The Full Landscape of Robust Mean Testing: Sharp Separations between Oblivious and Adaptive Contamination

Clément Canonne* University of Sydney
Samuel B. Hopkins† MIT
Jerry Li‡ Microsoft Research
Allen Liu§ MIT
Shyam Narayanan¶ MIT

July 21, 2023

Abstract

We consider the question of Gaussian mean testing, a fundamental task in high-dimensional distribution testing and signal processing, subject to adversarial corruptions of the samples. We focus on the relative power of different adversaries, and show that, in contrast to the common wisdom in robust statistics, there exists a strict separation between adaptive adversaries (strong contamination) and oblivious ones (weak contamination) for this task. Specifically, we resolve both the information-theoretic and computational landscapes for robust mean testing. In the exponential-time setting, we establish the tight sample complexity of testing $\mathcal{N}(0, I)$ against $\mathcal{N}(\alpha v, I)$, where $\|v\|_2 = 1$, with an ε-fraction of adversarial corruptions, to be

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)\right),$$

while the complexity against adaptive adversaries is

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^4}\right)\right),$$

which is strictly worse for a large range of vanishing ε, α. To the best of our knowledge, ours is the first separation in sample complexity between the strong and weak contamination models.

In the polynomial-time setting, we close a gap in the literature by providing a polynomial-time algorithm against adaptive adversaries achieving the above sample complexity $\tilde{\Theta}(\max(\sqrt{d}/\alpha^2, d\varepsilon^2/\alpha^4))$, and a low-degree lower bound (which complements an existing reduction from planted clique) suggesting that all efficient algorithms require this many samples, even in the oblivious-adversary setting.

*clement.canonne@sydney.edu.au. Supported by an ARC DECRA (DE230101329) and an unrestricted gift from Google Research.

†samhop@mit.edu. Supported by NSF Award No. 2238080 and MLA@CSAIL

‡jerrl@microsoft.com.

§cliu568@mit.edu Supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship.

¶shyamsn@mit.edu. Supported by an NSF Graduate Fellowship and a Google Fellowship.

Contents

1 Introduction **1**
- 1.1 Types of Adversaries . . . 2
- 1.2 Our Results . . . 3
- 1.3 Related Work . . . 5
- 1.4 Overview of Techniques . . . 6

2 Preliminaries and Notation **12**
- 2.1 Basic Definitions . . . 12
- 2.2 Useful Probabilistic Tools and Inequalities . . . 12
- 2.3 Simplification of Alternative Hypothesis . . . 15
- 2.4 Notation . . . 16

3 Reducing to “Friendly" Oblivious Contaminations **16**
- 3.1 Structure of Obliviously Contaminated Samples . . . 17
- 3.2 Oblivious Filtering via Sample Splitting . . . 18

4 Mean Testing Robustly Against Oblivious Adversaries **29**
- 4.1 Setup and Algorithm . . . 29
- 4.2 Consequences of Assumption 2 . . . 30
- 4.3 The Null Case: Mean . . . 33
- 4.4 The Null Case: Variance . . . 33
- 4.5 The Alternative Case: Variance . . . 36
- 4.6 Proof of Lemma 4.3 . . . 39

5 Lower bound in the Huber model **40**
- 5.1 Main Lower Bound . . . 40

6 Improved Lower Bound against Oblivious Adversaries **45**
- 6.1 Lower bound instance . . . 45
- 6.2 Likelihood Ratio Computation . . . 45
- 6.3 Final Computation . . . 49

7 The Sample Complexity under Strong Contamination **51**

8 Polynomial-Time Algorithm **55**
- 8.1 Regularity conditions . . . 55
- 8.2 Filtering preliminaries . . . 58
- 8.3 Additional preliminaries . . . 60
- 8.4 The filtering algorithm for $n \leq d$ . . . 61
- 8.5 The filtering algorithm for $n > d$ . . . 63
- 8.6 Bounding row sums . . . 65
- 8.7 Putting it all together . . . 67

9 Computational Lower Bound **68**

A Mathematica code to verify the computation from Section 6.3 **73**

1 Introduction

Among all high-dimensional distribution testing (i.e., hypothesis testing) problems, *Gaussian mean testing* is one of the most basic, with connections to signal processing where it corresponds to *signal detection* under white noise. Given n independent samples $X_1, \dots, X_n \in \mathbb{R}^d$, the goal is to decide between two hypotheses:

> $\mathbf{H}_0$: $X_1, \dots, X_n$ were drawn from $\mathcal{N}(0, I)$, an origin-centered identity-covariance Gaussian.
> $\mathbf{H}_1$: $X_1, \dots, X_n$ were drawn from $\mathcal{N}(\mu, I)$ for some vector μ with $\|\mu\|_2 \geq \alpha$.

The following simple tester uses only $\Theta(\sqrt{d}/\alpha^2)$ samples, the information-theoretic optimum: reject the null iff the norm of the empirical mean $\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_2$ is larger than some well-chosen threshold. The number of samples scales as the square root of the dimension: in contrast, $\Theta(d/\alpha^2)$ samples (linear in the dimension) are needed to *learn* the mean μ of a Gaussian $\mathcal{N}(\mu, I)$ up to ℓ_2 error α. This d-vs-$\sqrt{d}$ gap is a prime example of a core theme in the literature on distribution testing: testing requires fewer samples than learning.

This simple tester is not robust to even a small fraction of adversarially corrupted samples. Concretely, suppose that an ε-fraction of the samples $X_1, \dots, X_n$ are chosen by a malicious adversary. Even after preprocessing the dataset by removing obvious outliers – say, X_i such that $\|X_i\|_2 \gg \mathbb{E}\|X_i\|_2 \approx \sqrt{d}$ – the simple tester with $\Theta(\sqrt{d}/\alpha^2)$ samples can be fooled by just a single corrupted sample.

Robust distribution testing has been extensively studied in robust statistics (the sub-field of statistics dealing with adversarially-corrupted data) [DKS17; DK23], and yet basic questions about robust mean testing remain open. Most importantly: *what is the sample-optimal robust mean tester?* As we show, the answer to this question is intimately intertwined with another unanswered question in robust statistics: *how much does it matter if the adversary sees the uncorrupted portion of the dataset?*

We find the latter question interesting for (at least) two reasons. First, it is a foundational question about the power of statistical adversaries – since modeling assumptions can have a strong effect on algorithm design, it is important to understand the consequences of basic assumptions. We are not the first to ask the question from this perspective; see also recent work of Blanc, Lange, Malik, and Tan [BLMT22]. Second, the question is pertinent to *data poisoning attacks* in machine learning [DGJ+21; GTX+22], where an adversary injects a small amount of malicious training data into a machine learning pipeline. Such attacks can be feasible in practice and hence are a significant concern [KNL+20]. If an *oblivious* adversary is strictly less powerful than an *adaptive* one, then keeping the training data secret is a potential (partial) defense against data poisoning.

It turns out that oblivious and adaptive adversaries have equal power for robust mean testing's close (and intensely studied [DK23]) cousin, robust mean *estimation*.[1] Here, the goal is to estimate μ up to ℓ_2 error α – in both adaptive and oblivious cases this requires $\Theta(\frac{d}{\alpha^2})$ samples. Indeed, this appears to be the case for a range of robust estimation problems, including covariance estimation and linear regression. This suggests a conventional wisdom in robust statistics: adaptivity does not buy statistical adversaries additional power.

Returning to robust mean testing, recent work by Narayanan [Nar22] shows that the sample complexity of robust mean testing against an adaptive adversary is $\tilde{\Theta}(\max(\sqrt{d}/\alpha^2, d\varepsilon^2/\alpha^4))$.[2] This brings us to:

Main Question: *What is the optimal robust mean tester against an **oblivious** adversary? Are the sample*

[1] Here we mean that the *sample complexity* of robust mean estimation is insensitive to details of the adversary's power. However, some separations are known, for instance between *additive* versus *additive and subtractive* adversaries in the polynomial-time setting [DKK+18]. See Section 1.3 for further discussion.

[2] Narayanan's work focuses on differentially private mean testing, but this result can be extracted using known reductions between robustness and privacy.

complexities of testing against adaptive and oblivious adversaries the same, as they are in robust estimation?

We answer this question by showing that the common wisdom – being resilient to stronger adversaries comes essentially "for free" – does *not* extend to mean testing, where being robust against an oblivious adversary is strictly easier than against a fully adaptive one (Theorem 1.4)! In fact, we resolve (up to log factors) the sample complexity of robust Gaussian mean testing in the presence of an oblivious adversary, by designing a new robust mean tester and proving a nearly-matching information-theoretic lower bound.

To make the landscape even more interesting, we also show that this separation vanishes when one requires the tester to be *computationally efficient*. We first give a polynomial-time (in fact, quadratic time) variant of Narayanan's tester, and then we obtain a lower bound against a large class of efficient algorithms ("low-degree algorithms") which shows a matching sample complexity against both oblivious and adaptive adversaries (Theorem 1.7). (This complements a reduction from planted clique by Brennan and Bresler [BB20] which also suggests that efficient algorithms require $\frac{d\varepsilon^2}{\alpha^4}$ samples even against oblivious adversaries.) One consequence is a new statistical-computational gap for robust mean testing against an oblivious adversary.

In order to discuss our results in more detail, we describe in the next section the standard adversarial corruption models we consider in our work, and how they relate. Then we state our results and provide an overview of the new techniques and ideas that underlie our proofs and algorithms.

1.1 Types of Adversaries

We focus on two main types of adversarial corruptions: namely, the *adaptive* (strong) and *oblivious* corruption models. These have a long history in Statistics and Algorithmic Robust Statistics; see [DK19; DK23] for a more thorough discussion. In what follows, we assume that the corruption rate ε is provided to the algorithm. Note that this is without loss of generality, as, given d, α, and the expressions of the sample complexities, the algorithm can compute the largest value of ε it can tolerate for a given number n of samples.

The first corruption model allows an *adaptive* adversary to look at the samples, and choose an ε-fraction of them to alter arbitrarily. Which subset of the samples was corrupted is unknown to the algorithm.

Definition 1.1 (Strong contamination model)**.** In the strong contamination model, n i.i.d. samples $X'_1, \dots, X'_n$ are drawn from the underlying unknown distribution $\mathcal{D}$. The adversary, upon observing $X'_1, \dots, X'_n$, chooses εn indices $i_1, \dots, i_{\varepsilon n}$ and values $X''_{i_1}, \dots, X''_{i_{\varepsilon n}}$. The algorithm then receives the sequence $X_1, \dots, X_n$, where $X_{i_j} = X''_{i_j}$ for all $j \in [\varepsilon n]$, and $X_i = X'_i$ otherwise. Crucially, both the εn indices and the values X''_i can depend on the "uncorrupted" samples $X'_1, \dots, X'_n$.

In contrast, in the *oblivious* contamination model, the adversary must commit to which fraction of the samples it will corrupt, and how, *before* observing the actual realization of the samples. (It is, however, allowed knowledge of both the specification of the algorithm and the underlying distribution.)

Definition 1.2 (Oblivious contamination model)**.** The adversary chooses εn indices $i_1, \dots, i_{\varepsilon n}$ and values $X''_{i_1}, \dots, X''_{i_{\varepsilon n}}$. Then n i.i.d. samples $X'_1, \dots, X'_n$ are drawn from the underlying unknown distribution $\mathcal{D}$, and the algorithm is provided with the sequence $X_1, \dots, X_n$, as in Definition 1.1.

This definition does allow the corrupted samples to be chosen in a correlated fashion; however, they cannot depend on the realizations of the uncorrupted points themselves. This oblivious model can be further weakened, leading to what is known as the *Huber contamination model* where the corrupted data points themselves must be chosen independently of each other:

Definition 1.3 (Huber contamination model)**.** In the Huber contamination model, the adversary chooses a corruption distribution $\tilde{\mathcal{D}}$ (possibly a function of the algorithm and underlying unknown distribution $\mathcal{D}$). Then n i.i.d. samples $X_1, \ldots, X_n$ are drawn from the mixture $(1-\varepsilon)\mathcal{D}+\varepsilon\tilde{\mathcal{D}}$, and provided to the algorithm.

While the focus of our work is on the adaptive and oblivious contamination models, some of our lower bounds apply even to the weaker Huber contamination model.

