq \gamma K^4\|\mathbf{A}\|_F\right] \leq 2e^{-c\gamma^{1/3}}$ for *all* $\gamma \geq 0$, where $c = \min(c', 1/2C^{1/3})$. This follows because $2e^{-\min(c', 1/2C^{1/3})\gamma^{1/3}} > 1$ for any $\gamma \leq C$, so the bound is vacuously true for small values of γ.

We start by applying triangle inequality and AM-GM inequality to give:

$$\|\mathbf{e}\|_2^2 \leq 2\|\mathbf{g} \odot \mathbf{Ag}\|_2^2 + 2\|\mathrm{diag}(\mathbf{A})\|_2^2 \leq 2\|\mathbf{g} \odot \mathbf{Ag}\|_2^2 + 2\|\mathbf{A}\|_F^2, \tag{A.1}$$

so we focus on bounding $\|\mathbf{g} \odot \mathbf{Ag}\|_2^2$. Following the proof of Lemma 4.1, we have that:

$$\|\mathbf{g} \odot \mathbf{Ag}\|_2^2 = \sum_{i=1}^d \left(\sum_{j=1}^d A_{ij}g_ig_j\right)^2 = \sum_{i=1}^d\sum_{j=1}^d A_{ij}g_ig_j\sum_{k=1}^d A_{ik}g_ig_k = \sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d A_{ij}A_{ik}g_ig_ig_jg_k$$

Let $\mathbf{G}$ be a diagonal matrix containing $\mathbf{g}$ on its diagonal and let $\tilde{\mathbf{A}} = \mathbf{GAG}$. The matrix $\tilde{\mathbf{A}}$ has entries equal to $\tilde{A}_{ij} = A_{ij}g_ig_j$. Morever, note that, since each g_i is assumed to by symmetric, it is identically distributed to g_ir_i where $r_1, \ldots r_n$ are independent Rademacher random variables. So we equivalently have that:

$$\|\mathbf{g} \odot \mathbf{Ag}\|_2^2 = \sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d A_{ij}A_{ik}g_ig_ig_jg_kr_jr_kr_i^2 = \sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d \tilde{A}_{ij}\tilde{A}_{ik}r_jr_k = \mathbf{r}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\mathbf{r}. \tag{A.2}$$

We conclude that the quantity $\|\mathbf{g} \odot \mathbf{Ag}\|_2^2$ is exactly equal to Hutchinson's estimator (implemented with Rademacher random variables) applied to the matrix $\mathbf{B} = \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$. As such, we expect that $\|\mathbf{g} \odot \mathbf{Ag}\|_2^2$ will tightly concentrate around $\mathrm{tr}(\mathbf{B}) = \|\tilde{\mathbf{A}}\|_F^2$. So the main challenge becomes to bound $\|\tilde{\mathbf{A}}\|_F^2$, which itself is a random variable.

Fortunately, we can bound this quantity by again making a connection to trace estimation. We have that:

$$\|\tilde{\mathbf{A}}\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbf{A}_{ij}^2 (g_i)^2 (g_j)^2 = \mathbf{g}^{2^T}(\mathbf{A}\circ\mathbf{A})\mathbf{g}^2,$$

where $\mathbf{g}^2$ denotes the vector obtained by squaring each entry of $\mathbf{g}$. Then, let $\bar{\mathbf{g}} = \mathbf{g}^2 - \mathbf{1}$ where $\mathbf{1}$ is an all ones vector. Note that $\mathbb{E}[\bar{\mathbf{g}}] = \mathbf{0}$ and since $\mathbf{g}$ is sub-Gaussian, $\mathbf{g}^2$ is a sub-exponential random variable with parameter K^2, and thus $\bar{\mathbf{g}}$ is sub-exponential with parameter $c'K^2$ for fixed constant c' (see [22], Exercise 2.7.10). We have that:

$$\begin{aligned}\mathbf{g}^{2^T}(\mathbf{A}\circ\mathbf{A})\mathbf{g}^2 = (\bar{\mathbf{g}}+\mathbf{1})^T(\mathbf{A}\circ\mathbf{A})(\bar{\mathbf{g}}+\mathbf{1}) &= \bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}} + 2\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1} + \mathbf{1}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1} \\ &= \bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}} + 2\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1} + \|\mathbf{A}\|_F^2. \end{aligned} \tag{A.3}$$

