

# Appendices

## Contents

<b>Appendices</b>	<b>1</b>
<b>A Basic Facts about Gaussian Distributions</b>	<b>2</b>
<b>B An Improved Analysis of NA-Hutch++</b>	<b>3</b>
<b>C Lower Bounds</b>	<b>8</b>
C.1 Case 1: Lower Bound for Small $\epsilon$ . . . . .	8
C.2 Case 2: Lower Bound for Every $\epsilon$ . . . . .	9

## A Basic Facts about Gaussian Distributions

Let  $\mathcal{N}(\mu, \sigma^2)$  denote a Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Let  $\chi^2(n)$  denote a  $\chi^2$  distribution with  $n$  degrees of freedom. Our analysis extensively uses the following facts about Gaussian and  $\chi^2$  distributions:

**Definition A.1** (Gaussian and Wigner Random Matrices). *We let  $\mathbf{G} \sim \mathcal{N}(n)$  denote an  $n \times n$  random Gaussian matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. We let  $\mathbf{W} \sim \mathcal{W}(n) = \mathbf{G} + \mathbf{G}^T$  denote an  $n \times n$  Wigner matrix, where  $\mathbf{G} \sim \mathcal{N}(n)$ .*

**Fact A.1** ( $\chi^2$  Tail Bound (**Lemma 1** of [1])). *Let  $Z \sim \chi^2(n)$ . Then for any  $x > 0$ ,*

$$\begin{aligned}\Pr[Z \geq n + 2\sqrt{nx} + 2x] &\leq e^{-x} \\ \Pr[Z \leq n - 2\sqrt{nx}] &\leq e^{-x}\end{aligned}$$

**Fact A.2** (Rotational Invariance). *Let  $\mathbf{R} \in \mathbb{R}^{n \times n}$  be an orthonormal matrix. Let  $\mathbf{g} \in \mathbb{R}^n$  be a random vector with i.i.d.  $\mathcal{N}(0, 1)$  entries. Then  $\mathbf{R}\mathbf{g}$  has the same distribution as  $\mathbf{g}$ .*

**Fact A.3** (Upper Gaussian Tail Bound). *Let  $Z \sim \mathcal{N}(0, \sigma^2)$  be a univariate Gaussian random variable. Then for any  $t > 0$ ,*

$$\Pr[Z \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

**Fact A.4** (Lower Gaussian Tail Bound). *Letting  $Z \sim \mathcal{N}(0, 1)$  be a univariate Gaussian random variable, for any  $t > 0$ ,*

$$\Pr[Z \geq t] \geq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{t} \exp(t^2/2)$$

**Lemma A.2** (Concentration of Singular Values of a Gaussian Random Matrix (**Eq. 2.3** of [2])). *Let  $\mathbf{G} \sim \mathcal{N}(n)$ , and  $s_{\max}(\mathbf{G})$  denote the maximum singular value of  $\mathbf{G}$ . Then  $\forall t \geq 0$ ,*

$$\Pr[s_{\max}(\mathbf{G}) \leq 2\sqrt{n} + t] \geq 1 - 2\exp(-t^2/2)$$

**Fact A.5** (KL Divergence Between Multivariate Gaussian Distributions (**Eq. 8** of [3], or Section 9 of [4])). *Let  $\mathcal{P} \sim \mathcal{N}(\mu_1, \Sigma_1)$  and  $\mathcal{Q} \sim \mathcal{N}(\mu_2, \Sigma_2)$  be two  $k$ -dimensional multivariate normal distributions. The Kullback-Leibler divergence between  $\mathcal{P}$  and  $\mathcal{Q}$  is*

$$\mathcal{D}_{KL}(\mathcal{P} \parallel \mathcal{Q}) = \frac{1}{2} \left\{ (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{\det(\Sigma_1)}{\det(\Sigma_2)} - k \right\}$$

**Fact A.6** (Conditioning Increases KL Divergence (**Theorem 2.2 - 5** of [5])). *Let  $\mathcal{P}_{Y|X}$ ,  $\mathcal{Q}_{Y|X}$  be two conditional probability distributions over spaces  $X \in \mathcal{X}$  and  $Y \in \mathcal{Y}$ , let  $\mathcal{P}_Y = \mathcal{P}_{Y|X} \mathcal{P}_X$  and  $\mathcal{Q}_Y = \mathcal{Q}_{Y|X} \mathcal{P}_X$ . Then,*

$$\mathcal{D}_{KL}(\mathcal{P}_Y \parallel \mathcal{Q}_Y) \leq \mathcal{D}_{KL}(\mathcal{P}_{Y|X} \parallel \mathcal{Q}_{Y|X} \mid \mathcal{P}_X) := \int \mathcal{D}_{KL}(\mathcal{P}_{Y|X=x} \parallel \mathcal{Q}_{Y|X=x}) d\mathcal{P}_X$$

**Fact A.7** (KL Divergence Data Processing Inequality (Page 18 of [6])). *For any function  $f$  and random variables  $X$  and  $Y$  on the same probability space, it holds that*

$$\mathcal{D}_{KL}(f(X) \parallel f(Y)) \leq \mathcal{D}_{KL}(X \parallel Y)$$

## B An Improved Analysis of NA-Hutch++

In this section, we give an improved analysis of NA-Hutch++, showing that the query complexity of NA-Hutch++ can be improved from  $O(\log(1/\delta)/\epsilon)$ , as shown in [7], to  $O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$  on PSD (positive semidefinite) input matrices  $\mathbf{A}$ , to get a  $(1 \pm \epsilon)$  approximation to  $\text{tr}(\mathbf{A})$  with probability  $1 - \delta$ . The NA-Hutch++ algorithm is duplicated here for convenience as follows:

---

**Algorithm 1** NA-Hutch++ [7]: Stochastic trace estimation with **non-adaptive** matrix-vector queries

---

- 1: **Input:** Matrix-vector multiplication oracle for PSD matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ . Number  $m$  of queries.
  - 2: **Output:** Approximation to  $\text{tr}(\mathbf{A})$ .
  - 3: Fix constants  $c_1, c_2, c_3$  such that  $c_1 < c_2$  and  $c_1 + c_2 + c_3 = 1$ .
  - 4: Sample  $\mathbf{S} \in \mathbb{R}^{n \times c_1 m}$ ,  $\mathbf{R} \in \mathbb{R}^{n \times c_2 m}$ , and  $\mathbf{G} \in \mathbb{R}^{n \times c_3 m}$ , with i.i.d.  $\mathcal{N}(0, 1)$  entries.
  - 5:  $\mathbf{Z} = \mathbf{AR}$ ,  $\mathbf{W} = \mathbf{AS}$
  - 6: **return**  $t = \text{tr}((\mathbf{S}^T \mathbf{Z})^\dagger (\mathbf{W}^T \mathbf{Z})) + \frac{1}{c_3 m} (\text{tr}(\mathbf{G}^T \mathbf{A} \mathbf{G}) - \text{tr}(\mathbf{G}^T \mathbf{Z} (\mathbf{S}^T \mathbf{Z})^\dagger \mathbf{W}^T \mathbf{G}))$ .
- 

**Roadmap.** Recall that NA-Hutch++ splits its matrix-vector queries between computing an  $O(1)$ -approximate rank- $k$  approximation  $\tilde{\mathbf{A}}$  and performing Hutchinson's estimate on the residual matrix  $\mathbf{A} - \tilde{\mathbf{A}}$ . The key to an improved query complexity of NA-Hutch++ is on the analysis of the size of random Gaussian sketching matrices  $\mathbf{S}$ ,  $\mathbf{R}$  in Algorithm 1 that one needs to get an  $O(1)$ -approximate rank- $k$  approximation  $\tilde{\mathbf{A}}$  in the Frobenius norm. To get the desired rank- $k$  approximation, we need  $\mathbf{S}$  and  $\mathbf{R}$  to satisfy two properties: 1) subspace embedding as in **Lemma 3.3** and 2) approximate matrix product for orthogonal subspaces as in **Lemma 3.4**. Specifically, we show in **Lemma 3.4** that choosing  $\mathbf{S}$  and  $\mathbf{R}$  to be of size  $O(k + \log(1/\delta))$  suffices to get the second property with probability  $1 - \delta$ .

