

# Private High-Dimensional Hypothesis Testing

Shyam Narayanan

SHYAMSN@MIT.EDU

Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA

Editors: Po-Ling Loh and Maxim Raginsky

## Abstract

We provide improved differentially private algorithms for identity testing of high-dimensional distributions. Specifically, for  $d$ -dimensional Gaussian distributions with known covariance  $\Sigma$ , we can test whether the distribution comes from  $\mathcal{N}(\mu^*, \Sigma)$  for some fixed  $\mu^*$  or from some  $\mathcal{N}(\mu, \Sigma)$  with total variation distance at least  $\alpha$  from  $\mathcal{N}(\mu^*, \Sigma)$  with  $(\epsilon, 0)$ -differential privacy, using only

$$\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \epsilon^{2/3}} + \frac{1}{\alpha \cdot \epsilon}\right)$$

samples if the algorithm is allowed to be computationally inefficient, and only

$$\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \epsilon}\right)$$

samples for a computationally efficient algorithm. We also provide a matching lower bound showing that our computationally inefficient algorithm has optimal sample complexity. We also extend our algorithms to various related problems, including mean testing of Gaussians with bounded but unknown covariance, uniformity testing of product distributions over  $\{-1, 1\}^d$ , and tolerant testing. Our results improve over the previous best work of [Canonne et al. \(2020b\)](#) for both computationally efficient and inefficient algorithms, and even our computationally efficient algorithm matches the optimal *non-private* sample complexity of  $O\left(\frac{\sqrt{d}}{\alpha^2}\right)$  in many standard parameter settings. In addition, our results show that, surprisingly, private identity testing of  $d$ -dimensional Gaussians can be done with fewer samples than private identity testing of discrete distributions over a domain of size  $d$  ([Acharya et al., 2018b](#)), which refutes a conjectured lower bound of [Canonne et al. \(2020b\)](#).

**Keywords:** Differential Privacy, Hypothesis Testing, Multivariate Distributions

## 1. Introduction

Hypothesis testing is one of the oldest and most widely studied problems in statistical inference, and is critical in research for nearly every scientific branch. Simply put, hypothesis testing asks, given a collection of  $N$  data points  $\mathbf{X} = \{X^{(1)}, \dots, X^{(N)}\}$ , whether the data points come from a distribution proposed by a null hypothesis  $\mathcal{H}_0$  or by an alternative hypothesis  $\mathcal{H}_1$ .

Classical hypothesis testing often focuses on univariate or discrete distributions. However, these distributions may be insufficient in many applications. For instance, if we wish to test a hypothesis on patient data, each patient may have numerous features such as those corresponding to vitals, organ function, metabolic rate, presence or severity of diseases, etc. Consequently, it is crucial to develop hypothesis testing procedures for high-dimensional distributions.

In many practical applications of hypothesis testing, the data may reveal highly sensitive information about an individual. For instance, the data may include whether a patient has a certain disease, or has participated in a crime or embarrassing activity. For this reason, an increasingly

important challenge that has arisen in hypothesis testing to ensure that the test is not only accurate but also preserves the privacy of the individuals that contribute data. The notion of privacy we wish to guarantee is commonly called *differential privacy*, which was first formulated by [Dwork et al. \(2006\)](#). Differential privacy has emerged as the leading notion of privacy both in theory and practice, and has been utilized to measure privacy by companies such as Apple ([Team, 2017](#)), Google ([Erlingsson et al., 2014](#)), and Microsoft ([Ding et al., 2017](#)), as well as the US Census Bureau ([Dajani et al., 2017](#)). Informally, differential privacy provably ensures that changing one piece of the data does not affect the output of the algorithm significantly. Hence, an adversary cannot reconstruct any single user’s data based on the algorithm’s output, thereby ensuring that each user’s data is secure.

In this paper, we study *differentially private* hypothesis testing for *high-dimensional* distributions. We primarily study two major classes of distributions: multivariate Gaussians and Boolean product distributions. Specifically, we consider the following problems, as well as certain generalizations of them, and improve over the previous state-of-the-art results by [Canonne et al. \(2020b\)](#).

1. How many samples from a multivariate Gaussian distribution  $\mathcal{N}(\mu, \Sigma)$ , where  $\mu \in \mathbb{R}^d$  is the mean vector and  $\Sigma \in \mathbb{R}^{d \times d}$  is the covariance matrix, are necessary to privately determine whether  $\mu = \mu^*$  for some fixed  $\mu^* \in \mathbb{R}^d$  or  $\mu$  is “far” from  $\mu^*$ ?
2. How many samples from a product distribution  $\mathcal{D}$  over  $\{-1, 1\}^d$  are necessary to privately determine whether  $\mathcal{D}$  equals some fixed product distribution  $\mathcal{D}^*$  or  $\mathcal{D}$  is “far” from  $\mathcal{D}^*$ ?

### 1.1. Private Hypothesis Testing

In this subsection, we describe the basics of differential privacy and differentially private hypothesis testing. First, we define the notion of neighboring datasets and differential privacy.

**Definition 1 (Neighboring Datasets)** *Let  $\mathcal{X}$  be some domain, and let  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)}) \in \mathcal{X}^N$  and  $\mathbf{X}' = (X'^{(1)}, \dots, X'^{(N)}) \in \mathcal{X}^N$  be two datasets of size  $N$  from  $\mathcal{X}$ . Then, we say that  $\mathbf{X}$  and  $\mathbf{X}'$  are neighboring datasets if there is at most one value of  $1 \leq i \leq N$  such that  $X^{(i)} \neq X'^{(i)}$ .*

**Definition 2 (Differential Privacy ([Dwork et al., 2006](#)))** *Let  $0 \leq \epsilon, \delta \leq 1$ . A randomized algorithm  $\mathcal{A} : \mathcal{X}^N \rightarrow \mathcal{O}$  is said to be  $(\epsilon, \delta)$ -differentially private (DP) if for any two neighboring datasets  $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^N$  and any subset  $O \subset \mathcal{O}$ ,*

$$\mathbb{P}[\mathcal{A}(\mathbf{X}') \in O] \leq e^\epsilon \cdot \mathbb{P}[\mathcal{A}(\mathbf{X}) \in O] + \delta. \tag{1}$$

We next describe hypothesis testing. A hypothesis  $\mathcal{H}$  represents a class of distributions over  $\mathcal{X}$ , which may consist of either a single distribution or a family of distributions with certain constraints. In distribution testing, we are given two hypotheses  $\mathcal{H}_0$  and  $\mathcal{H}_1$  that are *disjoint*, meaning no distribution is in both classes. We are then given  $N$  i.i.d. samples from  $\mathcal{D}$ , where  $\mathcal{D}$  is a distribution from  $\mathcal{H}_0 \cup \mathcal{H}_1$ , and our goal is to determine whether  $\mathcal{D} \in \mathcal{H}_0$  or  $\mathcal{D} \in \mathcal{H}_1$ .

We now give a formal definition for *hypothesis testing*, and then we formally define *private hypothesis testing*.

**Definition 3** *Fix  $N \in \mathbb{N}$  as the number of samples, and let  $\mathcal{A} : (\mathbb{R}^d)^N \rightarrow \{0, 1\}$  be an algorithm that takes as input  $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ . Given disjoint hypotheses  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , and parameters  $0 \leq \epsilon, \delta \leq 1$ , we say that an algorithm  $\mathcal{A}$  is a hypothesis testing algorithm that can distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  if:*

- For all distributions  $\mathcal{D} \in \mathcal{H}_0$ , if each  $X^{(i)}$  is drawn i.i.d. from  $\mathcal{D}$ , then for  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)})$ ,  $\mathbb{P}[\mathcal{A}(\mathbf{X}) = 0] \geq \frac{2}{3}$ , where the probability is over both the samples  $X^{(1)}, \dots, X^{(N)} \leftarrow \mathcal{D}$  and the randomness of the algorithm  $\mathcal{A}$ .
- For all distributions  $\mathcal{D}' \in \mathcal{H}_1$ , if each  $X^{(i)}$  is drawn i.i.d. from  $\mathcal{D}'$ , then  $\mathbb{P}[\mathcal{A}(\mathbf{X}) = 1] \geq \frac{2}{3}$ .

**Definition 4** Fix  $N \in \mathbb{N}$  as the number of samples, and let  $\mathcal{A} : (\mathbb{R}^d)^N \rightarrow \{0, 1\}$  be an algorithm that takes as input  $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ . Given disjoint hypotheses  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , and parameters  $0 \leq \varepsilon, \delta \leq 1$ , we say that an algorithm  $\mathcal{A}$  can  $(\varepsilon, \delta)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  if:

- $\mathcal{A}$  can distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .
- $\mathcal{A}$  is  $(\varepsilon, \delta)$ -DP (where  $\mathcal{X} = \mathbb{R}^d$  and  $\mathcal{O} = \{0, 1\}$ ).

Note that the privacy must hold for any neighboring datasets  $\mathbf{X}, \mathbf{X}' \in (\mathbb{R}^d)^N$ , even if they are not drawn from any distribution.

## 1.2. Our Results

In all of the problems we investigate, our goal is to devise an algorithm  $\mathcal{A}$  that can  $(\varepsilon, \delta)$ -privately distinguish between a null hypothesis  $\mathcal{H}_0$  and an alternative hypothesis  $\mathcal{H}_1$ , where the number of samples  $N$  is as small as possible. We also are interested in constructing such an algorithm that is efficient, meaning that the runtime is polynomial in the number of samples  $N$  and the dimension  $d$ .

**Identity testing of Gaussians with known covariance:** The main and perhaps simplest problem we study in this paper is *private identity testing* of a multivariate Gaussian with known covariance matrix  $\Sigma$ . In identity testing, the goal is to distinguish between a null hypothesis that only consists of  $N$  samples drawn i.i.d. from a single hypothesis distribution  $\mathcal{N}(\mu^*, \Sigma)$ , and an alternative hypothesis consisting of  $N$  samples drawn i.i.d. from  $\mathcal{N}(\mu, \Sigma)$ , where the mean  $\mu$  is “far” from  $\mu^*$ . In this work, we improve over work by [Canonne et al. \(2020b\)](#) by improving their sample complexity both in the case of inefficient and efficient algorithms, and by providing an optimal lower bound to complement these results.

First, we state our sample complexity upper bound where we allow for an inefficient algorithm.

**Theorem 5 (Inefficient Upper Bound)** Fix  $\mu^* \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^{d \times d}$  as a known positive definite covariance matrix. Also, fix parameters  $0 < \alpha, \varepsilon \leq \frac{1}{2}$ . Then, there exists an algorithm that, using

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon} \right)$$

samples<sup>1</sup>, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$ , which solely consists of  $\mathcal{N}(\mu^*, \Sigma)$ , and  $\mathcal{H}_1$ , which consists of all distributions  $\mathcal{N}(\mu, \Sigma)$  for  $\mu$  with  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \geq \alpha$ .

Expressed more simply, our goal is to privately distinguish between the mean of a multivariate Gaussian being some fixed  $\mu^*$  and the mean being far away from  $\mu^*$ , given some number of samples from the multivariate Gaussian. Our notion of far away depends on the quantity

1. We use  $N = \tilde{O}(F)$  to mean that there exists a fixed constant  $C$  such that  $N = O(F(\log F)^C)$ .

$\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)}$ , which is also called the *Mahalanobis distance*  $d_\Sigma(\mu^*, \mu)$ . While this choice may appear more confusing than simply using  $\ell_2$  (a.k.a. Euclidean) distance, it is more practical than  $\ell_2$  distance as it scales properly with linear transformations of multivariate Gaussian distributions. When  $\Sigma = I$ , the identity matrix, this is equivalent to  $\|\mu - \mu^*\|$ , the Euclidean distance between  $\mu$  and  $\mu^*$ . In addition, whenever the Mahalanobis distance is smaller than 1, it is asymptotically equivalent to the total variation distance  $d_{\text{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu^*, \Sigma))$ .

Our result improves over the previous best inefficient algorithm of [Canonne et al. \(2020b\)](#), which had a sample complexity of  $N = \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/2}}{\alpha\sqrt{\varepsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$ . Importantly, we remove the dependence on  $\frac{d^{1/2}}{\alpha\sqrt{\varepsilon}}$  that was present in ([Canonne et al., 2020b](#)), which provides a strict improvement whenever  $\varepsilon < \alpha^2$  and  $\frac{1}{d} < \alpha^2 \cdot \varepsilon$ . We remark that the inefficient algorithm of ([Canonne et al., 2020b](#)) has a mild inaccuracy, and an important part of our result involves fixing the previous proof.

Next, we show that [Theorem 5](#) is tight, even if the algorithm is allowed to be  $(0, \varepsilon)$ -differentially private as opposed to  $(\varepsilon, 0)$ -differentially private<sup>2</sup>. Specifically, we prove the following:

**Theorem 6 (Lower Bound)** *Let all notation be as in [Theorem 5](#). Then, any algorithm that can  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  must have sample complexity at least*

$$N = \Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right).$$

This improves over the previous lower bound of  $\Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{1}{\alpha \cdot \varepsilon}\right)$  ([Canonne et al., 2020b](#)), which combines the non-private lower bound of  $\Omega\left(\frac{d^{1/2}}{\alpha^2}\right)$  ([Canonne et al., 2020a](#)) and the private lower bound of  $\Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$  for testing 1-dimensional distributions ([Acharya et al., 2018b](#)). We remark that the previous lower bound was technically shown only for testing Boolean product distributions, but it extends to multivariate Gaussians easily.

Because the algorithm we devise for [Theorem 5](#) has very slow runtime, a natural question is how many samples are necessary if the algorithm must run in polynomial time in  $N$  and  $d$ . Indeed, we show the following result, that only needs a slightly larger number of samples but runs efficiently.

**Theorem 7 (Efficient Upper Bound)** *Let all notation be as in [Theorem 5](#). Then, there exists an algorithm that, using*

$$N = \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon}\right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  in time polynomial in  $N$  and  $d$ .*

This improves over the previous best polynomial-time algorithm of [Canonne et al. \(2020b\)](#), which required  $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/2}}{\alpha \cdot \varepsilon}\right)$  samples. This algorithm matches even the optimal *non-private* algorithm as long as  $\frac{\alpha}{d^{1/4}} \leq \varepsilon$ , and strictly improves over the previous best efficient algorithm whenever  $\varepsilon < \alpha$ . It also strictly improves even over the previous best *inefficient* private algorithm if

---

2. We remark that in the case of private *hypothesis testing*,  $(\varepsilon, \delta)$ -DP and  $(\varepsilon + \delta, 0)$ -DP are known to be *asymptotically equivalent* for any  $\varepsilon, \delta < \frac{1}{2}$  ([Acharya et al., 2018b](#)). In general, however, it is harder to achieve  $(\varepsilon + \delta, 0)$ -DP.

$\frac{1}{\sqrt{d}} < \varepsilon < \alpha^2$ . In addition, if we consider the dependence on  $d$  as the bottleneck and only consider the terms dependent on  $\sqrt{d}$ , the number of samples needed is only roughly  $\sqrt{d} \cdot \frac{1}{\alpha^2}$ , matching the optimal *non-private* sample complexity! In contrast, [Canonne et al. \(2020b\)](#) required roughly  $\sqrt{d} \cdot \left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$  for efficient algorithms, and  $\sqrt{d} \cdot \left(\frac{1}{\alpha^2} + \frac{1}{\alpha\sqrt{\varepsilon}}\right)$  for inefficient algorithms.

**Generalizations to other distributions:** Our results above generalize to related hypothesis testing problems, such as hypothesis testing for Gaussians with unknown covariance, hypothesis testing for Boolean Product distributions, and tolerant identity testing. We describe the results informally here, and provide more formal statements in the Appendix.

First, we show that our results on privately testing multivariate Gaussians with known covariance can be extended to Gaussians with unknown but bounded covariance. One caveat is that we are no longer able to distinguish between  $\mu = \mu^*$  and  $\mu, \mu^*$  being far in Mahalanobis distance, as the Mahalanobis distance depends on the unknown matrix  $\Sigma$ . Instead, we distinguish between  $\mu = \mu^*$  and  $\mu, \mu^*$  being far in  $\ell_2$  distance, if we are promised that  $\Sigma$  has bounded spectral norm.

**Theorem 8 (Bounded but Unknown Covariance, Informal)** *Let  $\mathcal{H}_0$  consist of  $\mathcal{N}(\mu^*, \Sigma)$  over all covariance matrices with bounded spectral norm  $\|\Sigma\|_2 \leq 1$ , and  $\mathcal{H}_1$  consist of  $\mathcal{N}(\mu, \Sigma)$  over all covariance matrices  $\|\Sigma\|_2 \leq 1$  and  $\mu : \|\mu - \mu^*\| \geq \alpha$ . Then, to distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , the same upper and lower bounds as in Theorems 5, 6, and 7 hold.*

Next, we show that our results on testing multivariate Gaussians also extend to identity testing for “balanced” Boolean product distributions over  $\{-1, 1\}^d$ , i.e., distributions where each coordinate is independent (but not necessarily identically distributed). Namely, we can privately test whether a product distribution is some fixed  $\mathcal{P}^*$  or has total variation distance far from  $\mathcal{P}^*$ , as long as the expectation of  $\mathcal{P}^*$  is between  $-1/2$  and  $1/2$  in each coordinate<sup>3</sup>.

**Theorem 9 (Product Distributions, Informal)** *Fix  $\mu^* \in [-1/2, 1/2]^d$ , and suppose that  $\mathcal{H}_0$  consists only of the product distribution  $\mathcal{P}^*$  over  $\{-1, 1\}^d$  with mean  $\mu^*$ , and  $\mathcal{H}_1$  consists of all product distributions  $\mathcal{P}$  over  $\{-1, 1\}^d$  such that  $d_{TV}(\mathcal{P}, \mathcal{P}^*) \geq \alpha$ . Then, to distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , the same upper and lower bounds as in Theorems 5, 6, and 7 all hold for any  $\mu^* \in [-1/2, 1/2]^d$ .*

Theorem 9 implies bounds for private *uniformity* testing of Boolean product distributions, since the uniform distribution over  $\{-1, 1\}^d$  is a product distribution with mean  $\mathbf{0}$ . Theorem 9 improves over both the previous best upper bound of  $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/2}}{\alpha\varepsilon^{1/2}} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$  for inefficient algorithms and  $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/2}}{\alpha\varepsilon}\right)$  for efficient algorithms ([Canonne et al., 2020b](#)). In addition, it also improves over the best lower bound of  $\Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$  ([Canonne et al., 2020b](#)). Note that the previous best upper and lower bounds for privately testing balanced Boolean product distributions and multivariate Gaussian distributions match, as do our bounds.

We remark that identity testing of “unbalanced” Boolean product distributions, i.e., where  $\mu^*$  is not promised to be in  $[-1/2, 1/2]^d$ , is not always achievable with the same number of samples. Indeed, [Canonne et al. \(2020b\)](#) showed that if the null hypothesis distribution  $\mathcal{P}^*$  is sufficiently unbalanced, there is a sample complexity lower bound of  $\Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/2}}{\alpha\sqrt{\varepsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$ .

3.  $-1/2$  and  $1/2$  can be replaced by any constants bounded away from  $-1$  and  $1$ .

Finally, we show that our results on private identity testing of Gaussians and Boolean product distributions extend to private *tolerant* identity testing. In this setting, we allow for some slack in the null hypothesis, and have to distinguish between the mean  $\mu$  being far from  $\mu^*$  versus *close* to  $\mu^*$ , as opposed to just equaling  $\mu^*$ . Tolerant testing is useful as it provides meaningful guarantees even if the underlying distribution is very close to, but does not perfectly satisfy, the null distribution.

**Theorem 10 (Tolerant Hypothesis Testing, Informal)** *Theorems 5, 6, and 7 all hold if we replace the null hypothesis  $\mathcal{H}_0$  with all distributions  $\mathcal{N}(\mu, \Sigma)$  such that  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \leq \frac{\alpha}{2}$ . Likewise, for any  $\mu^* \in [-1/2, 1/2]^d$  and for  $\mathcal{P}^*$  the product distribution with mean  $\mu^*$ , Theorem 9 holds even if  $\mathcal{H}_0$  consists of all product distributions  $\mathcal{P}$  with  $d_{TV}(\mathcal{P}, \mathcal{P}^*) \leq \frac{\alpha}{C}$  and  $\mathcal{H}_1$  consists of product distributions  $\mathcal{P}$  with  $d_{TV}(\mathcal{P}, \mathcal{P}^*) \geq \alpha$ , for a sufficiently large constant  $C$ .*

One surprising consequence of our algorithms is that the number of samples we require for private identity testing of  $d$ -dimensional Gaussians and private uniformity testing of  $d$ -dimensional Boolean product distributions is in fact *smaller than* the number of samples needed for private uniformity testing of a discrete distribution over just  $d$  elements. Indeed, private uniformity testing of a discrete distribution requires  $\Theta\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\sqrt{\varepsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$  samples (Acharya et al., 2018b), whereas for  $d$ -dimensional product distributions and Gaussians, we are able to remove the dependence on  $\frac{\sqrt{d}}{\alpha\sqrt{\varepsilon}}$ . Hence, we refute a conjecture of Canonne et al. (2020b), which postulates that private uniformity testing of discrete distributions over  $[d]$  and of product distributions over  $\{-1, 1\}^d$  have asymptotically equivalent sample complexities.

### 1.3. Related Work

Ignoring privacy constraints, hypothesis testing from a statistical point of view dates back nearly a century, notably to Neyman and Pearson (1933). Hypothesis testing has become a popular area of study in theoretical computer science more recently (where it is also called *distribution testing*), starting with (Goldreich and Ron, 2000; Batu et al., 2000), and with a large body of subsequent literature over the past two decades (see, for instance, (Canonne, 2020) for a survey of the distribution testing field). While much of the work has focused on discrete or univariate distributions, there has recently been significant work in the multivariate setting as well. This work on multivariate hypothesis testing has come both from statistical (Hotelling, 1931; Srivastava and Du, 2008; Chen and Qin, 2010; Cai and Ma, 2013; Javanmard and Montanari, 2014; Ramdas et al., 2016) and computational (Alon et al., 2007; Rubinfeld and Xie, 2010; Acharya et al., 2015; Daskalakis and Pan, 2017; Acharya et al., 2018a; Daskalakis et al., 2018; Gheissari et al., 2018; Bezáková et al., 2020; Canonne et al., 2020a; Diakonikolas and Kane, 2021) perspectives.

*Differentially private* hypothesis testing began with work by Vu and Slavkovic (2009); Uhler et al. (2013), and has seen significant work during the past decade. Apart from the work by Canonne et al. (2020b), which this paper primarily improves over, perhaps the work most closely related to ours is that of Cai et al. (2017); Aliakbarpour et al. (2018); Acharya et al. (2018b), which study the problem of private identity testing (as well as closeness testing) of discrete distributions. Various other problems in private hypothesis testing have also been studied, including testing simple hypotheses (Cummings et al., 2018; Canonne et al., 2019), selection from a discrete set of multiple hypotheses (Bun et al., 2019), goodness-of-fit and independence testing (Wang et al., 2015; Gaboardi et al., 2016; Rogers and Kifer, 2017; Kakizaki et al., 2017; Aliakbarpour et al., 2019),

ANOVA testing (Campbell et al., 2018; Swanberg et al., 2019), and nonparametric hypothesis testing (Couch et al., 2019). Finally, hypothesis testing has also been studied with respect to local differential privacy (Duchi et al., 2013; Gaboardi and Rogers, 2018; Sheffet, 2018; Acharya et al., 2019; Gopi et al., 2020; Lam-Weil et al., 2022).

Apart from private *hypothesis testing* of high-dimensional multivariate distributions, there has also been work on private *learning* of Gaussians and multivariate distributions (Karwa and Vadhan, 2018; Kamath et al., 2019a,b; Biswas et al., 2020; Kamath et al., 2020; Cai et al., 2021; Wang and Xu, 2021). Notably, however, learning a distribution requires a linear dependence on the dimension  $d$ , whereas hypothesis testing only requires a square-root dependence on the dimension.

#### 1.4. Roadmap

In Section 2, we give a technical outline for Theorems 5 through 10. All formal proofs are deferred to the appendix. In Appendix A, we define notation and prove some preliminary results. In Appendix B, we prove several important concentration bounds. In Appendix C, we prove Theorem 5. In Appendix D, we prove Theorem 6. In Appendix E, we prove Theorem 7. Finally, in Appendix F, we prove Theorems 8, 9, and 10.

## 2. Technical Overview

In this section, we provide an outline for each of the theorems we prove. We first describe the non-private algorithm for testing the mean of a known-covariance Gaussian. We then outline Theorems 7, 5, and 6 (in that order), and then outline how we can generalize these results to prove Theorems 8, 9, and 10. We view Theorems 7 and 6 as our most interesting results from a technical perspective.

If the covariance  $\Sigma$  is known, we may scale and shift so that WLOG  $\mu^* = \mathbf{0}$  and  $\Sigma = I$  is the identity covariance matrix in  $d$  dimensions. So, our goal is to determine whether  $\mu = 0$  or  $\|\mu\| \geq \alpha$ . For simplicity, we will assume that the alternative hypothesis is  $\|\mu\| = \alpha$  as opposed to  $\|\mu\| \geq \alpha$ .

**Non-Private Hypothesis Testing:** The optimal non-private algorithm (Srivastava and Du, 2008; Canonne et al., 2020a) is incredibly simple: given  $N$  samples  $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ , it just takes the sum of the samples,  $\bar{X} = X^{(1)} + \dots + X^{(N)}$ , and computes the statistic  $T = \|\bar{X}\|^2$ . Based on how large  $T$  is, the algorithm decides whether  $\mu = 0$  or  $\|\mu\| \geq \alpha$ . It is simple to show that if each  $X^{(i)}$  is drawn i.i.d. from  $\mathcal{N}(0, I)$ , then  $\mathbb{E}[T] = N \cdot d$  and  $\text{Var}[T] = O(N^2 \cdot d)$ . Conversely, if each  $X^{(i)}$  is drawn i.i.d. from  $\mathcal{N}(\mu, I)$  where  $\|\mu\| = \alpha$ , then  $\mathbb{E}[T] = N \cdot d + \alpha^2 \cdot N^2$  and  $\text{Var}[T] = O(N^2 \cdot d + N^3 \cdot \alpha^2)$ .

For the statistic  $T$  to successfully distinguish between the two hypotheses, by Chebyshev's inequality, the square of difference in the means must significantly exceed the variances. As the difference in means is  $\alpha^2 \cdot N^2$  and the variances are  $O(N^2 \cdot d + N^3 \cdot \alpha^2)$ , it suffices to choose  $N$  so that  $(\alpha^2 \cdot N^2)^2 \geq \Omega(N^2 \cdot d + N^3 \cdot \alpha^2)$ . This is equivalent to  $N \geq \Omega(\sqrt{d}/\alpha^2)$ , so for  $N \geq C\sqrt{d}/\alpha^2$  for a sufficiently large constant  $C$ , we will be able to distinguish between  $\mu = 0$  and  $\|\mu\| = \alpha$ .

**Theorem 7:** We note that our algorithm, while somewhat based on the non-private hypothesis testing result, deviates significantly from Canonne et al. (2020b) and other work on private hypothesis testing. For that reason, we do not describe the previous techniques of Canonne et al. (2020b).