1.2 Our Results

Our main result settles the sample complexity of robust mean testing under oblivious contamination, and establishes a strict separation between oblivious and adaptive contamination models. In what follows, $\tilde{O}, \tilde{\Theta}, \tilde{\Omega}$ hide polylogarithmic factors in the argument, and we always assume[3] $\alpha \le O(1)$ and $\varepsilon \le \alpha/(\log n)^{O(1)}$ (except in Theorem 1.6), which is information-theoretically necessary, up to the factor $(\log n)^{O(1)}$.

Theorem 1.4 (Obliviously-robust mean testing (Informal; see Theorems 4.1, 5.1 and 6.1))**.** *In the* oblivious *contamination model, there is a mean tester which is robust to ε-contamination, which uses*

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)\right), \tag{1}$$

samples in the oblivious *contamination model, and this is information-theoretically tight up to logarithmic factors. Moreover, $\tilde{\Omega}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}\right)\right)$ samples are needed even in the weaker Huber contamination model.*

We offer a little interpretation of the (surprisingly complex) expression (1). If d dominates the other parameters, *i.e.*, $d \gg 1/\mathrm{poly}(\alpha), 1/\mathrm{poly}(\varepsilon)$, then $\frac{d\varepsilon^3}{\alpha^4}$ is the dominant term. But if $d, 1/\alpha, 1/\varepsilon$ are within small polynomial factors, any of the four terms in (1) can dominate.

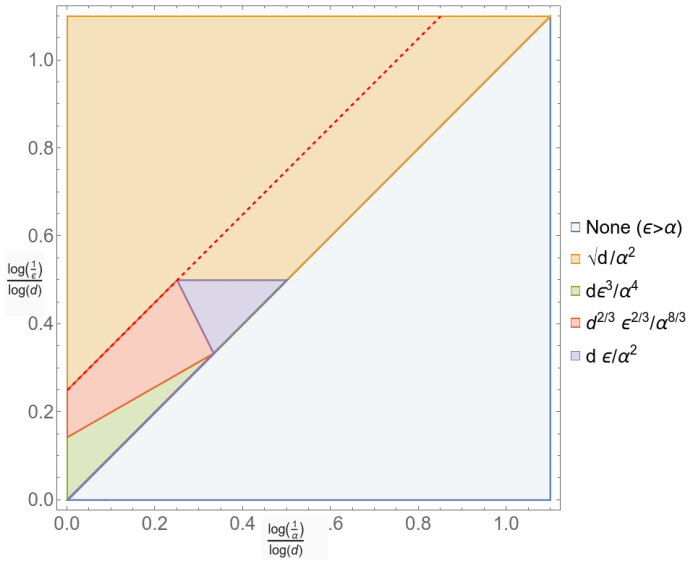

Figure 1: The various phases of the sample complexity of robust mean testing in the oblivious contamination model, as stated in Theorem 1.4: each area of this plot corresponds to which term of the sample complexity dominates, as a function of d, ε, α. The separation between adaptive and oblivious contamination occurs at the red dashed line (to the right, the oblivious sample complexity is strictly smaller). The lower half corresponds to $\alpha < \varepsilon$, where testing is information-theoretically impossible.

To see that Theorem 1.4 implies a strict separation between the oblivious and adaptive models, we recall:

[3] We note that, for identity-covariance Gaussian distributions, mean ℓ_2 distance α corresponds (for small α) to total variation (TV) distance $\Theta(\alpha)$. Thus, ℓ_2 mean testing corresponds to TV testing, which motivates the regime $\alpha \ll 1$ as of particular interest.

Theorem 1.5 ([Nar22], see also Theorem 7.1)**.** *In the* adaptive *contamination model, the optimal sample complexity of ε-robust mean testing is*

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^4}\right)\right) \tag{2}$$

The sample complexity (1) is strictly smaller than (2) for a range of vanishing ε, α, *e.g.*, with $\varepsilon = \Omega\left(\frac{\alpha}{d^{1/4}}\right)$.

For completeness, in Section 7 we show explicitly how to obtain Theorem 1.5 by combining Narayanan's result on differentially-private mean testing with known robust-privacy equivalence results (as in e.g. [GH22; HKMN22; AUZ23]). We further conjecture that a similar separation holds between the oblivious and Huber contamination models; to establish such a separation, it would be enough to prove a (non-efficient) $\tilde{O}(\max(\sqrt{d}/\alpha^2, d\varepsilon^3/\alpha^4))$ sample complexity upper bound in the latter, which in light of Theorem 1.4 would be nearly tight. We leave this as an interesting open problem.

A subtle difference between our strong and oblivious contamination models concerns which "good" samples are *removed* by the adversary. In the strong model, the adversary chooses adaptively which of the good samples to remove, whereas the oblivious adversary can only choose good samples to remove at random. Thus, the oblivious adversary could be equivalently defined as merely *adding* samples and doing no removals at all. One might ask whether the separation in sample complexities we establish between adaptive and oblivious adversaries actually arises from the ability of the adaptive adversary to remove samples, rather than from adaptivity itself.[4] We show in Section 7 that the lower bound of Theorem 1.5 actually holds even against adaptive adversaries that may only *add* data points, meaning that the sample complexity separation between adaptive and oblivious adversaries really is caused by the difference in addaptivity for the *added* samples. This extension to additive-only adaptive adversaries also readily follows from results proven in [Nar22].

Turning now to efficient algorithms, we provide the first polynomial-time algorithm which nearly matches the optimal sample complexity in the adaptive model. Prior to our work, the best polynomial-time approach was to learn the mean using $O(d/\alpha^2)$ samples, or to apply a polynomial-time algorithm of Narayanan [Nar22] which works only when $\varepsilon \leq \alpha \cdot d^{-1/4}$.

Theorem 1.6 (Adaptively-robust efficient mean testing (Informal; see Theorem 8.1))**.** *In the adaptive contamination model, there is a* quadratic-time *algorithm for ε-robust mean testing with sample complexity $\tilde{O}(\max(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^4}))$, as long as $\alpha \geq O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

This computationally efficient analogue of Theorem 1.5 raises the question of whether a similar analogue of Theorem 1.4 is possible. (The tester described in Theorem 1.4 relies on a computationally inefficient "filtering step"; see Section 1.4). Our next result shows strong evidence that this is not possible, and that the separation between adaptive and oblivious contamination models vanishes when restricting oneself to computationally efficient algorithms.

Theorem 1.7 (Computational lower bound (Informal; see Theorem 9.1))**.** *In the oblivious contamination model, any ε-robust* low-degree *mean testing algorithm in the Huber contamination model has sample complexity*

$$\Omega\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^4}\right)\right). \tag{3}$$

[4] For instance, one could consider an oblivious adversary which is allowed to replace the good distribution $\mathcal{D}$ with $\mathcal{D}$ conditioned on any event of probability $1-\varepsilon$, thus obliviously "removing" part of $\mathcal{D}$. We thank Guy Blanc for pointing this out.

Theorem 1.7 complements a reduction from planted clique [BB20] which suggests that $n^{\Omega(\log n)}$ time is required to beat $\frac{d\varepsilon^2}{\alpha^4}$ samples, even in the Huber model. The quantitative version of our result (Theorem 9.1) suggests something stronger (albeit for a restricted class of algorithms, rather than via reduction) – namely, that $\exp(n^{\Omega(1)})$ time is needed to use $(\frac{d\varepsilon^2}{\alpha^4})^{1-\Omega(1)}$ samples, even in the Huber model. We hope that our results, by uncovering a richer landscape in robust statistics than previously known and showing that the choice of contamination setting is much less innocuous than commonly believed, will spark interest in revisiting these modelling assumptions for various other tasks.

1.3 Related Work

Gaussian Mean Testing. Gaussian mean testing is known in statistics as the Gaussian sequence model [Erm91; Bar02; IIS03]; the understanding that it is possible to use fewer samples than dimensions appears relatively recent [SD08]. A recent influential work, [DKS17], records the sample-optimal mean tester and the "folklore" $\Omega(\sqrt{d}/\alpha^2)$ lower bound, and initiates the study of the complexity of *robust* mean testing. More recent work focuses on variants such as mean testing under a sparsity assumption [GC22], testing with unknown covariance [CCK+21; DKP23], testing subject to differential privacy [CKM+20; Nar22], robustly testing the covariance [DK21], or (distributed) testing giving partial observations from each sample [ACT20; SVZ22].

(Algorithmic) Robust Statistics. Algorithmic robust statistics, especially in high dimensions, has experienced a recent renaissance following a range of algorithmic breakthroughs; see the book [DK23]. Robust mean *estimation* has played a fundamental role; the quest for efficient algorithms for robust mean estimation led to the invention of the *filter* technique [DKK+19].

Connection to (Differential) Privacy. A recent line of work [GH22; HKMN22; AUZ23] established a (two-way) correspondence between adversarially robust and differentially private algorithms for a range of tasks, a connection we use to obtain Theorem 1.5. Importantly, this correspondence applies to *adaptive* adversaries, and does not, to the best of our knowledge, differentiate between oblivious and adaptive adversaries.

Noise Models in Statistics and Learning. Many developments in computational learning theory have been guided by the mission to design algorithms which work in an array of noise models [BH20]. For instance, the statistical query model was invented to capture a class of PAC learning algorithms which tolerate *random classification noise* [Kea98]. A full survey is out of scope, but some highlights include *nasty noise*, which is essentially the adaptive contamination model we consider here [BEK02; DKS18], and Massart noise, which has led to exciting recent algorithmic advances [DGT19; DKMR22; NT22]. While *computational* separations are known between these noise models in classification settings (e.g., random classification noise is much easier to handle algorithmically than adversarial label noise), separations in sample complexity seem unlikely, because empirical risk minimization handles even the nastiest noise models.

Two works in particular study questions related to ours. First, [BLMT22] shows some equivalences between adaptive and oblivious adversaries up to polynomial factors in sample complexity, for restricted classes of algorithms (SQ) or adversaries (additive). [DKK+18; DKS17] together show a computational separation between what error α is achievable for robustly learning a high-dimensional Gaussian when the adversary can only add samples versus when they can add and remove samples. We emphasize that while previous work showed evidence for a computational gap, we believe ours is the first demonstration of an (unconditional) information-theoretic separation in a natural robust statistics setting.

1.4 Overview of Techniques

1.4.1 Exploiting Obliviousness to Robustly Test with Fewer Samples

Our Approach. We focus first on our main technical contribution, the mean tester from Theorem 1.4. To get an improved testing algorithm for oblivious contaminations (compared to adaptive contaminations), we need to exploit that the adversary must commit to the contaminated points before the remaining datapoints are drawn. A consequence is that the correlation between the sums of good points (G) and bad points (B) is comparable to independent random vectors of comparable norm:

$$\left\langle \frac{\sum_{i\in B} X_i}{\|\sum_{i\in B} X_i\|_2}, \frac{\sum_{i\in G} X_i}{\|\sum_{i\in G} X_i\|_2} \right\rangle \approx \frac{\pm 1}{\sqrt{d}}.$$

By contrast, an adaptive adversary can make this correlation as large as 1.

Hence, the only way the adversary can have a substantial effect on $\left\|\sum_{i\in[n]} X_i\right\|_2$ is by making $\|\sum_{i\in B} X_i\|_2$ larger than it would be for a set of εn good samples. Building on this idea, we can design a tester using $\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)\right)$ samples under (roughly) the additional assumption that the sum of every subset of the adversary's vectors has about the same norm it would if the samples were uncorrupted.

The second challenge is to remove this additional assumption. The standard approach in robust statistics to make bad samples "look like" good ones according to some tests (e.g. norms of sums of subsets of points) is to remove samples in subsets which violate those tests; this is often called "filtering". This risks removing about εn good samples as well, but in many settings this isn't an issue.

However, *removing any good samples after looking at all the samples potentially breaks obliviousness by introducing dependencies between good and bad samples!* We develop a novel *obliviousness-preserving* filtering technique. We (iteratively) split the samples into two subsets, U, V. Looking only at U, we devise a rule for which samples to keep and which to remove (keeping those contained in a certain intersection of halfspaces); then we apply this rule to V and show that it preserves obliviousness while ensuring that V now satisfies the assumption about sums of subsets of corrupted vectors. We turn now to a more detailed overview.