We bound $\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}}$ and $2\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1}$ separately, starting with the second. Let $\mathbf{a}_i$ denote the i^{th} row of $\mathbf{A}$ and note that $(\mathbf{A}\circ\mathbf{A})\mathbf{1}$ has i^{th} entry equal to $\|\mathbf{a}_i\|_2^2$. Since $\bar{\mathbf{g}}$ is mean 0, we can apply a Bernstein inequality for sub-exponential random variables ([22], Theorem 2.8.1) to the sum $\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1} = \sum_{i=1}^n \bar{g}_i\|\mathbf{a}_i\|_2^2$. We have that:

$$\Pr\left[|\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1}| \geq tK^2\right] \leq 2\exp\left(-c''\min\left(\frac{t^2}{\sum_{i=1}^n \|\mathbf{a}_i\|_2^4}, \frac{t}{\max_i \|\mathbf{a}_i\|_2^2}\right)\right),$$

where c'' is a fixed constant. Plugging in $t = \gamma\|\mathbf{A}\|_F^2$ and using that $\|\mathbf{A}\|_F^4 = \left(\sum_{i=1}^n \|\mathbf{a}_i\|_2^2\right)^2 \geq \sum_{i=1}^n \|\mathbf{a}_i\|_2^4$ and $\|\mathbf{A}\|_F^2 \geq \max_i \|\mathbf{a}_i\|_2^2$, we obtain the following bound fo any $\gamma \geq C$ for fixed constant C:

$$\Pr\left[|\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\mathbf{1}| \geq \gamma K^2\|\mathbf{A}\|_F^2\right] \leq 2e^{-c\gamma} \tag{A.4}$$

Next we bound the $\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}}$ term from Eq. (A.3) using a Hanson-Wright type inequality for sub-exponential random variables due to [10].[4] A similar bound is proven in [19].

LEMMA A.3. *(Proposition 1.1 from [10]) Let* $\mathbf{x}$ *be a random vector with i.i.d. mean* 0*, variance* σ^2 *random entries that are sub-exponential with parameter* E *and let* $\mathbf{M}$ *be any* $n \times n$ *matrix. For any* $t > 0$ *we have,*

$$\mathbb{P}\left(\left|\mathbf{x}^T\mathbf{M}\mathbf{x} - \sum_{i=1}^n \sigma^2 M_{ii}\right| \geq tE^2\right) \leq 2\exp\left(-c''\min\left(\frac{t^2}{\|\mathbf{M}\|_F^2}, \left(\frac{t}{\|\mathbf{M}\|_2}\right)^{1/2}\right)\right).$$

To apply Lemma A.3 to $\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}}$, first note that $\bar{\mathbf{g}}$'s entries have variance $\sigma^2 \leq CK^4$ for some fixed constant C because they are sub-exponential with parameter $c'K^2$. So we have that $\sum_{i=1}^n \sigma^2 M_{ii} \leq CK^4\sum_{i=1}^n A_{ii}^2 \leq CK^4\|\mathbf{A}\|_F^2$. Then plugging in $t = \frac{1}{c'^2}\gamma\|\mathbf{A}\|_F^2$ and using that $\|\mathbf{A}\|_F^4 = (\sum_{i,j} A_{ij}^2)^2 \geq \sum_{i,j} A_{ij}^4 = \|\mathbf{A}\circ\mathbf{A}\|_F^2 \geq \|\mathbf{A}\circ\mathbf{A}\|_2^2$, we have that:

$$\mathbb{P}\left(\left|\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}}\right| \geq (\gamma K^4 + CK^4)\|\mathbf{A}\|_F^2\right) \leq 2e^{-c\gamma^{1/2}},$$

for some fixed constant c and any $\gamma \geq 0$. Under our assumption that γ is larger than a fixed constant, we have that $CK^4 = O(\gamma K^4)$, so we can adjust the constant c to simplify the expression to

$$\mathbb{P}\left(\left|\bar{\mathbf{g}}^T(\mathbf{A}\circ\mathbf{A})\bar{\mathbf{g}}\right| \geq \gamma K^4\|\mathbf{A}\|_F^2\right) \leq 2e^{-c\gamma^{1/2}}. \tag{A.5}$$