After that, we show in **Lemma B.1** that if a sketching matrix  $\mathbf{S}$  satisfies the two properties mentioned above, with size  $O(k + \log(1/\delta))$ , one gets an  $O(1)$ -approximate low rank approximation with probability  $1 - \delta$  when solving a sketched version of the regression problem  $\min_{\mathbf{X}} \|\mathbf{S}^T (\mathbf{A} \mathbf{X} - \mathbf{B})\|_F$  for fixed matrices  $\mathbf{A}, \mathbf{B}$  with  $\text{rank}(\mathbf{A}) = k$ . **Lemma B.1** serves as an intermediate step to construct an  $O(1)$ -approximate rank- $k$  approximation  $\tilde{\mathbf{A}}$  with  $\mathbf{S}, \mathbf{R}$  having a size of only  $O(k + \log(1/\delta))$  in **Theorem 3.5**.

Finally, we combine **Theorem 3.2** from [7], which shows the trade-off between the rank  $k$  and the number  $l$  spent on estimating the small eigenvalues, and **Theorem 3.5**, which shows the number of non-adaptive queries one needs to get a desired rank- $k$  factor, to conclude in **Theorem 3.1** that NA-Hutch++ needs only  $O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$  non-adaptive queries, by setting  $k = \frac{\sqrt{\log(1/\delta)}}{\epsilon}$ .

**Lemma 3.3** (Subspace Embedding (Theorem 6 of [8])). *Given  $\delta \in (0, \frac{1}{2})$  and  $\epsilon \in (0, 1)$ , let  $\mathbf{S} \in \mathbb{R}^{r \times n}$  be a random matrix with i.i.d. Gaussian random variables  $\mathcal{N}(0, \frac{1}{r})$ . Then for any fixed  $d$ -dimensional subspace  $\mathbf{A} \in \mathbb{R}^{n \times d}$ , and for  $r = O((d + \log(\frac{1}{\delta}))/\epsilon^2)$ , the following holds with probability  $1 - \delta$  simultaneously for all  $x \in \mathbb{R}^d$ ,*

$$\|\mathbf{S} \mathbf{A} x\|_2 = (1 \pm \epsilon) \|\mathbf{A} x\|_2$$

**Lemma 3.4** (Approximate Matrix Product for Orthogonal Subspaces). *Given  $\delta \in (0, \frac{1}{2})$ , let  $\mathbf{U} \in \mathbb{R}^{n \times k}$ ,  $\mathbf{W} \in \mathbb{R}^{n \times p}$  be two matrices with orthonormal columns such that  $\mathbf{U}^T \mathbf{W} = \mathbf{0}$ ,  $p \geq \max(k, \log(1/\delta))$ ,  $\text{rank}(\mathbf{U}) = k$  and  $\text{rank}(\mathbf{W}) = p$ . Let  $\mathbf{S} \in \mathbb{R}^{r \times n}$  be a random matrix with i.i.d. Gaussian random variables  $\mathcal{N}(0, \frac{1}{r})$ . For  $r = O(k + \log(\frac{1}{\delta}))$ , the following holds with probability  $1 - \delta$ ,*

$$\|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{W}\|_F \leq O(1) \|\mathbf{W}\|_F$$

*Proof.* Let  $\mathbf{G} = \sqrt{r} \mathbf{U}^T \mathbf{S}^T \in \mathbb{R}^{k \times r}$  and  $\mathbf{H} = \sqrt{r} \mathbf{S} \mathbf{W} \in \mathbb{R}^{r \times p}$ . Since both  $\mathbf{U}$  and  $\mathbf{W}$  have orthonormal columns, both  $\mathbf{G}$  and  $\mathbf{H}$  are random matrices with i.i.d. Gaussian random variables  $\mathcal{N}(0, 1)$ . Furthermore, let  $\mathbf{g}_i, \forall i \in [k]$  denote the  $i$ -th row of  $\mathbf{G}$  and  $\mathbf{h}_j, \forall j \in [p]$  denote the  $j$ -th column of  $\mathbf{H}$ .

$$\begin{aligned} \|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{W}\|_F^2 &= \left\| \frac{1}{\sqrt{r}} \mathbf{G} \frac{1}{\sqrt{r}} \mathbf{H} \right\|_F^2 \\ &= \frac{1}{r^2} \sum_{i=1}^k \sum_{j=1}^p \langle \mathbf{g}_i, \mathbf{h}_j \rangle^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{r^2} \sum_{i=1}^k \sum_{j=1}^p \|g_i\|_2^2 \left\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{h}_j \right\rangle^2 \\
&= \frac{1}{r^2} \sum_{i=1}^k \|g_i\|_2^2 \left( \sum_{j=1}^p \left\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{h}_j \right\rangle^2 \right)
\end{aligned}$$

Since  $\|\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}\|_2 = 1$ ,  $\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{h}_j \rangle \sim \mathcal{N}(0, 1)$ . Thus,

$$\|U^T S^T S W\|_F^2 = \frac{1}{r^2} \sum_{i=1}^k \mathbf{c}_i \cdot \mathbf{d}_i$$

where  $\mathbf{c}_i \sim \chi^2(r)$ ,  $\mathbf{d}_i \sim \chi^2(p)$ ,  $\forall i \in [k]$ . Note that since  $\mathbf{W}$  has orthonormal columns,  $\|\mathbf{W}\|_F^2 = p$ .

The number  $r$  of rows our random sketch matrix  $\mathbf{S}$  needs in order to obtain an upper bound on the product of random Gaussian matrices  $\mathbf{S}U$  and  $\mathbf{S}W$ , up to a constant factor of  $\|\mathbf{W}\|_F$ , depends on the concentration of  $\mathbf{S}U$  and  $\mathbf{S}W$ . Specifically, to apply the  $\chi^2$  tail bound on some random variable  $\mathbf{v} \sim \chi^2(d)$  from Fact A.1 and to get that  $\mathbf{v}$  concentrates around  $O(1)d$  with probability  $1 - \delta$ , the degree  $d$  needs to be at least  $\log(1/\delta)$ . Since we require  $p = \text{rank}(\mathbf{W}) \geq \log(1/\delta)$ ,  $\mathbf{S}W$  is concentrated with high probability. The concentration of  $\mathbf{S}U$  depends on  $\text{rank}(U) = k$ . To upper bound  $\|(\mathbf{S}U)^T(\mathbf{S}W)\|_F$ , we consider two cases for  $k$ :

**Case I:** Consider the case when  $k \geq \log(\frac{1}{\delta})$ :

Since  $p \geq k \geq \log(\frac{1}{\delta})$ , by Fact A.1,  $\forall i \in [k]$ ,

$$\Pr[\mathbf{d}_i \leq O(1)p] \geq 1 - e^{-O(k)}$$

Since  $r = O(k + \log(1/\delta))$ , by Fact A.1,  $\forall i \in [k]$ ,

$$\Pr[\mathbf{c}_i \leq O(1)k] \geq 1 - e^{-O(k)}$$

By a union bound over  $2k$   $\chi^2$  random variables,

$$\Pr \left[ \sum_{i=1}^k \mathbf{c}_i \cdot \mathbf{d}_i \leq O(1)k^2p \right] \geq 1 - 2k \cdot e^{-O(k)}$$

Thus with probability  $1 - O(\delta)$ ,

$$\begin{aligned}
\|U^T S^T S W\|_F^2 &= \frac{1}{r^2} \sum_{i=1}^k \mathbf{c}_i \cdot \mathbf{d}_i \\
&\leq \frac{1}{r^2} O(1)k^2p \\
&= \frac{1}{r^2} O(1)k^2 \|\mathbf{W}\|_F^2
\end{aligned}$$

And so  $r = O(k + \log(1/\delta))$  gives  $\|U S^T S W\|_F \leq O(1)\|\mathbf{W}\|_F$  with probability  $1 - \delta$ .

**Case II:** Consider the case when  $k < \log(\frac{1}{\delta})$ .

Since  $p \geq \log(\frac{1}{\delta})$ , by Fact A.1,  $\forall i \in [k]$ ,

$$\Pr[\mathbf{d}_i \leq O(1)p] \geq 1 - e^{-O(\log(1/\delta))}$$

Since  $r = O(k + \log(1/\delta))$ , by Fact A.1,  $\forall i \in [k]$ ,

$$\Pr[\mathbf{c}_i \leq O(1)\log(1/\delta)] \geq 1 - e^{-O(\log(1/\delta))}$$

By a union bound over  $2k$   $\chi^2$  random variables, for  $k < \log(1/\delta)$

$$\Pr \left[ \sum_{i=1}^k \mathbf{c}_i \cdot \mathbf{d}_i \leq O(1)k \log(1/\delta)p \right] \geq 1 - 2k \cdot e^{-O(\log(1/\delta))}$$

Thus with probability  $1 - O(\delta)$ ,

$$\begin{aligned}\|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{W}\|_F^2 &= \frac{1}{r^2} \sum_{i=1}^k \mathbf{c}_i \cdot \mathbf{d}_i \\ &\leq \frac{1}{r^2} O(1) k \log(1/\delta) p \\ &= \frac{1}{r^2} O(1) k \log(1/\delta) \|\mathbf{W}\|_F^2\end{aligned}$$

Since  $k < \log(1/\delta)$ ,  $r = O(k + \log(1/\delta))$  in this case gives  $\|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{W}\|_F \leq O(1) \|\mathbf{W}\|_F$  with probability  $1 - \delta$ .