Background: Narayanan's Robust Tester. To understand quantitatively how we can exploit obliviousness of the adversary, we first review a robust mean tester which uses $\tilde{O}(\max(\sqrt{d}/\alpha^2, d\varepsilon^2/\alpha^4))$ samples in the strong contamination model, as long as $\varepsilon \ll \alpha$ (all of which is information-theoretically necessary).[5] Our polynomial-time algorithm is also an adaptation of the following robust tester.

As in many robust statistics settings, the overall scheme relies on finding a "good enough" subset of $(1-\varepsilon)n$ samples $S \subseteq [n]$, to then apply a non-robust algorithm on S – in this case, the simple tester based on $\|\sum_{i\in S} X_i\|_2^2$. For $X_1, \ldots, X_n \in \mathbb{R}^d$ which are clear from context and $T \subseteq [n]$, let $\mathrm{Sum}(T) = \sum_{i\in T} X_i$.

Definition 1.8 (Good Enough Subset (Informal))**.** For $X_1, \ldots, X_n \in \mathbb{R}^d$, we say $S \subseteq [n]$, $|S| = (1-\varepsilon)n$ is *good enough* if, for every $T \subseteq S$ with $|T| \le \varepsilon n$,

$$\|\mathrm{Sum}(T)\|_2^2 \le |T|d + \tilde{O}(\varepsilon^{1.5} n^{1.5}\sqrt{d} + \varepsilon^2 n^2) \text{ and } |\langle \mathrm{Sum}(S\setminus T), \mathrm{Sum}(T)\rangle| \le \tilde{O}(\varepsilon n^{1.5}\sqrt{d} + \varepsilon^2 n^2).$$

The choice of parameters in the definition guarantees that any subset of size $(1-\varepsilon)n$ of n independent samples from $\mathcal{N}(0, I)$ or $\mathcal{N}(\mu, I)$, for small-enough μ, is good enough with high probability. To see why

[5] A similar tester can be extracted from [Nar22]. While Narayanan's paper focuses on differentially private mean testing, the tester can be shown to be robust by virtue of its privacy guarantees; see Section 7. The tester we describe here is simpler than Narayanan's original tester, in part because we need only robustness, not privacy.

this holds intuitively, observe that if S consists of good samples only, then $|\langle \mathrm{Sum}(S \setminus T), \mathrm{Sum}(T)\rangle|$ is roughly distributed as $\mathcal{N}(0, \varepsilon n^2 d)$, and we need a union bound over $\approx n^{\varepsilon n}$ choices of T.

Definition 1.9 (Narayanan's tester)**.** Given n ε-contaminated samples, Narayanan's tester finds any good enough subset S and outputs $\mathbf{H}_0$ if $\|\mathrm{Sum}(S)\|_2^2 - (1-\varepsilon)nd \ll \alpha^2 n^2$ and $\mathbf{H}_1$ otherwise.

Analysis Sketch. Let $X_1, \dots, X_n$ be an ε-contaminated draw from either $\mathcal{N}(0, I)$ or $\mathcal{N}(\mu, I)$ for some $\|\mu\|_2 = \alpha$. Let $G \subseteq [n]$ be the uncorrupted samples. (For simplicity, in this overview we assume the adversary has only added samples; removed samples can be handled without much more difficulty.) Let $S \subseteq [n]$ be any good enough subset; we want to show $\|\mathrm{Sum}(S)\|_2^2 - (1-\varepsilon)d \geq \Omega(\alpha^2 n^2)$ in the alternative case, and $\|\mathrm{Sum}(S)\|_2^2 - (1-\varepsilon)d \ll \alpha^2 n^2$ in the null. First,

$$\mathbb{E}\|\mathrm{Sum}(G)\|_2^2 - (1-\varepsilon)d = \begin{cases} \mathbb{E}\sum_{i \neq j \in G} \langle X_i, X_j\rangle \approx \alpha^2 n^2 & \text{in the alternative case} \\ 0 & \text{in the null case} \end{cases}$$

and standard concentration arguments show that this holds with high probability so long as $n \gg \sqrt{d}/\alpha^2$. So we just have to show that $|\|\mathrm{Sum}(S)\|_2^2 - \|\mathrm{Sum}(G)\|_2^2| \ll \alpha^2 n^2$. This is doable using the following lemma.

Lemma 1.10 (Main Lemma for Narayanan's Tester)**.** *For any two good-enough subsets S, S' of $X_1, \dots, X_n \in \mathbb{R}^d$, $\left|\|\mathrm{Sum}(S)\|_2^2 - \|\mathrm{Sum}(S')\|_2^2\right| \ll \alpha^2 n^2$, so long as $n \gg d\varepsilon^2/\alpha^4$.*

Proof. We divide S into $S \cap S'$ and $S \setminus S'$ and S' into $S' \cap S$ and $S' \setminus S$, so we have

$$\begin{aligned}\|\mathrm{Sum}(S)\|_2^2 - \|\mathrm{Sum}(S')\|_2^2 = \|\mathrm{Sum}(S \cap S')\|_2^2 + 2\langle \mathrm{Sum}(S \cap S'), \mathrm{Sum}(S \setminus S')\rangle + \|\mathrm{Sum}(S \setminus S')\|_2^2 \\ - \|\mathrm{Sum}(S' \cap S)\|_2^2 - 2\langle \mathrm{Sum}(S' \cap S), \mathrm{Sum}(S' \setminus S)\rangle - \|\mathrm{Sum}(S' \setminus S)\|_2^2 .\end{aligned}$$

Now, $\|\mathrm{Sum}(S \cap S')\|_2^2 - \|\mathrm{Sum}(S' \cap S)\|_2^2 = 0$, and since $|S \setminus S'| = |S' \setminus S|$, also $|\|\mathrm{Sum}(S \setminus S')\|_2^2 - \|\mathrm{Sum}(S' \setminus S)\|_2^2| \leq \tilde{O}(\varepsilon^{1.5} n^{1.5}\sqrt{d} + \varepsilon^2 n^2)$, using good-enough-ness. By using good-enough-ness again, both $|\langle \mathrm{Sum}(S \cap S'), \mathrm{Sum}(S \setminus S')\rangle|$ and $|\langle \mathrm{Sum}(S' \cap S), \mathrm{Sum}(S' \setminus S)\rangle|$ are at most $\tilde{O}(\varepsilon n^{1.5}\sqrt{d} + \varepsilon^2 n^2)$. Since $\varepsilon \ll \alpha$, we have $\varepsilon^2 n^2 \ll \alpha^2 n^2$, and since $n \gg d\varepsilon^2/\alpha^4$, we have $\varepsilon n^{1.5}\sqrt{d} \ll \alpha^2 n^2$. □

This completes the analysis of Narayanan's tester. We record two important observations:

1. The reason that the tester requires $d\varepsilon^2/\alpha^4$ samples lies in the term $\langle \mathrm{Sum}(S \cap S'), \mathrm{Sum}(S \setminus S')\rangle$. Let's think of $S' = G$, the good samples, and S as some good-enough subset which contains around εn corrupted samples, $S \setminus G$. The adaptive adversary could choose the samples in $S \setminus G$ to make $\mathrm{Sum}(S \setminus G)$ too (anti)-correlated with $\mathrm{Sum}(S \cap G)$. There is a limit to how large he can make the (anti)correlation before S is no longer "good enough" – namely, he can make $\langle \mathrm{Sum}(S \cap G), \mathrm{Sum}(S \setminus G)\rangle$ as large as the largest inner product of the form $\langle \mathrm{Sum}(G \setminus T), \mathrm{Sum}(T)\rangle$ for $T \subseteq G$ with $|T| = \varepsilon n$, which is around $\varepsilon n^{1.5}\sqrt{d}$ by standard concentration.
2. Narayanan's tester requires finding a good-enough subset of $(1-\varepsilon)n$ samples; *prima facie* this requires exponential-time brute-force search, but we describe a polynomial-time variant of his approach later.

Using Only $d\varepsilon/\alpha^2$ Samples if the Adversary is Oblivious and Not "Too Big". Narayanan's tester is information-theoretically optimal (up to log factors) against adaptive adversaries. As our first taste of improved testing against an oblivious adversary, consider the following toy setup. Suppose the adversary is not only oblivious but also promises us that the εnd bad samples B will satisfy $\|\mathrm{Sum}(B)\|_2^2 \leq O(\varepsilon nd)$;

roughly, this constraints the adversary to add εn vectors of norm $\sqrt{d}$ which are approximately pairwise orthogonal. (If the adversary adds any vector of norm much larger, we can remove it before proceeding.) We will show how to test using $\sqrt{d}/\alpha^2 + d\varepsilon/\alpha^2$ samples, improving on Narayanan's tester for $\varepsilon \gg \alpha^2$.

We revisit the simple tester using just $\|\mathrm{Sum}([n])\|_2^2$. Dividing $[n]$ into good and corrupted samples G, B,

$$\|\mathrm{Sum}([n])\|_2^2 - nd = \Big(\|\mathrm{Sum}(G)\|_2^2 - (1-\varepsilon)nd\Big) + 2\left\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\right\rangle + \|\mathrm{Sum}(B)\|_2^2 - \varepsilon nd\,.$$

As usual, $\|\mathrm{Sum}(G)\|_2^2 - (1-\varepsilon)nd \geq \Omega(\alpha^2 n^2)$ in the alternative case and $\ll \alpha^2 n^2$ in the null; we want to show the remaining terms are $\ll \alpha^2 n^2$ in magnitude. Trivially, $|\|\mathrm{Sum}(B)\|_2^2 - \varepsilon nd| \leq O(\varepsilon nd) \ll \alpha^2 n^2$ when $d\varepsilon/\alpha^2 \ll n$, using our promise on $\|\mathrm{Sum}(B)\|_2^2$.

Now let's look at the term where we make the improvement over Narayanan's tester: $\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle$; we are looking to use obliviousness to beat the bound $\varepsilon n^{1.5}\sqrt{d}$. We fix $\mathrm{Sum}(B)$ and then sample the random vector $\mathrm{Sum}(G)$, which is distributed either as $\mathcal{N}(0, (1-\varepsilon)nI)$ or $\mathcal{N}((1-\varepsilon)n\mu, (1-\varepsilon)nI)$, meaning

$$\left\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\right\rangle \sim \begin{cases} \mathcal{N}\Big((1-\varepsilon)n\left\langle \mu, \mathrm{Sum}(B)\right\rangle, (1-\varepsilon)n\|\mathrm{Sum}(B)\|_2^2\Big) & \text{in the alternative case} \\ \mathcal{N}\Big(0, (1-\varepsilon)n\|\mathrm{Sum}(B)\|_2^2\Big) & \text{in the null case} \end{cases}.$$

So, $|\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle| \leq O(n\alpha \cdot \sqrt{\varepsilon nd} + n\sqrt{\varepsilon d}) \ll \alpha^2 n^2$, as $\|\mathrm{Sum}(B)\|_2^2 \leq O(\varepsilon nd)$ and $n \gg d\varepsilon/\alpha^2$. From this simple reasoning, we draw the following important conclusion:

If the adversary is oblivious and is constrained to add samples B which aren't "too big", then we can test using fewer samples than against an adaptive adversary.

This leads us to two key questions, whose answers form the main technical ingredients in our oblivious tester. Can we take an obliviously-corrupted dataset and remove samples in some way to ensure that in the resulting *filtered* dataset, the adversary has added samples B which aren't "too big", but do so in a way which doesn't introduce dependencies between good and bad samples which would break the obliviousness we're relying on? And, what is the right definition for "too big" – could a more refined definition lead to a tester using fewer than $d\varepsilon/\alpha^2$ samples?

Friendly Oblivious Adversaries and The Sum+Variance Tester. We will tackle the above questions in reverse order. We introduce a key definition:

Definition 1.11 (Informal, see Assumption 1)**.** A *friendly* oblivious adversary introduces $\{X_i\}_{i\in B}$ such that

1. For disjoint $S, T \subseteq B$ with $|S|, |T| \leq \varepsilon n$, $|\langle \mathrm{Sum}(S), \mathrm{Sum}(T)\rangle| \leq \tilde{O}(\sqrt{|S|\cdot|T|}\cdot(\sqrt{\varepsilon nd} + \varepsilon n))$.
2. For distinct $i, j \in B$, $|\langle X_i, X_j\rangle| \leq \tilde{O}(\sqrt{d})$, and for every $i \in B$, $\|X_i\|_2^2 = d \pm \tilde{O}(\sqrt{d})$.