Plugging in Eq. (A.4) and Eq. (A.5) to Eq. (A.3) an applying a union bound, we conclude that:

$$\Pr[|\mathbf{g}^{2^T}(\mathbf{A}\circ\mathbf{A})\mathbf{g}^2| \geq \gamma K^4\|\mathbf{A}\|_F^2 + 2\gamma K^2\|\mathbf{A}\|_F^2 + \|\mathbf{A}\|_F^2] \leq 4e^{-c\gamma^{1/2}}.$$

[4] The bound in [10] is stated for *symmetric* matrices, but it holds for all matrices without modification. In particular, for any $\mathbf{M}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\left(\frac{\mathbf{M}+\mathbf{M}^T}{2}\right)\mathbf{x}$, and by triangle inequality the symmetric matrix $\frac{\mathbf{M}+\mathbf{M}^T}{2}$ has Frobenius and spectral norm upper bounded by those of $\mathbf{A}$.

Since each entry of $\mathbf{g}$ has variance 1, K is greater than a fixed constant, so $\gamma K^4\|\mathbf{A}\|_F^2 + 2\gamma K^2\|\mathbf{A}\|_F^2 + \|\mathbf{A}\|_F^2 = O(\gamma K^4\|\mathbf{A}\|_F^2)$. So again adjusting constants, we can simplify the above expression to claim that for any $\gamma \geq 0$,

$$\Pr[\|\tilde{\mathbf{A}}\|_F^2 \geq \gamma K^4\|\mathbf{A}\|_F^2] \leq 2e^{-c\gamma^{1/2}}, \tag{A.6}$$

where we recall that $\|\tilde{\mathbf{A}}\|_F^2 = {\mathbf{g}^2}^T(\mathbf{A}\circ\mathbf{A})\mathbf{g}^2$.

We are now close to proving Lemma A.2. To do so, we need to bound $\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2$, which as discussed, is exactly equal to Hutchinson's estimator applied to the positive semi-definite matrix $\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$ – i.e., $\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2 = \mathbf{r}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\mathbf{r}$ where $\mathbf{r}$ is a Rademacher random vector. Let $\mathbf{B}$ denote $\mathbf{B} = \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$. It follows from the Hanson-Wright inequality (see [22], Theorem 6.2.1) that:

$$\Pr\left[\left|\mathbf{r}^T\mathbf{B}\mathbf{r} - \mathrm{tr}(\mathbf{B})\right| \geq \gamma\|\mathbf{B}\|_F\right] \leq 2e^{-c\gamma}.$$

Since $\mathbf{B}$ is PSD, we have that $\|\mathbf{B}\|_F \leq \mathrm{tr}(\mathbf{B})$ and further we have that $\mathrm{tr}(\mathbf{B}) = \|\tilde{\mathbf{A}}\|_F^2$. So we can apply triangle inequality to conclude that $\Pr\left[\mathbf{r}^T\mathbf{B}\mathbf{r} \geq (\gamma+1)\|\tilde{\mathbf{A}}\|_F^2\right] \leq 2e^{-c\gamma}$. Adjusting constants, it follows that for any γ,

$$\Pr\left[\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2 \geq \gamma\|\tilde{\mathbf{A}}\|_F^2\right] \leq 2e^{-c\gamma}. \tag{A.7}$$

We combine this bound with Eq. (A.6) to conclude that:

$$\Pr\left[\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2 \geq \gamma K^4\|\mathbf{A}\|_F^2\right] \leq 2e^{-c\gamma^{1/3}}. \tag{A.8}$$

To obtain Eq. (A.8), observe that to have $\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2 \geq \gamma K^4\|\mathbf{A}\|_F^2$ it must be that *either* $\|\tilde{\mathbf{A}}\|_F^2 \geq \gamma^{2/3}K^4\|\mathbf{A}\|_F^2$ or that $\|\mathbf{g}\odot\mathbf{A}\mathbf{g}\|_2^2 \geq \gamma^{1/3}\|\tilde{\mathbf{A}}\|_F^2$. By Eq. (A.6), the first event only happens with probability $\leq 2e^{-c\gamma^{1/3}}$ and by Eq. (A.7) the second only happens with probability $\leq 2e^{-c\gamma^{1/3}}$. Adjusting constants gives the equation.