Combining **Case I** and **Case II** allows us to conclude that for  $r = O(k + \log(1/\delta))$ ,  $\|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{W}\|_F \leq O(1) \|\mathbf{W}\|_F$  with probability  $1 - \delta$ .

□

**Lemma B.1** (Upper Bound on Regression Error). *Given  $\delta \in (0, \frac{1}{2})$ , let  $\mathbf{A}, \mathbf{B}$  be matrices that both have  $n$  rows and  $\text{rank}(\mathbf{A}) = k$ . Let  $\mathbf{S} \in \mathbb{R}^{n \times r}$  be a random matrix with i.i.d.  $\mathcal{N}(0, \frac{1}{r})$  Gaussian random variables. Let  $\tilde{\mathbf{X}} = \arg \min_{\mathbf{X}} \|\mathbf{S}^T (\mathbf{A} \mathbf{X} - \mathbf{B})\|_F$  and  $\mathbf{X}^* = \arg \min_{\mathbf{X}} \|\mathbf{A} \mathbf{X} - \mathbf{B}\|_F$ . For  $r = O(k + \log(1/\delta))$ , the following holds with probability  $1 - \delta$ ,*

$$\|\mathbf{A} \tilde{\mathbf{X}} - \mathbf{B}\|_F \leq O(1) \|\mathbf{A} \mathbf{X}^* - \mathbf{B}\|_F$$

*Proof.* Consider an orthonormal basis  $\mathbf{U}$  for the column span of  $\mathbf{A}$ . Let  $\tilde{\mathbf{Y}} = \arg \min_{\mathbf{Y}} \|\mathbf{S} \mathbf{U} \mathbf{Y} - \mathbf{S} \mathbf{B}\|_2$  and  $\mathbf{Y}^* = \arg \min_{\mathbf{Y}} \|\mathbf{U} \mathbf{Y} - \mathbf{B}\|_2$ . By the normal equations, the solutions to the two least squares problems are  $\tilde{\mathbf{Y}} = (\mathbf{S} \mathbf{U})^\dagger \mathbf{S} \mathbf{B}$  and  $\mathbf{Y}^* = \mathbf{U}^T \mathbf{B}$ .

We first show that  $\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F \leq O(1) \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F$ .

$$\begin{aligned}\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F^2 &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + \|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{U} \mathbf{Y}^*\|_F^2 \\ &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + \|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 && \text{(Since } \mathbf{U} \text{ has orthonormal columns)} \\ &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + \|(\mathbf{S} \mathbf{U})^\dagger \mathbf{S} \mathbf{B} - \mathbf{U}^T \mathbf{B}\|_F^2 \\ &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + \|(\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{B} - \mathbf{U}^T \mathbf{B}\|_F^2\end{aligned}$$

Since  $\mathbf{S}$  is a matrix with i.i.d.  $\mathcal{N}(0, \frac{1}{r})$  Gaussian random variables, by **Fact 3.3**, for any vector  $v \in \mathbb{R}^n$ , with probability  $1 - \delta$  and for some fixed constant  $\epsilon_1 \in (0, 1)$ ,  $\|\mathbf{S} \mathbf{U} v\|_2 = (1 \pm \epsilon_1) \|\mathbf{U} v\|_2$ . This implies the singular values of  $\mathbf{S} \mathbf{U}$  are in the range  $[1 - \epsilon_1, 1 + \epsilon_1]$ . Thus,

$$\begin{aligned}\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F^2 &\leq \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + O(1) \|(\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{B} - \mathbf{U}^T \mathbf{B}\|_F^2 \\ &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + O(1) \|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{B} - \mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{U} \mathbf{U}^T \mathbf{B}\|_F^2 \\ &= \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + O(1) \|\mathbf{U}^T \mathbf{S}^T \mathbf{S} (\mathbf{B} - \mathbf{U} \mathbf{Y}^*)\|_F^2\end{aligned}$$

Consider  $p = \text{rank}(\mathbf{U} \mathbf{Y}^* - \mathbf{B})$ . If  $p = O(k)$ , then  $\text{rank}(\mathbf{B}) = O(k)$ . For  $r = O(k)$ , we can use  $\mathbf{S}$  to reconstruct  $\mathbf{A}$  and  $\mathbf{B}$ . In this case,  $\tilde{\mathbf{X}} = \mathbf{X}^*$  and so  $\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F \leq O(1) \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F$ . If  $p = O(\log(1/\delta))$ , then  $\text{rank}(\mathbf{B}) = O(k + \log(1/\delta))$ . For  $r = O(k + \log(1/\delta))$ , we can again use  $\mathbf{S}$  to reconstruct  $\mathbf{A}$  and  $\mathbf{B}$  and get  $\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F \leq O(1) \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F$ .

Now consider  $p \geq \max(k, \log(1/\delta))$ . First note that  $\mathbf{B} - \mathbf{U} \mathbf{Y}^* = \mathbf{B} - \mathbf{U} \mathbf{U}^T \mathbf{B} = (\mathbf{I} - \mathbf{U} \mathbf{U}^T) \mathbf{B}$ , where  $\mathbf{U}$  has orthonormal columns and thus,  $\mathbf{U} \mathbf{U}^T$  is the projection matrix onto the column span  $\text{col}(\mathbf{U})$  of  $\mathbf{U}$ . We have  $(\mathbf{B} - \mathbf{U} \mathbf{Y}^*) \perp \text{col}(\mathbf{U})$ . Second, we can w.l.o.g. assume that  $\mathbf{U} \mathbf{Y}^* - \mathbf{B}$  has orthonormal columns; indeed, otherwise let  $\mathbf{U}' \mathbf{R}' = \mathbf{B} - \mathbf{U} \mathbf{Y}^*$  be the QR decomposition where  $\mathbf{U}'$  is an orthonormal basis for  $\text{col}(\mathbf{B} - \mathbf{U} \mathbf{Y}^*)$ . Then  $\|\mathbf{U}^T \mathbf{S}^T \mathbf{S} (\mathbf{B} - \mathbf{U} \mathbf{Y}^*)\|_F^2 = \|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{U}' \mathbf{R}'\|_F^2 = \|\mathbf{U}^T \mathbf{S}^T \mathbf{S} \mathbf{U}'\|_F^2$ .

Applying **Lemma 3.4**, with probability  $1 - O(\delta)$ ,

$$\|\mathbf{U} \tilde{\mathbf{Y}} - \mathbf{B}\|_F^2 \leq \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2 + O(1) \|\mathbf{U} \mathbf{Y}^* - \mathbf{B}\|_F^2$$

<sup>1</sup>† denotes the Moore-Penrose pseudoinverse

$$= O(1)\|\mathbf{U}\mathbf{Y}^* - \mathbf{B}\|_F^2$$

This concludes that  $\|\mathbf{U}\tilde{\mathbf{Y}} - \mathbf{B}\|_F \leq O(1)\|\mathbf{U}\mathbf{Y}^* - \mathbf{B}\|_F$ .