The parameters are chosen so that every pair of subsets S, T of *good* samples would satisfy these conditions.

To clarify why friendliness refines the "not too big" condition $\|\mathrm{Sum}(B)\|_2^2 \leq O(\varepsilon nd)$ from above, observe that subject to friendliness, for any $S \subseteq B$,

$$\|\mathrm{Sum}(S)\|_2^2 = |S|\cdot(d \pm \tilde{O}(\sqrt{d})) + O(\mathbb{E}_{S_1,S_2}\left\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\right\rangle) = d|S| + \tilde{O}(|S|\sqrt{\varepsilon nd} + |S|\sqrt{d})$$

where S_1, S_2 is a random partition of S. In particular, $\|\mathrm{Sum}(B)\|_2^2 = \varepsilon nd \pm o(\alpha^2 n^2)$ whenever $n \gg d\varepsilon^3/\alpha^4$.

Now we can introduce our robust mean tester which uses $\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4} + \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^2}$ samples (up to log factors) in the presence of a friendly oblivious adversary.

The Sum+Variance Tester (Algorithm 3): Given $X_1, \dots, X_n \in \mathbb{R}^d$, if $\|\mathrm{Sum}([n])\|_2^2 - nd \geq \Omega(\alpha^2 n^2)$, or if

$$\frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i, \mathrm{Sum}([n])\rangle - d}{\|\mathrm{Sum}([n])\|_2}\right)^2 \geq 1 + \Omega\left(\frac{\alpha^4 n}{\varepsilon d}\right),$$

return $\mathbf{H}_1$, otherwise return $\mathbf{H}_0$.

Analysis Sketch. For starters, we need to make sure that in the null case, $\|\mathrm{Sum}([n])\|_2^2 - nd \ll \alpha^2 n^2$. Splitting S into good samples G and corrupted samples B, we know $\|\mathrm{Sum}(G)\|_2^2 = (1-\varepsilon)nd \pm O(n\sqrt{d})$ and $|\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle| \leq O(n\sqrt{\varepsilon d})$ using standard concentration tools and obliviousness, and $\|\mathrm{Sum}(B)\|_2^2 = \varepsilon nd + \tilde{O}(\varepsilon^{1.5} n^{1.5}\sqrt{d} + \varepsilon n\sqrt{d})$ by friendliness. All together,

$$\|\mathrm{Sum}([n])\|_2^2 - nd = \|\mathrm{Sum}(G)\|_2^2 + 2\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle + \|\mathrm{Sum}(B)\|_2^2 - nd = \tilde{O}(n\sqrt{d} + \varepsilon^{1.5} n^{1.5}\sqrt{d})$$

which is at most $\alpha^2 n^2$ exactly when $n \gg \frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}$.

Ideally, we would show next that in the alternative case $\|\mathrm{Sum}([n])\|_2^2 - nd \gg \alpha^2 n^2$, but even a friendly, oblivious adversary can ensure this doesn't happen when $n \ll \frac{d\varepsilon}{\alpha^2}$. With knowledge of the vector μ, he can introduce samples $\{X_i\}_{i\in B}$ such that $\langle X_i, \mu\rangle \approx -\frac{\alpha^2}{\varepsilon}$, which introduces cancellations with $\mathbb{E}\,\mathrm{Sum}(G)$ that reduce $\|\mathrm{Sum}([n])\|_2^2$. Overall, he can ensure $\left|\|\mathrm{Sum}([n])\|_2^2 - nd\right| \ll \alpha^2 n^2$.

But now we encounter a typical theme in robust statistics: the adversary has had to introduce a small set of X_i's such that $\langle X_i, \mathrm{Sum}([n])\rangle$ is more negative than typical, thereby increasing the variance among $\{\langle X_i, \mathrm{Sum}([n])\rangle\}_{i\in[n]}$. For $i \in B$, we expect $\langle X_i, \mathrm{Sum}([n])\rangle$ to be $\frac{n\alpha^2}{\varepsilon}$ smaller than usual, so heuristically,

$$\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i, \mathrm{Sum}([n])\rangle - d}{\|\mathrm{Sum}([n])\|_2}\right)^2 \gtrsim \frac{1}{n}\cdot \varepsilon n \cdot \frac{\alpha^4 n^2}{\varepsilon^2 nd} = \frac{\alpha^4 n}{\varepsilon d},$$

where we used $\|\mathrm{Sum}([n])\|_2^2 \approx nd$. Adding the contribution from the samples in G gives us $1 + \Omega(\frac{\alpha^4 n}{\varepsilon d})$. We make this idea rigorous in Section 4.

Of course, outputting $\mathbf{H}_1$ when $\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i, \mathrm{Sum}([n])\rangle - d}{\|\mathrm{Sum}([n])\|_2}\right)^2 = 1 + \Omega(\frac{\alpha^4 n}{\varepsilon d})$ only makes sense if the adversary cannot make this happen in the null model. We show (Section 4.4) that no friendly oblivious adversary can make $\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i, \mathrm{Sum}([n])\rangle - d}{\|\mathrm{Sum}([n])\|_2}\right)^2 = 1 + \Omega(\frac{\alpha^4 n}{\varepsilon d})$ if $n \gg \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$.

Friendliness via Obliviousness-Preserving Filtering. We're still missing a key ingredient: how can we force an oblivious adversary to be friendly? Ensuring condition 2 of friendliness is straightforward. If we see any $|\|X_i\|_2^2 - d| \gg \sqrt{d}$, that X_i must have been introduced by the adversary and can be safely removed, and similarly if any pair i, j has $|\langle X_i, X_j\rangle| \gg \sqrt{d}$ then (by obliviousness) both X_i, X_j must be corrupted samples and can be removed. (We are using $\gg$ to hide logarithmic factors.)

But what about condition 1? A natural idea is to preprocess $X_1, \dots, X_n$ by removing any subsets S, T of size at most εn which violate condition 1. If we had a subset S which grossly violated 1 in the sense that $\|\mathrm{Sum}(S)\|_2^2 \geq 100\varepsilon nd$, we could conclude that S contains at least 99% bad samples. This might seem good enough – indeed, a common paradigm in robust statistics is *filtering*, removing samples in way which removes at least as many bad samples as good ones, since any such procedure can ultimately remove at most εn good samples. *However, removing any good samples after looking at all the samples, including the corrupted ones, creates dependencies between good and bad samples, thus breaking obliviousness!*

Sample-Splitting to Preserve Obliviousness. We introduce an *obliviousness-preserving* filter. We:

1. Randomly split $X_1, \dots, X_n$ into U and V.
2. Using only U, identify a set of unit vectors $v_1, \dots, v_\ell \in \mathbb{R}^d$.
3. For all $j \le \ell$, remove from V any X_i such that $|\langle X_i, v_j \rangle| \gg \sqrt{\log n}$, then return V.

The idea is that the returned V will (with high probability) be a set of samples corrupted by a friendly oblivious adversary. The threshold $\sqrt{\log n}$ is chosen so that with high probability no good sample is removed from V. This means that with high probability the scheme preserves obliviousness, since we could have gotten the same outcome by drawing the good samples in V only after performing filtering.[6]

The challenge is ensuring friendliness, which of course rests on the implementation of step 2. In this step, the basic idea is to find a family of subsets $T_1, \dots, T_\ell \subseteq U$ such that, for each $i \in [\ell]$,

- $|T_i| \ll \varepsilon n/(\log n)^{O(1)}$ (here $\ll$ hides constants; the $(\log n)^{O(1)}$ is crucial, as explained below), and
- if we choose $v_i = \mathrm{Sum}(T_i)/\|\mathrm{Sum}(T_i)\|_2$ and remove from U any X_j such that $\langle X_j, v_i \rangle \gg \sqrt{\log n}$, then U satisfies condition 1 of friendliness. If this happens, we'll say that $T_1, \dots, T_m$ "cleans" U.

We need to establish two things: first, that such a family $T_1, \dots, T_\ell$ which cleans U exists, and second, with high probability over the random split U, V, any $T_1, \dots, T_\ell \subseteq \{X_1, \dots, X_n\}$ which cleans U also cleans V. However, these are in tension. For the first, we would like to be able to choose the sets $T_1, \dots, T_\ell$ as large as possible, as this gives more flexibility in the choice of filtering directions and hence makes it easier to clean U. But, for the second, we need tight control over how many different choices of $T_1, \dots, T_\ell$ the cleaning algorithm could make, because we will need to make a union bound over all such choices; the smaller the sets $T_1, \dots, T_\ell$ have to be, the fewer choices there are.

Compression and Small Witnesses. The key idea to balance these concerns is to show that if S_1, S_2 violate θ-friendliness condition 1, then we can compress S_1 to a smaller set S_1' such that removing all $X_i \in S_2$ with $\left\langle X_i, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|_2} \right\rangle$ makes progress in cleaning U, which means we can add S_1' to our list of T_is. The following lemma shows this, as long as $S_1 \cup S_2$ already satisfy λ-friendliness for some $\lambda \gg \theta$ – we will be able to ensure that they already do via induction.

Lemma 1.12 (Small Witness Lemma, Basic Version of Lemma 3.12)**.** *Let $S_1, S_2 \subseteq \mathbb{R}^d$ have $|S_1|, |S_2| = \varepsilon n$ and $\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2) \rangle \ge \varepsilon n \cdot \sqrt{\theta d}$. Suppose $S_1 \cup S_2$ is λ-friendly, for some $\lambda \gg \theta$, and that there is some parameter $C > 0$ such that $|\langle X, X' \rangle| \le \theta\sqrt{d}/C$ and $\|X_i\|^2 = d \pm \theta\sqrt{d}/C$ for all $X, X' \in S_1 \cup S_2$. Then there is $S_1' \subseteq S_1$ with $|S_1'| \le \varepsilon n/C$ and $\Omega(\varepsilon n)$ vectors $X \in S_2$ such that $\left\langle X_i, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|_2} \right\rangle \ge \Omega\left(\sqrt{\frac{\theta}{C\varepsilon n}}\right)$.*

In Lemma 1.12, we think of $\theta \approx \varepsilon n (\log n)^{O(1)}$, so that $\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2) \rangle \ge \varepsilon n \sqrt{\theta d}$ is a violation of friendliness, and $C \approx (\log n)^{O(1)}$ so that S_1' is significantly smaller than S_1. Proving Lemma 1.12 is outside the scope of this overview, but the strategy is to first show that a large number of vectors in S_2 are correlated with $\mathrm{Sum}(S_1)$ (Claim 3.6), and then show this is preserved when we replace S_1 with a random subset $S_1' \subset S_1$. Lemma 1.12 shows that adding S_1' to the list of T_i's will result in removing $\Omega(\varepsilon n)$ vectors; this can only happen $O(1)$ times before all bad samples would be removed, so that we can think of $\ell = O(1)$.

Small Filters Generalize from U to V. Lastly, we need to establish that, if we find a short list of small $T_1, \dots, T_\ell$ which cleans U, then with high probability it also cleans V. Consider the set $\mathcal{T}$ of all possible $(T_1, \dots, T_\ell) \in \binom{n}{\varepsilon n/(\log n)^{O(1)}}^\ell$; note that $|\mathcal{T}| \le 2^{\varepsilon n/(\log n)^{O(1)}}$ because $\ell = O(1)$.

[6] In reality we will perform several rounds of obliviousness-preserving filtering, splitting V again into U', V' and so on; as rounds progress we ensure friendliness for pairs of subsets S, T of increasing size. We will ignore this detail in our technical overview.

Fixing some $(T_1, \dots, T_\ell) \in \mathcal{T}$, our goal is to show that with probability at least $1 - 2^{-\Omega(\varepsilon n)}$ over the random split U, V, if $T_1, \dots, T_\ell$ cleans U then it cleans $[n]$; then we can take a union bound over all of $\mathcal{T}$. By contrapositive, it is enough to show that, if after removing all X_i from $X_1, \dots, X_n$ such that $\langle \mathrm{Sum}(T_j), X_i \rangle \gg \sqrt{\log n}$, some subsets $S_1, S_2 \subseteq [n]$ remain which violate λ-friendliness, then with probability $1 - 2^{-\Omega(\varepsilon n)}$ the random set U also contains some S_1', S_2' which violate θ-friendliness, for some θ not too much less than λ. (This distinction between θ, λ is the origin of the two different friendliness levels in the small witness lemma.)