Before returning to the private setting, we note that  $\|\bar{X}\|^2$ , where  $\bar{X} = X^{(1)} + \dots + X^{(N)}$ , can be rewritten as  $\sum_{i=1}^N \sum_{j=1}^N \langle X^{(i)}, X^{(j)} \rangle$ . Because of this, if we write  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ , the

non-private algorithm can be rephrased as outputting 0 (null hypothesis) if  $\sum_{i,j} T_{i,j} < N \cdot d + \frac{\alpha^2 N^2}{2}$  and 1 (alternative hypothesis) if  $\sum_{i,j} T_{i,j} > N \cdot d + \frac{\alpha^2 N^2}{2}$ .

This motivates our private algorithm, which will attempt to compute  $T = \sum_{i=1}^N \sum_{j=1}^N T_{i,j}$  privately. However, we note that when a single data sample  $X^{(i)}$  changes, this affects  $T_{i,j}$  and  $T_{j,i}$  for all  $j$ . In other words, instead of preserving privacy when a single entry in the matrix  $\{T_{i,j}\}$  changes, we need to preserve privacy when an entire row and column in the matrix changes.

We will modify the matrix  $\mathbf{T}$  consisting of all the entries  $T_{i,j}$ , by subtracting  $d$  from each diagonal entry and then dividing the matrix by  $\tilde{O}(\sqrt{d})$  to get a new matrix  $\mathbf{V}$ . By applying classic concentration bounds, we can show that, assuming the original datapoints  $X^{(1)}, \dots, X^{(N)}$  are drawn i.i.d. from some  $\mathcal{N}(\mu, I)$ , with  $\|\mu\| \leq 1$ , then each entry in  $\mathbf{V}$  is bounded in the range  $[-1, 1]$  and each row/column sum of  $\mathbf{V}$  is bounded in magnitude by  $\sqrt{N}$ . Our goal will roughly be to distinguish between the sum of all the entries in  $\mathbf{V}$  being either in the range  $[-N, N]$  or  $[\gamma N^2 - N, \gamma N^2 + N]$ , for  $\gamma \approx \frac{\alpha^2}{\sqrt{d}}$ . In addition, we wish to perform this *privately*, where we consider two matrices  $\mathbf{V}, \mathbf{V}' \in \mathbb{R}^{N \times N}$  to be adjacent if they differ only in a single row or a single column.

If we may restrict ourselves to matrices where every row and column sum is at most  $\sqrt{N}$  in absolute value, the algorithm is quite simple. Let  $\bar{V} := \sum_{i=1}^N \sum_{j=1}^N V_{i,j}$ . By our restriction,  $\bar{V}$  cannot change by more than  $2\sqrt{N}$  if we only change a single row or column, so the statistic  $\bar{V} + \text{Lap}(\varepsilon^{-1} \cdot 2\sqrt{N})$  is  $(\varepsilon, 0)$ -differentially private. This means that as long as  $N, \varepsilon^{-1} \sqrt{N} \ll \gamma N^2$ , our algorithm will be accurate. This is equivalent to  $N \geq \Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}}\right)$ . The problem, however, is that we want differential privacy for arbitrary adjacent datasets  $X^{(1)}, \dots, X^{(N)}$ , for which we may not have the  $\sqrt{N}$  bound on the row and column sums of the corresponding matrix  $V$ . Even if each entry is bounded in the range  $[-1, 1]$ , we can increase  $\bar{V}$  by  $N$  in the worst case.

A common approach to fixing this is to “clip” each data point, i.e., replace a data point  $x$  with  $\max(x_{min}, \min(x_{max}, x))$  to keep  $x$  in the range  $[x_{min}, x_{max}]$ . This technique has been used in statistical data analysis since the early 20th century. In our case, a natural first attempt to clip the row and column sums. Indeed, we can rewrite  $\bar{V} = \frac{1}{2} \left[ \sum_{i=1}^N \sum_{j=1}^N V_{i,j} + \sum_{j=1}^N \sum_{i=1}^N V_{i,j} \right]$ , i.e.,  $\bar{V}$  is simply the average of the sum over all row sums and the sum over all column sums. Since we want each row sum and column sum to be bounded by the range  $[-\sqrt{N}, \sqrt{N}]$ , we can consider replacing our statistic  $\bar{V}$  with

$$G(\mathbf{V}) := \frac{1}{2} \left[ \sum_{i=1}^N g \left( \sum_{j=1}^N V_{i,j} \right) + \sum_{j=1}^N g \left( \sum_{i=1}^N V_{i,j} \right) \right], \quad (2)$$

where  $g(x) := \min(\sqrt{N}, \max(-\sqrt{N}, x))$  prevents each row/column sum from exceeding  $\sqrt{N}$  in absolute value. However, this still runs into the same problem as before. If we alter a row of  $V$  by increasing each entry in the row by 1, while the modified row sum does not increase by more than  $\sqrt{N}$  now, we still have that each column sum could potentially increase by 1, causing an overall increase by  $N$ .



To fix this, we consider an even stronger version of clipping, where after  $x$  exceeds  $\sqrt{N}$ , the function starts going back down again. Specifically, we instead consider the function

$$g(x) := \begin{cases} x & |x| \leq \sqrt{N} \\ 2\sqrt{N} - x & x \geq \sqrt{N} \\ -2\sqrt{N} - x & x \leq -\sqrt{N} \end{cases}.$$

Note that while  $g(x) = x$  in the range  $[-\sqrt{N}, \sqrt{N}]$ , for general  $x$  we have that  $g(x) \in [-x - O(\sqrt{N}), -x + O(\sqrt{N})]$ . Replacing  $g$  in our Equation (2) with the new function  $g$ , we now see what happens when we change a single row. If we increase every element in a row by 1, each column sum increases by 1, from which we would ideally hope that that  $g$  applied to each column sum increases by 1. Conversely, for the row that we update, we use the fact that  $g(x) \in [-x - O(\sqrt{N}), -x + O(\sqrt{N})]$  to say that in fact  $g$  applied to the row *decreases* by roughly  $N - O(\sqrt{N})$ . So, the overall change in the statistic  $G(\mathbf{V})$  is ideally  $O(\sqrt{N})$ , because the increase of each column sum by 1 cancels out with the decrease in  $g$  applied to the row sum.

The problem with this, however, is that if a column sum exceeds  $\sqrt{N}$  in absolute value,  $g$  applied to that column sum goes down instead. This is not an issue even if up to  $\sqrt{N}$  column sums exceed  $\sqrt{N}$  in absolute value, as even in this case, we have that  $\sqrt{N}$  column sums are decreasing by 1 instead of increasing by 1, so overall the statistic  $G(\mathbf{V})$  still does not change by more than  $O(\sqrt{N})$ . To fix this, we propose another private algorithm which detects and throws out matrices if too many row and column sums exceed  $\sqrt{N}$  in absolute value. We remark that if  $V$  comes from  $X^{(1)}, \dots, X^{(N)}$  drawn from either the null or alternative hypothesis, with high probability no row or column sum will exceed  $\sqrt{N}$  in absolute value, so we do not sacrifice accuracy with this algorithm.

We will consider a new threshold function  $f(x) = \max(0, \min(\frac{|x|}{\sqrt{N}} - 1, 1))$ : this function takes  $\frac{|x|}{\sqrt{N}} - 1$  and clips it to keep it in the range  $[0, 1]$ . Suppose we apply  $f$  to each row and column sum of  $V$ , i.e., we consider the statistic

$$F(\mathbf{V}) := \frac{1}{2} \left[ \sum_{i=1}^N f \left( \sum_{j=1}^N V_{i,j} \right) + \sum_{j=1}^N f \left( \sum_{i=1}^N V_{i,j} \right) \right].$$

If  $X^{(1)}, \dots, X^{(N)}$  actually came from the distribution, then  $f$  applied to each row and column is 0 with very high probability, as no row or column sum exceeds  $\sqrt{N}$  in absolute value. In addition, because  $f$  is capped by 0 and 1,  $f$  applied to a row sum doesn't change by more than 1 if we change the entire row. In addition, each column sum does not change by more than 1, so  $f$  does not change by more than  $1/\sqrt{N}$  for each column sum. So,  $F(\mathbf{V})$  does not change by more than  $\sqrt{N}$  for adjacent datasets. In addition, if more than  $2\sqrt{N}$  row/column sums of  $V$  exceed  $2\sqrt{N}$  in absolute value, then  $F(\mathbf{V}) \geq \sqrt{N}$ . It will be quite simple to utilize a Laplace Mechanism to privately reject any datasets with  $F(\mathbf{V})$  exceeding  $2\varepsilon^{-1}\sqrt{N}$ , which is an  $\varepsilon^{-1}$  factor greater than desired. However, with this weaker bound we can ensure  $G(\mathbf{V})$  does not change by more than  $\varepsilon^{-1}\sqrt{N}$  if a single row/column changes, which can be used to obtain a  $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{4/3}}\right)$ -sample upper bound. This is already an improvement in many regimes.

To improve upon this, we create a series of logarithmic threshold functions  $f_k$  for  $1 \leq k \leq O(\log N)$ , which have increasing thresholds. This will allow us to create statistics  $F_k(\mathbf{V})$  for each  $k$ , similar to  $F(\mathbf{V})$ . Roughly, we will show that in order for  $F_{k+1}$  to change significantly if we

change a single row/column, we require  $F_k$  to be large. From here, we can show that the relative change in  $F_{k+1}$  is much smaller than the relative change in  $F_k$ , unless  $F_k$  is sufficiently large that we could use a Laplace Mechanism to reject such a dataset. We can use these to privately reject any  $\mathbf{X}$  with  $F_{O(\log N)}(\mathbf{V}) \gg \tilde{O}(\sqrt{N})$ , which will allow for a better sample complexity bound for  $N$ .

**Theorem 5:** The proof of Theorem 5 is based on the corresponding result in [Canonne et al. \(2020b\)](#): while their result is not fully accurate, we show how to simultaneously fix their result and improve upon it.

We first sketch the ideas behind the computationally inefficient algorithm of [Canonne et al. \(2020b\)](#). The objective in [Canonne et al. \(2020b\)](#) is to create a map  $\hat{T}$  that sends any dataset  $\mathbf{X} \in (\mathbb{R}^d)^N$  to  $\mathbb{R}$  with two properties. The first property is that for any two adjacent datasets  $\mathbf{X}, \mathbf{X}'$ ,  $\hat{T}(\mathbf{X})$  and  $\hat{T}(\mathbf{X}')$  are “close” in value. The second is that  $\hat{T}(\mathbf{X})$  should almost always be “small” if  $\mathbf{X}$  is a sample of  $N$  i.i.d.  $\mathcal{N}(0, I)$  values, and  $\hat{T}(\mathbf{X})$  should almost always be “large” if  $\mathbf{X}$  is a sample of  $N$  i.i.d.  $\mathcal{N}(\mu, I)$  values, for any  $\mu$  with  $\|\mu\| = \alpha$ . (“Close”, “small”, and “large” can be effectively quantified.) By adding Laplace noise to  $\hat{T}(\mathbf{X})$  and determine if the output exceeds a certain threshold, one can privately distinguish between  $\mu = 0$  and  $\|\mu\| = \alpha$ .

In the case where the data points  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , we have strong concentration of the row and column sums of the corresponding matrix  $\mathbf{T} \in \mathbb{R}^{N \times N}$ , due to the independence of the data points. (Recall that  $\mathbf{T}$  is the matrix with  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ .) If we restrict ourselves exclusively to such datasets where the row and column sums of  $\mathbf{T}$  are properly bounded (call this set  $\mathcal{C}$ ), we can successfully obtain that any two adjacent datasets  $\mathbf{X}, \mathbf{X}'$  have relatively close values of  $T(\mathbf{X}) := \|\sum X^{(i)}\|^2$  and  $T(\mathbf{X}') = \|\sum X'^{(i)}\|^2$ . Hence,  $\hat{T}(\mathbf{X}) = T(\mathbf{X})$  would actually be a suitable choice if we could restrict ourselves to  $\mathcal{C}$ .

[Canonne et al. \(2020b\)](#) combines this observation with a theorem about Lipschitz extensions ([McShane, 1934](#)). The theorem by [McShane \(1934\)](#) states that if there exists a function  $T : \mathcal{C} \subset \mathcal{X}$  where  $\mathcal{X}$  is equipped with some metric (in our case  $\mathcal{X} = (\mathbb{R}^d)^N$  and the metric measures the number of different data points), and  $T$  is  $D$ -Lipschitz for some  $D$ , then there exists an extension  $\hat{T} : \mathcal{X} \rightarrow \mathbb{R}$  that is also  $D$ -Lipschitz. Expressed more simply in our setting, if we can ensure that  $|T(\mathbf{X}) - T(\mathbf{X}')| \leq K \cdot D$  for  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$  that differ in exactly  $K$  data points, then we can extend the function  $T$  to some  $\hat{T}$  which ensures that  $\hat{T}$  does not change by more than  $K$  on adjacent datasets. The smaller we can make  $D$  as a function of  $N$ , the smaller our sample complexity needs to be.

The main issue in [Canonne et al. \(2020b\)](#) is that they only prove that  $|T(\mathbf{X}) - T(\mathbf{X}')| \leq D$  for adjacent datasets  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$ . While this ostensibly ensures that  $|T(\mathbf{X}) - T(\mathbf{X}')| \leq K \cdot D$  for datasets  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$  that differ in at most  $K$  data points, this does not actually hold. For instance, if we change  $K$  of the data points in  $\mathbf{X}$  to make  $\mathbf{X}'$ , the intermediate datasets (obtained by changing the data points one at a time) may not be in  $\mathcal{C}$ . As a result, we in fact must prove that for all integers  $K$ ,  $|T(\mathbf{X}) - T(\mathbf{X}')| \leq K \cdot D$  for  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$  that differ in exactly  $K$  data points. To do this, we further restrict the class  $\mathcal{C} \subset (\mathbb{R}^d)^N$ , by showing a tight concentration of the norm of  $\sum_{i \in S} X^{(i)}$  for all subsets  $S \subset [N]$  of size  $K$  simultaneously, assuming each  $X^{(i)}$  was drawn i.i.d. from some  $\mathcal{N}(\mu, I)$ , and restricting  $\mathcal{C}$  to datasets that satisfy these tight concentration bounds. We then prove that for the restricted set  $\mathcal{C}$ , we obtain our desired Lipschitz property. Our concentration analysis is tighter than that of [Canonne et al. \(2020b\)](#), which provides a smaller Lipschitz parameter  $D$ , and therefore we obtain a reduced value of  $N$  as well. Hence, we are able to fix their inaccuracy as well as improve upon their result.

**Theorem 6:** Our starting point for the lower bound is a theorem of [Acharya et al. \(2018b\)](#), which relates a *coupling* of two distributions  $\mathcal{U}$  and  $\mathcal{V}$  over  $\mathcal{X}^N$  with privacy lower bounds. A coupling of  $\mathcal{U}$  and  $\mathcal{V}$  is a joint distribution over  $(\mathbf{X}, \mathbf{X}') \sim \mathcal{X}^N \times \mathcal{X}^N$  where the marginal of  $\mathbf{X}$  is  $\mathcal{U}$  and the marginal of  $\mathbf{X}'$  is  $\mathcal{V}$ . Specifically, they prove that if there exists a coupling over  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)})$  and  $\mathbf{X}' = (X'^{(1)}, \dots, X'^{(N)})$  where the expected number of  $i \leq N$  such that  $X^{(i)} \neq X'^{(i)}$  is at most  $O(1/\varepsilon)$ , then it is impossible to  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{U}$  and  $\mathcal{V}$ . This method has been used to provide privacy lower bounds in the discrete distribution setting [Acharya et al. \(2018b\)](#).

As noted by [Canonne et al. \(2020b\)](#), proving lower bounds for multivariate Gaussians is much more challenging than for discrete distributions, as the coupled distributions  $\mathcal{U}, \mathcal{V}$  must be generated as Gaussians with identity covariance, which will usually need strong independence guarantees in each coordinate. In contrast, proving similar lower bounds for distributions over a discrete domain  $\{1, 2, \dots, d\}$  do not require us to prove any independence guarantees.

Hence, to apply this result in our setting, two things are necessary. First, we need to decide the distributions  $\mathcal{U}$  and  $\mathcal{V}$ . Next, we need to establish a suitable coupling. The choice for  $\mathcal{U}$  is simple: it will just be the distribution over  $(\mathbb{R}^d)^N$  where each sample is i.i.d.  $\mathcal{N}(0, I)$ . For  $\mathcal{V}$ , we wish to find some distribution of mean vectors  $\mu$  with  $\|\mu\| \geq \alpha$ , and then sample  $N$  points from  $\mathcal{N}(\mu, I)$ . The distribution for the mean vectors we choose will roughly be  $\mu \sim \mathcal{N}(0, \frac{\alpha^2}{d} \cdot I)$ . While this does not ensure that  $\|\mu\| \geq \alpha$ , a simple concentration inequality ensures that  $\|\mu\| \geq \Omega(\alpha)$  with overwhelming probability, which will end up being sufficient.

Next, how do we establish a coupling between  $\mathcal{U}$  and  $\mathcal{V}$ ? The first trick we use is to rewrite the distributions  $\mathcal{U}, \mathcal{V}$  based on the mean vector  $\bar{X} = \frac{X^{(1)} + \dots + X^{(N)}}{N}$  of the  $N$  points in  $\mathcal{U}$  (or  $\mathcal{V}$ ), which we will write as  $a \cdot v$ , where  $a := \|\bar{X}\| \in \mathbb{R}_{>0}$  and  $v := \bar{X}/\|\bar{X}\|$  is a unit vector. Given  $\bar{X}$ , we consider each vector  $X^{(i)} = \bar{X} + y^{(i)} \cdot v + z^{(i)} = (a + y^{(i)}) \cdot v + z^{(i)}$ , where  $y^{(i)} \in \mathbb{R}$  and  $z^{(i)} \in \mathbb{R}^d$  is orthogonal to  $v$ . This means that we decompose  $X^{(i)} - \bar{X}$  into a component in the  $v$  direction and a component in the hyperplane orthogonal to  $v$ . (See [Figure 1](#) for an example of this decomposition.) The advantage of this is that one can show that for both the distribution  $\mathcal{U}$  and  $\mathcal{V}$ , the distribution of the vector  $v$  is uniform across the unit sphere, and the distribution of  $z^{(i)}$  are the same for both  $\mathcal{U}$  and  $\mathcal{V}$ .

Hence, we can reduce our coupling problem to a single dimension, by considering the distributions  $\{a + y^{(i)}\}_{i=1}^N$  and  $\{b + y^{(i)}\}_{i=1}^N$ , where  $a = \|\bar{X}\|$  for  $\mathbf{X} \sim \mathcal{U}$  and  $b = \|\bar{X}\|$  for  $\mathbf{X} \sim \mathcal{V}$ . Note that  $a, b$  are also random variables. One can show that  $a$  and  $b$  converge (in total variation distance) to roughly  $\mathcal{N}(\sqrt{\frac{d}{N}}, \frac{1}{2N})$  and  $\mathcal{N}(\sqrt{\frac{d}{N} + \alpha^2}, \frac{1}{2N})$ , respectively. To explain the intuition behind the Normality, since  $\bar{X}$  is a spherical Gaussian,  $a^2$  behaves exactly like a  $\chi^2$ -random variable with  $d$  degrees of freedom, as the mean of the distribution  $\mathcal{U}$  is the origin. Likewise,  $b^2$  behaves like an off-centered  $\chi^2$ -random variable with  $d$  degrees of freedom, as the mean of the distribution  $\mathcal{V}$  is not the origin but is somewhat close. Therefore,  $a$  behaves like a  $\chi$  random variable and  $b$  behaves like an off-centered  $\chi$  random variable, and as  $d$  grows, this converges to Normal. (We remark that the rate of convergence of  $a, b$  to Normal will not matter: it will just matter that for sufficiently large  $d$ , the total variation distances between  $a, b$  and their respective Normal distributions is less than, say, 0.01.) In addition, one can show that  $\{y^{(i)}\}$  are distributed as i.i.d. standard Normals minus their mean, which has distribution  $\mathcal{N}(0, \frac{1}{N})$ . Indeed, one can use this observation to show that if  $a, b$  were approximately Normal with variance  $\frac{1}{N}$  instead of  $\frac{1}{2N}$ , then  $\{a + y^{(i)}\}_{i=1}^N$  would converge in total variation distance to the distribution of  $N$  i.i.d. samples from  $\mathcal{N}(\sqrt{\frac{d}{N}}, 1)$ , and similarly,

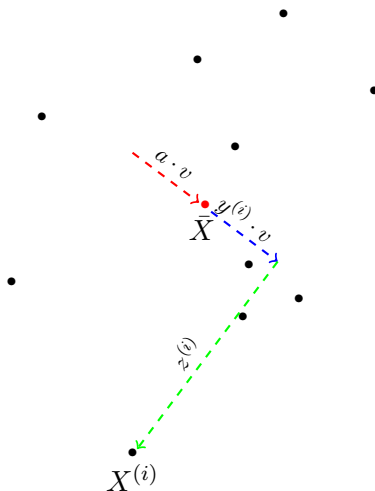


Figure 1: In this figure, we have a series of black points  $X^{(1)}, \dots, X^{(N)}$  and a red point representing their mean  $\bar{X}$ . We decompose an individual  $X^{(i)}$  as  $(a + y^{(i)}) \cdot v + z^{(i)}$ , such that  $a \cdot v$  equals  $\bar{X}$ ,  $y^{(i)} \cdot v$  goes in the same direction as  $\bar{X}$  from the origin, and  $z^{(i)}$  is perpendicular to  $v$ . Our reduction ignores the direction of  $v$  and the orthogonal component  $z^{(i)}$ , and focuses on the quantities  $a + y^{(i)}$ .

$\{b + y^{(i)}\}_{i=1}^N$  would converge in total variation distance to the distribution of  $N$  i.i.d. samples from  $\mathcal{N}(\sqrt{\frac{d}{N} + 2\alpha^2}, 1)$ .

To fix the issue that  $a, b$  have the wrong variance, we use the fact that the total variation distance between  $\mathcal{N}(0, \frac{1}{N})$  and  $\mathcal{N}(0, \frac{1}{2N})$  is less than 0.2. We use this to show that if could privately distinguish between  $\{a + y^{(i)}\}_{i=1}^N$  and  $\{b + y^{(i)}\}_{i=1}^N$  with 0.9 probability, then we could distinguish between  $N$  i.i.d. samples from  $\mathcal{N}(\sqrt{\frac{d}{N}}, 1)$  and from  $\mathcal{N}(\sqrt{\frac{d}{N} + 2\alpha^2}, 1)$  with a weaker 0.7 probability. Now that we have reduced it to univariate and independent samples, we are in a position to create a coupling, which will indeed get us the correct lower bound.

## 2.1. Generalizations

**Theorem 8:** In the case where the samples come from  $\mathcal{N}(\mu, \Sigma)$  where  $\Sigma$  is unknown, we run into the problem that  $T := \|X^{(1)} + \dots + X^{(N)}\|^2$  is no longer concentrated around  $Nd$  (if  $\mu = 0$ ) or  $Nd + \alpha^2 N^2$  (if  $\|\mu\| = \alpha$ ), since  $\Sigma$  is not necessarily the identity matrix. Instead, we will show that  $T$  is concentrated around  $N \cdot J$  (if  $\mu = 0$ ) or  $N \cdot J + \alpha^2 N^2$  (if  $\|\mu\| = \alpha$ ), where  $J = \text{Tr}(\Sigma)$  is unknown. Because  $J$  is unknown, we cannot attempt to directly privately estimate  $T$  and accept or reject the hypothesis based on whether  $T$  exceeds a threshold.

Instead, we will use the fact that given samples from  $\mathcal{N}(\mu, \Sigma)$ , we can generate samples from  $\mathcal{N}(0, \Sigma)$ . This is because if  $X, Y \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$  for any  $\mu \in \mathbb{R}^d$ , then  $\frac{X-Y}{\sqrt{2}} \sim \mathcal{N}(0, \Sigma)$ . Therefore, we can use this to privately estimate  $J$  and privately estimate  $T$ , and then accept or reject based on whether our estimate for  $T$  significantly exceeds our estimate for  $N \cdot J$ . Indeed, we can estimate both  $T$  and  $J$  similarly with the same sample complexity (up to constant factors) as in Theorem

5 and Theorem 7, using the fact that  $\Sigma$  is spectrally bounded by  $I$ . Since estimating  $T$  and  $J$  are sufficient for the testing problem, we can therefore generalize both Theorems 5 and 7.

**Theorem 9:** This case will be almost identical to the Gaussian case. For simplicity, we consider testing whether a product distribution  $\mathcal{P}(\mu)$  has mean  $\mu = 0$  or mean  $\mu$  with  $\|\mu\|_2 = \alpha$ . (Indeed, the  $\ell_2$  norm of  $\mu$  is asymptotically equal to the total variation distance between  $\mathcal{P}(\mu)$  and  $\mathcal{P}(0)$ .) For the upper bound, if we let  $T_{i,j} := \langle X^{(i)}, X^{(j)} \rangle$ , we are able to obtain the same concentration bounds as in the Gaussian case for each entry  $T_{i,j}$  and the sum of each row/column of the matrix  $\mathbf{T}$ , which is required in proving Theorem 7. In addition, we are able to obtain the same concentration bounds as in the Gaussian case for the sum  $\sum_{i \in S} X^{(i)}$  for all subsets  $S \subset [N]$ . From here, the rest of the upper bounds proceed in the same way as in the Gaussian case.

For the lower bound, there is a known reduction (Canonne et al., 2020b) from private identity testing of Gaussians to private uniformity testing of Boolean products, i.e., any upper bound for uniformity testing of Boolean products also holds for Gaussians. Therefore, our lower bound for identity testing of Gaussians implies a lower bound for uniformity testing of Boolean products.

**Theorem 10:** While in many scenarios tolerant identity testing is more difficult than standard identity testing, it will not be so in our case. For instance, suppose we wished to distinguish between i.i.d. samples from  $\mathcal{N}(\mu, I)$ , where either  $\|\mu\| = \frac{\alpha}{2}$  or  $\|\mu\| = \alpha$ . In the non-private setting, we can use the fact that the statistic  $T$  has expectation  $N \cdot d + \frac{\alpha^2}{4} \cdot N^2$  in the former setting, and expectation  $N \cdot d + \alpha^2 \cdot N^2$  in the latter setting. In other words, the difference between the mean of the respective statistics is still  $\Omega(\alpha^2 \cdot N^2)$ . Because of this, we can still do non-private hypothesis testing using the same number of samples. The private setting will work similarly, as our goal has been to output a private version of the statistic  $T$  (or some scaled version of it like  $\bar{V}$ ). Hence, we can apply the same algorithms as before, but change the threshold value accordingly, and still require the same number of samples up to an asymptotic factor.

## Acknowledgments

The author thanks Piotr Indyk for helpful feedback on this paper, as well as Clément Canonne for helpful discussions regarding the paper Canonne et al. (2020b).