Finally, consider the QR decomposition of  $\mathbf{A} = \mathbf{U}\mathbf{R}$  where  $\mathbf{U}$  is an orthonormal basis for the column span of  $\mathbf{A}$  and  $\mathbf{R}$  is an arbitrary matrix. Let  $\tilde{\mathbf{X}} = \arg \min_{\mathbf{X}} \|\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{B}\|_2$  and  $\mathbf{X}^* = \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_2$ . Note that

$$\begin{aligned} \min_{\mathbf{X}} \|\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{B}\|_F &= \min_{\mathbf{Y}} \|\mathbf{S}\mathbf{U}\mathbf{R}\mathbf{Y} - \mathbf{S}\mathbf{B}\|_F = \min_{\mathbf{Y}} \|\mathbf{S}\mathbf{U}\mathbf{Y} - \mathbf{S}\mathbf{B}\|_F \\ \min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F &= \min_{\mathbf{Y}} \|\mathbf{U}\mathbf{R}\mathbf{Y} - \mathbf{B}\|_F = \min_{\mathbf{Y}} \|\mathbf{U}\mathbf{Y} - \mathbf{B}\|_F \end{aligned}$$

Thus,

$$\|\mathbf{A}\tilde{\mathbf{X}} - \mathbf{B}\|_F = \|\mathbf{U}\tilde{\mathbf{Y}} - \mathbf{B}\|_F \leq O(1)\|\mathbf{U}\mathbf{Y}^* - \mathbf{B}\|_F = O(1)\|\mathbf{A}\mathbf{X}^* - \mathbf{B}\|_F$$

□

The following Theorem and its proof follows **Theorem 4.7** of [9], except that: 1) to get a rank  $k$  approximation to the matrix  $\mathbf{A}$ , the number of columns in the sketching matrices  $\mathbf{S}$  and  $\mathbf{R}$  was required to be  $m = O(k \log(\frac{1}{\delta}))$  in **Theorem 4.7** of [9]; 2)  $\mathbf{S}$  and  $\mathbf{R}$  in **Theorem 4.7** of [9] are random sign matrices. By applying **Lemma B.1**, we show that this number  $m$  can be reduced to  $O(k + \log(\frac{1}{\delta}))$ , and consider a specific application to PSD matrices.

**Theorem 3.5.** *Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  be an arbitrary PSD matrix. Let  $\mathbf{A}_k = \arg \min_{\text{rank-}k \mathbf{A}_k} \|\mathbf{A} - \mathbf{A}_k\|_F$  be the optimal rank- $k$  approximation to  $\mathbf{A}$  in Frobenius norm. If  $\mathbf{S} \in \mathbb{R}^{n \times m}$  and  $\mathbf{R} \in \mathbb{R}^{n \times cm}$  are random matrices with i.i.d.  $\mathcal{N}(0, 1)$  entries for some fixed constant  $c > 0$  with  $m = O(k + \log(1/\delta))$ , then with probability  $1 - \delta$ , the matrix  $\tilde{\mathbf{A}} = (\mathbf{A}\mathbf{R})(\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger (\mathbf{A}\mathbf{S})^T$  satisfies*

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq O(1)\|\mathbf{A} - \mathbf{A}_k\|_F$$

*Proof.* First, we consider  $\mathbf{S}$  to be a random matrix with i.i.d.  $\mathcal{N}(0, \frac{1}{m})$  entries and  $\mathbf{R}$  to be a random matrix with i.i.d.  $\mathcal{N}(0, \frac{1}{cm})$  entries.

Consider  $\tilde{\mathbf{X}} = \arg \min_{\mathbf{X}} \|\mathbf{S}^T \mathbf{A}\mathbf{R}\mathbf{X} - \mathbf{S}^T \mathbf{A}\|_F = (\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger \mathbf{S}^T \mathbf{A}$  and  $\mathbf{X}^* = \arg \min_{\mathbf{X}} \|\mathbf{A}\mathbf{R}\mathbf{X} - \mathbf{A}\|_F$ . By **Lemma B.1**, with probability  $1 - \delta$ ,

$$\|\mathbf{A}\mathbf{R}\tilde{\mathbf{X}} - \mathbf{A}\|_F \leq O(1)\|\mathbf{A}\mathbf{R}\mathbf{X}^* - \mathbf{A}\|_F$$

Now let  $\mathbf{A}_k = \arg \min_{\text{rank } k \mathbf{A}_k} \|\mathbf{A} - \mathbf{A}_k\|_F$  be the optimal rank- $k$  approximation to  $\mathbf{A}$ .

Consider  $\mathbf{X}_{opt} = \arg \min_{\mathbf{X}} \|\mathbf{X}\mathbf{A}_k - \mathbf{A}\|_F$  and  $\mathbf{X}' = \arg \min_{\mathbf{X}} \|\mathbf{X}\mathbf{A}_k\mathbf{R} - \mathbf{A}\mathbf{R}\|_F = (\mathbf{A}\mathbf{R})(\mathbf{A}_k\mathbf{R})^\dagger$ .

By **Lemma B.1** again, with probability  $1 - \delta$ ,

$$\begin{aligned} \|\mathbf{X}'\mathbf{A}_k - \mathbf{A}\|_F &= \|(\mathbf{A}\mathbf{R})(\mathbf{A}_k\mathbf{R})^\dagger \mathbf{A}_k - \mathbf{A}\|_F \\ &\leq O(1)\|\mathbf{X}_{opt}\mathbf{A}_k - \mathbf{A}\|_F = O(1)\|\mathbf{A} - \mathbf{A}_k\|_F \end{aligned}$$

This implies a good rank- $k$  approximation exists in the column span of  $\mathbf{A}\mathbf{R}$ . We now have with probability  $1 - \delta$ ,

$$\|\mathbf{A}\mathbf{R}\mathbf{X}^* - \mathbf{A}\|_F \leq \|(\mathbf{A}\mathbf{R})(\mathbf{A}_k\mathbf{R})^\dagger \mathbf{A}_k - \mathbf{A}\|_F \leq O(1)\|\mathbf{A} - \mathbf{A}_k\|_F$$

Thus by a union bound, with probability  $1 - 2\delta$ ,

$$\begin{aligned} \|\mathbf{A}\mathbf{R}(\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger \mathbf{S}^T \mathbf{A} - \mathbf{A}\|_F &= \|\mathbf{A}\mathbf{R}\tilde{\mathbf{X}} - \mathbf{A}\|_F \\ &\leq O(1)\|\mathbf{A}\mathbf{R}\mathbf{X}^* - \mathbf{A}\|_F \\ &\leq O(1)\|\mathbf{A} - \mathbf{A}_k\|_F \end{aligned}$$

Since we consider PSD  $\mathbf{A}$ ,  $\mathbf{S}^T \mathbf{A} = (\mathbf{A}\mathbf{S})^T$ . Let  $\tilde{\mathbf{A}} = (\mathbf{A}\mathbf{R})(\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger (\mathbf{A}\mathbf{S})^T$ , it follows that with probability  $1 - 2\delta$ ,

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq O(1)\|\mathbf{A} - \mathbf{A}_k\|_F$$

Let  $\mathbf{S}' = \sqrt{m}\mathbf{S}$  and  $\mathbf{R}' = \sqrt{cm}\mathbf{R}$  so that both  $\mathbf{S}'$  and  $\mathbf{R}'$  have i.i.d.  $\mathcal{N}(0, 1)$  entries. Notice that  $(\mathbf{A}\mathbf{R}')(\mathbf{S}'^T \mathbf{A}\mathbf{R}')^\dagger (\mathbf{A}\mathbf{S}')^T = (\mathbf{A}\mathbf{R})(\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger (\mathbf{A}\mathbf{S})^T$ . Thus  $\mathbf{S}$ ,  $\mathbf{R}$  can be chosen to both be random matrices with i.i.d.  $\mathcal{N}(0, 1)$  entries. The theorem follows after adjusting  $\delta$  by a constant factor. □

**Theorem 3.2** (Theorem 4 of [7]). Let  $\mathbf{A} \in \mathbb{R}^{d \times d}$  be PSD,  $\delta \in (0, \frac{1}{2})$ ,  $l \in \mathbb{N}, k \in \mathbb{N}$ . Let  $\tilde{\mathbf{A}}$  and  $\mathbf{\Delta}$  be any matrices with  $\text{tr}(\mathbf{A}) = \text{tr}(\tilde{\mathbf{A}}) + \text{tr}(\mathbf{\Delta})$  and  $\|\mathbf{\Delta}\|_F \leq O(1)\|\mathbf{A} - \mathbf{A}_k\|_F$  where  $\mathbf{A}_k = \arg \min_{\text{rank } k \mathbf{A}_k} \|\mathbf{A} - \mathbf{A}_k\|_F$ . Let  $H_l(\mathbf{M})$  denote Hutchinson's trace estimator with  $l$  queries on matrix  $\mathbf{M}$ . For fixed constants  $c, C$ , if  $l \geq c \log(\frac{1}{\delta})$ , then with probability  $1 - \delta$ ,  $Z = \text{tr}(\tilde{\mathbf{A}}) + H_l(\mathbf{\Delta})$ ,

$$|Z - \text{tr}(\mathbf{A})| \leq C \sqrt{\frac{\log(1/\delta)}{kl}} \cdot \text{tr}(\mathbf{A})$$

**Theorem 3.1.** Let  $\mathbf{A}$  be a PSD matrix. If **NA-Hutch++** is implemented with

$$m = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$$

matrix-vector multiplication queries, then with probability  $1 - \delta$ , the output of **NA-Hutch++**,  $t$ , satisfies  $(1 - \epsilon)\text{tr}(\mathbf{A}) \leq t \leq (1 + \epsilon)\text{tr}(\mathbf{A})$ .