For the latter, standard concentration arguments show that, with probability $1 - 2^{-\Omega(\varepsilon n)}$, the offending sets S_1, S_2 get split evenly between U and V, and this in turn is enough to show that some subsets of $U \cap S_1, U \cap S_2$ also violate friendliness.

1.4.2 Lower Bounds

Information-Theoretic Lower Bound for Obliviously-Robust Testing. Among our lower bounds, the greatest conceptual innovation lies in our proof that robust mean testing with an oblivious adversary requires $\tilde{\Omega}\Big(\min\Big(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\Big)\Big)$ samples. The remaining terms in the lower bound, $\frac{\sqrt{d}}{\alpha^2}$ and $\frac{d\varepsilon^3}{\alpha^4}$, come respectively from the complexity of non-robust mean testing and from a simpler argument using a Huber adversary, respectively. (The latter we describe below.)

To prove the lower bound, we will describe a distribution over mean vectors μ and adversarial vectors $\{X_i\}_{i \in B}$ such that the joint distribution of $\{X_i\}_{i \in B}$ together with $(1 - \varepsilon)n$ samples from $\mathcal{N}(\mu, I)$ is close in total variation to $\mathcal{N}(0, I)^{\otimes n}$. The key trick in designing this distribution is to *correlate*, but *not perfectly align*, $\mathrm{Sum}(B)$ with $-\mu$. Concretely, we:

1. Draw $X_i \sim \mathcal{N}(0, I)$ for $i \in B$.
2. Draw $\mu = -\beta \,\mathrm{Sum}(B) - z$, where $\beta = \beta(n, d, \varepsilon, \alpha) > 0$ is a suitable constant and $z \sim \mathcal{N}(0, \frac{\alpha^2}{d} I)$.

We show via direct calculation in Section 6 that the χ^2 divergence, and hence total variation distance, between these two distributions on sets of n samples is $o(1)$ so long as $n \ll \tilde{\Omega}\Big(\min\Big(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\Big)\Big)$. The trick above of sampling the corrupted samples $\{X_i\}_{i \in B}$ *before* drawing μ keeps these calculations tractable.

Information-Theoretic Lower Bound for Huber-Robust Testing. Our final information-theoretic lower bound shows that $\Omega(d\varepsilon^3/\alpha^4)$ samples are needed in the presence of a Huber adversary. Here we borrow the lower-bound instance from [DKS17] – the adversary just adds samples from $\mathcal{N}(-\beta \cdot \mu, I)$ for some well-chosen $\beta > 0$. We tighten the analysis of this instance from [DKS17] by using a *conditional* second moment (a.k.a. conditional χ^2 divergence) approach. ([DKS17] use a vanilla χ^2-divergence analysis of their lower bound instance; this method can prove at best a $d\varepsilon^4/\alpha^4$ lower bound, which they obtain.)

Low-Degree Lower Bound for Huber-Robust Testing. Finally, we show a *low-degree* lower bound in the Huber model (essentially equivalent to an SQ lower bound [BBH+20]) using the same instance from [DKS17]; this is a direct computation using now-standard techniques from [KWB22].

1.4.3 A Quadratic-Time Tester

Now we turn to our quadratic-time algorithm for robust mean testing against adaptive adversaries using $\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^2}{\alpha^4}$ (up to logarithmic factors) samples, matching Narayanan's tester. Up to logarithmic factors, our bound matches our low-degree lower bound mentioned above. Together, these bounds give strong evidence that computationally bounded algorithms must pay a factor of $\frac{d\varepsilon^2}{\alpha^4}$ in the sample complexity, and therefore cannot witness the improved rates described elsewhere in this paper, for any model of contamination. Recall that Narayanan's tester requires finding a good-enough subset (Definition 1.8). Since good-enough-ness

involves all subsets of εn samples, even checking whether some $S \subseteq [n]$ is good enough seems to require $n^{\varepsilon n}$ time.

Borrowing a technique from the robust *estimation*, we show that, at least for the good samples $G \subseteq [n]$, there's an efficiently-computable witness to their good-enough-ness. This witness is the top eigenvalue of the covariance matrix $\mathbb{E}_{i\sim G}(X_i - \mathbb{E}_{j\sim G} X_j)(X_i - \mathbb{E}_{j\sim G} X_j)^\top$, together with a uniform upper bound on the magnitude of the row-sums of the Gram matrix of $\{X_i : i \in G\}$.

For illustration here, consider the null case and imagine that $n \leq d$. Then it turns out to be nicer to consider the Gram matrix $M \in \mathbb{R}^{(1-\varepsilon)n \times (1-\varepsilon)n}$ with entries $M_{ij} = \langle X_i, X_j \rangle$; up to zeros it has the same eigenvalues as the covariance. Since $X_i \sim \mathcal{N}(0, I)$ for $i \in G$, we have $M = d \cdot I \pm O(\sqrt{nd})$. If 1_T is the $0/1$ indicator vector for $T \subseteq G$ with $|T| \leq \varepsilon n$, then $1_T^\top M 1_T$ certifies the first part of good-enough-ness:

$$\|\mathrm{Sum}(T)\|_2^2 = 1_T^\top M 1_T = d \cdot \|1_T\|_2^2 \pm O(\sqrt{nd}\|1_T\|_2^2) = |T|d + O(\varepsilon n^{1.5}\sqrt{d})\,.$$

For the second part, note that $\langle \mathrm{Sum}(G \setminus T), \sum(T) \rangle \approx \sum_{i\in T} \sum_{j\neq i} M_{ij}$ is roughly the row-sums of the (off-diagonals of the) matrix M for $i \in T$. Each row sum is at most $\tilde{O}(\sqrt{nd})$, so the sum is $\tilde{O}(\varepsilon n^{1.5}\sqrt{d})$.

These arguments (at least in the case $n \leq d$; $n > d$ is not very different) show that it is enough to find $S \subseteq [n]$ with $|S| = (1-\varepsilon)n$ and whose Gram matrix has eigenvalues $d \pm O(\sqrt{nd})$ and off-diagonal row-sums at most $\tilde{O}(\varepsilon n^{1.5}\sqrt{d})$. In Section 8 we design a filtering algorithm which does this by starting with $[n]$ and iteratively removing samples X_i with large projection onto too-large or small eigenvectors of the Gram matrix, or whose row-sum is too large, until all the row-sums and eigenvalues are as we desire.

2 Preliminaries and Notation

2.1 Basic Definitions

Given two distributions $\mathcal{D}_1, \mathcal{D}_2$, we recall the definitions of total variation distance and χ^2-divergence.

Definition 2.1 (Total Variation Distance)**.** Given two probability distributions $\mathcal{D}_1, \mathcal{D}_2$ over a measurable space $(\Omega, \mathcal{F})$, the *total variation distance* between $\mathcal{D}_1, \mathcal{D}_2$, denoted $\mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2)$, is $\sup_{A\in\mathcal{F}} |\mathcal{D}_1(A) - \mathcal{D}_2(A)|$.

Definition 2.2 (χ^2-divergence)**.** Given two distributions $\mathcal{D}_1, \mathcal{D}_2$ over a measurable space $(\Omega, \mathcal{F})$ with well-defined probability density functions p_1, p_2, the *χ^2-divergence* between $\mathcal{D}_1$ and $\mathcal{D}_2$, denoted $\mathrm{D}_{\chi^2}(\mathcal{D}_1 \| \mathcal{D}_2)$, is given by $\mathbb{E}_{X\sim\mathcal{D}_2}\left(\frac{p_1(X)}{p_2(X)} - 1\right)^2$. (Note that this is not symmetric in $\mathcal{D}_1, \mathcal{D}_2$.)

We recall the following standard relation between total variation distance and χ^2-divergence.

Fact 2.3. *For any distributions $\mathcal{D}_1, \mathcal{D}_2$ with probability density functions,* $\mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2)^2 \leq \frac{1}{4}\mathrm{D}_{\chi^2}(\mathcal{D}_1 \| \mathcal{D}_2)$.

2.2 Useful Probabilistic Tools and Inequalities

In this subsection, we recall several basic but useful concentration inequalities and moment bounds which we will rely on.

Gaussian Concentration. First, we note a well-known proposition regarding univariate Gaussians.

Fact 2.4. *For $X \sim \mathcal{N}(0,1)$ and $a, b \in \mathbb{R}$ such that $b < \frac{1}{2}$, we have that*

$$\mathbb{E}\left[e^{aX+bX^2}\right] = \frac{\exp\left(\frac{a^2}{2-4b}\right)}{\sqrt{1-2b}}.$$

In the special case $a = 0$, this becomes $\mathbb{E}\left[e^{bX^2}\right] = \frac{1}{\sqrt{1-2b}}$.

The following provides a generalization of Fact 2.4 to multivariate Gaussians: we include a proof for completeness.

Proposition 2.5. *Let $z \in \mathbb{R}^d$ be drawn from the Gaussian $\mathcal{N}(0, \delta^2 I)$. Then for parameters $a \geq 0$ and $s \in \mathbb{R}^d$, we have*

$$\mathbb{E}_z\left[\exp\left(-\frac{a}{2}\|z\|_2^2 - \langle s, z\rangle\right)\right] = \frac{1}{(a\delta^2+1)^{d/2}} \exp\left(\frac{\|s\|_2^2}{2(a+1/\delta^2)}\right)$$

Proof. We have

$$\mathbb{E}_z\left[\exp\left(-\frac{a}{2}\|z\|_2^2 - \langle s, z\rangle\right)\right] = \int \frac{1}{\delta^d(2\pi)^{d/2}} \exp\left(-\frac{a}{2}\|z\|_2^2 - \langle s, z\rangle - \frac{\|z\|_2^2}{2\delta^2}\right) dz\,.$$

To compute the integral, we can complete the square in the exponential and write it as

$$-\frac{1}{2}\left\|\sqrt{a+1/\delta^2}z - \frac{1}{\sqrt{a+1/\delta^2}}s\right\|_2 + \frac{\|s\|_2^2}{2(a+1/\delta^2)}$$

so

$$\mathbb{E}_z\left[\exp\left(-\frac{a}{2}\|z\|_2^2 - \langle s, z\rangle\right)\right] = \exp\left(\frac{\|s\|_2^2}{2(a+1/\delta^2)}\right)\left(\frac{1}{\sqrt{a+1/\delta^2}}\right)^d \frac{1}{\delta^d} = \frac{1}{\sqrt{a\delta^2+1}^d} \exp\left(\frac{\|s\|_2^2}{2(a+1/\delta^2)}\right)$$

as desired. □

Proposition 2.6. *Let $z \sim \mathcal{N}(0, I_n)$ be an n-dimensional Gaussian vector. Then, for any symmetric matrix M with all eigenvalues strictly greater than -1,*

$$\mathbb{E}[e^{-\frac{1}{2}\cdot z^\top M z}] = \det(I+M)^{-1/2}.$$

Proof. First, suppose that M is diagonal, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then, $-\frac{1}{2}z^\top M z = -\frac{1}{2}\cdot\sum \lambda_i z_i^2$. Since each coordinate of z_i is independent,

$$\mathbb{E}[e^{-\frac{1}{2}z^\top M z}] = \mathbb{E}\left[\prod_{i=1}^n e^{-\frac{1}{2}\lambda_i z_i^2}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{-\frac{1}{2}\lambda_i z_i^2}\right] = \prod_{i=1}^n \frac{1}{\sqrt{1+\lambda_i}} = \det(I+M)^{-1/2}\,,$$

using Fact 2.4 with $a = 0$. Finally, by rotational symmetry of Gaussians, the claim holds for all symmetric M. □

From Proposition 2.6, we have the following immediate corollary.

Corollary 2.7. *Let X, Y be a 2-dimensional multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix $\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. Then, if $a + c < \frac{1}{2}$, $\mathbb{E}[e^{X^2+Y^2}] = \frac{1}{\sqrt{(1-2a)(1-2c)-4b^2}}$*

Proof. We can write $e^{X^2+Y^2} = e^{x^\top \Sigma x} = e^{-\frac{1}{2}x^\top(-2\Sigma)x}$ for $x \sim \mathcal{N}(0, I)$. Since Σ is PSD and has trace less than $\frac{1}{2}$, its eigenvalues are both less than $\frac{1}{2}$, so -2Σ has all eigenvalues strictly more than -1. Hence, we can apply Proposition 2.6, noting that $\det(1 - 2\Sigma) = (1 - 2a)(1 - 2c) - 4b^2$, which completes the proof. □

Next, we note some basic facts about the norm and inner products of Gaussians.