## References

- Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems*, pages 3591–3599, 2015.
- Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9469–9481, 2018a.
- Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6879–6891, 2018b.
- Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. In *The 22nd International Conference on Artificial*

- Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2067–2076, 2019.
- Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In *35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 169–178, 2018.
- Maryam Aliakbarpour, Ilias Diakonikolas, Daniel Kane, and Ronitt Rubinfeld. Private testing of distributions via sample permutations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10877–10888, 2019.
- Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 496–505, 2007.
- Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000.
- Ivona Bezáková, Antonio Blanca, Zongchen Chen, Daniel Stefankovic, and Eric Vigoda. Lower bounds for testing graphical models: Colorings and antiferromagnetic ising models. *J. Mach. Learn. Res.*, 21:25:1–25:62, 2020.
- Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan R. Ullman. Coinpress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 1<sup>st</sup> edition, 2013.
- Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 156–167, 2019.
- Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv’it: Private and sample efficient identity testing. In *34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 635–644, 2017.
- T. Tony Cai and Zongming Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.
- T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Ann. Statist.*, 49(5):2825–2850, 2021.
- Zachary Campbell, Andrew Bray, Anna M. Ritz, and Adam Groce. Differentially private ANOVA testing. In *1st International Conference on Data Intelligence and Security, (ICDIS)*, pages 281–285, 2018.
- Clément L. Canonne. A survey on distribution testing. Your data is big. But is it blue? *Theory of Computing Library Graduate Surveys*, 9:1–100, 2020.

- Clément L. Canonne, Gautam Kamath, Audra McMillan, Adam D. Smith, and Jonathan R. Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 310–321, 2019.
- Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Trans. Inf. Theory*, 66(5):3132–3170, 2020a.
- Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan R. Ullman, and Lydia Zakyntinou. Private identity testing for high-dimensional distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.
- Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835, 2010.
- Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. Differentially private non-parametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 737–751, 2019.
- Rachel Cummings, Sara Krehbiel, Yajun Mei, Rui Tuo, and Wanrong Zhang. Differentially private change-point detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10848–10857, 2018.
- Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip LeLerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The modernization of statistical disclosure limitation at the u.s. Census Bureau, 2017. In *Presented at the September 2017 meeting of the Census Scientific Advisory Committee*, 2017.
- Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pages 697–703, 2017.
- Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1989–2007, 2018.
- Ilias Diakonikolas and Daniel M. Kane. The sample complexity of robust covariance testing. In *Conference on Learning Theory, (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 1511–1521, 2021.
- Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.
- John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 429–438. IEEE Computer Society, 2013.

- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284, 2006.
- Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 1054–1067, 2014.
- Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1612–1621, 2018.
- Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2111–2120, 2016.
- Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting ising models. *Electron. Commun. Probab.*, 23(76):1–12, 2018.
- Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, 7(20), 2000.
- Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1785–1816, 2020.
- Harold Hotelling. The generalization of student’s ratio. *Ann. Math. Statist.*, 2(3):360–378, 1931.
- Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, 2014.
- Kazuya Kakizaki, Kazuto Fukuchi, and Jun Sakuma. Differentially private chi-squared test by unit circle mechanism. In Doina Precup and Yee Whye Teh, editors, *34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1761–1770, 2017.
- Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan R. Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902, 2019a.
- Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan R. Ullman. Differentially private algorithms for learning mixtures of separated gaussians. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 168–180, 2019b.
- Gautam Kamath, Vikrant Singhal, and Jonathan R. Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235, 2020.



- Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 94, pages 44:1–44:9, 2018.
- Joseph Lam-Weil, Béatrice Laurent, and Jean-Michel Loubes. Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli*, 28(1):579–600, 2022.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.
- Edward J. McShane. Extension of range of functions. *Bull. Am. Math. Soc.*, 40(12):837–842, 1934.
- Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.
- Aaditya Ramdas, David Isenberg, Aarti Singh, and Larry A. Wasserman. Minimax lower bounds for linear independence testing. In *IEEE International Symposium on Information Theory (ISIT)*, pages 965–969, 2016.
- Ryan Rogers and Daniel Kifer. A new class of private chi-square hypothesis tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000, 2017.
- Ronitt Rubinfeld and Ning Xie. Testing non-uniform  $k$ -wise independent distributions over product spaces. In *37th International Colloquium on Automata, Languages and Programming (ICALP)*, 2010.
- Or Sheffet. Locally private hypothesis testing. In *35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 4612–4621, 2018.
- Muni S. Srivastava and Meng Du. A test for the mean vector with fewer observations than the dimension. *J. Multivar. Anal.*, 99(3):386–402, 2008.
- Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna M. Ritz, Adam Groce, and Andrew Bray. Improved differentially private analysis of variance. *Proc. Priv. Enhancing Technol.*, 2019(3): 310–330, 2019.
- Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017.
- Caroline Uhler, Aleksandra B. Slavkovic, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confidentiality*, 5(1), 2013.
- Duy Vu and Aleksandra B. Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops*, pages 138–143, 2009.
- Di Wang and Jinhui Xu. Differentially private high dimensional sparse covariance matrix estimation. *Theor. Comput. Sci.*, 865:119–130, 2021.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. *CoRR*, abs/1511.03376, 2015. URL <http://arxiv.org/abs/1511.03376>.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 222–227. IEEE Computer Society, 1977.

## Appendix A. Preliminaries

### A.1. Notation

We first note several pieces of notation that we use in our proofs.

- We use  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)})$  to represent a dataset of  $N$  points in  $\mathbb{R}^d$ . We use  $\bar{X}$  to denote either the sum or average of the data points in  $\mathbf{X}$ , depending on context.
- For two datasets  $\mathbf{X}, \mathbf{X}'$ , we use  $\rho(\mathbf{X}, \mathbf{X}')$  to represent the number of data points  $X^{(i)} \neq X'^{(i)}$ .
- We use  $\mathcal{N}(\mu, \Sigma)$  to represent the  $d$ -dimensional Multivariate Gaussian distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ . We use  $I$  to represent the  $d \times d$  identity matrix.
- We use  $\mathcal{P}(\mu)$  to represent the Boolean product distribution over  $\{-1, 1\}^d$  with expectation  $\mu$ .
- We use  $\|\cdot\|$  to represent the Euclidean ( $\ell_2$ ) norm of a vector, and  $\|\cdot\|_2$  to represent the spectral (operator) norm of a symmetric matrix.
- We use  $a = \pm b$  to mean  $-b \leq a \leq b$ , and  $a = b \pm c$  to mean  $b - c \leq a \leq b + c$ .
- We use a lazy  $\tilde{O}$  notation where  $\tilde{O}(f) = f \cdot (\log N)^{O(1)}$  for general  $f$ , even if  $f$  is sub-polynomial in  $N$ . Note that  $N$  will always be at least  $\sqrt{d}$ ,  $\alpha^{-1}$ , and  $\varepsilon^{-1}$ , so any logarithmic factors in  $d$ ,  $\alpha^{-1}$ , and  $\varepsilon^{-1}$  can be incorporated into the  $\tilde{O}$  notation.
- We use  $d_{\text{TV}}(\mathcal{D}, \mathcal{D}')$  to represent the total variation distance between two distributions  $\mathcal{D}$  and  $\mathcal{D}'$  over the same domain.

Finally, we will always assume that our privacy parameter  $\varepsilon$  and error parameter  $\alpha$  are at most a small constant (such as  $1/3$ ).

### A.2. Preliminary Propositions

We define a *subgaussian* random variable and recall some important properties of it.

**Definition 11** A 1-dimensional random variable  $X$  with mean  $\mu$  is subgaussian if there exists a constant  $C$  such that  $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-t^2/C^2}$  for all  $t \geq 0$ . We define the subgaussian norm of such a variable  $X$  to be the infimum over all  $C$  such that  $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-t^2/C^2}$  for all  $t \geq 0$ .

We note the following standard propositions about subgaussian random variables.

**Proposition 12** The univariate Normal distribution  $\mathcal{N}(\mu, \sigma^2)$  has subgaussian norm  $\sigma$ .

**Proposition 13** (*Boucheron et al., 2013*) If  $X_1, \dots, X_n$  are independent variables with each  $X_i$  having subgaussian norm  $\sigma_i$ , then the random variable  $X_1 + \dots + X_n$  has subgaussian norm at most  $O(\sqrt{\sum_{i=1}^n \sigma_i^2})$ .

Next, we note the following proposition linking the total variation distance between two multivariate Gaussians with covariance  $\Sigma$  and their Mahalanobis distance.

**Proposition 14** (*Folklore*) For two distributions  $\mathcal{N}(\mu, \Sigma)$  and  $\mathcal{N}(\nu, \Sigma)$ , recall that the Mahalanobis distance between these two distributions is  $\sqrt{(\mu - \nu)^T \Sigma^{-1} (\mu - \nu)}$ . Then,  $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\nu, \Sigma)) = \Theta(\min(\sqrt{(\mu - \nu)^T \Sigma^{-1} (\mu - \nu)}, 1))$ , i.e., the total variation distance and the Mahalanobis distance are asymptotically equivalent assuming the Mahalanobis distance does not exceed  $\Theta(1)$ .

We also have a similar proposition linking the total variation distance between two product distributions with their means.

**Proposition 15** (*Canonne et al., 2020b, Lemma 2.8*) Let  $\mathcal{P}(\mu), \mathcal{P}(\mu^*)$  be product distributions over  $\{-1, 1\}^d$ , where  $\mu, \mu^* \in [-1, 1]^d$ . Also, suppose that  $\mu^* \in [-1/2, 1/2]^d$ . Then,  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) = \Theta(\min(\|\mu - \mu^*\|, 1))$ .

Proposition 15 implies the following are equivalent: distinguishing between  $\mu^* = \mu$  and  $\|\mu^* - \mu\| \geq \alpha$ , and distinguishing between  $\mathcal{P}(\mu) = \mathcal{P}(\mu^*)$  and  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \geq \Theta(\alpha)$ , assuming that  $\mu^* \in [-1/2, 1/2]^d$ . Hence, in Theorem 9, it is sufficient to distinguish between  $\mathcal{P}^* := \mathcal{P}(\mu^*)$  and all distributions  $\mathcal{P}(\mu)$  with  $\|\mu - \mu^*\| \geq \alpha$ . Likewise, in Theorem 10, if  $C$  is sufficiently large, it is sufficient to distinguish between  $\mathcal{P}(\mu)$  with  $\|\mu - \mu^*\| \leq \frac{\alpha}{2}$  and  $\mathcal{P}(\mu)$  with  $\|\mu - \mu^*\| \geq \alpha$ .

Next, we recall a basic proposition about sufficient statistics of Normal distributions.

**Definition 16** Given a distribution  $\mathcal{D}(\theta)$  over  $\mathbb{R}^d$  parameterized by some  $\theta$  in a domain  $\Theta$ , and given samples  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{D}(\theta)$ , we say that a function  $T = T(X^{(1)}, \dots, X^{(N)})$  is a sufficient statistic for  $\theta$  if the distribution of  $X^{(1)}, \dots, X^{(N)}$  conditioned on  $T$  is the same as the distribution of  $X^{(1)}, \dots, X^{(N)}$  conditioned on  $T$  and  $\theta$ .

**Proposition 17** Let  $\Sigma$  be a fixed covariance matrix, and let  $\mathcal{N}(\mu, \Sigma)$  be parameterized only by  $\mu$ . Then, given samples  $X^{(1)}, \dots, X^{(N)}$ , the empirical mean  $\bar{X} = \frac{X^{(1)} + \dots + X^{(N)}}{N}$  is a sufficient statistic for  $\mu$ . In other words, if given  $N$  i.i.d. samples from  $\mathcal{N}(\mu, \Sigma)$ , the conditional distribution of  $X^{(1)}, \dots, X^{(N)}$  conditioned on  $\frac{X^{(1)} + \dots + X^{(N)}}{N}$  is independent of  $\mu$ .

**Corollary 18** Let  $X^{(1)}, \dots, X^{(N)}$  be distributed i.i.d. as  $\mathcal{N}(\mu, \Sigma)$ , and let  $\bar{X} = \frac{X^{(1)} + \dots + X^{(N)}}{N}$ . Then, for  $Z^{(1)}, \dots, Z^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$  and  $\bar{Z} = \frac{Z^{(1)} + \dots + Z^{(N)}}{N}$ , we have that  $X^{(1)}, \dots, X^{(N)}$  has the same distribution as  $\bar{X} + Z^{(1)} - \bar{Z}, \dots, \bar{X} + Z^{(N)} - \bar{Z}$ .

**Proof** It suffices to show they have the same distribution if we condition on  $\bar{X}$  and  $\bar{Z}$ .

By Proposition 17, if we know  $\bar{X}$ , it does not matter whether  $X^{(1)}, \dots, X^{(N)}$  were originally drawn from  $\mathcal{N}(\mu, \Sigma)$  or from  $\mathcal{N}(\mu', \Sigma)$  for some  $\mu' \neq \mu$ . So, we can instead pretend that  $X^{(1)}, \dots, X^{(N)}$  were drawn from  $\mathcal{N}(v, \Sigma)$ , for  $v = \bar{X} - \bar{Z}$ . In this case, by an additive shift of  $v$ , the conditional distribution of  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(v, \Sigma)$  conditioned on  $\bar{X}$  is the same as

the conditional distribution of  $Z^{(1)} + v, \dots, Z^{(N)} + v \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$  conditioned on  $\bar{Z}$ . Hence, by writing  $v = \bar{X} - \bar{Z}$ , we are done.  $\blacksquare$

Finally, we note the following concentration bound, due to [Laurent and Massart \(2000\)](#).

**Theorem 19** ([Laurent and Massart, 2000, Lemma 1](#)) *Let  $a_1, \dots, a_n$  be nonnegative reals, all bounded by some  $|a|_\infty > 0$ . Let  $|a|_2 = \sqrt{a_1^2 + \dots + a_n^2}$ . Let  $Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ , and let  $Z = \sum_{i=1}^n a_i Z_i^2$ . Then, for any positive real  $t$ , we have that*

$$\mathbb{P} \left( \left| Z - \sum_{i=1}^n a_i \right| \geq 2|a|_2 \sqrt{t} + 2|a|_\infty t \right) \leq 2e^{-t}.$$

### A.3. Reductions

In this subsection, we note several reductions between problems that will make things easier to prove.

First, we note the asymptotic equivalence between  $(\varepsilon, \delta)$ -DP and  $(\varepsilon + \delta, 0)$ -DP in the hypothesis testing case. Roughly, this follows because we assume the probability of acceptance/rejection has two-sided error. For instance, if we additively change the probability of acceptance from  $1/3$  to  $1/3 + \varepsilon$ , this is also a multiplicative  $1 + O(\varepsilon)$  change in the probability of acceptance, and a multiplicative  $1 - O(\varepsilon)$  change in the probability of rejection. Formally, we have the following:

**Proposition 20** ([Acharya et al., 2018b, Lemma 5](#)) *There is an  $(\varepsilon, \delta)$ -DP algorithm for a testing problem if and only if there is an  $(O(\varepsilon + \delta), 0)$ -DP algorithm for the same testing problem.*

Next, we recall [Propositions 14 and 15](#). From here, we note that it is asymptotically equivalent to distinguish between the the distribution being far from the null hypothesis distribution in total variation distance and the mean being far from the null hypothesis mean in Mahalanobis distance (in the known-covariance Gaussian case) or  $\ell_2$  distance (in the Boolean product case). As we saw in [Subsection A.2](#), in the Boolean product case the equivalence also extends to tolerant distribution testing.

Next, we note that by shifting the distribution, we may always assume WLOG (in the multivariate Gaussian cases) that  $\mu^* = 0$ . In addition, if the covariance is known, we may also assume WLOG that  $\Sigma = I$ , by multiplying all data points by the matrix  $\Sigma^{-1/2}$ .

We also note that in the Boolean product case, we may assume that  $\mu^* = 0$ . Indeed, we have the following lemma, due to [Canonne et al. \(2020b\)](#):

**Lemma 21** ([Canonne et al., 2020b, Lemma 5.1, rephrased](#)) *Suppose that  $\mathcal{P}(\mu^*)$  is a known Boolean product distribution, and we are given a sample  $X \sim \mathcal{P}(\mu)$  where  $\mu$  is unknown. Then, there exists a randomized transformation  $Y = f_{\mu^*}(X)$  such that  $Y \sim \mathcal{P}(\frac{\mu - \mu^*}{2})$ .*

Thus, up to changing  $\alpha$  by a factor of 2, we may assume WLOG that  $\mu^* = 0$  for identity testing of Boolean products, since  $\mu = \mu^*$  is equivalent to  $\frac{\mu - \mu^*}{2} = 0$  and  $\|\mu - \mu^*\| \geq \alpha$  is equivalent to  $\left\| \frac{\mu - \mu^*}{2} \right\| \geq \frac{\alpha}{2}$ . Note that this reduction also holds for the tolerant testing case as well.

Next, we note a proposition showing that testing known-covariance Gaussians is **easier** than Testing Product distributions. Because of this, the only lower bounds we need to show is for testing

known-covariance Gaussians. Of course, testing of unknown-covariance Gaussians and tolerant testing of Gaussians are harder problems, so the same lower bounds must hold, and the proposition we are about to state also implies that the same lower bounds hold for testing (and tolerant testing) of product distributions as well.

**Proposition 22** (*Canonne et al., 2020b, Theorem 3.1, rephrased*) *There exists a function  $F : \mathbb{R}^d \rightarrow \{-1, 1\}^d$  and an absolute constant  $c > 0$  such that the following holds. Suppose that  $X \sim \mathcal{N}(\mu, I)$ , where  $\mu$  is unknown (so,  $F$  does not depend on  $\mu$ ). Then:*

- *If  $\mu = 0$ , the distribution of  $F(X)$  is uniform on  $\{-1, 1\}^d$ .*
- *If  $\mu \neq 0$ , the distribution of  $F(X)$  is a product distribution  $\mathcal{P}(\mu')$  where  $d_{TV}(\mathcal{P}(0), \mathcal{P}(\mu')) \geq c \cdot \min(1, \|\mu\|)$ .*

Hence, we have a reduction from testing the mean of a Gaussian with identity covariance to testing uniformity of a product distribution, which implies that any lower bound for testing the mean of a identity covariance Gaussian implies the same lower bound for testing uniformity of a product distribution.

Finally, we note that in all of our cases, our alternative hypothesis may be phrased as  $\|\mu\| \geq \alpha$  (since we may assume  $\mu^* = 0$  and since we have provided an equivalence between total variation distance being large and  $\|\mu\|$  being large in the Boolean product case). In the following proposition, we show that it suffices to consider the weaker alternative hypothesis of  $\alpha \leq \|\mu\| \leq 2\alpha$ .

**Proposition 23** *Let  $0 < \alpha \leq 1$ , and fix some positive integer  $N$ , where  $N \geq \varepsilon^{-1}$ . Suppose  $\mathcal{A}$  is an algorithm that can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0 : N$  samples drawn i.i.d. from  $\mathcal{N}(0, I)$ , and  $\mathcal{H}_1 : N$  samples drawn i.i.d. from  $\mathcal{N}(\mu, I)$  where  $\alpha \leq \|\mu\| \leq 2\alpha$ . Then, using  $N' = O(N \log \frac{d}{\alpha} \log \log \frac{d}{\alpha})$  samples, we can  $(O(\varepsilon), 0)$ -privately distinguish between  $\mathcal{H}_0 : N$  samples drawn i.i.d. from  $\mathcal{N}(0, I)$ , and  $\mathcal{H}_1 : N$  samples drawn i.i.d. from  $\mathcal{N}(\mu, I)$  where  $\|\mu\| \geq \alpha$ .*

**Proof** First, we can learn the mean  $\mu$  with  $1 - o(1)$  probability up to error  $O(\sqrt{d})$  and with  $(0, \varepsilon)$ -privacy, by picking a random point among the sampled  $N' \geq N \geq \frac{1}{\varepsilon}$  points and outputting that as our initial guess  $\tilde{\mu}$ . If the estimate  $\tilde{\mu}$  is not  $O(\sqrt{d})$  in magnitude, we can instantly reject the null hypothesis and output 1.

Next, consider partitioning the points  $X^{(1)}, \dots, X^{(N')}$  into groups that each contain  $N$  points, and the groups are in turn partitioned into “supergroups” that each contain  $O(\log \log \frac{d}{\alpha})$ . There are a total of  $O(\log \frac{d}{\alpha})$  supergroups. Now, for each point  $X^{(j)}$  in the  $t$ th supergroup, we replace it with  $X'^{(j)} := (X^{(j)} + \sqrt{2^{2t-2} - 1} \cdot Z_j) / 2^{t-1}$ , where  $Z_j \sim \mathcal{N}(0, I)$  is independent for each point  $X^{(j)}$ . Note that if  $X^{(j)} \sim \mathcal{N}(\mu, I)$ , then  $X'^{(j)} \sim \mathcal{N}(\mu / 2^{t-1}, I)$ . Now, for each group, we run  $\mathcal{A}$  on the  $N$  points  $X'^{(j)}$  that are now in it. The output for each group will either be 0 or 1. Next, for each  $t$ , we compute the majority output for all groups in the  $t$ th supergroup to obtain a bit  $b_t$ , which is either 0 or 1. Our final algorithm returns 0 if and only if  $b_t = 0$  for all  $1 \leq t \leq O(\log \frac{d}{\alpha})$ .

First, suppose that the samples  $X^{(j)}$  are drawn i.i.d. from  $\mathcal{N}(0, I)$ . Then, the samples  $X'^{(j)}$  are also i.i.d. from  $\mathcal{N}(0, I)$ . This means that each time we run  $\mathcal{A}$  on some group, we output 0 with probability at least  $2/3$ , which means that the majority output in any supergroup will be 0 with probability at least  $1 - \frac{c}{\log d/\alpha}$  for some small constant  $c$ . So, each  $b_t$  equals 0 with probability at least  $1 - \frac{c}{\log d/\alpha}$ , which means that all of the  $b_t$ 's equal 0 with probability at least  $3/4$ . In addition,

our initial guess  $\tilde{\mu}$  has norm at most  $O(\sqrt{d})$  with  $1 - o(1)$  probability, so we output 0 as our final answer with probability at least  $3/4 - o(1) \geq 2/3$ .

Next, suppose that the samples are drawn i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $O(\sqrt{d}) \geq \|\mu\| \geq \alpha$ . Let  $t \geq 1$  be the unique integer such that  $2^{t-1} \cdot \alpha \leq \|\mu\| < 2^t \cdot \alpha$ . Note that  $t \leq O(\log \frac{d}{\alpha})$ . Then, every  $X^{(j)}$  in the  $t$ th supergroup is i.i.d.  $\mathcal{N}(\mu/2^{t-1}, I)$ , where  $\alpha \leq \|\mu/2^{t-1}\| \leq 2\alpha$ . So, in each group in the  $t$ th supergroup,  $\mathcal{A}$  outputs 1 with probability at least  $2/3$ , which means that  $b_t = 1$  with probability at least  $1 - \frac{c}{\log d/\alpha} \geq 2/3$ . We may have rejected the null hypothesis immediately, but that only helps us.

Finally, if the samples are drawn i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $\mu \geq O(\sqrt{d})$ , then we would have rejected the null hypothesis immediately with probability at least  $1 - o(1)$ . Hence, the overall algorithm is accurate.

To verify that the algorithm is private, note that replacing each  $X^{(j)}$  with  $X^{(j)}$  only depends on the value  $X^{(j)}$  and the position  $j$ . In addition, we are performing  $(\varepsilon, 0)$ -DP algorithms on each group, but the groups are distinct. So, any algorithm that only depends on the output of  $\mathcal{A}$  on each group is still  $(\varepsilon, 0)$ -DP. In addition, the first part of the algorithm of determining  $\tilde{\mu}$  is  $(0, \varepsilon)$ -DP, so the overall algorithm is  $(\varepsilon, \varepsilon)$ -DP by the weak composition theorem. Thus, we can modify the algorithm to be  $(O(\varepsilon), 0)$ -DP, by Proposition 20.  $\blacksquare$

**Remark 24** *We note that Proposition 23 extends easily to the cases of unknown but bounded covariance Gaussians, product distributions, and tolerant identity testing. Indeed, for the case of bounded covariance Gaussians, we note that we can replace each  $X^{(j)}$  in the  $t$ th supergroup with  $X^{(j)} := (X^{(j)} + \sqrt{2^{2t-2} - 1} \cdot Z_j)/2^{t-1}$  where  $Z_j \sim \mathcal{N}(0, I)$ , and then each  $X^{(j)}$  in the  $t$ th supergroup is independent and has distribution  $\mathcal{N}(\mu/2^t, (1 - 1/2^{2t-2}) \cdot I + (1/2^{2t-2}) \cdot \Sigma)$ , and note that if  $\|\Sigma\|_2 \leq 1$ , then  $\|(1 - 1/2^{2t-2}) \cdot I + (1/2^{2t-2}) \cdot \Sigma\|_2 \leq 1$  as well. For the case of product distributions, if each sample  $X^{(j)}$  were drawn from  $\mathcal{P}(\mu)$ , we can replace it with  $X^{(j)}$  where each coordinate of  $X_i^{(j)}$  is independently equal to  $X_i^{(j)}$  with probability  $\frac{1}{2} + \frac{1}{2^t}$  and is  $-X_i^{(j)}$  with probability  $\frac{1}{2} - \frac{1}{2^t}$ . Then, note that the distribution of  $X^{(j)}$  is precisely  $\mathcal{P}(\mu/2^{t-1})$ . Finally, in the case of tolerant testing, note that if  $\|\mu\| \leq \alpha/2$ , then  $\|\mu/2^{t-1}\| \leq \alpha/2^t \leq \alpha/2$  for all  $t \geq 1$ , so the same reduction in Proposition 23 holds.*

Therefore, in all of our theorems, it suffices to consider the null hypothesis of  $\mu = 0$  and the alternative hypothesis of  $\alpha \leq \|\mu\| \leq 2\alpha$ . (Recall that we can assume WLOG that  $\mu^* = 0$  always, and that  $\Sigma = I$  in the known-covariance Gaussian case.)

## Appendix B. Concentration Bounds

The main focus of this section is to state and prove various concentration bounds that are crucial in establishing accuracy of our differentially private algorithms. These bounds establish important properties of a matrix  $T$  where  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ , and each  $X^{(i)}$  is drawn i.i.d. either from some Gaussian  $\mathcal{N}(\mu, I)$  or some Boolean Product distribution  $\mathcal{P}(\mu)$ . We have 4 types of bounds.

1. Lemmas 25 and 26 provide bounds on the sum over all entries in  $T$ .
2. Lemmas 27 and 28 provide bounds on individual terms of the matrix  $T_{i,j}$ .
3. Proposition 29 and Lemma 30 provide bounds on the sum of a row or column of  $T$ .

4. Proposition 31 to Lemma 36 provide bounds on the sum of submatrices of  $T$ .

We will combine these bounds together in Theorem 37.

We will also prove related results (Propositions 38 and 39, and Theorem 40) for when each sample  $X^{(i)}$  is drawn i.i.d. from some Gaussian  $\mathcal{N}(\mu, \Sigma)$ , where  $\|\Sigma\|_2 \leq 1$  is unknown. This will be important for generalizing our upper bound results in Theorem 8.

Finally, we will also prove two results relating to univariate Normal and chi variables (Propositions 41 and 42), which will be important in establishing our sample complexity lower bound.

**Lemma 25** (*Canonne et al., 2020b*) Let  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , and let  $\bar{X} = \sum_{i=1}^N X^{(i)}$ . Then,

- $\mathbb{E} [\|\bar{X}\|^2] = Nd + N^2 \cdot \|\mu\|^2$ .
- $\text{Var} [\|\bar{X}\|^2] = O(N^2d + N^3 \cdot \|\mu\|^2)$ .