*Proof.* Set  $k = l = O(\frac{\sqrt{\log(1/\delta)}}{\epsilon})$ .

Consider  $\tilde{\mathbf{A}} = (\mathbf{A}\mathbf{R})(\mathbf{S}^T \mathbf{A}\mathbf{R})^\dagger (\mathbf{A}\mathbf{S})^T$ , where  $\mathbf{S} \in \mathbb{R}^{n \times s}$ ,  $\mathbf{R} \in \mathbb{R}^{n \times r}$  are both random matrices with i.i.d.  $\mathcal{N}(0, 1)$  entries, and  $\mathbf{\Delta} = \mathbf{A} - \tilde{\mathbf{A}}$ .

By **Theorem 3.5**, for  $s = r = O(k + \log(1/\delta)) = O(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta))$ , with probability  $1 - \delta$ ,

$$\|\mathbf{\Delta}\|_F \leq O(1) \cdot \|\mathbf{A} - \mathbf{A}_k\|_F$$

Thus for the output of **NA-Hutch++**,  $t$ , by **Theorem 3.2** and a union bound, with probability  $1 - 2\delta$ ,

$$|t - \text{tr}(\mathbf{A})| \leq \epsilon \cdot \text{tr}(\mathbf{A})$$

The total number of non-adaptive queries **NA-Hutch++** needs is

$$m = s + r + l = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right).$$

□

## C Lower Bounds

In this section, we show that a query complexity of  $O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$  is tight for any non-adaptive trace estimation algorithm, up to a  $O(\log \log(1/\delta))$  factor, stated in **Theorem 4.1**. The analysis considers two separate cases: for small  $\epsilon$ , we show the term  $O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$  is tight in Section C.1, and for any  $\epsilon$ , we show the term  $O(\log(1/\delta))$  is tight up to a  $O(\log \log(1/\delta))$  factor in Section C.2. When combined, these two lower bounds handle arbitrary  $\epsilon$ , since the latter lower bound dominates precisely when the former lower bound does not apply.

Our hard distribution consists of shifted Wigner matrices and exploits the symmetry and concentration properties of the Gaussian ensemble.

**Theorem 4.1** (Lower Bound for Non-Adaptive Queries). *Let  $\epsilon \in (0, 1)$ . Any algorithm that accesses a real PSD matrix  $\mathbf{A}$  through matrix-vector multiplication queries  $\mathbf{A}\mathbf{q}_1, \mathbf{A}\mathbf{q}_2, \dots, \mathbf{A}\mathbf{q}_m$ , where  $\mathbf{q}_1, \dots, \mathbf{q}_m$  are real-valued, non-adaptively chosen vectors, requires*

$$m = \Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$$

queries to output an estimate  $t$  such that with probability at least  $1 - \delta$ ,  $(1 - \epsilon)\text{tr}(\mathbf{A}) \leq t \leq (1 + \epsilon)\text{tr}(\mathbf{A})$ .

*Proof of Theorem 4.1.* For small  $\epsilon = O(1/\sqrt{\log(1/\delta)})$ , note that the first term  $\frac{\sqrt{\log(1/\delta)}}{\epsilon}$  dominates.

**Theorem 4.2** (see Section C.1) shows any algorithm needs  $\Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$  non-adaptive queries in this case.

For  $\epsilon > 1/\sqrt{\log(1/\delta)}$ , note that the second term  $\log(1/\delta)$  dominates. **Theorem 4.3** (see Section C.2) shows any algorithm needs  $\Omega\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$  non-adaptive queries for any  $\epsilon \in (0, 1)$ .

The two cases combined imply an  $\Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$  lower bound.  $\square$

### C.1 Case 1: Lower Bound for Small $\epsilon$

Suppose that we draw a matrix  $\mathbf{G} \in \mathbb{R}^{n \times n}$  from the Gaussian distribution and try to learn the entries of the matrix via matrix-vector queries. After a few queries, it turns out that the conditional distribution of the remaining matrix is also Gaussian-distributed, no matter how the queries are chosen. This nice property allows concise reasoning for lower bounding the remaining uncertainty of the matrix, even after seeing a few query results.

**Lemma C.1.** (Conditional Distribution [Lemma 3.4 of [10]]) *Let  $\mathbf{G} \sim \mathcal{N}(n)$  be as in Definition A.1 and suppose our matrix is  $\mathbf{W} = (\mathbf{G} + \mathbf{G}^\top)/2$ . Suppose we have any sequence of vector queries,  $\mathbf{v}_1, \dots, \mathbf{v}_T$ , along with responses  $\mathbf{w}_i = \mathbf{W}\mathbf{v}_i$ . Then, conditioned on our observations, there exists a rotation matrix  $\mathbf{V}$ , independent of  $\mathbf{w}_i$ , such that*

$$\mathbf{V}\mathbf{W}\mathbf{V}^\top = \begin{bmatrix} Y_1 & Y_2^\top \\ Y_2 & \widetilde{\mathbf{W}} \end{bmatrix}$$

where  $Y_1, Y_2$  are deterministic and  $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{G}} + \widetilde{\mathbf{G}}^\top)/2$ , where  $\widetilde{\mathbf{G}} \sim \mathcal{N}(n - T)$ .

**Theorem 4.2** (Lower Bound for Small  $\epsilon$ ). *For any PSD matrix  $\mathbf{A}$  and all  $\epsilon = O(1/\sqrt{\log(1/\delta)})$ , any algorithm that succeeds with probability at least  $1 - \delta$  in outputting an estimate  $t$  such that  $(1 - \epsilon)\text{tr}(\mathbf{A}) \leq t \leq (1 + \epsilon)\text{tr}(\mathbf{A})$ , requires*

$$m = \Omega(\sqrt{\log(1/\delta)}/\epsilon)$$

matrix-vector queries.

*Proof.* By standard minimax arguments, it suffices to construct a hard distribution for any deterministic algorithm.

Consider  $\mathbf{G} \sim \mathcal{N}(n)$  for  $n = \Omega(\log(1/\delta))$ . From concentration of the singular values of large Gaussian matrices (Lemma A.2), with probability at least  $1 - \delta/10$  we have  $\|\mathbf{G}\|_{op} \leq C\sqrt{n}$  for some absolute constant  $C$ .

Therefore, consider the family of matrices  $\mathbf{W} = \mathbf{I} + \frac{1}{2C\sqrt{n}}(\mathbf{G} + \mathbf{G}^\top)$ . From our bound on  $\|\mathbf{G}\|_{op}$ , with probability at least  $1 - \delta/10$ ,  $\mathbf{W}$  is positive semi-definite and symmetric. Furthermore, since  $\text{tr}(\mathbf{G}) \sim N(0, n)$ , we see that  $\text{tr}(\mathbf{W}) \leq 2n$  with probability at least  $1 - \delta/10$ .

We set the multiplicative error to  $\epsilon = \frac{\sqrt{\log(1/\delta)}}{n}$  and it suffices to show that if we see only  $n/2$  queries, we can compute  $\text{tr}(\mathbf{W})$  up to additive error at best  $c\sqrt{\log(1/\delta)}$  with probability at least  $1 - \delta$ , for some  $c = \Omega(1)$ . By Lemma C.1, we see that conditioned on the queries, our matrix  $\mathbf{W}$  can be decomposed into a determined part and a Gaussian submatrix  $\widetilde{\mathbf{W}} = \frac{1}{2C\sqrt{n}}(\widetilde{\mathbf{G}} + \widetilde{\mathbf{G}}^\top)$ , where  $\widetilde{\mathbf{G}} \sim \mathcal{N}(n/2)$ .