Fact 2.8. *Consider n points $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn from a Gaussian $\mathcal{N}(\mu, I)$ where $\|\mu\| \le O(1)$. Then, with probability $1 - \delta$, we have for all $i \in [n]$*

$$d - 10\left(\sqrt{\log(n/\delta)d} + \log(n/\delta)\right) \le \|X_i\|^2 \le d + 10\left(\sqrt{\log(n/\delta)d} + \log(n/\delta)\right).$$

In light of Fact 2.8, it suffices to consider when all of the points, including the contaminations (in any of the models) have $\|X_i\|^2 = d \pm O(\sqrt{d} \cdot \operatorname{poly}\log(d, n))$, since we can simply remove all points whose norm is too large or too small: with high probability these points are all contaminated.

Fact 2.9. *Let z_1, z_2 be Gaussians $\mathcal{N}(0, I)$ in $\mathbb{R}^d$. Then, for any $C \le O(\sqrt{d})$, $\mathbb{P}(|\langle z_1, z_2\rangle| \ge C\sqrt{d}) \le 2e^{-\Omega(C^2)}$.*

Proof. First, with probability at least $e^{-\Omega(d)}$, $\|z_2\|_2 \le 2\sqrt{d}$, by Proposition 2.11. Then, conditioned on the norm of z_2, $\langle z_1, z_2\rangle$ has distribution $\mathcal{N}(0, 1) \cdot \|z_2\|_2$, which is at most $C \cdot \|z_2\|_2$ with probability at least $2e^{-C^2/2}$. Hence, $\mathbb{P}(|\langle z_1, z_2\rangle| \ge C\sqrt{d}) \le 2e^{-C^2/8} + e^{-\Omega(d)}$. □

We will also make use of the *Hanson-Wright inequality*.

Lemma 2.10 (Hanson–Wright)**.** *Given an $n \times n$ matrix $A \in \mathbb{R}^{n\times n}$ and an n-dimensional Gaussian vector $Z \sim \mathcal{N}(0, I)$, for any $t \ge 0$,*

$$\mathbb{P}\left(\left|Z^\top AZ - \mathbb{E}[Z^\top AZ]\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_{op}}\right)\right),$$

for some absolute constant $c > 0$. This implies that, with very high probability, $\left|Z^\top AZ - \mathbb{E}[Z^\top AZ]\right| \le \tilde{O}(\|A\|_F)$.

The Hanson-wright inequality with $A = I$, the $n \times n$-dimensional identity vector, immediately implies the following.

Proposition 2.11. *Let $z_1, \ldots, z_n$ be i.i.d. Gaussians. Then, for any $t \ge 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n z_i^2 - n\right| \ge t\right) \le 2\exp\left(-c \cdot \min\left(\frac{t^2}{n}, t\right)\right).$$

We note one more result about Gaussian samples, which follows from well-known facts about sufficient statistics. The following result says that if we know the mean $\bar{X}$ of some Gaussian samples X_i drawn as $\mathcal{N}(\mu, I)$, the posterior distribution of the deviations $X_i - \bar{X}$ does not depend on the mean μ.

Proposition 2.12 ([Nar22, Corollary 18])**.** *For any $\mu \in \mathbb{R}^d$, let $X_1, \ldots, X_n$ be distributed i.i.d. as $\mathcal{N}(\mu, I)$, and let $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, independent of $(X_1, \ldots, X_n)$, and $\bar{Z} = \frac{1}{n}(Z_1 + \cdots + Z_n)$, we have that $X_1, \ldots, X_n$ has the same distribution as $\bar{X} + Z_1 - \bar{Z}, \ldots, \bar{X} + Z_n - \bar{Z}$.*

Hypergeometric distributions. Next, we will require some bounds on Hypergeometric distributions. First, we define a Hypergeometric distribution.

Definition 2.13. For $n \in \mathbb{N}$ and $0 \le k_1, k_2 \le n$, a Hypergeometric distribution $\mathrm{HGeom}(n, k_1, k_2)$ is the distribution of the random variable Y generated as follows. Fix a set $[n]$ of size n, and let S, T be independent random subsets of $[n]$ of size k_1, k_2 respectively. Then, output $Y = |S \cap T|$.

It is well-known that $\mathrm{HGeom}(n, k_1, k_2)$ has expectation $\frac{k_1 \cdot k_2}{n}$. We will also use the following Bernstein-type inequality for Hypergeometric distributions.

Lemma 2.14 ([GW17, Corollary 1])**.** *Suppose that $k_1, k_2 \le \frac{1}{2} \cdot n$, and $X \sim \mathrm{HGeom}(n, k_1, k_2)$. Then, for all $\lambda > 0$,*

$$\mathbb{P}\left(\sqrt{k_1} \cdot \left(\frac{X}{k_1} - \frac{k_2}{n}\right) > \lambda\right) \le \exp\left(-\frac{\lambda^2/2}{(k_2/n) + \lambda/(3\sqrt{k_1})}\right).$$

The following is a direct corollary of Lemma 2.14.

Corollary 2.15. *Suppose that $k_1 = k_2 = \varepsilon n$, and $X \sim \mathrm{HGeom}(n, k_1, k_2)$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\frac{X}{n} - \varepsilon^2 > t\right) \le \exp\left(-\min\left(\frac{t^2 \cdot n}{4\varepsilon^2}, \frac{t \cdot n}{4}\right)\right).$$

We also will utilize the following subgaussian concentration bound for Hypergeometric distributions.

Proposition 2.16 ([Ska13])**.** *If $X \sim \mathrm{HGeom}(n, k_1, k_2)$, then for any $t \ge 0$, $\mathbb{P}[X \ge \mathbb{E}[X] + t \cdot k_1] \le e^{-2t^2 \cdot k_1}$, and $\mathbb{P}[X \ge \mathbb{E}[X] - t \cdot k_1] \le e^{-2t^2 \cdot k_1}$.*

2.3 Simplification of Alternative Hypothesis

We recall that we wish to distinguish between the null hypothesis where $\mu = 0$ and the alternative hypothesis where $\|\mu\|_2 \ge \alpha$. In this subsection, we briefly explain why it suffices to consider a slightly weaker alternative hypothesis of $\alpha \le \|\mu\|_2 \le 2\alpha$. This reduction is very similar to one used in [Nar22, Proposition 23]. We will only describe the reduction for oblivious robust testing, as we will not (directly) need the reduction in the adaptive case.

Proposition 2.17. *Let $0 < \alpha \le O(1)$. Suppose $\mathcal{A}$ is an algorithm that can distinguish between n ε-obliviously contaminated samples from $\mathcal{N}(0, I)$ and n ε-obliviously contaminated samples from $\mathcal{N}(\mu, I)$ where $\|\mu\|_2 \in [\alpha, 2\alpha]$, with probability at least 0.9. Then, there exists an algorithm $\mathcal{A}'$ that can distinguish between $n \cdot \mathrm{polylog}(n, d, \frac{1}{\alpha})$ $\frac{\varepsilon}{\mathrm{polylog}(n,d,\frac{1}{\alpha})}$-obliviously contaminated samples from $\mathcal{N}(0, I)$ and $n \cdot \mathrm{polylog}(n/d)$ $\frac{\varepsilon}{\mathrm{polylog}(n,d,\frac{1}{\alpha})}$-obliviously contaminated samples from $\mathcal{N}(\mu, I)$, where $\|\mu\|_2 \ge \alpha$, with probability at least 0.9.*

Proof. Suppose our dataset of $n \cdot \mathrm{poly}\log(n, d, \frac{1}{\alpha})$ points is called X, which we split into groups $X^{(r,t)}$, where $1 \le r \le R = O(\log nd)$ and $1 \le t \le T = O(\log \frac{n}{\alpha})$, and where $|X^{(r,t)}| = n$. Also, let $X^{(t)} = \bigcup_r X^{(r,t)}$. We can consider conditioning on the location and value of each corrupted point, and then consider drawing the uncorrupted points. Then, if X is $\frac{\varepsilon}{O(\log^2(nd/\alpha))}$-obliviously corrupted, each $X^{(r,t)}$ is ε-obliviously corrupted. Also, conditioned on the indices and values of the corrupted points, the uncorrupted points in each $X^{(r,t)}$ are independent. (For the rest of the proof, we will think of the corrupted indices/values as fixed.)

First, we show an amplification result that on each $X^{(t)}$, we can distinguish between $\mu = 0$ and $\|\mu\|_2 \in [\alpha, 2\alpha]$, with failure probability at most $\frac{1}{nd}$ (instead of failure probability 0.1). For each $X^{(r,t)}$, because the corruptions are oblivious to the data, the probability of $\mathcal{A}$ outputting the right answer on the group $X^{(r,t)}$ is at least 0.9, and is independent across groups (after the above conditioning). So by a Chernoff bound, $\mathcal{A}$ will output the right answer on at least $0.8 \cdot R$ groups with probability at least $1 - \frac{1}{nd}$, for any fixed t. So, the algorithm should simply output the majority across all r.

Next, the same result holds if the alternative hypothesis is $\|\mu\|_2 \in [\alpha \cdot 2^{t-1}, \alpha \cdot 2^t]$ for any $t \geq 1$. To see why, replace each $X_i \in X^{(r,t)}$ with $X_i' := (X_i + \sqrt{2^{2t} - 1} \cdot Z_i)/2^t$, where $Z_j \sim \mathcal{N}(0, I)$ is independent for each X_i. If $X_i \sim \mathcal{N}(\mu, I)$, then $X_i' \sim \mathcal{N}(\mu/2^t, I)$. Moreover, $\{X_i'\}$ is still ε-obliviously corrupted, because $\{Z_i\}$ is chosen as i.i.d. Gaussians independent of the samples.

The algorithm $\mathcal{A}'$ thus works as follows. For each $1 \leq t \leq O(\log(n/d))$, we test on $X^{(t)}$ whether $\mu = 0$ or $\|\mu\|_2 \in [\alpha \cdot 2^{t-1}, \alpha \cdot 2^t]$, with failure probability at most $\frac{1}{nd}$. $\mathcal{A}'$ rejects if any of these tests on $X^{(t)}$ rejects for some $1 \leq t \leq T$. Under the null hypothesis, with at least 0.99 probability, no test will reject. However, if $\|\mu\|_2 \in [\alpha, 10\sqrt{d}]$, then there exists some $0 \leq t \leq O(\log(d/\alpha))$ such that $\|\mu\|_2 \in [\alpha \cdot 2^t, \alpha \cdot 2^{t+1}]$, so the test on $X^{(t)}$ will reject. Finally, we use an additional $O(1)$ randomly chosen points to robustly test whether $\|\mu\|_2 \geq 10\sqrt{d}$, with 0.99 success probability. □

2.4 Notation

We record here several notational conventions.

- In what follows, $\alpha > 0$ is the distance parameter, $\varepsilon \in [0, 1]$ is the corruption rate, d denotes the dimension, and n is the number of samples. In the remainder of the paper, we assume $\alpha \leq O(1)$. In addition, one must have $\varepsilon \leq \alpha$.
- We use $\tilde{O}$, $\tilde{\Omega}$, $\tilde{\Theta}$ to hide polylogarithmic factors in the argument.
- Given a distribution $\mathcal{D}$, we use $p_{\mathcal{D}}(\cdot)$ to represent the corresponding PDF (whenever it is well-defined).
- Throughout this paper, for a set of vectors S, we will use the shorthand $\mathrm{Sum}(S)$ to denote the sum of the vectors in S, *i.e.*, $\mathrm{Sum}(S) := \sum_{x \in S} x$.

Throughout the remainder of the paper, we will assume $\alpha \geq 0.99^d$. We will also assume the desired failure probability $\delta \geq 0.99^d$. Thus, we can also assume that the number of samples $n \leq 100d \log(1/\delta)/\alpha^2 \leq (1.1)^d$ since that would suffice to learn the distribution to accuracy 0.1α.[7]

3 Reducing to "Friendly" Oblivious Contaminations

The first key step in our oblivious upper bound is arguing that we can reduce to when the contaminated points are reasonably behaved. Formally, we want to argue that it suffices to consider a friendly oblivious contamination defined as follows.