**Lemma 26** (*Canonne et al., 2020b*) Let  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{P}(\mu)$ , where  $\|\mu\| \leq 1$ , and let  $\bar{X} = \sum_{i=1}^N X^{(i)}$ . Then,

- $\mathbb{E} [\|\bar{X}\|^2] = Nd + \Theta(N^2 \cdot \|\mu\|^2)$ .
- $\text{Var} [\|\bar{X}\|^2] = O(N^2d + N^3 \cdot \|\mu\|^2)$ .

**Lemma 27** Let  $X, Y \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , where  $\|\mu\| \leq 1$ . Then, with probability at least  $1 - \frac{1}{N^3}$ ,  $\|X\|^2 = d \pm O(\sqrt{d} \cdot \log N)$ . Likewise, with probability at least  $1 - \frac{1}{N^3}$ ,  $\langle X, Y \rangle = \pm O(\sqrt{d} \cdot \log N)$ .

**Proof** By rotational symmetry of Gaussians, we may assume WLOG that  $\mu = (\beta, 0, 0, \dots, 0)$ , where  $\beta = \|\mu\|$ . Then,  $\|X\|^2 = (\beta + Z_1)^2 + Z_2^2 + \dots + Z_d^2$ , where each  $Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . First, note that  $|Z_1| \leq O(\sqrt{\log N})$  with at least  $1 - \frac{1}{2N^3}$  probability, in which case  $(\beta + Z_1)^2 = O(\log N)$ . Next,  $Z_2^2 + \dots + Z_d^2 \sim \chi_{d-1}^2$ . By Theorem 19, we have that  $\mathbb{P}(|\chi_k^2 - k| \geq 2\sqrt{kt} + 2t) \leq 2e^{-t}$  for all  $t \geq 0$ , so by setting  $k = d-1$  and  $t = O(\log N)$ , we have that  $\mathbb{P}(|\chi_{d-1}^2 - (d-1)| \geq O(\sqrt{d \log N} + \log N)) \leq \frac{1}{4N^3}$ . Therefore, we have that  $\mathbb{P}(\|\|X\|^2 - d\| \geq O(\sqrt{d \log N} + \log N)) \leq 1 - \frac{1}{N^3}$ .

Next,  $\langle X, Y \rangle = (\beta + Z_1)(\beta + Z'_1) + Z_2 Z'_2 + \dots + Z_d Z'_d$ , where each  $Z_i, Z'_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . We have  $|Z_1|, |Z'_1| \leq O(\sqrt{\log N})$  with at least  $1 - \frac{1}{2N^3}$  probability, in which case  $(\beta + Z_1)(\beta + Z'_1) = O(\log N)$ . Next,

$$Z_2 Z'_2 + \dots + Z_d Z'_d = \frac{1}{4} [(Z_2 + Z'_2)^2 + \dots + (Z_d + Z'_d)^2 - (Z_2 - Z'_2)^2 + \dots + (Z_d - Z'_d)^2],$$

and using the fact that  $Z_i + Z'_i, Z_i - Z'_i$  are i.i.d.  $\mathcal{N}(0, 2)$  variables, this equals  $\frac{1}{2}[A - B]$  where  $A, B \stackrel{i.i.d.}{\sim} \chi_{d-1}^2$ . As we have already seen,  $\mathbb{P}(|A - (d-1)| \geq O(\sqrt{d \log N} + \log N)) \leq \frac{1}{4N^3}$  and  $\mathbb{P}(|B - (d-1)| \geq O(\sqrt{d \log N} + \log N)) \leq \frac{1}{4N^3}$ , so  $\mathbb{P}(|A - B| \geq O(\sqrt{d \log N} + \log N)) \leq \frac{1}{2N^3}$ . Therefore, writing  $\langle X, Y \rangle = \frac{1}{2}(A - B) + (\beta + Z_1)(\beta + Z'_1)$ , we have that  $\mathbb{P}(\langle X, Y \rangle \geq O(\sqrt{d \log N} + \log N)) \leq \frac{1}{N^3}$ .  $\blacksquare$

**Lemma 28** *Let  $\mu \in [-1, 1]^d$  be such that  $\|\mu\| \leq 1$ . Let  $X, Y \stackrel{i.i.d.}{\sim} \mathcal{P}(\mu)$ . Then,  $\|X\|^2 = d$  with probability 1, and with probability at least  $1 - \frac{1}{N^3}$ ,  $\langle X, Y \rangle = \pm O(\sqrt{d \cdot \log N})$ .*

**Proof** Since  $X \in \{-1, 1\}^d$ , we always have  $\|X\|^2 = d$ . So, the first part of the lemma is immediate.

For the second part of the lemma, we have  $\langle X, Y \rangle = \sum_{i=1}^d X_i Y_i$ . For each  $i \in [d]$ ,  $X_i Y_i$  has mean  $\mathbb{E}[X_i] \cdot \mathbb{E}[Y_i] = \mu_i^2$ . In addition,  $X_i Y_i$  is bounded, so it has variance  $O(1)$ . By Hoeffding's inequality, with probability at least  $1 - \frac{1}{N^3}$ ,  $\sum_{i=1}^d X_i Y_i$  is not more than  $O(\sqrt{d \log N})$  away from its expectation, which is  $\sum_{i=1}^d \mu_i^2 = \|\mu\|^2 = O(1)$ . This proves the second part of the lemma.  $\blacksquare$

**Proposition 29** *Fix a vector  $v \in \mathbb{R}^d$  of magnitude 1 and a vector  $\mu \in [-1, 1]^d$ , and let  $K \geq 1$  be a fixed integer. Let  $X^{(1)}, \dots, X^{(K)}$  be distributed either as i.i.d.  $d$ -dimensional Gaussians  $\mathcal{N}(\mu, I)$  or as i.i.d. product distributions  $\mathcal{P}(\mu)$ , and define  $\bar{X} := X^{(1)} + \dots + X^{(K)}$ . Then,  $Y := \langle \bar{X} - K \cdot \mu, v \rangle$  is a subgaussian random variable with subgaussian norm at most  $O(\sqrt{K})$ , meaning that  $\mathbb{P}[|Y| \geq t] \leq 2e^{-\Omega(t^2/K)}$  for all  $t \geq 0$ .*

**Proof** Suppose  $\bar{X}$  is distributed as the sum of  $K$  i.i.d.  $d$ -dimensional Gaussians  $\mathcal{N}(\mu, I)$ . Then,  $\bar{X} \sim \mathcal{N}(K \cdot \mu, K \cdot I)$ . Then,  $\bar{X} - K \cdot \mu \sim \mathcal{N}(0, K \cdot I)$ , so  $\langle \bar{X} - K \cdot \mu, v \rangle \sim \mathcal{N}(0, K)$ . Hence it has subgaussian norm at most  $\sqrt{K}$ .

Suppose  $\bar{X}$  is distributed as the sum of  $K$  i.i.d.  $d$ -dimensional product distributions over  $\{-1, 1\}^d$  with mean  $\mu$ . Then,  $\bar{X}_i - K \cdot \mu_i$  is the sum of  $K$  i.i.d. mean 0 variables bounded in magnitude by 2. Hence, each of these bounded variables has subgaussian norm  $O(1)$ , so  $\bar{X}_i - K \cdot \mu_i$  has subgaussian norm  $O(\sqrt{K})$  by Proposition 13. Now, note that  $\langle \bar{X} - K \cdot \mu, v \rangle = \sum_{i=1}^d v_i (\bar{X}_i - K \cdot \mu_i)$ , and the random variables  $\bar{X}_i - K \cdot \mu_i$  have mean 0 and are independent. Therefore, by Proposition 13,  $\sum_{i=1}^d v_i (\bar{X}_i - K \cdot \mu_i)$  has squared subgaussian norm at most  $O(K) \cdot \sum_{i=1}^d v_i^2 = O(K)$ , so  $\langle \bar{X} - K \cdot \mu, v \rangle = \sum_{i=1}^d v_i (\bar{X}_i - K \cdot \mu_i)$  has subgaussian norm at most  $O(\sqrt{K})$ .  $\blacksquare$

**Lemma 30** *Let  $X^{(1)}, \dots, X^{(N)}$  be distributed either as i.i.d.  $d$ -dimensional Gaussians  $\mathcal{N}(\mu, I)$  or as i.i.d. product distributions  $\mathcal{P}(\mu)$ , where  $\|\mu\| \leq 1$ . Then,  $\langle X^{(1)}, X^{(1)} + \dots + X^{(N)} \rangle = d \pm \tilde{O}(\sqrt{Nd} + \|\mu\| \cdot N)$  with probability at least  $1 - \frac{3}{N^3}$ .*

**Proof** First, note that by Lemmas 27 and 28, we have that  $\|X^{(1)}\|^2 = d \pm O(\sqrt{d} \cdot \log N)$  with probability at least  $1 - \frac{1}{N^3}$ . Next, since  $X^{(1)}$  is independent of  $X^{(2)}, \dots, X^{(N)}$ , by setting  $K = N - 1$ ,  $v = X^{(1)} / \|X^{(1)}\|$ , and  $t = O(\sqrt{(N-1) \log N})$  in Proposition 29, we have that with probability at least  $1 - \frac{1}{N^3}$ ,  $|\langle X^{(1)}, X^{(2)} + \dots + X^{(N)} - (N-1)\mu \rangle| \leq O(\sqrt{N \log N}) \cdot \|X^{(1)}\|$ . In addition,

we know that  $\|X^{(1)}\| \leq O\left(\sqrt{d + \sqrt{d} \log N}\right) = O(\sqrt{d \log N})$  with probability at least  $1 - \frac{1}{N^3}$ .

Finally, we note that in the Gaussian setting,  $\langle X^{(1)}, \mu \rangle$  is distributed as  $\mathcal{N}(0, \|\mu\|^2)$ , and in the product distribution testing,  $\langle X^{(1)}, \mu \rangle$  is distributed as  $\sum_{i=1}^d X_i^{(1)} \mu_i$ , which has subgaussian norm  $O\left(\sqrt{\sum_{i=1}^d \mu_i^2}\right) = O(\|\mu\|)$ . Thus, with probability at least  $1 - \frac{1}{N^3}$ ,  $\langle X^{(1)}, \mu \rangle = O(\sqrt{\log N} \cdot \|\mu\|)$ , for both the Gaussian and product settings.

To summarize, we have that

$$\langle X^{(1)}, X^{(1)} + \dots + X^{(N)} \rangle = \|X^{(1)}\|^2 + \langle X^{(1)}, X^{(2)} + \dots + X^{(N)} - (N-1)\mu \rangle + (N-1) \cdot \langle X^{(1)}, \mu \rangle,$$



which with probability at least  $1 - \frac{3}{N^3}$  is

$$d \pm O\left(\sqrt{d} \log N + \sqrt{N \log N} \cdot \sqrt{d \log N} + N \sqrt{\log N} \cdot \|\mu\|\right) = d \pm \tilde{O}\left(\sqrt{Nd} + \|\mu\| \cdot N\right).$$

■

**Proposition 31** *Let  $Z \sim \mathcal{N}(0, 1)$  be a standard univariate Gaussian. Then, for any  $-\frac{1}{10} \leq \gamma \leq \frac{1}{10}$ ,  $\mathbb{E}[e^{\gamma \cdot Z^2}] \leq 1 + \gamma + O(\gamma^2)$ .*

**Proof** By integrating the PDF of the standard normal, we have that

$$\begin{aligned} \mathbb{E}\left[e^{\gamma \cdot Z^2}\right] &= \int_{-\infty}^{\infty} e^{\gamma \cdot x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1-2\gamma)x^2/2} dx \\ &= \frac{1}{\sqrt{1-2\gamma}} = 1 + \gamma + O(\gamma^2) \end{aligned}$$

for all  $\gamma \in [-1/10, 1/10]$ .

■

**Proposition 32** *Let  $K \geq 1$  be an integer, and  $Z' = \frac{1}{\sqrt{K}} \sum_{i=1}^K Z'_i$ , where each  $Z'_i$  is an i.i.d. random variable with support in  $[-1, 1]$ . Finally, let  $Z = Z' - \mathbb{E}[Z']$ , and let  $\eta = \text{Var}(Z) = \text{Var}(Z')$ . Then, for any  $-\frac{1}{10} \leq \gamma \leq \frac{1}{10}$ ,  $\mathbb{E}[e^{\gamma \cdot Z^2}] \leq 1 + \eta \cdot \gamma + O(\gamma^2)$ .*

**Proof** First, by Hoeffding's inequality, we have that for any  $t \geq 0$ ,  $\mathbb{P}(|Z| \geq t) \leq 2e^{-t^2/2}$ . Therefore,

$$\begin{aligned} \mathbb{E}\left[e^{1/3 \cdot Z^2}\right] &= \int_0^{\infty} \mathbb{P}\left(e^{1/3 \cdot Z^2} \geq t\right) dt \\ &= \int_0^{\infty} \mathbb{P}\left(|Z| \geq \sqrt{3 \ln t}\right) dt \\ &= \int_0^{\infty} \min\left(1, 2e^{-3 \ln t/2}\right) dt \\ &\leq 1 + \int_1^{\infty} \frac{2}{t^{3/2}} dt = O(1). \end{aligned}$$

Now, for all  $x \in \mathbb{R}$ , note that  $e^x \leq 1 + x + x^2 e^{|x|}$ , which is immediate by the Taylor series expansion of  $x$ . Hence, for any  $-\frac{1}{10} \leq \gamma \leq \frac{1}{10}$ ,  $e^{\gamma Z^2} \leq 1 + \gamma Z^2 + \gamma^2 Z^4 e^{|\gamma| Z^2}$ . Further, we can write  $x \leq e^x$  for all  $x$ , which means  $Z^2/10 \leq e^{Z^2/10}$ , so  $Z^2 \leq 10e^{Z^2/10}$ . Therefore,

$$e^{\gamma Z^2} \leq 1 + \gamma Z^2 + \gamma^2 Z^4 e^{|\gamma| Z^2} \leq 1 + \gamma Z^2 + 10^2 \gamma^2 e^{(|\gamma|+1/5)Z^2}.$$

Now, using the fact that  $\mathbb{E}[Z^2] = \text{Var}(Z) = \eta$ , we have that

$$\begin{aligned} \mathbb{E}\left[e^{\gamma \cdot Z^2}\right] &\leq \mathbb{E}\left[1 + \gamma Z^2 + 10^2 \gamma^2 e^{(|\gamma|+1/5)Z^2}\right] \\ &= 1 + \eta \cdot \gamma + O(\gamma^2), \end{aligned}$$

since  $|\gamma| + \frac{1}{5} \leq \frac{1}{3}$ .

■

**Proposition 33** *Let  $X^{(1)}, \dots, X^{(K)} \sim \mathcal{N}(0, 1)$  be i.i.d.  $d$ -dimensional Gaussians. Then, for any sufficiently large  $C$ ,  $\|X^{(1)} + \dots + X^{(K)}\|^2 = Kd \pm C(K\sqrt{Kd} + K^2)$  with failure probability at most  $e^{-(C/20) \cdot K}$ .*

**Proof** Let  $\hat{X} = X^{(1)} + \dots + X^{(K)}$ . Note that each coordinate  $\hat{X}_i$  has distribution  $\mathcal{N}(0, K)$ . Therefore, for any  $\gamma \in [-0.1, 0.1]$ ,  $\mathbb{E}[e^{(\gamma/K) \cdot \hat{X}_i^2}] \leq 1 + \gamma + O(\gamma^2) = e^{\gamma + O(\gamma^2)}$ . Therefore, since each coordinate of  $\hat{X}$  is independent, by Proposition 31,

$$\mathbb{E} \left[ e^{(\gamma/K) \cdot \|\hat{X}\|^2} \right] = \prod_{i=1}^d \mathbb{E}[e^{(\gamma/K) \cdot \hat{X}_i^2}] \leq e^{d\gamma + O(d\gamma^2)}. \quad (3)$$

If we set  $0 < \gamma < 0.1$ , this means that for any  $C \geq 0$ ,

$$\begin{aligned} \mathbb{P}(\|\hat{X}\|^2 \geq Kd + C(K\sqrt{Kd} + K^2)) &= \mathbb{P} \left( e^{(\gamma/K) \cdot \|\hat{X}\|^2} \geq e^{(\gamma/K) \cdot (Kd + CK\sqrt{Kd} + CK^2)} \right) \\ &\leq e^{d\gamma + O(d\gamma^2) - (\gamma/K) \cdot (Kd + CK\sqrt{Kd} + CK^2)} \\ &\leq e^{-C \cdot |\gamma| \sqrt{Kd} - C \cdot |\gamma| \cdot K + O(d\gamma^2)} \end{aligned}$$

Likewise, if we set  $-0.1 < \gamma < 0$ , this means that for any  $C \geq 0$ ,

$$\begin{aligned} \mathbb{P}(\|\hat{X}\|^2 \leq Kd - C(K\sqrt{Kd} + K^2)) &= \mathbb{P} \left( e^{(\gamma/K) \cdot \|\hat{X}\|^2} \geq e^{(\gamma/K) \cdot (Kd - CK\sqrt{Kd} - CK^2)} \right) \\ &\leq e^{d\gamma + O(d\gamma^2) - (\gamma/K) \cdot (Kd - CK\sqrt{Kd} - CK^2)} \\ &\leq e^{-C \cdot |\gamma| \sqrt{Kd} - C \cdot |\gamma| \cdot K + O(d\gamma^2)}. \end{aligned}$$

In the case that  $K \leq d/100$ , we set  $\gamma = \pm \sqrt{K/d}$  to obtain that  $e^{-C \cdot |\gamma| \sqrt{Kd} - C \cdot |\gamma| \cdot K + O(d\gamma^2)} \leq e^{-C \cdot K + O(K)} \leq e^{-(C/2) \cdot K}$ , assuming  $C$  is sufficiently large. In the case that  $K \geq d/100$ , we set  $\gamma = \pm \frac{1}{10}$  to obtain that  $e^{-C \cdot |\gamma| \sqrt{Kd} - C \cdot |\gamma| \cdot K + O(d\gamma^2)} \leq e^{-C/10 \cdot K + O(d/100)} \leq e^{-C/20 \cdot K}$ , assuming  $C$  is sufficiently large. Hence, we have that

$$\mathbb{P} \left( \|\hat{X}\|^2 = Kd \pm C(K\sqrt{Kd} + K^2) \right) \geq 1 - 2e^{-C/20 \cdot K}. \quad \blacksquare$$

**Proposition 34** *Let  $X^{(1)}, \dots, X^{(K)} \sim \mathcal{N}(0, 1)$  be i.i.d. from  $\mathcal{P}(\mu)$ , where  $\mu \in [-1, 1]^d$ . Then, for any sufficiently large  $C$ ,  $\|X^{(1)} + \dots + X^{(K)} - K \cdot \mu\|^2 = (d - \|\mu\|^2)K \pm C(K\sqrt{Kd} + K^2)$  with failure probability at most  $e^{-(C/20) \cdot K}$ .*

**Proof** The proof is almost identical to that of Proposition 33. Now, note that  $X_i^{(k)}$  for each  $k \in [K]$  is i.i.d., bounded in magnitude by 1, and has mean  $\mu_i$ . Therefore, the  $i$ th coordinate of  $\hat{X} := X^{(1)} + \dots + X^{(K)} - K \cdot \mu$  has a distribution  $\sqrt{K} \cdot Z$ , where  $Z$  is captured by the distribution in Proposition 32, and  $\text{Var}(Z) = 1 - \mu_i^2$ . Therefore, for any  $\gamma \in [-0.1, 0.1]$ ,  $\mathbb{E}[e^{(\gamma/K) \cdot \hat{X}_i^2}] \leq 1 + (1 - \mu_i^2)\gamma + O(\gamma^2) = e^{(1 - \mu_i^2)\gamma + O(\gamma^2)}$ . Therefore, since each coordinate of  $\hat{X}$  is independent,

$$\mathbb{E} \left[ e^{(\gamma/K) \cdot \|\hat{X}\|^2} \right] = \prod_{i=1}^d \mathbb{E}[e^{(\gamma/K) \cdot \hat{X}_i^2}] \leq \prod_{i=1}^d e^{(1 - \mu_i^2)\gamma + O(\gamma^2)} = e^{(d - \|\mu\|^2)\gamma + O(d\gamma^2)}. \quad (4)$$

The remainder of the proof is identical, except with replacing  $Kd + C(K\sqrt{Kd} + K^2)$  with  $K(d - \|\mu\|^2) + C(K\sqrt{Kd} + K^2)$  and  $Kd - C(K\sqrt{Kd} + K^2)$  with  $K(d - \|\mu\|^2) - C(K\sqrt{Kd} + K^2)$ . ■

**Lemma 35** *Let  $X^{(1)}, \dots, X^{(K)}$  be distributed i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $\|\mu\| \leq 1$ . Then, with probability at least  $1 - O(N^{-2K})$ , we have that  $\|X^{(1)} + \dots + X^{(K)}\|^2 = Kd \pm \tilde{O}(K\sqrt{Kd} + K^2)$ .*

**Proof** By Proposition 33, if we set  $C$  as a sufficiently large multiple of  $\log N$ , then with probability at least  $1 - N^{-2K}$ ,  $\|X^{(1)} + \dots + X^{(K)} - K\mu\|^2 = Kd \pm O(\log N) \cdot (K\sqrt{Kd} + K^2)$ . In addition, since  $X^{(1)} + \dots + X^{(K)} - K\mu \sim \mathcal{N}(0, K)$ , we have that for any fixed vector  $\mu$ ,  $\langle X^{(1)} + \dots + X^{(K)} - K\mu, \mu \rangle \sim \mathcal{N}(0, K \cdot \|\mu\|^2)$ , which with probability at least  $1 - 2N^{-2K}$  does not exceed  $O(\sqrt{K \cdot \|\mu\|^2} \cdot \sqrt{K \log N}) = O(K\sqrt{\log N})$  in absolute value.

Therefore, with probability at least  $1 - O(N^{-2K})$ , we have

$$\begin{aligned} \|X^{(1)} + \dots + X^{(K)}\|^2 &= \|X^{(1)} + \dots + X^{(K)} - K\mu\|^2 + 2K\langle X^{(1)} + \dots + X^{(K)} - K\mu, \mu \rangle + K^2\|\mu\|^2 \\ &= Kd \pm \tilde{O}(K\sqrt{Kd} + K^2) \pm 2K \cdot \tilde{O}(K) \pm K^2 \\ &= Kd \pm \tilde{O}(K\sqrt{Kd} + K^2). \end{aligned}$$

■

**Lemma 36** *Let  $X^{(1)}, \dots, X^{(K)}$  be distributed i.i.d. from  $\mathcal{P}(\mu)$ , where  $\|\mu\| \leq 1$ . Then, with probability at least  $1 - O(N^{-2K})$ , we have that  $\|X^{(1)} + \dots + X^{(K)}\|^2 = Kd \pm \tilde{O}(K\sqrt{Kd} + K^2)$ .*

**Proof** By Proposition 34, if we set  $C$  as a sufficiently large multiple of  $\log N$ , then with probability at least  $1 - N^{-2K}$ ,  $\|X^{(1)} + \dots + X^{(K)} - K\mu\|^2 = K(d - \|\mu\|^2) \pm O(\log N) \cdot (K\sqrt{Kd} + K^2)$ . However, note that  $\|\mu\|^2 \leq 1$ , so  $K(d - \|\mu\|^2) = Kd - O(K)$ . In addition, for any fixed vector  $\mu$ , by Proposition 29,  $\langle X^{(1)} + \dots + X^{(K)} - K\mu, \mu \rangle$ , with probability at least  $1 - 2N^{-2K}$ , does not exceed  $O(\|\mu\| \cdot \sqrt{K} \cdot \sqrt{K \log N}) = O(K\sqrt{\log N})$ .

Therefore, with probability at least  $1 - O(N^{-2K})$ , we have

$$\begin{aligned} \|X^{(1)} + \dots + X^{(K)}\|^2 &= \|X^{(1)} + \dots + X^{(K)} - K\mu\|^2 + 2K\langle X^{(1)} + \dots + X^{(K)} - K\mu, \mu \rangle + K^2\|\mu\|^2 \\ &= Kd - O(K) \pm \tilde{O}(K\sqrt{Kd} + K^2) \pm 2K \cdot \tilde{O}(K) \pm K^2 \\ &= Kd \pm \tilde{O}(K\sqrt{Kd} + K^2). \end{aligned}$$

■

Combining these concentration bounds together, we obtain the following theorem.

**Theorem 37** *Let  $\alpha \leq \frac{1}{2}$ , and let  $\mu \in \mathbb{R}^d$  satisfy  $\|\mu\| \leq 2\alpha$ . Let  $X^{(1)}, \dots, X^{(N)}$  be drawn i.i.d. either according to  $\mathcal{N}(\mu, I)$  or according to  $\mathcal{P}(\mu)$ . Finally, let  $\mathbf{T}$  be the matrix with entries  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ . Then, with probability at least 0.99, the following all hold simultaneously.*

1. We have  $\sum_{i=1}^N \sum_{j=1}^N T_{i,j} = Nd + N^2\|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$
2. For all  $i$ ,  $T_{i,i} = d \pm \tilde{O}(\sqrt{d})$  and for all  $i \neq j$ ,  $T_{i,j} = \pm \tilde{O}(\sqrt{d})$ .

3. For all  $i$ ,  $\sum_{j=1}^N T_{i,j} = d \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$ , and for all  $j$ ,  $\sum_{i=1}^N T_{i,j} = d \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$ .
4. For all subsets  $S \subset [N]$ , if  $|S| = K$ , then  $\sum_{i,j \in S} T_{i,j} = Kd \pm \tilde{O}(K\sqrt{Kd} + K^2)$ .

**Proof** Part 1 of the theorem is immediate from Lemmas 25 and 26, if we use Chebyshev's inequality to make the failure probability 0.005. Part 2 is immediate from Lemmas 27 and 28, and since we need this to be true for all pairs  $(i, j)$ , the failure probability is at most  $O(1/N^3) \cdot N^2 = O(1/N)$ .

To see why Part 3 follows from Lemma 30, note that by symmetry, the result of Lemma 30 also holds for  $\sum_{j=1}^N T_{i,j} = \langle X^{(i)}, X^{(1)} + \dots + X^{(N)} \rangle$ . Since we need this to be true for all  $i$ , the failure probability is at most  $O(1/N^3) \cdot N = O(1/N^2)$ . In addition, since  $\mathbf{T}$  is a symmetric matrix, we have that the sum of the entries in row  $i$  and the sum of the entries in column  $i$  are the same, so we also have that for all  $j$ ,  $\sum_{i=1}^N T_{i,j} = d \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$ .

Finally, Part 4 follows from Lemmas 35 and 36, since they imply that  $\|\sum_{i \in S} X^{(i)}\|^2 = \sum_{i,j \in S} T_{i,j} = Kd \pm \tilde{O}(K\sqrt{Kd} + K^2)$  with failure probability  $O(N^{-2K})$ . Union bounding this over all subsets of size  $K$  in  $[N]$  and over all choices of  $K$  still means the failure probability is at most  $\sum_{K=1}^N O(N^{-2K}) \cdot \binom{N}{K} = \sum_{K=1}^N O(N^{-K}) = O(1/N)$ .