Therefore, our conditional distribution of the trace of  $\mathbf{W}$  is, up to a deterministic shift, the same as the distribution of  $\widetilde{\mathbf{W}}$ , which is simply a Gaussian with variance  $1/C^2$ . Since we must determine a Gaussian of constant variance up to an additive error of  $c\sqrt{\log(1/\delta)}$  with probability at least  $1 - \delta$ , we conclude that  $c = \Omega(1)$ .  $\square$

## C.2 Case 2: Lower Bound for Every $\epsilon$

We give a general  $\Omega\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$  lower bound, that holds for every  $\epsilon \in (0, 1)$ , on the query complexity for non-adaptive trace estimation algorithms stated in Theorem 4.3. The proof of Theorem 4.3 is via a reduction to a distribution testing problem in Problem 4.4, whose hardness (in terms of query complexity) is shown in Lemma 4.5.

**Theorem 4.3** (Lower Bound on Non-adaptive Queries for PSD Matrices). *Let  $\epsilon \in (0, 1)$ . Any algorithm that accesses a real, PSD matrix  $\mathbf{A}$  through matrix-vector queries  $\mathbf{A}\mathbf{q}_1, \mathbf{A}\mathbf{q}_2, \dots, \mathbf{A}\mathbf{q}_m$ , where  $\mathbf{q}_1, \dots, \mathbf{q}_m$  are real-valued non-adaptively chosen vectors, requires*

$$m = \Omega\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$$

to output an estimate  $t$  such that with probability at least  $1 - \delta$ ,  $(1 - \epsilon)\text{tr}(\mathbf{A}) \leq t \leq (1 + \epsilon)\text{tr}(\mathbf{A})$ .

*Proof.* The proof is via reduction to a distribution testing problem stated in Problem 4.4. Given a real, PSD input matrix  $\mathbf{A}$ , let  $\mathcal{A}$  be an algorithm that uses  $m$  non-adaptive matrix-vector queries and outputs a trace estimation  $t$  of  $\mathbf{A}$  such that for some  $\epsilon \in (0, 1)$ , with probability at least  $1 - \delta$ ,  $(1 - \epsilon)\text{tr}(\mathbf{A}) \leq t \leq (1 + \epsilon)\text{tr}(\mathbf{A})$ .

Consider  $n = \log(1/\delta)$ . Let  $Z_i, \forall i \in [n]$  be the  $i$ -th diagonal entry of  $\mathbf{W} \sim \mathcal{W}(n) = \mathbf{G} + \mathbf{G}^\top$  as in Definition A.1. Note that  $\mathbf{G}$  has i.i.d.  $\mathcal{N}(0, 1)$  entries, and that the diagonal of  $\mathbf{G}$  and  $\mathbf{G}^\top$  are the same. This implies  $Z_i \sim \mathcal{N}(0, 4)$ .

Since the  $Z_i$  are i.i.d.,

$$\text{tr}(\mathbf{W}) = \sum_{i=1}^n Z_i \sim \mathcal{N}(0, 4n) = \mathcal{N}(0, 4 \log(1/\delta))$$

By Fact A.3,

$$\begin{aligned} \Pr[\text{tr}(\mathbf{W}) \geq 2\sqrt{2} \log(1/\delta)] &\leq \delta \\ \Pr[\text{tr}(\mathbf{W}) \leq -2\sqrt{2} \log(1/\delta)] &\leq \delta \end{aligned}$$

For a unit vector  $\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathbb{R}^n$ ,

$$\text{tr}\left(\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \frac{\mathbf{g}^\top}{\|\mathbf{g}\|_2}\right) = \left\| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\|_2^2 = 1$$

Let  $\mathbf{B}$  be the random matrix generated from distribution  $\mathcal{P}$  or  $\mathcal{Q}$  in Problem 4.4. First, we claim that with probability at least  $1 - 4\delta$ ,  $\mathbf{B}$  is a PSD matrix. Note that  $C \log^{3/2}(\frac{1}{\delta}) \cdot \frac{1}{\|\mathbf{g}\|_2^2} \mathbf{g}\mathbf{g}^\top$  is PSD.

Thus it suffices to show  $\mathbf{W} + 6\sqrt{\log(\frac{1}{\delta})}\mathbf{I}$  is PSD with high probability.

By **Lemma A.2**, with probability  $1 - 2\delta$ ,

$$\|\mathbf{G}\|_{op} \leq 3\sqrt{\log(1/\delta)}$$

By the triangle inequality and a union bound, with probability  $1 - 4\delta$ ,

$$\|\mathbf{W}\|_{op} = \|\mathbf{G} + \mathbf{G}^T\|_{op} \leq 6\sqrt{\log(1/\delta)}$$

This implies  $\mathbf{W} + 6\sqrt{\log(\frac{1}{\delta})}\mathbf{I}$  is PSD with probability  $1 - 4\delta$ .

If  $\mathbf{B} \sim \mathcal{P}$ , with probability at least  $1 - \delta$ ,

$$\begin{aligned} \text{tr}(\mathbf{B}) &= C \log^{3/2}(1/\delta) + \text{tr}(\mathbf{W}) + 6 \log^{3/2}(1/\delta) \\ &\geq (C + 6) \log^{3/2}(1/\delta) - 2\sqrt{2} \log(1/\delta) \end{aligned}$$

If  $\mathbf{B} \sim \mathcal{Q}$ , with probability at least  $1 - \delta$ ,

$$\text{tr}(\mathbf{B}) = \text{tr}(\mathbf{W}) + 6 \log^{3/2}(\log(1/\delta)) \leq 2\sqrt{2} \log(1/\delta) + 6 \log^{3/2}(1/\delta)$$

Consider the trace estimation algorithm  $\mathcal{A}$  and let the output  $t = \mathcal{A}(\mathbf{B})$ . Consider the constant  $C > \frac{10(1+\epsilon)}{1-\epsilon} - 6$ . If  $\mathbf{B} \sim \mathcal{P}$ , with probability at least  $1 - 2\delta$ ,

$$\begin{aligned} t &\geq (1 - \epsilon) \text{tr}(\mathbf{B}) \\ &\geq (1 - \epsilon) \left( (C + 6) \log^{3/2}(1/\delta) - 2\sqrt{2} \log(1/\delta) \right) \\ &> 6(1 + \epsilon) \log^{3/2}(1/\delta) \end{aligned}$$

If  $\mathbf{B} \sim \mathcal{Q}$ , with probability at least  $1 - 2\delta$ ,

$$\begin{aligned} t &\leq (1 + \epsilon) \text{tr}(\mathbf{B}) \\ &\leq (1 + \epsilon) \left( 6 \log^{3/2}(1/\delta) + 2\sqrt{2} \log(1/\delta) \right) \\ &< 6(1 + \epsilon) \log^{3/2}(1/\delta) \end{aligned}$$

In the worst case, if any of the instances generated from  $\mathcal{P}$  or  $\mathcal{Q}$  is non-PSD, our algorithm  $\mathcal{A}$  fails. Thus  $\mathcal{A}$  determines which distribution  $\mathbf{B}$  comes from with probability at least  $1 - 6\delta$ . By **Lemma 4.5**, this requires the number of matrix-vector queries  $\mathcal{A}$  uses to be  $m = \Omega\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$ . □

**Problem 4.4** (Hard PSD Matrix Distribution Test). *Given  $\delta \in (0, \frac{1}{2})$ , set  $n = \log(1/\delta)$ . Choose  $\mathbf{g} \in \mathbb{R}^n$  to be an independent random vector with i.i.d.  $\mathcal{N}(0, 1)$  entries. Consider two distributions:*

- *Distribution  $\mathcal{P}$  on matrices  $\left\{ C \log^{3/2}(\frac{1}{\delta}) \cdot \frac{1}{\|\mathbf{g}\|_2} \mathbf{g}\mathbf{g}^T + \mathbf{W} + 6\sqrt{\log(\frac{1}{\delta})}\mathbf{I} \right\}$ , for some fixed constant  $C > 1$ .*
- *Distribution  $\mathcal{Q}$  on matrices  $\left\{ \mathbf{W} + 6\sqrt{\log(\frac{1}{\delta})}\mathbf{I} \right\}$ .*

where  $\mathbf{W} \sim \mathcal{W}(n) = \mathbf{G} + \mathbf{G}^T$  as in Definition A.1. Let  $\mathbf{A}$  be a random matrix drawn from either  $\mathcal{P}$  or  $\mathcal{Q}$  with equal probability. Consider any algorithm which, for a fixed query matrix  $\mathbf{Q} \in \mathbb{R}^{n \times q}$ , observes  $\mathbf{A}\mathbf{Q}$ , and guesses if  $\mathbf{A} \sim \mathcal{P}$  or  $\mathbf{A} \sim \mathcal{Q}$  with success probability at least  $1 - \delta$ .