Definition 3.1. [(Friendly) Oblivious Contamination Model] We say $X_1, \ldots, X_n$ are obliviously ε-contaminated samples from a distribution $\mathcal{D}$ if they are drawn as follows: first $Y_1, \ldots, Y_{\varepsilon n}$ are chosen adversarially, then $Y_{\varepsilon n+1}, \ldots, Y_n \sim \mathcal{D}$ i.i.d., and finally $Y_1, \ldots, Y_n$ are randomly permuted to produce $X_1, \ldots, X_n$.

[7] Recall that we have a non-robust testing lower bound of $\Omega\big(\sqrt{d}/\alpha^2\big)$ and an efficient robust learning upper bound of $O\big(d/\alpha^2\big)$. In the regime $\alpha \leq 0.99^d$, however, $d/\alpha^2 = O(\sqrt{d}\log(1/\alpha)/\alpha^2) = \tilde{O}(d/\alpha^2)$, so the trivial upper and lower bounds match up to logarithmic factors. For the failure probability δ, note that we can amplify any constant success probability in both the oblivious and Huber contamination models by running multiple trials, at the cost of a multiplicative $O(\log(1/\delta))$ factor.

In the *friendly* oblivious contamination model, we additionally have the following assumption about the data:

Assumption 1. A dataset $X_1, \dots, X_n \in \mathbb{R}^d$ is κ-*friendly* if the following all hold:

1. For any disjoint subsets $S, T \subset [n]$ of sizes $k_1, k_2 \le \varepsilon \cdot n$,
$$\left| \left\langle \sum_{i \in S} X_i, \sum_{i \in T} X_i \right\rangle \right| \le \kappa \cdot (\sqrt{k_1 k_2} \cdot \max(\sqrt{\varepsilon n d}, \varepsilon n)).$$
2. For every distinct $i \neq j \in [n]$, $|\langle X_i, X_j \rangle| \le \kappa \cdot \sqrt{d}$.
3. For every $i \in [n]$, $\|X_i\|_2^2 = d \pm \kappa\sqrt{d}$.

In this definition, one should think of $\kappa = \mathrm{poly}(\log(n), \log(d))$.

Note that we need to make the reduction to friendly oblivious contamination while preserving the "obliviousness" of the contaminated points. Getting the first condition is the main difficulty (the latter two are relatively straight-forward in light of Fact 2.8 and Claim 3.3 below) as natural algorithms for filtering/removing points don't preserve this "obliviousness" and thus cannot be used. Nevertheless in this section, we show how to filter an arbitrary oblivious contamination on a dataset to a friendly oblivious contamination while preserving obliviousness. We will prove the following theorem.

Theorem 3.2 (Dealing with $\widetilde{O}(1)$-Friendly Contamination Suffices)**.** *Assume there exists an algorithm for robust mean testing in $\mathbb{R}^d$ under κ-friendly oblivious ε-contamination that uses $n = f(d, \alpha, \varepsilon)$ samples and succeeds with probability $p > 2/3$ where $\kappa = (10 \log(nd))^{2000}$. Assume $n \le (1.1)^d$. Then there exists an algorithm for robust mean testing in $\mathbb{R}^d$ under (arbitrary) oblivious $\varepsilon/2$-contaimination that succeeds with probability $p - 0.01$ and uses $n\,\mathrm{poly}(\log(nd))$ samples.*

3.1 Structure of Obliviously Contaminated Samples

We begin by proving a few basic structural properties that hold with high probability for an obliviously contaminated dataset. First, we show that the inner product between any two points that are not both contaminated must be small.

Claim 3.3. *Consider a set $S = \{X_1, \dots, X_n\}$ of n points in $\mathbb{R}^d$ that are drawn from $N(\mu, I)$ and then ε-contaminated in the oblivious contamination model. Let $R \subset S$ be the subset of contaminated points. Also assume $\|\mu\| \le 1$ and $n \le (1.1)^d$. Then for any $0.99^d < \delta < 1$, with probability $1 - \delta$, we have that for all $X_i \in S \backslash R, X_j \in S$ with $i \neq j$,*
$$\frac{|\langle X_i, X_j \rangle|}{\|X_i\| \, \|X_j\|} \le \frac{10 \log(n/\delta)}{\sqrt{d}}$$

Proof. Since the contamination is oblivious, we can imagine fixing the index j first and then drawing X_i. We can write $X_i = \mu + v$ where $v \sim N(0, I)$. We have with probability $1 - \delta/(2n^2)$
$$|\langle X_i, X_j / \|X_j\| \rangle| = |\langle \mu, X_j / \|X_j\| \rangle + \langle v, X_j / \|X_j\| \rangle| \le |1 + 5\sqrt{\log(n/\delta)}|$$
where in the last step we simply noted that $\langle v, X_j / \|X_j\| \rangle$ is distributed as a standard Gaussian and the desired inequality follows from standard tail bounds. Also by Fact 2.8, $\|X_i\| \ge \sqrt{d/2}$ with probability at least $1 - \delta/(2n)$ and combining this with the above gives
$$\frac{|\langle X_i, X_j \rangle|}{\|X_i\| \, \|X_j\|} \le \frac{10 \log(n/\delta)}{\sqrt{d}} .$$

Union bounding the failure probability over all i, j we are done. $\square$

We also have the following bound on the number of uncontaminated points with large projection onto any direction determined by a small subset of datapoints.

Claim 3.4. *Consider a set $S = \{X_1, \ldots, X_n\}$ of n points in $\mathbb{R}^d$ that are drawn from $N(\mu, I)$ and then ε-contaminated in the oblivious contamination model. Let $R \subset S$ denote the contaminated points. Also assume $\|\mu\| \leq 1$ and $n \leq (1.1)^d$. Then for any $0.99^d < \delta < 1$, with probability $1 - \delta$, we have the following property: for any subset $T \subset S$,*

$$\left| \left\{ X_i \in S \backslash R \;\middle|\; |\langle X_i, \mathrm{Sum}(T)/\left\|\mathrm{Sum}(T)\right\|\rangle| \geq 10\sqrt{\log(n/\delta)} \right\} \right| \leq 2|T| \,.$$

Proof. We consider a fixed set T and then union bound over all possible choices of T. For a fixed set T, we can imagine fixing the points $X_i \in T$ first and then drawing the remaining points $X_i \in S \backslash (R \cup T) \sim N(\mu, I)$ afterwards. It suffices to upper bound the probability that more than $|T|$ of these points satisfy

$$|\langle X_i, \mathrm{Sum}(T)/\left\|\mathrm{Sum}(T)\right\|\rangle| \geq 10\sqrt{\log(n/\delta)} \,.$$

This probability can be upper bounded by

$$(\delta/n)^{10|T|} \cdot n^{|T|} \leq (\delta/n)^{9|T|}$$

and then union bounding over all possible choices of T gives the desired statement. $\square$

3.2 Oblivious Filtering via Sample Splitting

Recall that our approach to prove Theorem 3.2 will be to "obliviously" filter the dataset, removing some of the contaminated points, so that the remaining data is friendly. In light of Fact 2.8 and Claim 3.3, it is not difficult to enforce the latter two conditions for friendliness since we can simply remove points whose norm is too large or too small and also remove pairs of points whose inner product is too large. The main difficulty lies in enforcing the first condition and this is our focus for the remainder of this section.

It will be convenient to make the following definition.

Definition 3.5. Let S be a set of vectors in $\mathbb{R}^d$. For parameters λ, m, k, we say that S is (λ, m, k)-balanced if for all pairs of disjoint subsets $S_1, S_2 \subset S$ with $|S_1|, |S_2| \leq m$ and $|S_1||S_2| \leq k$, we have

$$|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle| \leq \sqrt{\lambda |S_1||S_2| d}$$

Roughly, it will suffice to ensure that our dataset is balanced for $\lambda \sim \kappa^2 \varepsilon n$, $m \sim \varepsilon n$ and $k \sim m^2$ [8]. We will do this by iteratively doubling k i.e. going from $(\lambda, m, k/2)$-balanced to (λ, m, k)-balanced. At a high-level the way we do this while maintaining obliviousness of the contaminations is as follows. We randomly split the dataset into two parts A, B and only look at A to construct some filter that "cleans" A i.e. makes it (λ, m, k)-balanced. We then argue that with high probability, this filter must clean B and we simply apply it to B and iterate on the remaining data (throwing away A). Crucially, this sample splitting preserves the obliviousness of the contaminations because the filters are constructed independently of the

[8] For most of this section, we will work in the regime $\varepsilon n \,\mathrm{poly}(\log(nd)) < d$. We will show a reduction when we finally prove Theorem 3.2 that allows us to reduce to this case.

uncontaminated data since we can view the uncontaminated points in B as being drawn after running our algorithm on A. See Algorithm 1 and Algorithm 2 for more specific details.

We first need to prove a few basic properties. If a set of vectors S is $(\lambda, m, k/2)$-balanced and not (λ, m, k)-balanced, then there must be some disjoint subsets $S_1, S_2 \subset S$ with $k/2 \leq |S_1||S_2| \leq k$ that witness this i.e.

$$|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle| \geq \sqrt{\lambda |S_1||S_2| d}\,.$$

The above statement says that on average, vectors in S_2 have large inner product with $\mathrm{Sum}(S_1)$. In the next claim, we prove that this is actually the case for a large subset of S_2.

Claim 3.6. *Let S_1, S_2 be two disjoint sets of vectors in $\mathbb{R}^d$. Let k, m be some parameters such that $|S_1|, |S_2| \leq m$. Assume that $S_1 \cup S_2$ is $(\lambda, m, k/2)$-balanced. Then if $|S_1| \cdot |S_2| \leq k$ and*

$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle \geq \sqrt{\theta |S_1||S_2| d}$$

for some parameter $\theta \leq \lambda$, then there is a subset $S_2' \subset S_2$ with $|S_2'| \geq \frac{\theta |S_2|}{8\lambda}$ such that for all $v \in S_2'$,

$$\langle v, \mathrm{Sum}(S_1)\rangle \geq \frac{1}{4}\sqrt{\frac{\theta |S_1| d}{|S_2|}}$$

Proof. Let T be the set of all vectors $v \in S_2$ such that

$$\langle v, \mathrm{Sum}(S_1)\rangle \geq 2\sqrt{\frac{\lambda^2 |S_1|}{\theta |S_2|} d}\,.$$

If T has size larger than $\theta |S_2|/(4\lambda)$, then by taking T' to be a random subset of T of size $\theta |S_2|/(4\lambda)$, we would get

$$\mathbb{E}\left[\langle \mathrm{Sum}(T'), \mathrm{Sum}(S_1)\rangle\right] \geq \frac{\theta |S_2|}{4\lambda} \cdot 2\sqrt{\frac{\lambda^2 |S_1|}{\theta |S_2|} d} = \frac{1}{2}\sqrt{\theta |S_1||S_2| d} \geq \sqrt{\lambda |T'||S_1| d}$$

Hence, the above deterministically happens for some $T' \subset T$ of size $\theta |S_2|/(4\lambda)$, which contradicts the assumption that $S_1 \cup S_2$ is $(\lambda, m, k/2)$-balanced. Thus, we must actually have $|T| \leq \theta |S_2|/(4\lambda)$, and

$$\langle \mathrm{Sum}(T), \mathrm{Sum}(S_1)\rangle \leq \sqrt{\lambda |T||S_1| d} \leq \frac{1}{2}\sqrt{\theta |S_1||S_2| d}\,.$$

In particular, this means that

$$\langle \mathrm{Sum}(S_2 \backslash T), \mathrm{Sum}(S_1)\rangle \geq \frac{1}{2}\sqrt{\theta |S_1||S_2| d}\,.$$

Next, let R be the set of all vectors $v \in S_2$ such that

$$\langle v, \mathrm{Sum}(S_1)\rangle \leq \frac{1}{4}\sqrt{\frac{\theta |S_1| d}{|S_2|}}\,.$$

We have that

$$\langle \mathrm{Sum}(S_2 \backslash (T \cup R)), \mathrm{Sum}(S_1)\rangle \geq \frac{1}{4}\sqrt{\theta |S_1||S_2| d}\,.$$

Thus, by the construction of R, T, we conclude that the number of vectors $v \in S_2$ such that

$$\langle v, \mathrm{Sum}(S_1)\rangle \geq \frac{1}{4}\sqrt{\frac{\theta|S_1|d}{|S_2|}}$$

is at least

$$\frac{\langle \mathrm{Sum}(S_2 \backslash (T \cup R)), \mathrm{Sum}(S_1)\rangle}{2\sqrt{\frac{\lambda^2|S_1|}{\theta|S_2|}d}} \geq \frac{\theta|S_2|}{8\lambda}\,. \qquad \square$$

With the above equipped, we can now show that if $S_1, S_2 \subset S$ are two sets of samples that violate (λ, m, k)-balancedness, then if we split S into two parts A, B, with all but exponentially small (in $\min(|S_1|, |S_2|)$) probability, both parts A, B will witness a violation for slightly smaller values of λ, m, k.