Hence, the overall failure probability is at most  $0.005 + O(1/N) \leq 0.01$ .  $\blacksquare$

**Proposition 38** *Let  $X \sim \mathcal{N}(0, \Sigma)$ , where  $\|\Sigma\|_2 \leq 1$ . In addition, let  $J := \text{Tr}(\Sigma)$ . Then, for any sufficiently large  $C$ ,  $\|X\|^2 = J \pm (\sqrt{Cd} + C)$  with probability at least  $1 - e^{-C/10}$ .*

**Proof** By rotating, we may assume WLOG that  $\Sigma$  is diagonal. Then, note that  $\|X\|^2 = \sum_{i=1}^d \Sigma_{ii} Z_i^2$ , where each  $Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . We apply Theorem 19, where  $n = d$ , and  $a_i = \Sigma_{ii}$  (so  $J = \sum_{i=1}^n a_i$ ). Note that  $|a|_\infty = \|\Sigma\|_2 \leq 1$  and  $|a|_2 \leq \sqrt{d}$ . Thus, we have that  $\mathbb{P}\left(\left|\|X\|^2 - J\right| \geq \sqrt{Cd} + C\right) \leq 2e^{-C/4} \leq e^{-C/10}$  if  $C$  is sufficiently large.  $\blacksquare$

**Proposition 39** *Let  $X \sim \mathcal{N}(0, \Sigma)$ , where  $\|\Sigma\|_2 \leq 1$ . Then, for any vector  $\mu$  of magnitude at most  $2\alpha$ ,  $\langle \mu, \Sigma \rangle$  has subgaussian norm at most  $O(\alpha)$ .*

**Proof** Note that  $\langle \mu, \Sigma \rangle$  is a Gaussian with variance  $\mu^T \Sigma \mu \leq \|\mu\|^2 \cdot \|\Sigma\|_2 \leq 4\alpha^2$ . So, the subgaussian norm is at most  $O(\alpha)$ .  $\blacksquare$

**Theorem 40** *Let  $\alpha \leq \frac{1}{2}$ , and let  $\mu \in \mathbb{R}^d$  satisfy  $\|\mu\| \leq 2\alpha$ . Let  $X^{(1)}, \dots, X^{(N)}$  be drawn i.i.d. either according to  $\mathcal{N}(\mu, \Sigma)$ , where  $\|\Sigma\|_2 \leq 1$  is unknown. Finally, let  $\mathbf{T}$  be the matrix with entries  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ . Then, with probability at least 0.99, the following all hold simultaneously.*

1. We have  $\sum_{i=1}^N \sum_{j=1}^N T_{i,j} = N \cdot J + N^2 \|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$
2. For all  $i$ ,  $T_{i,i} = d \pm \tilde{O}(\sqrt{d})$  and for all  $i \neq j$ ,  $T_{i,j} = \pm \tilde{O}(\sqrt{d})$ .
3. For all  $i$ ,  $\sum_{j=1}^N T_{i,j} = d \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$ , and for all  $j$ ,  $\sum_{i=1}^N T_{i,j} = J \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$ .
4. For all subsets  $S \subset [N]$ , if  $|S| = K$ , then  $\sum_{i,j \in S} T_{i,j} = KJ \pm \tilde{O}(K\sqrt{Kd} + K^2)$ .

**Proof** For part 1 of the theorem, note that  $\sum_{i=1}^N \sum_{j=1}^N T_{i,j} = \|\bar{X}\|^2$ , where  $\bar{X} = \sum_{i=1}^N X^{(i)} \sim \mathcal{N}(N \cdot \mu, N \cdot \Sigma)$ . We can write  $\bar{X} = N \cdot \mu + \sqrt{N} \cdot X$ , where  $X \sim \mathcal{N}(0, \Sigma)$ . Then,  $\|\bar{X}\|^2 = N^2 \|\mu\|^2 + 2N^{3/2} \cdot \langle \mu, X \rangle + N \cdot \|X\|^2$ . But we know that since  $\langle \mu, X \rangle$  has subgaussian norm at most  $O(\alpha)$ , so doesn't exceed  $O(\alpha)$  in absolute value with probability at least 0.999. In addition, by Proposition 38, we have that  $\|X\|^2 = J \pm O(\sqrt{d})$  with probability at least 0.999. Together, this implies that  $\|\bar{X}\|^2 = N \cdot J + N^2 \|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$ .

For part 2 of the theorem, note that  $T_{i,i}$  has distribution  $\|\mu + X\|^2$ , where  $\|\mu\| \leq 1$  and  $X \sim \mathcal{N}(0, \Sigma)$ . In addition, note that  $\|\mu + X\|^2 = \|\mu\|^2 + 2\langle \mu, X \rangle + \|X\|^2$ . By Proposition 38,  $\|X\|^2 = J \pm \tilde{O}(\sqrt{d})$  with probability at least  $1 - \frac{1}{N^3}$ , and by Proposition 39,  $\langle \mu, X \rangle = \pm \tilde{O}(1)$  with probability at least  $1 - \frac{1}{N^3}$ . Finally,  $\|\mu\|^2 \leq 1$ . Overall, this implies that with probability at least  $1 - \frac{2}{N^2}$ ,  $T_{i,i} = J \pm \tilde{O}(\sqrt{d})$  for all  $i$ .

For  $i \neq j$ , note that  $T_{i,j}$  has distribution  $\langle \mu + X, \mu + X' \rangle$ , where  $X, X' \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ . So,  $\langle \mu + X, \mu + X' \rangle = \|\mu\|^2 + \langle \mu, X \rangle + \langle \mu, X' \rangle + \langle X, X' \rangle$ . We already know that  $\|\mu\|^2, \langle \mu, X \rangle, \langle \mu, X' \rangle$  are all in the range  $\pm \tilde{O}(1)$  with probability  $1 - \frac{2}{N^3}$ . In addition,  $\langle X, X' \rangle = \frac{1}{4} [(X + X')^2 - (X - X')^2]$ , but since  $X, X'$  are independent, this means  $X + X', X - X'$  are both  $\mathcal{N}(0, 2 \cdot \Sigma)$ , which means  $\|X + X'\|^2, \|X - X'\|^2$  are both in the range  $2J \pm \tilde{O}(\sqrt{d})$  with probability at least  $1 - \frac{2}{N^3}$ . Overall, we have that for all  $i \neq j$ ,  $T_{i,j} = \langle \mu + X, \mu + X' \rangle = \pm \tilde{O}(\sqrt{d})$  with probability at least  $1 - O(\frac{1}{N})$ .

For part 3 of the theorem, by symmetry it suffices to show that  $\sum_{j=1}^N T_{i,j} = d \pm \tilde{O}(\sqrt{Nd} + \alpha \cdot N)$  for any fixed  $i$  with probability at least  $1 - O(\frac{1}{N^2})$ . Note that  $\sum_{j \neq i} T_{i,j} = \langle X^{(i)}, \sum_{j \neq i} X^{(j)} \rangle$  and  $\sum_{j \neq i} X^{(j)}$  has distribution  $\mathcal{N}((N-1) \cdot \mu, (N-1) \cdot \Sigma)$ . Hence,  $\sum_{j \neq i} T_{i,j}$  has distribution  $\langle \mu + X, (N-1)\mu + \sqrt{N-1}X' \rangle$ , where  $X, X' \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ . We can write this as  $(N-1)\|\mu\|^2 + (N-1)\langle \mu, X \rangle + \sqrt{N-1}\langle \mu, X' \rangle + \sqrt{N-1} \cdot \langle X, X' \rangle$ . We know that  $(N-1)\|\mu\|^2 = O(N \cdot \alpha^2)$ . Also, we already saw that with probability at least  $1 - O(\frac{1}{N^2})$ ,  $\langle \mu, X \rangle, \langle \mu, X' \rangle = \pm \tilde{O}(\alpha)$ , and  $\langle X, X' \rangle = \tilde{O}(\sqrt{d})$ . So overall, with failure probability at most  $O(\frac{1}{N^2})$ ,  $\sum_{j \neq i} T_{i,j} = \pm \tilde{O}(\alpha \cdot N + \sqrt{Nd})$ . Finally, since  $T_{i,i} = J \pm \tilde{O}(\sqrt{d})$  with probability at least  $1 - O(\frac{1}{N^3})$ , we have that  $\sum_{j=1}^N T_{i,j} = J \pm \tilde{O}(\alpha \cdot N + \sqrt{Nd})$  with the desired failure probability.

For part 4, note that for any fixed subset  $S \subset [N]$  of size  $K$ ,  $\sum_{i,j \in S} T_{i,j} = \|\sum_{i \in S} X^{(i)}\|^2$ , and  $\sum_{i \in S} X^{(i)} = K \cdot \mu + \sqrt{K} \cdot X$ , for  $X \sim \mathcal{N}(0, \Sigma)$ . So,  $\sum_{i,j \in S} T_{i,j}$  has distribution  $K^2 \|\mu\|^2 + 2K\sqrt{K}\langle X, \mu \rangle + K \cdot \|X\|^2$ . Note that  $K^2 \|\mu\|^2 \leq K^2$  always, and since  $\langle X, \mu \rangle$  has subgaussian norm at most  $O(1)$  by Proposition 39, we have that  $2K\sqrt{K}\langle X, \mu \rangle = \pm \tilde{O}(K^2)$  with probability at least  $1 - N^{-2K}$ . Finally, by Proposition 38, we have that  $K \cdot \|X\|^2 = K \cdot [J \pm \tilde{O}(\sqrt{Kd} + K)]$  with probability at least  $1 - N^{-2K}$ . Hence, with probability at least  $1 - 2N^{-2K}$ , we have that  $\sum_{i,j \in S} T_{i,j} = K \cdot J \pm \tilde{O}(K\sqrt{Kd} + K^2)$ . Taking a union bound over all  $1 \leq K \leq N$  and all  $\binom{N}{K} \leq N^K$  sets  $S$  of size  $K$ , we have that this holds for all  $S$  with probability at least  $1 - O(\frac{1}{N})$ . ■

Finally, we also prove a result relating the Normal distribution and the  $\chi_d$  distribution, as well as a result relating Normal distributions with different scalings.

**Proposition 41** *As  $d \rightarrow \infty$ , we have that  $\lim_{d \rightarrow \infty} d_{TV} \left( \chi_d, \mathcal{N}(\sqrt{d}, \frac{1}{2}) \right) = 0$ .*

**Proof** It is well known that  $\chi_d$  has PDF

$$\begin{aligned} \frac{1}{2^{(d/2)-1}\Gamma(d/2)} \cdot x^{d-1}e^{-x^2/2} &= \left(1 \pm O\left(\frac{1}{d}\right)\right) \cdot \frac{1}{2^{(d/2)-1} \cdot \left(\frac{d}{2e}\right)^{d/2} \cdot \sqrt{\frac{2\pi}{d/2}}} \cdot x^{d-1}e^{-x^2/2} \\ &= \left(1 \pm O\left(\frac{1}{d}\right)\right) \cdot \frac{e^{d/2}}{d^{(d-1)/2} \cdot \sqrt{\pi}} \cdot x^{d-1}e^{-x^2/2}, \end{aligned}$$

where the first equality follows by Stirling's formula and the second equality follows by rearrangement. For  $x = \sqrt{d} + \theta$  for any  $\theta$  bounded by a constant, this equals

$$\begin{aligned} &\left(1 \pm O\left(\frac{1}{d}\right)\right) \cdot \frac{e^{d/2}}{d^{(d-1)/2} \cdot \sqrt{\pi}} \cdot \sqrt{d}^{d-1} \left(1 + \frac{\theta}{\sqrt{d}}\right)^{d-1} \cdot e^{-(\sqrt{d}+\theta)^2/2} \\ &= \left(1 \pm O\left(\frac{1}{d}\right)\right) \cdot \frac{e^{d/2}}{\sqrt{\pi}} \cdot e^{(\theta/\sqrt{d}-\theta^2/2d) \cdot d \pm O(1/\sqrt{d})} \cdot e^{-(\sqrt{d}+\theta)^2/2} \\ &= \left(1 \pm O\left(\frac{1}{\sqrt{d}}\right)\right) \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-\theta^2}, \end{aligned}$$

where the first equality follows by setting  $x = \sqrt{d} + \theta$ , the second equality follows using a Taylor expansion for  $\ln(1+x)$  for  $x = \frac{\theta}{\sqrt{d}}$  and noting that  $\theta = \pm O(1)$ , and the third equality follows by rearrangement.

Hence, for any fixed constant  $C$ , the PDF of  $\chi_d$  at  $\sqrt{d}+\theta$ , for any  $|\theta| \leq C$ , equals  $\left(1 \pm O\left(\frac{1}{\sqrt{d}}\right)\right) \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-\theta^2}$ . For any two distributions  $\mathcal{F}, \mathcal{G}$  over with PDFs  $f(x), g(x)$ , it is well known that  $d_{\text{TV}}(\mathcal{F}, \mathcal{G}) \leq \int_{-\infty}^{\infty} \max(f(x) - g(x), 0)$ . Hence, by setting  $f(x)$  as the PDF of  $\mathcal{N}(0, 1/2)$  and  $g(x)$  as the PDF of  $\chi_d$ , we have that

$$d_{\text{TV}}\left(\mathcal{N}\left(0, \frac{1}{2}\right), \chi_d\right) \leq O\left(\frac{1}{\sqrt{d}}\right) \cdot \int_{-C}^C \frac{1}{\sqrt{\pi}} e^{-\theta^2} + \int_{x \notin [-C, C]} f(x) = O\left(\frac{1}{\sqrt{d}} + e^{-C^2}\right),$$

which means that  $\limsup_{d \rightarrow \infty} d_{\text{TV}}\left(\chi_d, \mathcal{N}\left(\sqrt{d}, \frac{1}{2}\right)\right) \leq O(e^{-C^2})$ . As this is true for all  $C$ , the result is immediate. ■

**Proposition 42** *We have that  $d_{\text{TV}}\left(\mathcal{N}\left(0, \frac{1}{2}\right), \mathcal{N}(0, 1)\right) \leq \frac{1}{6}$ .*

**Proof** We can compute the total variation distance as

$$\frac{1}{2} \cdot \int_{-\infty}^{\infty} \left| \frac{1}{\sqrt{\pi}} \cdot e^{-x^2} - \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \right| dx \leq 0.1661,$$

which is at most  $\frac{1}{6}$ . ■

## Appendix C. Proof of Theorem 5

In this section, we prove Theorem 5, which establishes a (tight) sample complexity upper bound of  $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right)$  for identity testing of Gaussians with known covariance. We also conclude with a brief section describing the necessity of changing the construction used by [Canonne et al. \(2020b\)](#) in order to make the construction fully correct.

First, we note by Proposition 23, we may assume that the null hypothesis  $\mathcal{H}_0$  is that we are given  $N$  samples i.i.d. from  $\mathcal{N}(0, I)$ , and the alternative hypothesis  $\mathcal{H}_1$  is that we are given  $N$  samples i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $\alpha \leq \|\mu\| \leq 2\alpha$ .

The major tool we use in developing our sample-optimal but inefficient upper bound is the McShane-Whitney Extension theorem, which we now state.

**Theorem 43** (*McShane, 1934*) *Let  $\mathcal{X}$  be any metric space with a metric  $\rho$ , and let  $\mathcal{C} \subset \mathcal{X}$ . Suppose there is a function  $F : \mathcal{C} \rightarrow \mathbb{R}$  and a positive real number  $D$  such that  $F$  is  $D$ -Lipschitz on  $\mathcal{C}$ , i.e., for all  $x, y \in \mathcal{C}$ ,  $|F(x) - F(y)| \leq D \cdot \rho(x, y)$ . Then, there exists an extension  $\hat{F} : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\hat{F}(x) = F(x)$  for all  $x \in \mathcal{C}$ , and  $\hat{F}$  is  $D$ -Lipschitz on all of  $\mathcal{X}$ , meaning that for all  $x, y \in \mathcal{X}$ ,  $|\hat{F}(x) - \hat{F}(y)| \leq D \cdot \rho(x, y)$ .*

We will apply Theorem 43 with  $\mathcal{X} = (\mathbb{R}^d)^N$ , i.e.,  $\mathcal{X}$  is the set of all datasets of size  $N$  in  $\mathbb{R}^d$ . The distance metric  $\rho$  that we use measures the number of differing rows, i.e., for  $\mathbf{X} = \{X^{(1)}, \dots, X^{(N)}\}$  and  $\mathbf{X}' = \{X'^{(1)}, \dots, X'^{(N)}\}$ , we define  $\rho(\mathbf{X}, \mathbf{X}') = |\{i : X^{(i)} \neq X'^{(i)}\}|$ . For a dataset  $\mathbf{X} \in (\mathbb{R}^d)^N$ , we associate with it the matrix  $\mathbf{T} \in \mathbb{R}^{N \times N}$  with entries  $T_{i,j} := \langle X^{(i)}, X^{(j)} \rangle$ . We will define  $\mathcal{C} := \mathcal{C}_\alpha \subset (\mathbb{R}^d)^N$  to be the set of datasets  $\{X^{(1)}, \dots, X^{(N)}\}$  that satisfy Properties 2, 3, and 4 listed in Theorem 37. More precisely, we choose some sufficiently large  $L = \text{poly}(\log N)$ , and define  $\mathcal{C}$  to be the set of all datasets  $\mathbf{X} \in (\mathbb{R}^d)^N$  such that

- For all  $i \in [N]$ ,  $T_{i,i} = d \pm L \cdot \sqrt{d}$ , and for all  $i \neq j \in [N]$ ,  $T_{i,j} = \pm L \cdot \sqrt{d}$ .
- For all  $i$ ,  $\sum_{j=1}^N T_{i,j} = d \pm L(\sqrt{Nd} + \alpha N)$ , and for all  $j$ ,  $\sum_{i=1}^N T_{i,j} = d \pm L(\sqrt{Nd} + \alpha N)$ .
- For all subsets  $S \subset [N]$ , if  $|S| = K$ , then  $\sum_{i,j \in S} T_{i,j} = K \cdot \left[ d \pm L \cdot (\sqrt{Kd} + K) \right]$ .

We prove the following lemma.

**Lemma 44** *Let  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$  be datasets that differ in  $K \leq \frac{C}{\varepsilon}$  rows, for some fixed  $C$ . Let  $\bar{X} = \sum_{i=1}^N X^{(i)}$ , and  $\bar{X}' = \sum_{i=1}^N X'^{(i)}$ . Then,*

$$\|\bar{X}\|^2 - \|\bar{X}'\|^2 \leq 6K \cdot L \cdot \left( \sqrt{Nd} + \alpha N + \frac{C}{\varepsilon} \right).$$

**Proof** Write  $\bar{X} = Y + Z$  and  $\bar{X}' = Y + Z'$ , where  $Y$  is the sum of the identical rows in  $\mathbf{X}$  and  $\mathbf{X}'$ ,  $Z$  is the sum of the rows in  $\mathbf{X}$  that are not the same in  $\mathbf{X}'$ , and  $Z'$  is the sum of the rows in  $\mathbf{X}'$  that are not the same in  $\mathbf{X}$ . Let  $\mathbf{T}$  be the matrix with  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ , and  $\mathbf{T}'$  be the matrix with  $T'_{i,j} = \langle X'^{(i)}, X'^{(j)} \rangle$ . Let  $S \subset [N]$  represent the indices of the rows in  $\mathbf{X}$  and  $\mathbf{X}'$  that differ. Then, note that  $\langle Z, \bar{X} \rangle$  equals the sum of the entries  $T_{i,j}$  with  $i \in S, j \in [N]$  (and similarly  $\langle Z', \bar{X}' \rangle$  is the sum of the entries  $T'_{i,j}$  with  $i \in S, j \in [N]$ ). Also,  $\|Z\|^2$  is the sum of the entries  $T_{i,j}$  with  $i, j \in S$  (and similarly  $\|Z'\|^2$  is the sum of the entries  $T'_{i,j}$  with  $i \in S$ ).

Next, note that  $\|\bar{X}\|^2 - \|\bar{X}'\|^2 = 2\langle Z, \bar{X} \rangle - 2\langle Z', \bar{X}' \rangle - (\|Z\|^2 - \|Z'\|^2)$ . Then, note that  $\langle Z, \bar{X} \rangle = K \cdot (d \pm L(\sqrt{Nd} + \alpha \cdot N))$ , since it is the sum of the entries of  $K$  rows in  $T$ . Likewise,  $\langle Z', \bar{X}' \rangle = K \cdot (d \pm L(\sqrt{Nd} + \alpha \cdot N))$ . In addition, since  $\|Z\|^2$  is the sum of the entries  $T_{i,j}$  with  $i, j \in S$ , and since  $|S| = K$ , we have that  $\|Z\|^2 = Kd \pm L(K\sqrt{Kd} + K^2)$ . Similarly, we also have  $\|Z'\|^2 = Kd \pm L(K\sqrt{Kd} + K^2)$ .

Therefore,

$$\begin{aligned} \|\bar{X}\|^2 - \|\bar{X}'\|^2 &= 2\langle Z, \bar{X} \rangle - 2\langle Z', \bar{X}' \rangle - (\|Z\|^2 - \|Z'\|^2) \\ &= \pm K \cdot 4L \cdot (\sqrt{Nd} + \alpha N) \pm 2L \cdot K \cdot (\sqrt{Kd} + K) \\ &= \pm 6KL \cdot \left( \sqrt{Nd} + \alpha N + \frac{C}{\varepsilon} \right), \end{aligned}$$

where the last line follows because  $K \leq \frac{C}{\varepsilon}$  and  $K \leq N$ .  $\blacksquare$

Let  $C$  be a sufficiently large constant, and let  $\Delta := 6L \left( \sqrt{Nd} + \alpha N + \frac{C}{\varepsilon} \right)$ . Define  $D := \frac{\varepsilon}{C} \cdot \alpha^2 \cdot N^2$ , and suppose that  $\Delta \leq D$ . We now consider the statistic function  $\tilde{T} : \mathcal{C} \rightarrow \mathbb{R}$  that sends a dataset  $\mathbf{X}$  to  $\max(0, \min(\|\bar{X}\|^2 - Nd, \alpha^2 \cdot N^2))$ , where  $\bar{X} = X^{(1)} + \dots + X^{(N)}$ . Note that  $\tilde{T}(\mathbf{X})$  is simply  $\|\bar{X}\|^2 - Nd$  but clipped by 0 and  $\alpha^2 N^2$ .

We first show that  $\tilde{T} : \mathcal{C} \rightarrow \mathbb{R}$  is  $D$ -Lipschitz, where we recall that the distance metric  $\rho$  over  $\mathcal{C}$  is the number of differing rows.

**Lemma 45** *Assume that  $\Delta \leq D$ . Then, for any integer  $K$  and any datasets  $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$  that differ in  $K$  rows,  $|\tilde{T}(\mathbf{X}) - \tilde{T}(\mathbf{X}')| \leq K \cdot D$ .*

**Proof** First, assume that  $K \geq C \cdot \varepsilon^{-1}$ , where we recall that  $C$  is a large constant and that  $D = \frac{\varepsilon}{C} \cdot \alpha^2 \cdot N^2$ . Then, since  $\tilde{T}(\mathbf{X})$  and  $\tilde{T}(\mathbf{X}')$  are bounded below by 0 and bounded above by  $\alpha^2 \cdot N^2 = \varepsilon^{-1} \cdot C \cdot D \leq K \cdot D$ , we indeed have that  $|\tilde{T}(\mathbf{X}) - \tilde{T}(\mathbf{X}')| \leq K \cdot D$ .

Next, assume that  $K \leq C \cdot \varepsilon^{-1}$ . Then, by Lemma 44, we have that  $|\|\bar{X}\|^2 - \|\bar{X}'\|^2| \leq K \cdot \Delta$ . Since  $\tilde{T}(\mathbf{X})$  simply takes  $\|\bar{X}\|^2$ , subtracts  $Nd$ , and then prevents the statistic from being less than 0 or more than  $\alpha^2 \cdot N^2$  (and similar for  $\tilde{T}(\mathbf{X}')$ ), we will also have that  $|\tilde{T}(\mathbf{X}) - \tilde{T}(\mathbf{X}')| \leq K \cdot \Delta$ . Since we are assuming that  $\Delta \leq D$ , this also implies that  $|\tilde{T}(\mathbf{X}) - \tilde{T}(\mathbf{X}')| \leq K \cdot D$ .  $\blacksquare$

Hence, by Theorem 43, there exists an extension  $\hat{T}$  of  $\tilde{T}$  from  $(\mathbb{R}^d)^N \rightarrow \mathbb{R}$ , such that for any datasets  $\mathbf{X}, \mathbf{X}' \in (\mathbb{R}^d)^N$  that differ in  $K$  rows,  $|\hat{T}(\mathbf{X}) - \hat{T}(\mathbf{X}')| \leq K \cdot D$ . Our final (computationally inefficient) algorithm will be the following. We compute  $\hat{T}(\mathbf{X}) + \text{Lap}(D/\varepsilon)$ . If this exceeds  $\frac{\alpha^2 \cdot N^2}{2}$ , we output 1, and otherwise we output 0. (Note that this algorithm is computationally inefficient because computing the extension  $\hat{T}$  is computationally inefficient.)

We are now ready to prove Theorem 5.

**Proof** of Theorem 5: First, we note that this algorithm is  $(\varepsilon, 0)$ -differentially private because  $\hat{T}(\mathbf{X})$  has sensitivity at most  $D$  whenever we change a single row of  $\mathbf{X}$ , by the properties of our Lipschitz extension.

To verify accuracy, note that if each row of  $\mathbf{X}$  is drawn i.i.d. from  $\mathcal{N}(0, I)$ , then with probability at least 0.99,  $\mathbf{X} \in \mathcal{C}$  and  $\|\bar{X}\|^2 \leq Nd \pm CN\sqrt{d}$  by Theorem 37, assuming  $C$  is sufficiently large. Hence, with probability at least 0.99,  $\hat{T}(\mathbf{X}) = \tilde{T}(\mathbf{X}) \leq CN\sqrt{d}$ , which means that with probability at least 0.9,  $\hat{T}(\mathbf{X}) + \text{Lap}(D/\varepsilon) \leq CN\sqrt{d} + 10 \cdot \frac{D}{\varepsilon}$ , which we want to be at most  $\frac{\alpha^2 \cdot N^2}{2}$ . Indeed,



if  $C$  is sufficiently large, then  $10 \cdot \frac{D}{\varepsilon} \leq \frac{1}{4} \cdot \alpha^2 N^2$ , so we just need that  $\Delta \leq D = \frac{\varepsilon}{C} \cdot \alpha^2 \cdot N^2$  and  $CN\sqrt{d} \leq \frac{1}{4} \cdot \alpha^2 \cdot N^2$ .

Next, if each row of  $\mathbf{X}$  is drawn i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $\alpha \leq \|\mu\| \leq 2\alpha$ , then with probability at least 0.99,  $\mathbf{X} \in \mathcal{C}$  and  $\|\bar{X}\|^2 \geq Nd + N^2\alpha^2 - C(N\sqrt{d} + \alpha N\sqrt{N})$  by Theorem 37. Hence, with probability at least 0.99,  $\hat{T}(\mathbf{X}) = \tilde{T}(\mathbf{X}) \geq N^2\alpha^2 - C(N\sqrt{d} + \alpha N\sqrt{N})$ , which means that with probability at least 0.9,  $\hat{T}(\mathbf{X}) + \text{Lap}(D/\varepsilon) \geq N^2\alpha^2 - C(N\sqrt{d} + \alpha N\sqrt{N}) - 10 \cdot \frac{D}{\varepsilon}$ , which we want to be at least  $\frac{\alpha^2 \cdot N^2}{2}$ . If  $C$  is sufficiently large, then  $10 \cdot \frac{D}{\varepsilon} \leq \frac{1}{4} \cdot \alpha^2 N^2$ , so we just need that  $\Delta \leq D$  and that  $C(N\sqrt{d} + \alpha N\sqrt{N}) \leq \frac{1}{4} \cdot \alpha^2 \cdot N^2$ .