**Lemma 4.5** (Hardness of Problem 4.4). *Given  $\delta \in (0, \frac{1}{2})$ . Consider a non-adaptively chosen query matrix  $\mathbf{Q} \in \mathbb{R}^{n \times q}$  on input  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , as in **Problem 4.4**, where  $n = \log(1/\delta)$ . If  $q = o\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$ , no algorithm can solve **Problem 4.4** with success probability  $1 - \delta$ .*

*Proof.* We claim that without loss of generality, we only need to consider  $\mathbf{Q}$  to be the first  $q$  standard basis vectors, i.e.,  $\mathbf{Q} = \mathbf{E}_q = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_q]$ . First note that we only need to consider query matrix  $\mathbf{Q}$  with orthonormal columns, since for general  $\mathbf{Q}$ , letting  $\mathbf{Q} = \mathbf{U}\mathbf{R}$  be the QR decomposition of  $\mathbf{Q}$ , we can reconstruct  $\mathbf{A}\mathbf{Q}$  from  $(\mathbf{A}\mathbf{U})\mathbf{R}$ . Next, let  $\bar{\mathbf{Q}} \in \mathbb{R}^{n \times (n-q)}$  be the orthonormal basis for null( $\mathbf{Q}$ ). Define an orthonormal matrix  $\mathbf{R} = [\mathbf{Q}, \bar{\mathbf{Q}}] \in \mathbb{R}^{n \times n}$ . By **Fact A.2**,  $\mathbf{W}\mathbf{E}_q$  has the same distribution as  $\mathbf{W}\mathbf{R}\mathbf{E}_q = \mathbf{W}\mathbf{Q}$ . Similarly,  $(C \log(\frac{1}{\delta}) \cdot \frac{1}{\|\mathbf{g}\|_2} \mathbf{g}\mathbf{g}^T + \mathbf{W})\mathbf{E}_q$  has the same distribution as

$(C \log(\frac{1}{\delta}) \cdot \frac{1}{\|\mathbf{g}\|_2^2} \mathbf{g}\mathbf{g}^T + \mathbf{W})\mathbf{Q}$ . Therefore, we only need to consider the case when the queries are the first  $q$  standard basis vectors.

Consider the two possible observed distributions from **Problem 4.4**: 1) distribution  $\mathcal{P}'$ , which has  $(C \log(\frac{1}{\delta}) \cdot \frac{1}{\|\mathbf{g}\|_2^2} \mathbf{g}\mathbf{g}^T + \mathbf{W} + 2\sqrt{\log(1/\delta)}\mathbf{I})\mathbf{Q}$  for fixed constant  $C > 1$ , and 2) distribution  $\mathcal{Q}'$  which has  $(\mathbf{W} + 2\sqrt{\log(1/\delta)}\mathbf{I})\mathbf{Q}$ .

We argue that if the number  $q$  of queries is too small, then the total variation distance between  $\mathcal{P}'$  and  $\mathcal{Q}'$ , conditioned on an event  $\mathcal{E}$  with probability at least  $\delta$ , is upper bounded by a small constant. This will imply that no algorithm can succeed with probability at least  $1 - \delta$ . We upper bound the total variation distance between  $\mathcal{P}'$  and  $\mathcal{Q}'$  via the Kullback–Leibler (KL) divergence between  $\mathcal{P}'$  and  $\mathcal{Q}'$  and then apply Pinsker's inequality.

Consider the following event on over the randomness of  $\mathbf{g}$ :  $\mathcal{E} = \left\{ \mathbf{g} : \frac{1}{\|\mathbf{g}\|_2^2} \|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{50C^2 n^3} \right\}$ . Note that  $\mathbf{g}^T \mathbf{Q} = [\langle \mathbf{g}, \mathbf{e}_1 \rangle, \langle \mathbf{g}, \mathbf{e}_2 \rangle, \dots, \langle \mathbf{g}, \mathbf{e}_q \rangle] = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_q]$ , i.e., the first  $q$  coordinates of  $\mathbf{g}$ . First, we show that  $\Pr[\mathcal{E}] = \Omega(\delta)$ .

Since  $\mathbf{g}_i \sim \mathcal{N}(0, 1)$ , by **Fact A.4**, for the  $i$ -th entry of  $\mathbf{g}^T \mathbf{Q}$ ,  $\forall i \in [q]$ ,

$$\Pr[|\mathbf{g}_i| \leq \frac{1}{10C \cdot n\sqrt{q}}] = \Omega\left(\frac{1}{n\sqrt{q}}\right)$$

which implies for a single entry,

$$\Pr[\mathbf{g}_i^2 \leq \frac{1}{100C^2 \cdot n^2 q}] = \Omega\left(\frac{1}{n\sqrt{q}}\right)$$

Since all  $q$  queries are independent, for all entries  $i \in [q]$ ,

$$\Pr[\|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2}] = \Omega\left(\left(\frac{1}{n\sqrt{q}}\right)^q\right) = \Omega\left(\exp\left(-\frac{q}{2} \ln(n^2 q)\right)\right)$$

Consider the following conditional probability,

$$\begin{aligned} & \Pr\left[\|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2} \wedge \|\mathbf{g}\|_2^2 \geq \frac{n}{2}\right] \\ &= \Pr\left[\|\mathbf{g}\|_2^2 \geq \frac{n}{2} \mid \|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2}\right] \cdot \Pr\left[\|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2}\right] \end{aligned}$$

Assume  $q < \frac{n}{2}$  and let  $\mathbf{g}_{(q+1):n}$  denote the  $q+1$ -th to the  $n$ -th entry of  $\mathbf{g}$ . Note that all entries of  $\mathbf{g}$  are independent and  $\|\mathbf{g}_{(q+1):n}\|_2^2 \sim \chi^2(d)$  with degree  $d > \frac{n}{2}$ . By **Fact A.1**, since  $\|\mathbf{g}\|_2^2 \geq \|\mathbf{g}_{(q+1):n}\|_2^2$ ,

$$\Pr\left[\|\mathbf{g}\|_2^2 \geq \frac{n}{2} \mid \|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2}\right] = \Omega(1)$$

Thus,

$$\begin{aligned} \Pr\left[\frac{1}{\|\mathbf{g}\|_2^2} \|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{50C^2 n^3}\right] &\geq \Pr\left[\|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{100C^2 \cdot n^2} \wedge \|\mathbf{g}\|_2^2 \geq \frac{n}{2}\right] \\ &\geq \Omega(1) \cdot \Omega\left(\exp\left(-\frac{q}{2} \ln(n^2 q)\right)\right) \end{aligned}$$

Assume we only have a small number  $q = o\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$  of queries. Then,

$$\Pr[\mathcal{E}] = \Pr\left[\frac{1}{\|\mathbf{g}\|_2^2} \|\mathbf{g}^T \mathbf{Q}\|_2^2 \leq \frac{1}{50C^2 \cdot n^3}\right] \geq 10\delta \quad (1)$$

Note that  $n = \log(1/\delta)$ , and so

$$\Pr[\mathcal{E}] = \Pr\left[C^2 \log^3\left(\frac{1}{\delta}\right) \frac{\|\mathbf{g}^T \mathbf{Q}\|_2^2}{\|\mathbf{g}\|_2^2} \leq \frac{1}{50}\right] \geq 10\delta$$

Next, note that it suffices to show that the probability of success conditioned on  $\mathcal{E}$  is less than  $1/3$ . This implies our result since  $\mathcal{E}$  occurs with probability at least  $10\delta$ , implying that our probability of failure is indeed  $\Omega(\delta)$ . Therefore, we focus on showing that the probability of success conditioned on  $\mathbf{g} \in \mathcal{E}$  is small via standard information theoretic arguments with KL divergence bounds.