Finally, we return to equation Eq. (A.1), combining it with Eq. (A.8) to conclude that:

$$\Pr\left[\|\mathbf{e}\|_2^2 \geq (2\gamma K^4+1)\|\mathbf{A}\|_F^2\right] \leq 2e^{-c\gamma^{1/3}}.$$

Again, since K is greater than a fixed constant, we have that $\gamma K^4 = \Omega(1)$ and adjusting constants yields Lemma A.2. ∎

A.3 Completing the Proof We are finally ready to prove Lemma A.1, which we do by taking advantage of the symmetry of $\mathbf{d}_1^m(\mathbf{A}) - \mathbf{d}_2^m$. Our proof uses a standard version of McDiamard's inequality (see e.g. [22], Theorem 2.9.1), which we state below:

FACT A.1. (MCDIARMID'S INEQUALITY) *Let $x_1,\ldots,x_m \in \mathcal{X}_1\times\ldots\times\mathcal{X}_m$ be independent random variables from domains $\mathcal{X}_1,\ldots,\mathcal{X}_m$. Let $f:\mathcal{X}_1\times\ldots\times\mathcal{X}_m\to\mathbb{R}$ be any function such that for each coordinate i and all realizations of $x_1,\ldots,x_m$, we have a difference bound of $\max_{\tilde{x}_i\in\mathcal{X}_i}|f(x_1,\ldots,x_i,\ldots,x_m) - f(x_1,\ldots,\tilde{x}_i,\ldots,x_m)| \leq c_i$. Then for any $t>0$,*

$$\Pr\left[|f(\mathbf{x}_1,\ldots,\mathbf{x}_m) - \mathbb{E}\,f(\mathbf{x}_1,\ldots,\mathbf{x}_m)| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^m c_i^2}}.$$

Proof. [Proof of Lemma A.1] For $z\in 1,\ldots,n$ and $i\in 1,2$, let $\mathbf{e}_{z,i}$ be a random variable distributed as $\mathbf{g}\odot\mathbf{A}\mathbf{g} - \mathrm{diag}(\mathbf{A})$. Let $\mathbf{w}_z = \mathbf{e}_{z,1} - \mathbf{e}_{z,2}$ and let $r_1,\ldots,r_m$ be i.i.d Rademacher random variables. By the symmetry of each $\mathbf{w}_z$, we can write:

$$\mathbf{d}_1^m - \mathbf{d}_2^m = \frac{1}{m}\sum_{z=1}^m r_z\mathbf{w}_z.$$

We condition on the random choice of $\mathbf{w}_1,\ldots,\mathbf{w}_z$ and apply McDiamard's inequality. Specifically, by triangle inequality, the function $f(r_1,\ldots,r_m) = \left\|\frac{1}{m}\sum_{z=1}^m r_z\mathbf{w}_z\right\|_2 = \|\mathbf{d}_1^m(\mathbf{A}) - \mathbf{d}_2^m\|_2$ can change by at most $2\|\mathbf{w}_z\|_2/m$ if we change the input r_z. So by Fact A.1, we have that:

$$\Pr\left[\left|\|\mathbf{d}_1^m - \mathbf{d}_2^m\|_2 - \mathbb{E}\left[\|\mathbf{d}_1^m - \mathbf{d}_2^m\|_2\right]\right| \geq t\right] \leq 2e^{-\frac{m^2t^2}{2\sum_{z=1}^m\|\mathbf{w}_z\|_2^2}}.$$