Hence, our algorithm is both accurate and private as long as  $N$  is sufficiently large so that  $\Delta = 6L \left( \sqrt{Nd} + \alpha N + \frac{C}{\varepsilon} \right) \leq \frac{\varepsilon}{C} \cdot \alpha^2 \cdot N^2$  and  $C(N\sqrt{d} + \alpha N\sqrt{N}) \leq \frac{1}{4} \alpha^2 \cdot N^2$ . Since  $C$  is a fixed large constant and  $L$  is polylogarithmic in  $N$ , it suffices that

$$N \geq \tilde{O} \left( \frac{\sqrt{d}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon} \right).$$

This completes the proof. ■

### C.1. On changing the construction of [Canonne et al. \(2020b\)](#)

An important part of both our argument and the argument of [Canonne et al. \(2020b\)](#) for the inefficient testing algorithm was that the statistic  $F$  on the subset  $\mathcal{C}$  is  $D$ -Lipschitz. The previous work of [Canonne et al. \(2020b\)](#) made the subtle mistake of assuming that it suffices for any two adjacent datasets  $\mathcal{X}, \mathcal{X}'$  in  $\mathcal{C}$  to satisfy  $|F(\mathcal{X}) - F(\mathcal{X}')| \leq D$ . In reality, we need that for any datasets  $\mathcal{X}, \mathcal{X}' \in \mathcal{C}$  that differ in  $K$  positions (for any  $K \leq N$ ),  $|F(\mathcal{X}) - F(\mathcal{X}')| \leq D \cdot K$ . It is reasonable to believe these definitions are equivalent: indeed, if  $\mathcal{X}, \mathcal{X}' \in \mathcal{C}$  differ by  $K$  positions, then we can construct a series of adjacent datasets  $\mathcal{X} =: \mathcal{X}^{(0)}, \mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \dots, \mathcal{X}^{(N)} := \mathcal{X}'$  where  $\mathcal{X}^{(i-1)}$  and  $\mathcal{X}^{(i)}$  are adjacent. A natural guess is that if  $\mathcal{X}, \mathcal{X}' \in \mathcal{C}$ , then one can ensure that each  $\mathcal{X}^{(i)}$  is also in  $\mathcal{C}$ . If so, then a triangle inequality implies the statistic  $F$  is  $D$ -Lipschitz on  $\mathcal{C}$ , as long as adjacent datasets in  $\mathcal{C}$  have the statistic differ by at most  $D$ .

Unfortunately, it may be impossible to make the intermediate datasets  $\mathcal{X}^{(i)}$  be in  $\mathcal{C}$  for the choice of  $\mathcal{C}$  used in [Canonne et al. \(2020b\)](#). Indeed, the choice of  $\mathcal{C}$  used in [Canonne et al. \(2020b\)](#) was the set of data  $\mathbf{X} = \{X^{(1)}, \dots, X^{(N)}\}$  (where each  $X^{(i)} \in \mathbb{R}^d$ ) such that  $\|X^{(i)}\|^2 \leq \Delta$  and  $|\langle X^{(i)}, X^{(1)} + \dots + X^{(N)} \rangle| \leq \Delta$  for some appropriate choice of  $\Delta$ . However, even if  $N = 2$ ,  $d = 3$ , and  $K = 2$ , we lose our desired property. For instance, define  $a = (1, 0, 1)$ ,  $b = (-1, 0, 1)$ ,  $c = (0, 1, 1)$ , and  $d = (0, -1, 1)$  to be points in  $\mathbb{R}^d$ . Let  $\mathcal{C}$  be the set of pairs  $\{X^{(1)}, X^{(2)}\}$  such that  $\|X^{(i)}\|^2 \leq 2$  and  $|\langle X^{(i)}, X^{(1)} + X^{(2)} \rangle| \leq 2$ , for both  $i = 1$  and  $i = 2$ . It is clear that the sets  $\{a, b\}$  and  $\{c, d\}$  are in  $\mathcal{C}$ , and they differ in  $K = 2$  positions. However, any method of 2 steps used to convert  $\{a, b\}$  to  $\{c, d\}$  cannot stay in  $\mathcal{C}$  throughout, because one can see that none of  $\{a, c\}$ ,  $\{a, d\}$ ,  $\{b, c\}$ , or  $\{b, d\}$  are in  $\mathcal{C}$ .

## Appendix D. Proof of Theorem 6

In this section, we prove Theorem 6. In other words, we show that any  $(0, \varepsilon)$ -differentially private algorithm that can successfully distinguish between samples from  $\mathcal{N}(0, I)$  and  $\mathcal{N}(\mu, I)$  for  $\|\mu\| \geq \alpha$

requires at least  $\Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3}\cdot\varepsilon^{2/3}} + \frac{1}{\alpha\cdot\varepsilon}\right)$  samples. We reiterate that in the case of differentially private *hypothesis testing*, the sample complexity required for  $(\varepsilon, 0)$ -DP and  $(0, \varepsilon)$ -DP are known to be *asymptotically equivalent* for any  $\varepsilon < \frac{1}{2}$  (Acharya et al., 2018b).

First, we note that even a non-private algorithm that can successfully distinguish between  $N$  samples from  $\mathcal{N}(0, I)$  and  $\mathcal{N}(\mu, I)$  where  $\|\mu\| \geq \alpha$  requires  $N \geq \Omega\left(\frac{d^{1/2}}{\alpha^2}\right)$  (Srivastava and Du, 2008; Canonne et al., 2020a). Hence, it suffices to show that a  $(0, \varepsilon)$ -private algorithm that can successfully distinguish between  $N$  samples from  $\mathcal{N}(0, I)$  and  $\mathcal{N}(\mu, I)$  where  $\|\mu\| \geq \alpha$  requires  $N \geq \Omega\left(\frac{d^{1/3}}{\alpha^{4/3}\cdot\varepsilon^{2/3}}\right)$  and  $N \geq \Omega\left(\frac{1}{\alpha\cdot\varepsilon}\right)$ . We remark that in the related scenario of privately testing a binary discrete distribution, the  $N \geq \Omega\left(\frac{1}{\alpha\cdot\varepsilon}\right)$  has already been established (Acharya et al., 2018b). The proof in the Multivariate Gaussian case will not be significantly different (in fact, this part of the lower bound holds even for a univariate Gaussian), but we include the proof for completeness.

First, we will generalize our previous notion of privately distinguishing between two hypotheses of i.i.d. samples to privately distinguishing between hypotheses of possibly non-i.i.d. samples. Now, we will consider the problem of privately distinguishing between two distributions  $\mathcal{U}$  and  $\mathcal{V}$  over  $\mathcal{X}^N$ . Here,  $\mathcal{U}$  and  $\mathcal{V}$  need not be i.i.d. samples. For an algorithm  $\mathcal{A} : \mathcal{X}^N \rightarrow \{0, 1\}$  to  $(\varepsilon, \delta)$ -privately distinguish between two distributions  $\mathcal{U}$  and  $\mathcal{V}$  over  $\mathcal{X}^N$  (in our case  $\mathcal{X} = \mathbb{R}^d$ ), it must hold that:

- $\mathcal{A}$  is  $(\varepsilon, \delta)$ -differentially private.
- If  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)}) \sim \mathcal{U}$ , then  $\mathbb{P}(\mathcal{A}(\mathbf{X}) = 0) \geq \frac{2}{3}$ .
- If  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)}) \sim \mathcal{V}$ , then  $\mathbb{P}(\mathcal{A}(\mathbf{X}) = 1) \geq \frac{2}{3}$ .

To prove our lower bound on the number of samples  $N$ , it suffices to prove a lower bound for distinguishing between  $\mathcal{U}$  and  $\mathcal{V}$  where  $\mathcal{U}$  is the distribution of  $N$  i.i.d. samples from  $\mathcal{N}(0, I)$ , and  $\mathcal{V}$  is generated by first taking some distribution  $\mathcal{D}$  supported only on points  $\mu \in \mathbb{R}^d$  with  $\|\mu\|_2 \geq \alpha$ , and then sampling  $X^{(1)}, \dots, X^{(N)} \sim \mathcal{N}(\mu, I)$ , where  $\mu \sim \mathcal{D}$ . To see why, suppose there existed an algorithm  $\mathcal{A}$  that could privately distinguish between  $\mathcal{H}_0$  of all samples being i.i.d.  $\mathcal{N}(0, I)$  and  $\mathcal{H}_1$  of all samples being i.i.d.  $\mathcal{N}(\mu, I)$  for any  $\|\mu\| \geq \alpha$ . Then,  $\mathcal{A}$  would be  $(\varepsilon, \delta)$ -DP, and if  $\mathbf{X} \sim \mathcal{U}$ , then  $\mathbb{P}(\mathcal{A}(\mathbf{X}) = 0) \geq \frac{2}{3}$ . Finally, if  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)}) \sim \mathcal{V}$ , then  $\mathbb{P}(\mathcal{A}(\mathbf{X}) = 1) \geq \frac{2}{3}$ . This is true because we can condition on the choice of  $\mu \sim \mathcal{D}$ , and we always have that  $\mathbb{P}(\mathcal{A}(\mathbf{X}) = 1) \geq \frac{2}{3}$  regardless of what choice of  $\mu$  we picked.<sup>4</sup>

Next, we note the following theorem which will be crucial in establishing our lower bounds.

**Theorem 46** (Acharya et al., 2018b, Theorem 11, rephrased) *Let  $\mathcal{X}$  represent the domain that samples are coming from, and assume  $\mathcal{U}$  and  $\mathcal{V}$  be distributions over  $\mathcal{X}^N$  (which may not necessarily be i.i.d. samples), such that there is a coupling between  $\mathcal{U}, \mathcal{V}$  such that  $\mathbb{E}_{\mathbf{X} \sim \mathcal{U}, \mathbf{X}' \sim \mathcal{V}}[\rho(\mathbf{X}, \mathbf{X}')] \leq D$ . Then, if there existed an algorithm that could  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{H}_0 : \mathbf{X} \sim \mathcal{U}$  and  $\mathcal{H}_1 : \mathbf{X} \sim \mathcal{V}$ , then  $D \geq c_1/\varepsilon$  for some constant  $c_1 > 0$ .*

As a corollary, we have the following result.

**Corollary 47** *Again, let  $\mathcal{X}$  represent the domain, and let  $\mathcal{U}, \mathcal{V}$  be distributions over  $\mathcal{X}^N$ . Suppose that there exist distributions  $\mathcal{U}', \mathcal{V}'$  such that  $d_{TV}(\mathcal{U}, \mathcal{U}'), d_{TV}(\mathcal{V}, \mathcal{V}') \leq \frac{1}{4}$ . In addition, suppose*

4. This idea of choosing a distribution for  $\mu$  is quite standard, and is similar to Yao's Minimax Principle (Yao, 1977).

there is a coupling between  $\mathcal{U}', \mathcal{V}'$  such that  $\mathbb{E}_{\mathbf{X} \sim \mathcal{U}', \mathbf{X}' \sim \mathcal{V}'}[\rho(\mathbf{X}, \mathbf{X}')] \leq D$ . Then, if there existed an algorithm that could  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{H}_0 : \mathbf{X} \sim \mathcal{U}$  and  $\mathcal{H}_1 : \mathbf{X} \sim \mathcal{V}$ , then  $D \geq c_2/\varepsilon$  for some constant  $c_2 > 0$ .

**Proof** Suppose there exists an algorithm  $\mathcal{A}$  that could  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{H}_0 : \mathbf{X} \sim \mathcal{U}$  and  $\mathcal{H}_1 : \mathbf{X} \sim \mathcal{V}$ . This means that if  $\mathbf{X} \sim \mathcal{U}$ ,  $\mathcal{A} = 0$  with probability at least  $2/3$ , and if  $\mathbf{X} \sim \mathcal{V}$ ,  $\mathcal{A} = 1$  with probability at least  $2/3$ . By repeating the algorithm  $O(1)$  times and taking the majority, we can improve the value  $2/3$  to  $11/12$ , though the algorithm is now  $(0, C\varepsilon)$ -DP for some constant  $C$ . But then, since  $d_{\text{TV}}(\mathcal{U}, \mathcal{U}') \leq 1/4$ , this means that  $|\mathbb{P}_{\mathbf{X} \sim \mathcal{U}}(\mathcal{A}(\mathbf{X}) = 0) - \mathbb{P}_{\mathbf{X} \sim \mathcal{U}'}(\mathcal{A}(\mathbf{X}) = 0)| \leq 1/4$  for any algorithm  $\mathcal{A}$ , which means that  $\mathbb{P}_{\mathbf{X} \sim \mathcal{U}'}(\mathcal{A}(\mathbf{X}) = 0) \geq \frac{11}{12} - \frac{1}{4} \geq \frac{2}{3}$ . By a symmetric argument,  $\mathbb{P}_{\mathbf{X} \sim \mathcal{V}'}(\mathcal{A}(\mathbf{X}) = 1) \geq \frac{2}{3}$ . Therefore, we can apply Theorem 46 to conclude that  $D \geq c_1/(C \cdot \varepsilon) = c_2/\varepsilon$  for  $c_2 = c_1/C$ .  $\blacksquare$

Before we prove the main part of the lower bound (i.e., that  $N \geq \Omega\left(\frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}}\right)$ ), we first show how Theorem 46 implies an  $\Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$ -lower bound. We first have the following proposition.

**Proposition 48** *Let  $\mathcal{U}$  represent the distribution of selecting  $\mathbf{X} = \{X^{(1)}, \dots, X^{(N)}\} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I)$ , and let  $\mathcal{V}$  represent the distribution of selecting  $\mathbf{X}' = \{X'^{(1)}, \dots, X'^{(N)}\} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , where  $\mu = (\alpha, 0, 0, \dots, 0) \in \mathbb{R}^d$ , and  $0 < \alpha \leq 1$ . Then, there exists a coupling between  $\mathcal{U}, \mathcal{V}$  such that  $\mathbb{E}_{\mathbf{X} \sim \mathcal{U}, \mathbf{X}' \sim \mathcal{V}}[\rho(\mathbf{X}, \mathbf{X}')] \leq O(\alpha) \cdot N$ .*

**Proof** We start by considering univariate Normal distributions. Consider drawing the PDF of  $\mathcal{N}(0, 1)$  and of  $\mathcal{N}(\alpha, 1)$ . Let  $S$  be the region of points  $(x, y) \in \mathbb{R}^2$ , where  $y$  is nonnegative but below the PDF of  $\mathcal{N}(0, 1)$  at  $x$ . Likewise, let  $S'$  be the region of points  $(x, y) \in \mathbb{R}^2$ , where  $y$  is nonnegative but below the PDF of  $\mathcal{N}(\alpha, 1)$  at  $x$ . Both  $S, S'$  have area 1 by definition of PDF, and the intersection of  $S, S'$  has area  $1 - d_{\text{TV}}(\mathcal{N}(0, 1), \mathcal{N}(\alpha, 1)) = 1 - \Theta(\alpha)$ .

Now, we consider the following method of generating samples  $X^{(1)}, \dots, X^{(N)}$  and  $X'^{(1)}, \dots, X'^{(N)}$  over  $\mathbb{R}^d$ . For each  $i$ , pick a point  $(x, y)$  uniformly at random in  $S$ , and let  $X_1^{(i)} = x$ . If  $(x, y) \in S'$ , then we define  $(x', y') = (x, y)$ . Otherwise, we choose  $(x', y')$  uniformly at random from  $S' \setminus S$ . Finally, we let  $X_1'^{(i)} = x'$ . Finally, we draw  $X_2^{(i)} = X_2'^{(i)}, \dots, X_n^{(i)} = X_n'^{(i)}$  each as an i.i.d.  $\mathcal{N}(0, 1)$ .

First, note that each  $X^{(i)}$  is independent and has the correct distribution. Also, note that each  $X'^{(i)}$  is independent. To check the distribution of  $X'^{(i)}$ , note that we select a point  $(x, y)$  in  $S \cap S'$  with probability  $1 - d_{\text{TV}}(\mathcal{N}(0, 1), \mathcal{N}(\alpha, 1)) = 1 - \Theta(\alpha)$ , which equals the area of  $S \cap S'$ . Hence, with probability equal to this area, we select a uniform point  $(x', y')$  in  $S \cap S'$ , and otherwise, we select a uniform point  $(x', y')$  in  $S' \setminus S$ , so the overall distribution of  $(x, y)$  is just uniform from  $S'$ . Hence,  $X_1'^{(i)}$  has the distribution  $\mathcal{N}(\alpha, 1)$ , which means that  $X'^{(i)}$  has the correct distribution.

The final thing we verify is that  $\mathbb{E}[\rho(\mathbf{X}, \mathbf{X}')] = O(\alpha) \cdot N$ . Indeed, note that  $X^{(i)} = X'^{(i)}$  whenever we selected  $(x, y) \in S \cap S'$ , so  $\mathbb{P}(X^{(i)} \neq X'^{(i)})$  is at most 1 minus the area of  $S \cap S'$ , which is  $O(\alpha)$ . Hence, by linearity of expectation, we have that  $\mathbb{E}[\rho(\mathbf{X}, \mathbf{X}')] = O(\alpha) \cdot N$ .  $\blacksquare$

Note that we immediately get an  $\Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$  sample complexity lower bound, even in 1 dimension. Indeed, to determine whether we have  $N$  samples from  $\mathcal{N}(0, I)$  or from  $\mathcal{N}(\mu, I)$ , by combining Proposition 48 and Theorem 46, we need that  $O(\alpha) \cdot N \geq c_1/\varepsilon$ , which means that  $N \geq \Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$ .

We now set up our main lower bound. Let  $\mathcal{D}$  be the distribution  $\mathcal{N}(0, \frac{2\alpha^2}{d} \cdot I)$  conditioned on the magnitude being at least  $\alpha$ . Note that  $\mathcal{D}$  has no support in the region  $\{x : \|x\| \leq \alpha\}$ . Let  $\mathcal{V}$  be the distribution over  $\{X^{(1)}, \dots, X^{(N)}\}$  where we first draw  $\mu \sim \mathcal{D}$  and then draw each  $X^{(i)}$  from  $\mathcal{N}(\mu, I)$ . Also, let  $\mathcal{U}$  be the distribution over  $\{X^{(1)}, \dots, X^{(N)}\}$  where we draw each  $X^{(i)}$  from  $\mathcal{N}(0, I)$ . Our goal will be to show the following: unless  $N \geq \Omega\left(\frac{d^{1/3}}{\varepsilon^{2/3} \cdot \alpha^{4/3}}\right)$ , there exist distributions  $\mathcal{U}', \mathcal{V}'$  over  $(\mathbb{R}^d)^N$  such that  $d_{\text{TV}}(\mathcal{U}, \mathcal{U}'), d_{\text{TV}}(\mathcal{V}, \mathcal{V}') \leq \frac{1}{4}$ , and a coupling of  $\mathcal{U}', \mathcal{V}'$  such that  $\mathbb{E}_{\mathbf{X} \sim \mathcal{U}', \mathbf{X}' \sim \mathcal{V}'}[\rho(\mathbf{X}, \mathbf{X}')] < c_2/\varepsilon$ .

We will also need the following two basic propositions.

**Proposition 49** *Let  $X^{(1)}, \dots, X^{(N)} \sim \mathcal{N}(0, I)$  and  $\bar{X} = \frac{X^{(1)} + \dots + X^{(N)}}{N}$ . Then,  $\|\bar{X}\| \sim \frac{1}{\sqrt{N}} \cdot \chi_d$ , where  $\chi_d$  is the square root of a chi-square variable with  $d$  degrees of freedom.*

**Proof** This is immediate from the fact that  $\bar{X} \sim \mathcal{N}(0, \frac{1}{N} \cdot I) = \frac{1}{\sqrt{N}} \cdot \mathcal{N}(0, I)$ . Thus,  $\|\bar{X}\|^2 = \frac{1}{N} \cdot \sum_{i=1}^d Z_i^2$ , where each  $Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ , which completes the proof.  $\blacksquare$

**Proposition 50** *Let  $X^{(1)}, \dots, X^{(N)} \sim \mathcal{N}(\mu, I)$ , where  $\mu \sim \mathcal{N}(0, \frac{2\alpha^2}{d} I)$ , and let  $\bar{X} = \frac{X^{(1)} + \dots + X^{(N)}}{N}$ . Then,  $\|\bar{X}\| \sim \sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d$ .*

**Proof** This is immediate from the fact that  $\bar{X} = \mu + \mathcal{N}(0, \frac{1}{N} \cdot I) = \mathcal{N}(0, \frac{2\alpha^2}{d} + \frac{1}{N})$ .  $\blacksquare$

We now prove the following result.

**Theorem 51** *Let  $\mathcal{D}, \mathcal{U}, \mathcal{V}$  be as defined previously. Suppose that  $d \geq C$  and  $\frac{d}{\alpha^2} \geq C \cdot N$  for some sufficiently large constant  $C$ . Then, unless  $N \geq \Omega\left(\frac{d^{1/3}}{\varepsilon^{2/3} \cdot \alpha^{4/3}}\right)$ , there exist distributions  $\mathcal{U}', \mathcal{V}'$  over  $(\mathbb{R}^d)^N$  such that  $d_{\text{TV}}(\mathcal{U}, \mathcal{U}') \leq 0.25$ ,  $d_{\text{TV}}(\mathcal{V}, \mathcal{V}') \leq 0.25$ , along with a coupling of  $\mathcal{U}', \mathcal{V}'$  such that  $\mathbb{E}_{\mathbf{X} \sim \mathcal{U}', \mathbf{X}' \sim \mathcal{V}'}[\rho(\mathbf{X}, \mathbf{X}')] \leq \frac{c_2}{\varepsilon}$ .*

**Proof** We first show how to sample  $X^{(1)}, \dots, X^{(N)}$  in a different manner that ends up with the same distribution as  $\mathcal{U}$ . By Corollary 18, the distribution of  $X^{(1)}, \dots, X^{(N)}$  is the same as the distribution of  $\bar{X} + Z^{(1)} - \bar{Z}, \bar{X} + Z^{(2)} - \bar{Z}, \dots, \bar{X} + Z^{(N)} - \bar{Z}$ , where  $Z^{(1)}, \dots, Z^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I)$  and  $\bar{Z} = \frac{Z^{(1)} + \dots + Z^{(N)}}{N}$ . By Proposition 49 and the rotational symmetry of Gaussians, we can sample  $\bar{X} = a \cdot v$ , where  $v$  is drawn uniformly from the unit sphere, and  $a \sim \sqrt{\frac{1}{N}} \cdot \chi_d$ . Again using the rotational symmetry of Gaussians, we can write  $Z^{(i)} = y^{(i)} \cdot v + z^{(i)}$ , where we draw each  $y^{(i)}$  i.i.d. from the univariate distribution  $\mathcal{N}(0, 1)$  and each  $z^{(i)}$  i.i.d. from the distribution  $\mathcal{N}(0, I - vv^T)$ , or equivalently, from the distribution  $\mathcal{N}(0, I)$  but projected onto the subspace orthogonal to  $v$ . Hence, if we write  $y'^{(i)} := y^{(i)} - \frac{y^{(1)} + \dots + y^{(N)}}{N}$  and  $z'^{(i)} := z^{(i)} - \frac{z^{(1)} + \dots + z^{(N)}}{N}$  for all  $i \in [N]$ , we obtain the following equivalent sampling procedure to sampling from  $\mathcal{U}$ . First, sample  $v$  uniformly from the unit sphere and  $a \sim \sqrt{\frac{1}{N}} \cdot \chi_d$ , and compute  $\bar{X} = a \cdot v$ . Next, sample each  $y'^{(i)}, z'^{(i)}$  as above, and let  $X^{(i)} = (a + y'^{(i)}) \cdot v + z'^{(i)}$  for all  $i \in [N]$ .

Next, we replace the distribution  $a \sim \sqrt{\frac{1}{N}} \cdot \chi_d$  with  $a' \sim \mathcal{N}\left(\sqrt{\frac{d}{N}}, \frac{1}{N}\right)$ . By Proposition 41, if  $d$  is sufficiently large, then  $d_{\text{TV}}(\chi_d, \mathcal{N}(\sqrt{d}, 1/2)) \leq 0.01$ , and by Proposition 42,

$d_{\text{TV}}(\mathcal{N}(\sqrt{d}, 1/2), \mathcal{N}(\sqrt{d}, 1)) \leq \frac{1}{6}$ . Therefore, if  $d$  is sufficiently large, then  $d_{\text{TV}}(\chi_d, \mathcal{N}(\sqrt{d}, 1)) \leq 0.01 + \frac{1}{6} \leq 0.2$ . By scaling, we have that  $d_{\text{TV}}\left(\sqrt{\frac{1}{N}} \cdot \chi_d, \mathcal{N}\left(\sqrt{\frac{d}{N}}, \frac{1}{N}\right)\right) \leq 0.2$ . Thus, if we let  $\mathcal{U}'$  be the distribution where we sample  $v$  uniformly from the unit sphere,  $a' \sim \mathcal{N}\left(\sqrt{\frac{d}{N}}, \frac{1}{N}\right)$ , and  $X^{(i)} = (a' + y'^{(i)}) \cdot v + z'^{(i)}$  for all  $i$  (where  $y'^{(i)}, z'^{(i)}$  are sampled as before), we have that  $d_{\text{TV}}(\mathcal{U}, \mathcal{U}') \leq 0.2$ .

Now, note that  $a'$  has the same distribution as  $\sqrt{\frac{d}{N}} + \bar{w}$ , where  $\bar{w} \sim \mathcal{N}(0, \frac{1}{N})$  is the distribution of averaging  $N$  i.i.d.  $\mathcal{N}(0, 1)$  variables. Hence, by Corollary 18, the tuple  $\{a' + y'^{(i)}\}_{i=1}^N$  has the same overall distribution as  $\{\sqrt{\frac{d}{N}} + y^{(i)}\}_{i=1}^N$ , where we recall that  $y^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . So, to generate a sample from  $\mathcal{U}'$  we can instead sample  $X^{(i)} = \left(\sqrt{\frac{d}{N}} + y^{(i)}\right) \cdot v + z'^{(i)}$  for all  $i$ , where each  $y^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ .

We similarly generate a method of sampling from  $\mathcal{V}'$  with  $d_{\text{TV}}(\mathcal{V}', \mathcal{V})$  small. First, we modify  $\mathcal{D}$  to  $\mathcal{D}_1$  by removing the conditioning on the magnitude being at least  $\alpha$ . Note that if  $\mu \sim \mathcal{D}_1$ ,  $\|\mu\|^2 \sim \frac{2\alpha^2}{d} \cdot \chi_d^2$ . Thus, basic concentration of  $\chi_d^2$  distributions, if  $d$  is sufficiently large then  $\|\mu\|^2 \geq \alpha^2$  with probability at least 0.99. Hence,  $d_{\text{TV}}(\mathcal{D}, \mathcal{D}_1) \leq 0.01$ . This also implies that if we let  $\mathcal{V}_1$  be the distribution over  $\{X^{(1)}, \dots, X^{(N)}\}$  where we first select  $\mu \sim \mathcal{D}_1$  and then draw each  $X^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , then  $d_{\text{TV}}(\mathcal{V}, \mathcal{V}_1) \leq 0.01$ .