Conditioning on event  $\mathcal{E}$ , we now upper bound the KL divergence between  $\mathcal{P}'$  and  $\mathcal{Q}'$  conditioned on a fixed  $\mathbf{g} \in \mathcal{E}$ . Since both distributions come from symmetric matrices, we remove the redundant random variables from observed random matrices from  $\mathcal{P}'$ ,  $\mathcal{Q}'$  and consider only the lower triangular portion, so that both have dimensions  $l = n + (n - 1) + \dots + (n - (q - 1))$ . Note that these redundant random variables in the upper triangular portion can be removed without increasing the KL divergence, since they are perfectly correlated with its counterpart variable in the lower triangular region, which we show as follows:

Consider two lists  $L_{\mathcal{P}'}, L_{\mathcal{Q}'}$  of  $l$  random variables, corresponding to a vectorization of the observed lower triangular part of the random matrices from  $\mathcal{P}'$  and  $\mathcal{Q}'$ . Consider also a function  $f$ , which duplicates parts of the random variables in  $L_{\mathcal{P}'}$  and  $L_{\mathcal{Q}'}$ , such that  $f(L_{\mathcal{P}'})$  and  $f(L_{\mathcal{Q}'})$  reconstruct the original observed matrix of size  $n \times q$  from  $\mathcal{P}'$  and  $\mathcal{Q}'$ , respectively. Then, by the data processing inequality of KL divergence from **Fact A.7**,

$$\mathcal{D}_{KL}(\mathcal{P}' \parallel \mathcal{Q}') = \mathcal{D}_{KL}(f(L_{\mathcal{P}'}) \parallel f(L_{\mathcal{Q}'})) \leq \mathcal{D}_{KL}(L_{\mathcal{P}'} \parallel L_{\mathcal{Q}'})$$

From now on, we assume that  $\mathcal{P}', \mathcal{Q}'$  are lower triangular. The KL divergence between  $\mathcal{P}'|\mathbf{g}$  and  $\mathcal{Q}'|\mathbf{g}$  considering the lower triangular part can be calculated since they are both multivariate Gaussians with the same covariance matrix (of rank  $l$ ). The KL divergence thus only depends on the difference between the mean  $\Delta\mu$  of the two multivariate Gaussians (see **Fact A.5**), which is the lower triangular part contained in  $C \log^{3/2}(\frac{1}{\delta}) \frac{\mathbf{g}\mathbf{g}^T}{\|\mathbf{g}\|_2^2} \mathbf{Q}$ . Furthermore, since all redundant variables are removed, the distribution on the remaining variables is dimension-independent, with variance 2 from the randomness of  $\mathbf{W}$ .

Let  $\widetilde{\mathbf{M}} = [\mathbf{m}_1, \dots, \mathbf{m}_q]$  be the observed lower triangular parts of  $\Delta\mu$ , where  $\mathbf{m}_i \in \mathbb{R}^{n-i+1}, \forall i \in [q]$ . Let  $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_q]$  where  $\mathbf{q}_i \in \mathbb{R}^n, \forall i \in [q]$  be the queries. By **Fact A.5**, for any  $\mathbf{g} \in \mathcal{E}$  (an event of probability at least  $10\delta$ ),

$$\begin{aligned} \mathcal{D}_{KL}(\mathcal{P}'|\mathbf{g} \parallel \mathcal{Q}'|\mathbf{g}) &\leq \mathcal{D}_{KL}(L_{\mathcal{P}'|\mathbf{g}} \parallel L_{\mathcal{Q}'|\mathbf{g}}) \\ &\leq \sum_{i=1}^q \|C \log^{3/2}(\frac{1}{\delta}) \mathbf{m}_i\|_2^2 \\ &\leq C^2 \log^3(\frac{1}{\delta}) \sum_{i=1}^q \|\frac{\mathbf{g}\mathbf{g}^T}{\|\mathbf{g}\|_2^2} \mathbf{q}_i\|_2^2 \\ &= C^2 \log^3(\frac{1}{\delta}) \sum_{i=1}^q \langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{q}_i \rangle^2 \\ &= C^2 \log^3(\frac{1}{\delta}) \frac{\|\mathbf{g}^T \mathbf{Q}\|_2^2}{\|\mathbf{g}\|_2^2} \\ &\leq \frac{1}{50} \end{aligned}$$

By **Fact A.6**, since conditioning (on  $\mathbf{g}$ ) increases KL divergence between  $\mathcal{P}'$  and  $\mathcal{Q}'$ , let  $f(\mathbf{g})$  be the conditional probability density of  $\mathbf{g}$  on  $\mathcal{E}$ . Then,

$$\mathcal{D}_{KL}(\mathcal{P}' \parallel \mathcal{Q}') \leq \int_{\mathbf{g}} \mathcal{D}_{KL}(\mathcal{P}'|\mathbf{g} \parallel \mathcal{Q}'|\mathbf{g}) f(\mathbf{g}) d\mathbf{g} \leq \mathcal{D}_{KL}(\mathcal{P}'|\mathbf{g} \parallel \mathcal{Q}'|\mathbf{g}) = \frac{1}{50}$$

By Pinsker's inequality, given  $\mathcal{E}$  happens,

$$\mathcal{D}_{TV}(\mathcal{P}' \parallel \mathcal{Q}') \leq \sqrt{\frac{1}{2} \mathcal{D}_{KL}(\mathcal{P}' \parallel \mathcal{Q}')} = \sqrt{\frac{1}{100}} < \frac{1}{3}$$

If the total variation distance between any two distributions  $\mathcal{P}'$  and  $\mathcal{Q}'$  is at most  $\delta$ , then any algorithm that distinguishes between  $\mathcal{P}'$  and  $\mathcal{Q}'$  can succeed with probability at most  $\frac{1}{2} + \frac{\delta}{2}$ .

---

<sup>2</sup>For two arbitrary distributions  $\mathcal{P}'$  and  $\mathcal{Q}'$ , let the total variation distance between them be  $\mathcal{D}_{TV}(\mathcal{P}' \parallel \mathcal{Q}') =$

Since  $\mathcal{D}_{TV}(\mathcal{P}' \parallel \mathcal{Q}') \leq \frac{1}{3}$  in our case, this implies that any algorithm for distinguishing  $\mathcal{P}'$  and  $\mathcal{Q}'$  can succeed with probability at most  $\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{3} = \frac{2}{3}$ , and so fails with probability  $> \frac{1}{3}$ . Since  $\Pr[\mathcal{E}] \geq 10\delta$ , the overall failure probability of an algorithm for distinguishing  $\mathcal{P}$  from  $\mathcal{Q}$  is thus  $10\delta \cdot \frac{1}{3} > \delta$ . This implies that to achieve success probability at least  $1 - \delta$ ,  $q = \Omega\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$ .

□

---

$\sup_{\mathcal{E}} |\mathcal{P}'(\mathcal{E}) - \mathcal{Q}'(\mathcal{E})| = \delta$ , where  $\mathcal{E}$  is an event. Consider an algorithm  $\mathcal{A}$  that distinguishes samples from  $\mathcal{P}'$  or  $\mathcal{Q}'$ , and an arbitrary sample  $\mathbf{x}$ . Let  $\mathcal{E} = \Pr[\mathcal{A}(\mathbf{x}) = \mathcal{P}', \mathbf{x} \sim \mathcal{P}']$ . If  $\mathcal{A}$  succeeds with probability  $\geq \frac{1}{2} + \frac{\delta}{2}$ , then this implies  $\Pr[\mathcal{A}(\mathbf{x}) = \mathcal{P}', \mathbf{x} \sim \mathcal{P}'] \geq \frac{1}{2} + \frac{\delta}{2}$ , and  $\Pr[\mathcal{A}(\mathbf{x}) = \mathcal{P}', \mathbf{x} \sim \mathcal{Q}'] \geq \frac{1}{2} + \frac{\delta}{2} - \delta = \frac{1}{2} - \frac{\delta}{2}$ . This also implies  $\Pr[\mathcal{A}(\mathbf{x}) = \mathcal{Q}', \mathbf{x} \sim \mathcal{Q}'] \leq 1 - (\frac{1}{2} - \frac{\delta}{2}) = \frac{1}{2} + \frac{\delta}{2}$ , which means the success probability  $\mathcal{A}$  is at most  $\frac{1}{2} + \frac{\delta}{2}$ .

## References

- [1] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by modelselection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.
- [2] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.
- [3] Joram Soch and Carsten Allefeld. Kullback-leibler divergence for the normal-gamma distribution, 2016.
- [4] Derivations for linear algebra and optimization. [https://stanford.edu/~jduchi/projects/general\\_notes.pdf](https://stanford.edu/~jduchi/projects/general_notes.pdf).
- [5] Lecture notes on information theory. <http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf>.
- [6] Lecture notes for statistics 311/electrical engineering 377. <https://web.stanford.edu/class/stats311/lecture-notes.pdf>.
- [7] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation, 2020.
- [8] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.
- [9] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 205–214, New York, NY, USA, 2009. Association for Computing Machinery.
- [10] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for pca via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259, 2018.