By triangle inequality, we have that $\mathbb{E}\left[\|\mathbf{d}_1^m - \mathbf{d}_2^m\|_2\right] \leq \mathbb{E}\left[\|\mathbf{d}_1^m\|_2\right] + \mathbb{E}\left[\|\mathbf{d}_2^m\|_2\right]$. Moreover, by Lemma 4.1, and the fact that $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ for any random variable X, we have that $\mathbb{E}\left[\|\mathbf{d}_1^m\|_2\right] \leq cK^2\|\mathbf{A}\|_F/\sqrt{m}$ for a fixed constant c. So plugging in $t = \frac{1}{m}\sqrt{2\gamma\sum_{z=1}^m \|\mathbf{w}_z\|_2^2}$, overall we conclude that for any $\gamma \geq 0$,

$$\Pr\left[\|\mathbf{d}_1^m - \mathbf{d}_2^m\|_2 \geq \sqrt{\frac{1}{m}}\left(K^2\|\mathbf{A}\|_F + \sqrt{2\gamma\sum_{z=1}^m \|\mathbf{w}_z\|_2^2/m}\right)\right] \leq 2e^{-\gamma}. \tag{A.9}$$

With Eq. (A.9) in place, we are left to bound $\sum_{z=1}^m \|\mathbf{w}_z\|_2^2$. By triangle inequality, this sum can be upper bounded $2\sum_{z=1}^m \|\mathbf{e}_{z,1}\|_2^2 + \|\mathbf{e}_{z,2}\|_2^2$. By Lemma A.2, each $\mathbf{e}_{z,i}$ satisfies $\Pr\left[\|\mathbf{e}_{z,i}\|_2^2 \geq \gamma K^4\|\mathbf{A}\|_F^2\right] \leq 2e^{-c\gamma^{1/3}}$. So, following the characterization of generalized subexponential random variables from [19] (see Proposition 5.1 in that work), we conclude that for a constant c, $\|\mathbf{e}_{z,i}\|_2^2$ is an α-subexponential random variable[5] for $\alpha = 1/3$, with parameter $cK^4\|\mathbf{A}\|_F^2$. Applying Lemma A.3 from [10], we have that $\|\mathbf{e}_{z,i}\|_2^2 - \mathbb{E}[\|\mathbf{e}_{z,i}\|_2^2]$ is also $\frac{1}{3}$ subexponential with parameter $c'K^4\|\mathbf{A}\|_F$. We can then apply Corollary 1.4 from [10] to conclude that for all $\beta \geq 0$,

$$\Pr\left(\left|\sum_{z=1}^m \sum_{i=1,2} \|\mathbf{e}_{z,i}\|_2^2 - \mathbb{E}\left[\|\mathbf{e}_{z,i}\|_2^2\right]\right| \geq m \cdot \beta K^4\|\mathbf{A}\|_F^2\right) \leq 2e^{-c\min\left(\beta m, \beta^{1/3}m^{1/3}\right)}.$$

By Lemma 4.1 we have that $\mathbb{E}\left[\|\mathbf{e}_{z,i}\|_2^2\right] \leq \frac{C}{2}K^4\|\mathbf{A}\|_F^2$ for all i, z and a constant C. So, applying triangle inequality, adjusting constants, and recalling that $\sum_{z=1}^m \|\mathbf{w}_z\|_2^2 \leq 2\sum_{z=1}^m \sum_{i=1,2} \|\mathbf{e}_{z,i}\|_2^2$, we conclude that:

$$\Pr\left(\sum_{z=1}^m \|\mathbf{w}_z\|_2^2 \geq m \cdot (1+\beta)K^4\|\mathbf{A}\|_F^2\right) \leq 2e^{-c\beta^{1/3}m^{1/3}}. \tag{A.10}$$

Combining Eq. (A.10) with Eq. (A.9) and again adjusting constants we have that for constants C, c,

$$\Pr\left[\|\mathbf{d}_1^m - \mathbf{d}_2^m\|_2 \geq \sqrt{\frac{1}{m}}\left(1 + \sqrt{\gamma(1+\beta)}\right)K^2\|\mathbf{A}\|_F\right] \leq 2e^{-\gamma} + 2e^{-c\beta^{1/3}m^{1/3}}.$$

The right hand side of the inequality is $\leq \delta$ as long as $\gamma \geq \log(4/\delta)$ and $\beta \geq \frac{\log^3(4/\delta)/c^3}{m}$. Plugging in and adjusting constants proves Lemma A.1.

□

[5] Note that this is different from a subexponential random variable with parameter α, as in Definition 2.2. An α-subexponential random variable as defined by [10, 19] has slower asymptotic tail decay than a standard subexponential random variable when $\alpha < 1$.