Next, by Proposition 50 and a similar calculation as in the  $\mathcal{U}$  case, we can generate a sample from  $\mathcal{V}_1$  by writing  $X^{(i)} = (b + y'^{(i)}) \cdot v + z'^{(i)}$ , where  $v, y'^{(i)}, z'^{(i)}$  are generated in the same manner as previously, but  $b$  is distributed as  $\sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d$ . Since  $d_{\text{TV}}(\chi_d, \mathcal{N}(\sqrt{d}, 1)) \leq 0.2$ , we have that by scaling,

$$d_{\text{TV}}\left(\sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d, \mathcal{N}\left(\sqrt{\frac{d}{N} + 2\alpha^2}, \frac{1}{N} + \frac{2\alpha^2}{d}\right)\right) \leq 0.2.$$

Using the fact that  $\frac{1}{N} + \frac{2\alpha^2}{d} \leq \frac{1+\frac{2}{C}}{N}$  assuming that  $\frac{d}{\alpha^2} \geq C \cdot N$ , and letting  $C$  be sufficiently large, we have that

$$d_{\text{TV}}\left(\sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d, \mathcal{N}\left(\sqrt{\frac{d}{N} + 2\alpha^2}, \frac{1}{N}\right)\right) \leq 0.21.$$

Therefore, we can replace the distribution  $b \sim \sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d$  with  $b' \sim \mathcal{N}\left(\sqrt{\frac{d}{N} + 2\alpha^2}, \frac{1}{N}\right)$ . From here, we consider the distribution  $\mathcal{V}'$  over  $\{X'^{(1)}, \dots, X'^{(N)}\}$  where we draw each  $X'^{(i)} \sim (b' + y'^{(i)}) \cdot v + z'^{(i)}$ . Since the total variation distance between  $\sqrt{\frac{1}{N} + \frac{2\alpha^2}{d}} \cdot \chi_d$  and  $\mathcal{N}\left(\sqrt{\frac{d}{N} + 2\alpha^2}, \frac{1}{N}\right)$  is at most 0.21, we have that  $d_{\text{TV}}(\mathcal{V}_1, \mathcal{V}') \leq 0.21$ . Thus,  $d_{\text{TV}}(\mathcal{V}, \mathcal{V}') \leq d_{\text{TV}}(\mathcal{V}, \mathcal{V}_1) + d_{\text{TV}}(\mathcal{V}_1, \mathcal{V}') \leq 0.01 + 0.21 \leq 0.22$ . In addition, as in the  $\mathcal{U}$  case, we have that samples from  $\{b' + y'^{(i)}\}_{i=1}^N$  has the same distribution as samples from  $\left\{\sqrt{\frac{d}{N} + 2\alpha^2} + y^{(i)}\right\}$ , where each  $y^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . So, we can draw samples from  $\mathcal{V}'$  by sampling  $v$  uniformly from the unit sphere, sampling  $y^{(i)}, z'^{(i)}$  as before, and letting  $X^{(i)} = \left(\sqrt{\frac{d}{N} + 2\alpha^2} + y^{(i)}\right) \cdot v + z'^{(i)}$  for all  $i$ .

To finish, we have to establish a coupling between  $\mathcal{U}'$  and  $\mathcal{V}'$ . To do this, we sample the same  $v$  uniformly on the sphere for both  $\mathcal{U}'$  and  $\mathcal{V}'$ , and the same choices of  $z^{(i)}$  for each  $i \in [N]$  for both  $\mathcal{U}'$  and  $\mathcal{V}'$ . Next, we note that we can couple the samples of  $\sqrt{\frac{d}{N}} + y^{(i)}$  and the samples of  $\sqrt{\frac{d}{N} + 2\alpha^2} + y^{(i)}$ , using the fact that  $\sqrt{\frac{d}{N} + 2\alpha^2} - \sqrt{\frac{d}{N}} = \frac{\sqrt{d}}{\sqrt{N}} \cdot \left( \sqrt{1 + \frac{2\alpha^2 N}{d}} - 1 \right) = \frac{\sqrt{d}}{\sqrt{N}} \cdot O\left(\frac{2\alpha^2 N}{d}\right) = O\left(\frac{\alpha^2 \sqrt{N}}{\sqrt{d}}\right)$ . Therefore, by setting  $\gamma = \sqrt{\frac{d}{N} + 2\alpha^2} - \sqrt{\frac{d}{N}} = O\left(\frac{\alpha^2 \sqrt{N}}{\sqrt{d}}\right)$  and applying Proposition 48 in the one-dimensional case, we have that there exists a coupling between  $\mathcal{U}'$  and  $\mathcal{V}'$  that differ in  $O\left(\frac{\alpha^2 \sqrt{N}}{\sqrt{d}} \cdot N\right) \leq c_3 \left(\frac{\alpha^2 \cdot N^{3/2}}{\sqrt{d}}\right)$  samples for some constant  $c_3$ . We wish to show that this is at most  $\frac{c_2}{\varepsilon}$ , unless  $\Omega\left(\frac{d^{1/3}}{\varepsilon^{2/3} \cdot \alpha^{4/3}}\right)$ . But note that if  $c_3 \left(\frac{\alpha^2 \cdot N^{3/2}}{\sqrt{d}}\right) \geq \frac{c_2}{\varepsilon}$ , then  $N \geq \left(\frac{c_2}{c_3}\right)^{2/3} \cdot \frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}}$ , as desired. This completes the proof.  $\blacksquare$

We are now ready to prove Theorem 6.

**Proof** of Theorem 6: First, we already know that  $N \geq \Omega\left(\frac{\sqrt{d}}{\alpha^2}\right)$  even in the non-private setting, and that  $N \geq \Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$  using Proposition 48. So, it suffices to show that  $N \geq \Omega\left(\frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}}\right)$ .

Now, if  $d \geq C$  and  $\frac{d}{\alpha^2} \geq C \cdot N$  for some sufficiently large constant  $C$ , we will apply Theorem 51. Consider the null hypothesis  $\mathcal{H}_0$  where the distribution  $\{X^{(1)}, \dots, X^{(N)}\}$  is generated by sampling each  $X^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I)$ , and the alternative hypothesis  $\mathcal{H}_1$  where the distribution is generated by sampling  $\mu \sim \mathcal{D}$  and  $\{X^{(1)}, \dots, X^{(N)}\}$  is generated by sampling each  $X^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ . Since  $\mathcal{D}$  only has support on  $\|\mu\|_2 \geq \alpha$ , this is a valid alternative hypothesis. In this case, by Corollary 47 and Theorem 51, we must have that if an algorithm can  $(0, \varepsilon)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , then  $N \geq \Omega\left(\frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}}\right)$ .

Finally, we consider the cases where  $d \leq C$  or  $\frac{d}{\alpha^2} \leq C \cdot N$ . If  $d \leq C$ , then  $d = O(1)$ , which means that  $\frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}} \leq O\left(\frac{1}{\alpha \cdot \varepsilon} + \frac{\sqrt{d}}{\alpha^2}\right)$ . This is true because when  $d = O(1)$ ,  $\frac{1}{\alpha^{4/3} \varepsilon^{2/3}}$  is a weighted geometric average of  $\frac{1}{\alpha \cdot \varepsilon}$  and  $\frac{1}{\alpha^2}$ . So, the lower bound of  $\Omega\left(\frac{1}{\alpha \cdot \varepsilon} + \frac{\sqrt{d}}{\alpha^2}\right)$  implies the lower bound of  $\Omega\left(\frac{d^{1/3}}{\alpha^{4/3} \varepsilon^{2/3}}\right)$  if  $d = O(1)$ . If  $\frac{d}{\alpha^2} \leq C \cdot N$ , then  $N \geq \Omega\left(\frac{d}{\alpha^2}\right)$ . In addition, we also know that  $N \geq \Omega\left(\frac{1}{\alpha \cdot \varepsilon}\right)$ . It is simple to verify that  $\frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}}$  is a weighted geometric average of  $\frac{d}{\alpha^2}$  and  $\frac{1}{\alpha \cdot \varepsilon}$ , so we therefore also have  $N \geq \Omega\left(\frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}}\right)$ . This completes the proof.  $\blacksquare$

## Appendix E. Proof of Theorem 7

In this section, we prove Theorem 7. Again, we may apply Proposition 23, and assume WLOG the null hypothesis  $\mathcal{H}_0$  is that we are given  $N$  samples i.i.d. from  $\mathcal{N}(0, I)$ , and the alternative hypothesis  $\mathcal{H}_1$  is that we are given  $N$  samples i.i.d. from  $\mathcal{N}(\mu, I)$ , where  $\alpha \leq \|\mu\| \leq 2\alpha$ . In addition, due to Proposition 20, it suffices to show  $(0, O(\varepsilon))$ -DP rather than  $(\varepsilon, 0)$ -DP.

The main technical contribution in our proof is the following theorem, which provides an algorithm that can test whether the sum of the entries of a matrix is large or small, but the algorithm's output does not change significantly even if we alter an entire row or column of the matrix.

**Theorem 52** Fix parameters  $\gamma \leq 1$  and  $L \geq 1$ . Suppose that  $\gamma \cdot N^2 \geq C \left( \frac{L \cdot \log N}{\varepsilon} + \frac{\log^2 N}{\varepsilon^2} \right)$  for some sufficiently large constant  $C$ . Then, there exists an algorithm  $\mathcal{A}$  that runs in  $\text{poly}(N)$  time on matrices in  $[-1, 1]^{N \times N}$  and outputs either 0 or 1, with the following properties.

- For any matrices  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  that only differ in a single row (or only differ in a single column),  $|\mathbb{P}[\mathcal{A}(\mathbf{V}) = 0] - \mathbb{P}[\mathcal{A}(\mathbf{V}') = 0]| \leq 2\varepsilon$ .
- For any matrix  $\mathbf{V} \in [-1, 1]^{N \times N}$  that has every row and column sum at most  $L$  in absolute value, and with sum of all entries at most  $\frac{\gamma \cdot N^2}{4}$ , then  $\mathbb{P}[\mathcal{A}(\mathbf{V}) = 0] \geq 0.99$ .
- For any matrix  $\mathbf{V} \in [-1, 1]^{N \times N}$  that has every row and column sum at most  $L$  in absolute value, and with sum of all entries at least  $\frac{3\gamma \cdot N^2}{4}$ , then  $\mathbb{P}[\mathcal{A}(\mathbf{V}) = 1] \geq 0.99$ .

**Proof** Consider setting a threshold  $\tau := L + \frac{16K}{\varepsilon}$ , where  $K := \log_2 N$ . Now, consider the following set of functions  $\{f_k\}_{k=1}^K$ . For each  $k \in [N]$ , we define  $f_k(x) := \max\left(\min\left(1, \frac{|x|}{\tau} - k\right), 0\right)$ , i.e., it takes the function  $\frac{|x|}{\tau} - k$  and prevents it from exceeding 1 or becoming negative.

Now, given a matrix  $\mathbf{V} \in [-1, 1]^{N \times N}$  and for each  $k \in [K]$ , we define the statistic

$$F_k(\mathbf{V}) := \sum_{i=1}^N f_k\left(\sum_{j=1}^N V_{i,j}\right) + \sum_{j=1}^N f_k\left(\sum_{i=1}^N V_{i,j}\right).$$

We prove the following fact.

**Proposition 53** Consider two matrices  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  that only differ either in a single row or in a single column, and consider some  $k \geq 2$ . Then,

$$|F_k(\mathbf{V}) - F_k(\mathbf{V}')| \leq 1 + \frac{4}{\tau} \cdot F_{k-1}(\mathbf{V}).$$

**Proof** By symmetry of  $F_k$  with respect to rows and columns, assume WLOG that  $\mathbf{V}, \mathbf{V}'$  only differ in a single row  $i$ . Then, we can bound the absolute difference between  $F_k(\mathbf{V})$  and  $F_k(\mathbf{V}')$  by considering only how the  $i$ th row changes, and how each column changes at the  $i$ th row position. As a result, we have

$$|F_k(\mathbf{V}) - F_k(\mathbf{V}')| \leq \left| f_k\left(\sum_{j=1}^N V_{i,j}\right) - f_k\left(\sum_{j=1}^N V'_{i,j}\right) \right| + \sum_{j=1}^N \left| f_k\left(\sum_{i'=1}^N V_{i',j}\right) - f_k\left(\sum_{i'=1}^N V'_{i',j}\right) \right|.$$

We note that since the range of  $f_k$  is bounded in the range  $[0, 1]$ , the first part of the summand above is at most 1. To bound the second part of the summand, we note that for all  $j \in [N]$ ,  $\sum_{i'=1}^N V_{i',j}$  and  $\sum_{i'=1}^N V'_{i',j}$  differ by at most 2, since only a single row changes and each entry remains in  $[-1, 1]$ . In this case, the difference between  $f_k\left(\sum_{i'=1}^N V_{i',j}\right)$  and  $f_k\left(\sum_{i'=1}^N V'_{i',j}\right)$  is at most  $\frac{2}{\tau}$ , and in fact only changes if either  $\sum_{i'=1}^N V_{i',j}$  or  $\sum_{i'=1}^N V'_{i',j}$  is at least  $k \cdot \tau$  in absolute value. This means that  $\sum_{i'=1}^N V_{i',j}$  is at least  $k \cdot \tau - 2$  in absolute value, which means that  $f_{k-1}\left(\sum_{i'=1}^N V_{i',j}\right) \geq 1 - \frac{2}{\tau} \geq \frac{1}{2}$ . Overall, this means that for all  $j \in [N]$ ,

$$\left| f_k\left(\sum_{i'=1}^N V_{i',j}\right) - f_k\left(\sum_{i'=1}^N V'_{i',j}\right) \right| \leq \frac{4}{\tau} \cdot f_{k-1}\left(\sum_{i'=1}^N V_{i',j}\right).$$

Combining everything, we have that if  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  differ in a single row, then

$$\begin{aligned} |F_k(\mathbf{V}) - F_k(\mathbf{V}')| &\leq \left| f_k\left(\sum_{j=1}^N V_{i,j}\right) - f_k\left(\sum_{j=1}^N V'_{i,j}\right) \right| + \sum_{j=1}^N \left| f_k\left(\sum_{i'=1}^N V_{i',j}\right) - f_k\left(\sum_{i'=1}^N V'_{i',j}\right) \right| \\ &\leq 1 + \sum_{j=1}^N \frac{4}{\tau} f_{k-1}\left(\sum_{i'=1}^N V_{i',j}\right) \\ &\leq 1 + \frac{4}{\tau} \cdot F_{k-1}(\mathbf{V}). \end{aligned}$$

■

Now, let  $\eta := \frac{8 \cdot K}{\varepsilon \cdot \tau}$ , and define  $F(\mathbf{V}) := \sum_{k=1}^K \lfloor \frac{F_k(\mathbf{V})}{2+2N \cdot \eta^{k-1}} \rfloor$ . Recall that  $\tau \geq \frac{16K}{\varepsilon}$ , which means that  $\frac{16K}{\tau\varepsilon} \leq 1$ , and  $\eta \leq \frac{1}{2}$ . We now show the following.

**Proposition 54** *Suppose that  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  only differ either in a single row or in a single column. If  $F(\mathbf{V}) \leq K/\varepsilon$ , then  $|F(\mathbf{V}) - F(\mathbf{V}')| \leq K$ .*

**Proof** Again, by symmetry of  $F_k$  with respect to rows and columns, assume WLOG that  $\mathbf{V}, \mathbf{V}'$  only differ in a single row  $i$ . It suffices to show that for all  $k \in [K]$ ,  $|F_k(\mathbf{V}) - F_k(\mathbf{V}')| \leq 2 + 2N \cdot \eta^{k-1}$ . This would imply that  $\lfloor \frac{F_k(\mathbf{V})}{2+2N \cdot \eta^{k-1}} \rfloor$  and  $\lfloor \frac{F_k(\mathbf{V}')}{2+2N \cdot \eta^{k-1}} \rfloor$  differ by at most 1 for all  $k \in [K]$ , which is sufficient.

For  $k = 1$ , we have that  $F_1(\mathbf{V}), F_1(\mathbf{V}')$  are trivially between 0 and  $2N$ , and since  $2+2N \cdot \eta^{k-1} = 2 + 2N$ , the claim is immediate.

For  $k \geq 2$ , note that  $|F_k(\mathbf{V}) - F_k(\mathbf{V}')| \leq 1 + \frac{4}{\tau} \cdot F_{k-1}(\mathbf{V})$  by Proposition 53. In addition, since  $F(\mathbf{V}) \leq \frac{K}{\varepsilon}$ , we have that  $F_{k-1}(\mathbf{V}) \leq (2 + 2N \cdot \eta^{k-2}) \cdot (\frac{K}{\varepsilon} + 1) \leq (2 + 2N \cdot \eta^{k-2}) \cdot \frac{2K}{\varepsilon}$ . Therefore,

$$|F_k(\mathbf{V}) - F_k(\mathbf{V}')| \leq 1 + \frac{4}{\tau} \cdot (2 + 2N \cdot \eta^{k-2}) \cdot \frac{2K}{\varepsilon} \leq 1 + \frac{16K}{\tau\varepsilon} + 2N \cdot \frac{8K}{\tau\varepsilon} \cdot \eta^{k-2} \leq 2 + 2N \cdot \eta^{k-1}.$$

■

Next, we consider the piecewise linear function

$$g(x) := \begin{cases} x & |x| \leq (K+2) \cdot \tau \\ 2(K+2) \cdot \tau - x & x \geq (K+2) \cdot \tau \\ -2(K+2) \cdot \tau - x & x \leq -(K+2) \cdot \tau \end{cases}$$

and define

$$G(\mathbf{V}) := \sum_{i=1}^N g\left(\sum_{j=1}^N V_{i,j}\right) + \sum_{j=1}^N g\left(\sum_{i=1}^N V_{i,j}\right).$$

First note that for any real numbers  $x, y$ , that  $|(g(x) - g(y)) - (y - x)| \leq 4(K+2) \cdot \tau$ . In addition, note that  $g$  is a 1-Lipschitz function, meaning  $|g(x) - g(y)| \leq |x - y|$  for all  $x, y \in \mathbb{R}$ . Given this, we can prove the following.



**Proposition 55** *Suppose that  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  only differ either in a single row or in a single column. If  $F(\mathbf{V}) \leq K/\varepsilon$ , then  $|G(\mathbf{V}) - G(\mathbf{V}')| \leq 4(K+2)\tau + \frac{48K}{\varepsilon}$ .*

**Proof** Again, by symmetry of  $G$  with respect to rows and columns, assume WLOG that  $\mathbf{V}, \mathbf{V}'$  only differ in a single row  $i$ . To compute  $G(\mathbf{V}) - G(\mathbf{V}')$ , note that we only have to consider how the  $i$ th row changes, and how each column sum changes due to the  $i$ th row position. For each  $j \in [N]$ , define  $a_j := V'_{i,j} - V_{i,j}$ . Then,  $\sum_{j=1}^N V'_{i,j} - \sum_{j=1}^N V_{i,j} = \sum_{j=1}^N a_j$ , which means that  $g\left(\sum_{j=1}^N V'_{i,j}\right) - g\left(\sum_{j=1}^N V_{i,j}\right) = -\sum_{j=1}^N a_j \pm 4(K+2)\tau$ .

Next, we consider how each column  $j$  changes. We note that for any fixed  $j \in [N]$ ,  $\sum_{i'=1}^N V'_{i',j} - \sum_{i'=1}^N V_{i',j} = a_j$ . In addition, if  $\left|\sum_{i'=1}^N V_{i',j}\right| \leq (K+1) \cdot \tau$ , then  $g(x) = x$  for  $x = \sum_{i'=1}^N V_{i',j}$  and  $x = \sum_{i'=1}^N V'_{i',j}$ , since  $\left|\sum_{i'=1}^N V'_{i',j}\right| \leq (K+1) \cdot \tau + |a_j| \leq (K+1) \cdot \tau + 2 \leq (K+2) \cdot \tau$ . Therefore, if  $\left|\sum_{i'=1}^N V_{i',j}\right| \leq (K+1) \cdot \tau$ , then  $g\left(\sum_{i'=1}^N V'_{i',j}\right) - g\left(\sum_{i'=1}^N V_{i',j}\right) = a_j$ . Alternatively, because  $g$  is 1-Lipschitz, we still have that  $\left|g\left(\sum_{i'=1}^N V'_{i',j}\right) - g\left(\sum_{i'=1}^N V_{i',j}\right)\right| \leq 2$ , which means that  $g\left(\sum_{i'=1}^N V'_{i',j}\right) - g\left(\sum_{i'=1}^N V_{i',j}\right) = a_j \pm 4$ .

Now, note that since  $F(\mathbf{V}) \leq K/\varepsilon$ , this means that  $F_K(\mathbf{V}) \leq (2 + 2N \cdot \eta^{K-1}) \cdot \left(\frac{K}{\varepsilon} + 1\right) \leq (2 + 2N \cdot \eta^{K-1}) \cdot \frac{2K}{\varepsilon}$ . Since  $\eta \leq \frac{1}{2}$  and  $K = \log_2 N$ , this means that  $F_K(\mathbf{V}) \leq \frac{12K}{\varepsilon}$ . This, in turn, implies that the number of columns  $j$  such that  $\left|\sum_{i'=1}^N V_{i',j}\right| > (K+1) \cdot \tau$  is at most  $\frac{12K}{\varepsilon}$ . So, we have that

$$\sum_{j=1}^N \left[ g\left(\sum_{i'=1}^N V'_{i',j}\right) - g\left(\sum_{i'=1}^N V_{i',j}\right) \right] = \left[ \sum_{j=1}^N a_j \right] \pm 4 \cdot \frac{12K}{\varepsilon}.$$

So, overall, we have that

$$G(\mathbf{V}) - G(\mathbf{V}') = -\sum_{j=1}^N a_j \pm 4(K+2)\tau + \sum_{j=1}^N a_j \pm 4 \cdot \frac{12K}{\varepsilon} = \pm \left[ 4(K+2)\tau + \frac{48K}{\varepsilon} \right].$$

■

The algorithm  $\mathcal{A}$  will work as follows. First, we compute  $F(\mathbf{V})$ . With probability  $\min\left(1, F(\mathbf{V}) \cdot \frac{\varepsilon}{K}\right)$ , we will automatically output 1. Alternatively, if we have not yet output 1, we compute  $\tilde{G}(\mathbf{V}) := G(\mathbf{V}) + \text{Lap}\left(\frac{4(K+2)\tau}{\varepsilon} + \frac{48K}{\varepsilon^2}\right)$ . If  $\tilde{G}(\mathbf{V})$  is more than  $\gamma \cdot N^2$ , we output 1; otherwise we output 0.

We will prove that this algorithm is private up to changing an entire row or column of  $\mathbf{V}$ , and then we prove that this algorithm is accurate. Together, these complete the proof of Theorem 52.

**Proposition 56** *For any matrices  $\mathbf{V}, \mathbf{V}' \in [-1, 1]^{N \times N}$  that only differ on a single row (or only differ on a single column),  $|\mathbb{P}[\mathcal{A}(\mathbf{V}) = 0] - \mathbb{P}[\mathcal{A}(\mathbf{V}') = 0]| \leq 2\varepsilon$ .*

**Proof** First, note that if both  $F(\mathbf{V}) \geq \frac{K}{\varepsilon}$  and  $F(\mathbf{V}') \geq \frac{K}{\varepsilon}$ , then  $\mathcal{A}$  automatically outputs 1 with probability 1 for both  $\mathbf{V}$  and  $\mathbf{V}'$ .

So, we assume that either  $F(\mathbf{V}) < \frac{K}{\varepsilon}$  or  $F(\mathbf{V}') < \frac{K}{\varepsilon}$ . In this case, Proposition 54 tells us that  $|F(\mathbf{V}) - F(\mathbf{V}')| \leq K$ , which means that the probability of automatically outputting 1 before even computing  $\tilde{G}(T)$  does not change by more than  $\varepsilon$ . Next, we note that by Proposition

55,  $|G(\mathbf{V}) - G(\mathbf{V}')| \leq 4(K+2)\tau + \frac{48K}{\varepsilon}$ . Hence, by applying a Laplace mechanism scaled by  $\frac{1}{\varepsilon} \cdot (4(K+2)\tau + \frac{48K}{\varepsilon})$ , we have that the probability of  $\tilde{G}(\mathbf{V})$  being more than  $\gamma \cdot N^2$  and the probability of  $\tilde{G}(\mathbf{V}')$  being more than  $\gamma \cdot N^2$  do not differ by more than  $\varepsilon$ . Overall, we have that  $|\mathbb{P}[\mathcal{A}(\mathbf{V}) = 0] - \mathbb{P}[\mathcal{A}(\mathbf{V}') = 0]| \leq 2\varepsilon$ , as desired.  $\blacksquare$

**Proposition 57** *Suppose that  $\mathbf{V} \in [-1, 1]^{N \times N}$  has every row and column sum at most  $L$  in absolute value. Also, suppose that  $\gamma \cdot N^2 \geq C \left( \frac{L \cdot \log N}{\varepsilon} + \frac{\log^2 N}{\varepsilon^2} \right)$ . Then, if the sum of the entries of  $\mathbf{V}$  is at most  $\frac{\gamma \cdot N^2}{4}$ , then  $\mathbb{P}[\mathcal{A}(T) = 0] \geq 0.99$ , and if the sum of the entries of  $\mathbf{V}$  is at least  $\frac{3\gamma \cdot N^2}{4}$ , then  $\mathbb{P}[\mathcal{A}(T) = 1] \geq 0.99$ .*

**Proof** Note that since every row and column sum is at most  $L \leq \tau$  in absolute value, we have that for all  $k \in [K]$ ,  $f_k(\sum_{j=1}^N V_{i,j}) = 0$  for all rows  $i$ , and  $f_k(\sum_{i=1}^N V_{i,j}) = 0$  for all columns  $j$ . Therefore,  $F_k(\mathbf{V}) = 0$  for all  $k \in [K]$ , so with probability 1 we compute  $\tilde{G}(T)$ .

Likewise, we have that  $g(\sum_{j=1}^N V_{i,j}) = \sum_{j=1}^N V_{i,j}$  for all rows  $i$ , and  $g(\sum_{i=1}^N V_{i,j}) = \sum_{i=1}^N V_{i,j}$  for all columns  $j$ . Therefore,

$$G(\mathbf{V}) = \sum_{i=1}^N \sum_{j=1}^N V_{i,j} + \sum_{j=1}^N \sum_{i=1}^N V_{i,j} = 2 \cdot \sum_{i=1}^N \sum_{j=1}^N V_{i,j}.$$

In other words,  $G(\mathbf{V})$  is just twice the sum of the entries of  $\mathbf{V}$ .

So, if the sum of all entries in  $\mathbf{V}$  is at most  $\frac{\gamma N^2}{4}$ , then  $G(\mathbf{V}) \leq \frac{\gamma N^2}{2}$ . So, assuming that  $\gamma \cdot N^2 \geq 10 \left( \frac{4(K+2)\tau}{\varepsilon} + \frac{48K}{\varepsilon^2} \right)$ , we have that the probability that  $\tilde{G}(\mathbf{V}) \geq \gamma N^2$  is at most  $e^{-5} \leq 0.01$ . Likewise, if the sum of all entries in  $\mathbf{V}$  is at least  $\frac{3\gamma N^2}{4}$ , then  $G(\mathbf{V}) \geq \frac{3\gamma N^2}{2}$ . So, assuming that  $\gamma \cdot N^2 \geq 10 \left( \frac{4(K+2)\tau}{\varepsilon} + \frac{48K}{\varepsilon^2} \right)$ , we have that the probability that  $\tilde{G}(\mathbf{V}) \leq \gamma N^2$  is at most  $e^{-5} \leq 0.01$ . But since  $\tau = L + \frac{16K}{\varepsilon}$ , it suffices for

$$\gamma \cdot N^2 \geq 10 \left( \frac{4(K+2)L}{\varepsilon} + \frac{64(K+2)K}{\varepsilon^2} + \frac{48K}{\varepsilon^2} \right) = O \left( \frac{L \cdot \log N}{\varepsilon} + \frac{\log^2 N}{\varepsilon^2} \right).$$

Thus,  $\mathcal{A}$  succeeds with at least 0.99 probability either if the sum of the entries is at least  $\frac{3\gamma N^2}{4}$  or if the sum of the entries is at most  $\frac{\gamma N^2}{4}$ .  $\blacksquare$

We now return to the scenario of having  $N$  samples  $X^{(1)}, \dots, X^{(N)}$ . We first note the following corollary of Theorem 37.

Given  $N$  samples  $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ , we define the matrix  $\mathbf{T} \in \mathbb{R}^{N \times N}$  such that  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$  for all  $i, j \in [N]$ . Next, we modify the matrix to create a matrix  $\mathbf{V}$  such that  $V_{i,i} = \frac{T_{i,i} - d}{R}$  for all  $i \in [N]$  and  $V_{i,j} = \frac{T_{i,j}}{R}$  for all  $i \neq j \in [N]$ , where  $R = \tilde{O}(\sqrt{d})$  is a sufficiently large poly-logarithmic multiple of  $\sqrt{d}$ . As a direct corollary of Theorem 37, we have the following.

**Corollary 58** *Suppose that  $X^{(i)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I_d)$  for all  $i \in [N]$ , where  $0 \leq \|\mu\|_2 \leq 2\alpha$ . Then, with probability at least 0.99,*

- The sum of the entries of  $\mathbf{V}$  equals  $\frac{N^2 \cdot \|\mu\|_2^2}{R} \pm O\left(\frac{N\sqrt{d} + \alpha N\sqrt{N}}{R}\right)$ .
- All individual entries in  $\mathbf{V}$  are in the range  $[-1, 1]$ .
- The sum of each row and each column is at most  $\sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}}$  in absolute value.

**Proof** of Theorem 7: Our computationally efficient algorithm works as follows. First, sample  $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ , and construct the matrix  $\mathbf{V}$  as above. Now, let  $\gamma = \frac{\alpha^2}{R}$  and  $L = \sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}}$ . Finally, we apply the algorithm  $\mathcal{A}$  in Theorem 52 on  $\mathbf{V}$ .

Suppose that  $N^2 \cdot \alpha^2 \geq C(N\sqrt{d} + \alpha N\sqrt{N})$  for some sufficiently large constant  $C$  and that  $\gamma \cdot N^2 \geq C\left(\frac{L \log N}{\varepsilon} + \frac{\log^2 N}{\varepsilon^2}\right)$ . Then, for any  $\mathbf{X}, \mathbf{X}' \in (\mathbb{R}^d)^N$  that differ in a single row, the respective matrices  $\mathbf{V}, \mathbf{V}'$  differ in a single row and a single column, so  $|\mathbb{P}[\mathcal{A}(\mathbf{V}) = 0] - \mathbb{P}[\mathcal{A}(\mathbf{V}') = 0]| \leq 4\varepsilon$ , i.e., the algorithm is  $(0, 4\varepsilon)$ -differentially private. Next, if  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I)$ , then the sum of the entries of  $V$  is at most  $\frac{C}{4} \cdot \frac{N\sqrt{d} + \alpha N\sqrt{N}}{R} \leq \frac{1}{4} \cdot \frac{N^2 \cdot \alpha^2}{R} = \frac{1}{4} \cdot \gamma \cdot N^2$  with probability at least 0.99, which means the algorithm outputs 0 with probability at least 0.97. Finally, if  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ , for some  $\alpha \leq \|\mu\|_2 \leq 2\alpha$ , then the sum of the entries of  $V$  is at least  $\frac{N^2 \cdot \alpha^2}{R} - \frac{C}{4} \cdot \frac{N\sqrt{d} + \alpha N\sqrt{N}}{R} \geq \frac{N^2 \cdot \alpha^2}{R} - \frac{1}{4} \cdot \frac{N^2 \cdot \alpha^2}{R} = \frac{3}{4} \cdot \gamma \cdot N^2$ , which means the algorithm outputs 1 with probability at least 0.97.

Therefore, we just need  $N$  to satisfy  $N^2 \cdot \alpha^2 \geq C(N\sqrt{d} + \alpha N\sqrt{N})$  and  $\gamma \cdot N^2 \geq C\left(\frac{L \log N}{\varepsilon} + \frac{\log^2 N}{\varepsilon^2}\right)$ . Recalling that  $R = \tilde{O}(\sqrt{d})$  and that  $L = \sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}}$ , it is sufficient for

$$N = \tilde{O}\left(\frac{\sqrt{d}}{\alpha^2} + \frac{1}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{d^{1/4}}{\alpha \cdot \varepsilon} + \frac{1}{\alpha \cdot \varepsilon}\right) = \tilde{O}\left(\frac{\sqrt{d}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon}\right),$$

where we note that  $\frac{1}{\alpha^2}, \frac{1}{\alpha \cdot \varepsilon}$  are smaller terms, and that  $\frac{d^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}$  is a weighted geometric average of  $\frac{\sqrt{d}}{\alpha^2}$  and  $\frac{d^{1/4}}{\alpha \cdot \varepsilon}$  and therefore can be omitted.  $\blacksquare$

## Appendix F. Generalizations

In this section, we prove theorems 8, 9, and 10. We will also restate these theorems more formally while proving them. We note that the theorems proceed very similarly to Theorems 5, 6, and 7, so we will focus on the parts of these new theorems that differ, and simply sketch the remaining parts.

### F.1. Lower Bounds

We start with the easiest part, the lower bound. In the case of unknown covariance Gaussians, the lower bound we have in the known-covariance case immediately implies the same lower bound in the unknown-covariance case (where we are promised  $\|\Sigma\|_2 \leq 1$ ). Indeed, we can assume that  $\Sigma = I$  and the same lower bound applies.

For the case of identity testing for balanced product distributions, we remark that there is a known reduction (Canonne et al., 2020b, Theorem 3.1) which reduces the problem of identity testing of Gaussians with identity covariance to uniformity testing of product distributions over  $\{-1, 1\}^d$ . The reduction also holds for reducing to identity testing of product distributions over  $\{-1, 1\}^d$ ,

where the null hypothesis distribution  $\mathcal{P}_{\mu^*}$  satisfies  $|\mu_i^*| \leq 1 - \Omega(1)$  for all  $i \in [d]$ , i.e., if the null hypothesis distribution is balanced. Because this reduction shows that identity testing of Gaussians is easier than identity testing of balanced product distributions, it implies that any lower bound for identity testing of Gaussians implies the same lower bound for identity testing of balanced product distributions.

Finally, we note that tolerant testing only expands the possible null hypotheses, so any lower bounds for standard identity testing also imply the same lower bounds for the tolerant version.

Hence, we have the following theorem, which encompasses all lower bounds.

**Theorem 59** *Any  $(0, \varepsilon)$ -private algorithm that can distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  requires at least  $\Omega\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right)$  samples, for the following choices of  $\mathcal{H}_0$  and  $\mathcal{H}_1$ :*

1. **(Theorem 8, lower bound)**  $\mathcal{H}_0$  consists of  $\mathcal{N}(\mu^*, \Sigma)$  over all covariance matrices with bounded spectral norm  $\|\Sigma\|_2 \leq 1$ , and  $\mathcal{H}_1$  consists of  $\mathcal{N}(\mu, \Sigma)$  over all covariance matrices  $\|\Sigma\|_2 \leq 1$  and  $\mu : \|\mu - \mu^*\| \geq \alpha$ , where  $\mu^* \in \mathbb{R}^d$  is fixed.
2. **(Theorem 9, lower bound)**  $\mathcal{H}_0$  consists of the product distribution  $\mathcal{P}^*$  over  $\{-1, 1\}^d$  with mean  $\mu^*$ , and  $\mathcal{H}_1$  consists of all product distributions  $\mathcal{P}$  over  $\{-1, 1\}^d$  such that  $d_{TV}(\mathcal{P}, \mathcal{P}^*) \geq \alpha$ , where  $\mu^* \in [-1/2, 1/2]^d$  is fixed.
3. **(Theorem 10, lower bound for Gaussians)**  $\mathcal{H}_0$  consists of the Gaussians  $\mathcal{N}(\mu, \Sigma)$  over all  $\mu$  such that the Mahalanobis distance  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \leq \frac{\alpha}{2}$  and  $\mathcal{H}_1$  consists of  $\mathcal{N}(\mu, \Sigma)$  over all  $\mu$  such that  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \geq \alpha$ , where  $\mu^* \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^{d \times d}$  are fixed (and  $\Sigma$  is positive definite).
4. **(Theorem 10, lower bound for Products)**  $\mathcal{H}_0$  consists of all Boolean product distributions  $\mathcal{P}(\mu)$  such that  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \leq \frac{\alpha}{C}$  and  $\mathcal{H}_1$  consists of all  $\mathcal{P}(\mu)$  such that  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \geq \alpha$ , where  $\mu^* \in [-1/2, 1/2]^d$  is fixed and  $C$  is a sufficiently large constant.

## F.2. Upper Bounds: Unknown Covariance Case

We next consider scenario of identity mean testing where the samples come from a Gaussian with unknown but bounded covariance. In other words, we want to  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$ , where we are given  $N$  samples from  $\mathcal{N}(\mu^*, \Sigma)$  for some  $\|\Sigma\|_2 \leq 1$ , and  $\mathcal{H}_1$ , where we are given  $N$  samples from  $\mathcal{N}(\mu, \Sigma)$  for some  $\|\Sigma\|_2 \leq 1$  and  $\|\mu - \mu^*\|_2 \geq \alpha$ .

Specifically, we show the following theorem.

**Theorem 60** *Let  $\mu^* \in \mathbb{R}^d$  be fixed, and let  $\mathcal{H}_0$  consist of  $\mathcal{N}(\mu^*, \Sigma)$  over all covariance matrices with bounded spectral norm  $\|\Sigma\|_2 \leq 1$ , and  $\mathcal{H}_1$  consist of  $\mathcal{N}(\mu, \Sigma)$  over all covariance matrices  $\|\Sigma\|_2 \leq 1$  and  $\mu : \|\mu - \mu^*\| \geq \alpha$ . Then:*

1. **(Theorem 8, computationally inefficient case)** *There exists a computationally inefficient algorithm that, using*

$$N = \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

2. (**Theorem 8, computationally efficient case**) *There exists a computationally efficient algorithm that, using*

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon} \right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

Recall (by Proposition 23 and Remark 24) that we may think of  $\mathcal{H}_1$  as being given  $N$  i.i.d. samples from  $\mathcal{N}(\mu, \Sigma)$  for some  $\|\Sigma\|_2 \leq 1$  and  $\alpha \leq \|\mu - \mu^*\|_2 \leq 2\alpha$ . Also, by shifting, we assume WLOG that  $\mu^* = 0$ , but since  $\Sigma$  is unknown, we may not assume WLOG that  $\Sigma = I$ . Instead, we note that if we sample  $Y, Z \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$ , then even if  $\mu$  is unknown,  $W := \frac{Y-Z}{\sqrt{2}} \sim \mathcal{N}(0, \Sigma)$ . So, by using  $3N$  samples, we can generate  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$ , and  $W^{(1)}, \dots, W^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ . We start by considering the matrix  $\mathbf{U}$  where  $U_{i,j} = \langle X^{(i)}, X^{(j)} \rangle - \langle W^{(i)}, W^{(j)} \rangle$ . Assuming that  $\|\mu\| \leq 2\alpha$  where  $\alpha \leq \frac{1}{2}$ , we can apply Theorem 40 for some appropriate choice of  $L = \text{poly} \log(N)$  to obtain that with probability at least 0.98,

- $\sum_{i=1}^N \sum_{j=1}^N U_{i,j} = N^2 \|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$ .
- For all  $i, j \in [N]$  (including if  $i = j$ ),  $U_{i,j} = \pm L \cdot \sqrt{d}$ .
- For all  $i$ ,  $\sum_{j=1}^N U_{i,j} = \pm L(\sqrt{Nd} + \alpha \cdot N)$ , and for all  $j$ ,  $\sum_{i=1}^N U_{i,j} = \pm L(\sqrt{Nd} + \alpha \cdot N)$ .
- For all subsets  $S \subset [N]$ , if  $|S| = K$ , then  $\sum_{i,j \in S} U_{i,j} = K \cdot L \cdot (\sqrt{Kd} + K)$ .

Then, if we let  $\mathbf{T}$  be the matrix where add  $d$  to every diagonal entry in  $\mathbf{U}$ , we have that with probability at least 0.98,

- $\sum_{i=1}^N \sum_{j=1}^N T_{i,j} = Nd + N^2 \|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$ .
- For all  $i \neq j$ ,  $T_{i,j} = \pm L \cdot \sqrt{d}$ , and for all  $i$ ,  $V_{i,i} = d \pm L \cdot \sqrt{d}$ .
- For all  $i$ ,  $\sum_{j=1}^N T_{i,j} = d \pm L(\sqrt{Nd} + \alpha \cdot N)$ , and for all  $j$ ,  $\sum_{i=1}^N T_{i,j} = d \pm L(\sqrt{Nd} + \alpha \cdot N)$ .
- For all subsets  $S \subset [N]$ , if  $|S| = K$ , then  $\sum_{i,j \in S} T_{i,j} = K \cdot \left[ d \pm L \cdot (\sqrt{Kd} + K) \right]$ .

These are precisely the bounds we used for the matrix to distinguish between  $\mu = 0$  and  $\alpha \leq \|\mu\|_2 \leq 2\alpha$  in Section C. Hence, we can apply the same algorithm using this matrix  $\mathbf{T}$  instead. Indeed, we define  $\mathcal{C}$  to be the set of datasets  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in (\mathbb{R}^d)^{3N}$  that satisfy the second, third, and fourth properties above. Then, the same proof as in Lemmas 44 and 45 imply that for any datasets  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \mathcal{C}$  and  $(\mathbf{X}', \mathbf{Y}', \mathbf{Z}') \in \mathcal{C}$  that differ in  $K$  rows,  $|\tilde{T}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) - \tilde{T}(\mathbf{X}', \mathbf{Y}', \mathbf{Z}')| \leq K \cdot D$ , assuming that  $\Delta := 6L \left( \sqrt{Nd} + \alpha N + \frac{C}{\varepsilon} \right) \leq \frac{\varepsilon}{C} \cdot \alpha^2 \cdot N^2 =: D$ . Here,  $\tilde{T}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  is the statistic that adds up the entries of the corresponding matrix  $\mathbf{T} \in \mathbb{R}^{N \times N}$ , subtracts  $Nd$ , and clips the result by 0 from below and  $\alpha^2 \cdot N^2$  from above.

Therefore, we can create our  $D$ -Lipschitz extension  $\hat{T}$  of  $\tilde{T}$  that sends all datasets  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in (\mathbb{R}^d)^{3N}$  to  $\mathbb{R}$ . Finally, because we also have that the sum of all entries in  $\mathbf{T}$  equals  $Nd + N^2 \|\mu\|^2 \pm O(N\sqrt{d} + \alpha N\sqrt{N})$  with probability at least 0.98 if each  $X^{(i)}, Y^{(i)}, Z^{(i)}$  is drawn i.i.d. from

$\mathcal{N}(\mu, I)$ , we can apply the same algorithm as in Theorem 5 to obtain a private algorithm that succeeds with high probability using the same number of samples, up to a factor of 3. This completes the proof in the computationally inefficient case.

For the computationally efficient case, we let  $\mathbf{V}$  be the matrix where we divide each entry of  $\mathbf{U}$  by  $R$ , where  $R = \tilde{O}(\sqrt{d})$  is a sufficiently large polylogarithmic multiple of  $\sqrt{d}$ . Then, with probability at least 0.98,

- $\sum_{i=1}^N \sum_{j=1}^N V_{i,j} = \frac{N^2 \|\mu\|^2}{R} \pm O(N\sqrt{d} + \alpha N\sqrt{N})$ .
- For all  $i, j \in [N]$ ,  $V_{i,j} \in [-1, 1]$ .
- For all  $i$ ,  $\sum_{j=1}^N V_{i,j} = \pm(\sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}})$ , and for all  $j$ ,  $\sum_{i=1}^N V_{i,j} = \pm L(\sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}})$ .

These are the same guarantees as in Corollary 58. In addition, we again have that if we change exactly one entry among  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , this changes at most one row and one column. Because of this, we can apply Theorem 52 in the same way that we prove Theorem 7 to complete the proof in the computationally efficient case.

### F.3. Upper Bounds: Product Distributions

We establish the following theorem.

**Theorem 61** *Let  $\mu^* \in [-1/2, 1/2]^d$  be fixed, and let  $\mathcal{H}_0$  consist of  $\mathcal{P}(\mu^*)$ , and  $\mathcal{H}_1$  consist of  $\mathcal{P}(\mu)$  over all  $\mu$  with  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \geq \alpha$ . Then:*

1. **(Theorem 9, computationally inefficient case)** *There exists a computationally inefficient algorithm that, using*

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon} \right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

2. **(Theorem 9, computationally efficient case)** *There exists a computationally efficient algorithm that, using*

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon} \right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

The product distribution case will be remarkably similar to the identity-covariance Gaussian case. First, by Lemma 21, we may assume WLOG that  $\mu^* = 0$ . In addition, we can use Proposition 15 to state that for any distribution  $\mu$  with  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \geq \alpha$ , then  $\|\mu\|_2 \geq \frac{\alpha}{C_0}$  for some fixed absolute constant  $C_0$ . So, it suffices to distinguish between the null hypothesis of  $\mu = \mu^* = 0$  and  $\|\mu\| \geq \frac{\alpha}{C_0}$ . We can replace  $\frac{\alpha}{C_0}$  with  $\alpha$ , up to a constant factor loss in our sample complexity. We can again apply Remark 24 to say that  $\mathcal{H}_1$  is being given  $N$  i.i.d. samples from  $\mathcal{P}(\mu)$ , where  $\alpha \leq \|\mu\| \leq 2\alpha$ .

Now, we note that Theorem 37 works for both identity-covariance Gaussians and product distributions. Because of this, we obtain the same concentration bounds as in the known identity

covariance case. Hence, for the computationally inefficient case, Lemmas 44 and 45 can be trivially adapted to the product case, as can Theorem 5. Likewise, in the computationally efficient case, we can generate the matrix  $\mathbf{V}$  and apply Theorem 52 in the same way as we did in the identity covariance case.

#### F.4. Upper Bounds: Tolerant Testing

We start with the known-covariance Gaussian case. Specifically, we establish the following theorem.

**Theorem 62** *Let  $\mu^* \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^{d \times d}$  be fixed (with  $\Sigma$  positive definite). Let  $\mathcal{H}_0$  consist of  $\mathcal{N}(\mu, \Sigma)$  over all  $\mu$  with  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \leq \frac{\alpha}{2}$ , and  $\mathcal{H}_1$  consist of  $\mathcal{N}(\mu, \Sigma)$  over all  $\mu$  with  $\sqrt{(\mu - \mu^*)^T \Sigma^{-1} (\mu - \mu^*)} \geq \alpha$ . Then:*

1. **(Theorem 10, computationally inefficient Gaussian case)** *There exists a computationally inefficient algorithm that, using*

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon} \right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

2. **(Theorem 10, computationally efficient Gaussian case)** *There exists a computationally efficient algorithm that, using*

$$N = \tilde{O} \left( \frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon} \right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

We may assume WLOG that  $\mu^* = 0$  and  $\Sigma = I$ . First, we recall Lemma 25, which tells us that if  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$  and  $\bar{X} = X^{(1)} + \dots + X^{(N)}$ , then if  $\|\mu\| \leq \frac{\alpha}{2}$  then  $\mathbb{E}[\|\bar{X}\|^2] \leq Nd + N^2 \cdot \frac{\alpha^2}{4}$  and if  $\|\mu\| \geq \alpha$  then  $\mathbb{E}[\|\bar{X}\|^2] \geq Nd + N^2 \cdot \alpha^2$ . In addition, assuming  $\|\mu\| \leq 2\alpha$  and  $\alpha \leq 1$ , then  $\text{Var}[\|\bar{X}\|^2] = O(N^2d + N^3\alpha^2)$ .

Indeed, we will see that the same algorithms as in Theorem 5 and Theorem 7 will exactly apply in the tolerant case. Of course, the algorithms will still be  $(\varepsilon, 0)$ -differentially private and will still output 1 with probability at least  $2/3$  if the data is drawn i.i.d. from  $\mathcal{N}(\mu, I)$  with  $\alpha \leq \|\mu\| \leq 2\alpha$ . So, it suffices to show that the algorithm outputs 0 with probability at least  $2/3$  even if the data is drawn i.i.d. from  $\mathcal{N}(\mu, I)$  with  $\|\mu\| \leq \frac{\alpha}{2}$ .

To see why this holds in the computationally inefficient case, we recall that if  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$  for some  $\|\mu\| \leq \frac{\alpha}{2}$ , then with probability at least 0.99,  $\mathbf{X} = (X^{(1)}, \dots, X^{(N)}) \in \mathcal{C}$ . This implies that with probability at least 0.98,  $\hat{T}(\mathbf{X}) \leq N^2 \cdot \|\mu\|^2 + C(N\sqrt{d} + \alpha N\sqrt{N})$  for some sufficiently large constant  $C$  by Chebyshev's inequality. So, with probability at least 0.9, we have our final private estimate  $\hat{T}(\mathbf{X}) + \text{Lap}(D/\varepsilon) \leq N^2 \cdot \frac{\alpha^2}{4} + C(N\sqrt{d} + \alpha N\sqrt{N}) + 10 \cdot \frac{D}{\varepsilon}$ . Recalling that  $D = \frac{\varepsilon}{C} \cdot \alpha^2 N^2$ , this is less than  $N^2 \cdot \frac{\alpha^2}{2}$  as long as

$$N \geq \Omega \left( \frac{\sqrt{d}}{\alpha^2} \right).$$

But recall that we also needed  $\Delta = \tilde{O}(\sqrt{Nd} + \alpha N + \frac{C}{\varepsilon}) \leq D$ , and this holds as long as

$$N \geq \tilde{O}\left(\frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right).$$

Hence, we still obtain the same number of samples required for our inefficient upper bound.

In the computationally efficient case, we note that with probability at least 0.99, the matrix  $\mathbf{V}$  generated from  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, I)$  has the sum of all entries equal to  $\frac{N^2 \cdot \|\mu\|^2}{R} \pm O\left(\frac{N\sqrt{d} + \alpha N\sqrt{N}}{R}\right)$  by Corollary 58, where  $R = \tilde{O}(\sqrt{d})$ . Recalling that  $\|\mu\| \leq \frac{\alpha}{2}$  and  $\gamma = \frac{\alpha^2}{R}$ , this is at most  $\frac{\gamma \cdot N^2}{4} + O\left(\frac{N\sqrt{d} + \alpha N\sqrt{N}}{R}\right) \leq \frac{1.01 \cdot \gamma \cdot N^2}{4}$ , assuming that  $N \geq C\left(\frac{\sqrt{d}}{\alpha^2}\right)$  for some sufficiently large constant  $C$ . Hence, we can apply Theorem 52 (with a slight modification to replace  $\frac{\gamma \cdot N^2}{4}$  with  $\frac{1.01 \cdot \gamma \cdot N^2}{4}$ , for which the proof is nearly unaffected) to get that as long as  $\gamma \cdot N^2 \geq \tilde{O}\left(\frac{L}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$ , the algorithm works, where  $\gamma = \frac{\alpha^2}{\tilde{O}(\sqrt{d})}$  and  $L = \sqrt{N} + \alpha \cdot \frac{N}{\sqrt{d}}$ . As in the proof of Theorem 7, it suffices for  $N \geq \tilde{\Omega}\left(\frac{\sqrt{d}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon}\right)$ .

Next, we move to the Boolean Product case. Here, we establish the following theorem.

**Theorem 63** *Let  $\mu^* \in [-1/2, 1/2]^d$  and  $C$  be a sufficiently large constant. Let  $\mathcal{H}_0$  consist of  $\mathcal{P}(\mu)$  over all  $\mu$  with  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \leq \frac{\alpha}{C}$ , and  $\mathcal{H}_1$  consist of  $\mathcal{P}(\mu)$  over all  $\mu$  with  $d_{TV}(\mathcal{P}(\mu), \mathcal{P}(\mu^*)) \geq \alpha$ . Then:*

1. **(Theorem 10, computationally inefficient Product case)** *There exists a computationally inefficient algorithm that, using*

$$N = \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3} \cdot \varepsilon^{2/3}} + \frac{1}{\alpha \cdot \varepsilon}\right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

2. **(Theorem 10, computationally efficient Product case)** *There exists a computationally efficient algorithm that, using*

$$N = \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/4}}{\alpha \cdot \varepsilon}\right)$$

*samples, can  $(\varepsilon, 0)$ -privately distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .*

Again, we may assume WLOG that  $\mu^* = 0$ , by Lemma 21. In addition, by Proposition 15, we have that  $d_{TV}(\mathcal{P}(0), \mathcal{P}(\mu)) \in \left[\frac{\min(\|\mu\|, 1)}{C_0}, C_0 \cdot \min(\|\mu\|, 1)\right]$  for all  $\mu \in [-1, 1]^d$ , for some fixed absolute constant  $C_0$ . So, if we set  $C = 2C_0^2$ , it suffices to distinguish between  $\|\mu\| \leq \frac{\alpha}{C} \cdot C_0 = \frac{\alpha}{2C_0}$  and  $\|\mu\| \geq \frac{\alpha}{C_0}$ . By replacing  $\alpha$  with  $\frac{\alpha}{C_0}$ , up to an asymptotic factor it suffices to distinguish between  $\|\mu\|_2 \leq \frac{\alpha}{2}$  and  $\|\mu\|_2 \geq \alpha$  for all  $\alpha \leq \frac{1}{2}$ . Again, we may assume WLOG (by Proposition 23 and Remark 24) that we are distinguishing between  $\|\mu\|_2 \leq \frac{\alpha}{2}$  and  $\alpha \leq \|\mu\|_2 \leq 2\alpha$ .

At this point, the proof of Theorem 63 proceeds almost identically to the proof of Theorem 62. Indeed, we have the same concentration bounds for  $\|\bar{X}\|^2$  where  $\bar{X} = X^{(1)} + \dots + X^{(N)}$ , as well



as the same required bounds on the row and column sums, the individual entries, and the sums of submatrices in the matrix  $\mathbf{T}$  where  $T_{i,j} = \langle X^{(i)}, X^{(j)} \rangle$ , assuming that  $X^{(1)}, \dots, X^{(N)} \stackrel{i.i.d.}{\sim} \mathcal{P}(\mu)$ . Hence, the same proof as for Theorem 62 will apply to Theorem 63.