Lemma 3.7. *Let S_1, S_2 be two disjoint sets of vectors in $\mathbb{R}^d$. Let k, m be some parameters such that $|S_1|, |S_2| \leq m$. Assume that $S_1 \cup S_2$ is $(\lambda, m, k/2)$-balanced. Also assume that $|S_1| \cdot |S_2| \leq k$ and*

$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle \geq \sqrt{\lambda|S_1||S_2|d}\,.$$

Consider splitting S_1, S_2 each into two sets $S_{1,A}, S_{1,B}$ and $S_{2,A}, S_{2,B}$ respectively where each element is assigned to the first part independently with probability p. Then with probability $1 - 2^{-\frac{\min(|S_1|,|S_2|)p^3(1-p)^3}{10^{10}}}$, there are subsets $S'_{1,A}, S'_{2,A}, S'_{1,B}, S'_{2,B}$ with $S'_{1,A} \subset S_{1,A}$, $S'_{2,A} \subset S_{2,A}$, $S'_{1,B} \subset S_{1,B}$, $S'_{2,B} \subset S_{2,B}$ such that

$$\begin{aligned}
|S'_{1,A}| &= \frac{p^2|S_1|}{10^6} \\
|S'_{1,B}| &= \frac{(1-p)^2|S_1|}{10^6} \\
|S'_{2,A}| &= \frac{p^2|S_2|}{10^6} \\
|S'_{2,B}| &= \frac{(1-p)^2|S_2|}{10^6} \\
\langle \mathrm{Sum}(S'_{1,A}), \mathrm{Sum}(S'_{2,A})\rangle &\geq \frac{p^4\sqrt{\lambda|S_1||S_2|d}}{10^{13}} \\
\langle \mathrm{Sum}(S'_{1,B}), \mathrm{Sum}(S'_{2,B})\rangle &\geq \frac{(1-p)^4\sqrt{\lambda|S_1||S_2|d}}{10^{13}}
\end{aligned}$$

The proof of Lemma 3.7 relies on the claim below, which characterises what happens when we split one of the sets, say S_2 into two parts.

Claim 3.8. *Let S_1, S_2 be two disjoint sets of vectors in $\mathbb{R}^d$. Let k, m be some parameters such that $|S_1|, |S_2| \leq m$. Assume that $S_1 \cup S_2$ is $(\lambda, m, k/2)$-balanced. Also assume that $|S_1| \cdot |S_2| \leq k$ and*

$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle \geq \sqrt{\theta|S_1||S_2|d}$$

for some parameter $\theta \leq \lambda$. Now consider splitting S_2 into two pieces A, B where each element is independently assigned to A with probability p (and assigned to B otherwise). Then with probability

$1 - 2^{-\frac{p^2(1-p)^2\theta|S_2|}{10^2\lambda}}$, *there exist subsets* $A' \subset A$ *and* $B' \subset B$ *such that*

$$|A'| = \frac{p\theta|S_2|}{20\lambda}$$
$$|B'| = \frac{(1-p)\theta|S_2|}{20\lambda}$$
$$\langle \mathrm{Sum}(A'), \mathrm{Sum}(S_1)\rangle \geq \frac{p\theta}{10^2\lambda}\sqrt{\theta|S_1||S_2|d}$$
$$\langle \mathrm{Sum}(B'), \mathrm{Sum}(S_1)\rangle \geq \frac{(1-p)\theta}{10^2\lambda}\sqrt{\theta|S_1||S_2|d}$$

Proof. First, construct the set S_2' according to Claim 3.6. By Hoeffding's inequality, with probability $1 - 2^{-\frac{p^2(1-p)^2\theta|S_2|}{10^2\lambda}}$, we have $|S_2' \cap A| \geq p\theta|S_2|/(20\lambda)$ and $|S_2' \cap B| \geq (1-p)\theta|S_2|/(20\lambda)$. Let A', B' be arbitrary subsets of $S_2' \cap A, S_2' \cap B$ with sizes $p\theta|S_2|/(20\lambda)$ and $(1-p)\theta|S_2|/(20\lambda)$ respectively. Then by the properties of S_2' guaranteed by Claim 3.6 we have

$$\langle \mathrm{Sum}(A'), \mathrm{Sum}(S_1)\rangle \geq |A'| \cdot \frac{1}{4}\sqrt{\frac{\theta|S_1|d}{|S_2|}} \geq \frac{p\theta}{10^2\lambda}\sqrt{\theta|S_1||S_2|d}$$

and similar for $\langle \mathrm{Sum}(B'), \mathrm{Sum}(S_1)\rangle$, completing the proof. □

We can now prove Lemma 3.7 by applying Claim 3.8 twice.

Proof of Lemma 3.7. First consider when S_2 is split into $S_{2,A}$ and $S_{2,B}$ and apply Claim 3.8 with $\theta = \lambda$. This gives us sets $S_{2,A}^{(1)}$ and $S_{2,B}^{(1)}$ with

$$|S_{2,A}^{(1)}| = \frac{p|S_2|}{20}$$
$$|S_{2,B}^{(1)}| = \frac{(1-p)|S_2|}{20}$$
$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_{2,A}^{(1)})\rangle \geq \frac{p}{10^2}\sqrt{\lambda|S_1||S_2|d}$$
$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_{2,B}^{(1)})\rangle \geq \frac{(1-p)}{10^2}\sqrt{\lambda|S_1||S_2|d}\,.$$

Now we can apply Claim 3.8 again when splitting S_1 with $\theta = p\lambda/10^4$ to get $S_{1,A}'$ with

$$|S_{1,A}'| = \frac{p^2|S_1|}{10^6}$$
$$\langle \mathrm{Sum}(S_{1,A}'), \mathrm{Sum}(S_{2,A}^{(1)})\rangle \geq \frac{p^3}{10^8}\sqrt{\lambda|S_1||S_2|d}\,.$$

Now we can take $S_{2,A}'$ to be random subset of $S_{2,A}^{(1)}$ of size $p^2|S_2|/10^6$ and we have in expectation that

$$\langle \mathrm{Sum}(S_{1,A}'), \mathrm{Sum}(S_{2,A}')\rangle \geq \frac{p^4}{10^{13}}\sqrt{\lambda|S_1||S_2|d}$$

so in particular it holds for some choice of $S_{2,A}'$. We can construct $S_{2,B}'$ similarly. The overall failure probability over all applications of Claim 3.8 is at most $2^{-\frac{p^3(1-p)^3\min(|S_1|,|S_2|)}{10^{10}}}$ and this completes the proof. □

In light of Lemma 3.7, we know that when we split the set of samples S into two parts A, B, any pair of subsets S_1, S_2 that violates (λ, m, k)-balancedness creates a violation in both parts with (approximately) $\exp(-\min(|S_1|, |S_2|))$ failure probability. Now, we roughly proceed as follows. If the set of all possible filters considered by our algorithm has size less than $\exp(\min(|S_1|, |S_2|))$, then we can union bound and conclude that actually any filter that cleans A to be (λ', m', k')-balanced (for some slightly smaller λ', m', k') must actually clean S to be (λ, m, k)-balanced. Then it suffices to argue that there exists a filter in this set that actually cleans A. The full argument will be slightly more involved as we have to deal with different possibilities for $\min(|S_1|, |S_2|)$ separately.

We first need a few more basic observations.

Definition 3.9. Let S be a set of vectors in $\mathbb{R}^d$. We say that S is ρ-bounded if for all $v \in S$, $d - \sqrt{\rho d} \leq \|v\|^2 \leq d + \sqrt{\rho d}$ and for all distinct $u, v \in S$, $-\sqrt{\rho d} \leq \langle u, v \rangle \leq \sqrt{\rho d}$.

Claim 3.10. *Let $S \in \mathbb{R}^d$ be a set of vectors that is (λ, m, k)-balanced and λ-bounded. Then for all subsets $T \subset S$ with $|T| \leq \min(m, \sqrt{k})$, $\|\mathrm{Sum}(T)\|^2 \leq |T|d + 2|T|\sqrt{\lambda d}$.*

Proof. We can write

$$\|\mathrm{Sum}(T)\|^2 \leq |T|(d + \sqrt{\lambda d}) + \sum_{u,v \in T, u \neq v} \langle u, v \rangle .$$

Now consider a random partition of T into two sets T_1, T_2 where each element is assigned uniformly at random. Then

$$\sum_{u,v \in T, u \neq v} \langle u, v \rangle = 2\mathbb{E}[\langle \mathrm{Sum}(T_1), \mathrm{Sum}(T_2) \rangle] \leq |T|\sqrt{\lambda d}$$

where we used the assumption of (λ, m, k)-balancedness. Thus,

$$\|\mathrm{Sum}(T)\|^2 \leq |T|d + 2|T|\sqrt{\lambda d}$$

and we are done. □

Claim 3.11. *Let $S \subset \mathbb{R}^d$ be a set of vectors that is λ/m-bounded. Then it is (λ, m, m)-balanced.*

Proof. Consider disjoint subsets $S_1, S_2 \subset S$. Then by the assumption of λ/m-bounded,

$$|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2) \rangle| \leq |S_1||S_2|\sqrt{\lambda d/m} .$$

If $|S_1||S_2| \leq m$ then the above is at most $\sqrt{\lambda |S_1||S_2| d}$, completing the proof. □

Recall that one key point in the earlier sketch is that our algorithm can only enumerate over a (reasonably) small set of filters. Here we first show that if S_1, S_2 violate balancedness, then there exists a direction determined by a small subset S_1' with $|S_1'| \sim |S_1||S_2|/m$ such that filtering along this direction removes a large ($\sim |S_2|$) number of points. We can then aggregate multiple filtering directions for different choices of S_1' to construct our full filter. Note that bounding the sizes of the individual sets S_1' is the key for bounding the total number of possible filters being considered.

Lemma 3.12. *Let k, m, θ, λ, C be some parameters. Let S_1, S_2 be two disjoint sets of vectors in $\mathbb{R}^d$ with $|S_1|, |S_2| \leq m$, $10Cm \leq |S_1| \cdot |S_2| \leq k$, and*

$$\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2) \rangle \geq \sqrt{\theta |S_1||S_2| d} .$$

Also, assume that $S_1 \cup S_2$ is (λ, m, k)-balanced and $\theta/(10^5 Cm)$-bounded, where $\theta \le \lambda \le d$ and $C \ge 1$. Then, there exists a subset $S_1' \subset S_1$ with $|S_1'| \le |S_1||S_2|/(Cm)$ such that there are at least $\frac{\theta|S_2|}{80\lambda}$ vectors $v \in S_2$ such that

$$\left\langle v, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|} \right\rangle \ge \frac{\sqrt{\theta}}{16\sqrt{Cm}}$$

Proof. First we apply Claim 3.6 to S_2 to get a subset S_2' with the stated properties. Now consider any vector $v \in S_2'$. Consider drawing a random subset $S_1' \subset S_1$ of size $|S_1'| = |S_1||S_2|/(Cm)$ (note this is well defined because $|S_1||S_2| \ge 10Cm$ and $|S_1'| \le |S_1|$). First, we compute

$$\mathbb{E}_{S_1'}[\langle \mathrm{Sum}(S_1'), v\rangle] = \frac{|S_1'|}{|S_1|}\langle v, \mathrm{Sum}(S_1)\rangle \ge \frac{|S_1'|}{4}\sqrt{\frac{\theta d}{|S_1||S_2|}}$$

Next, we can compute the second moment

$$\begin{aligned}
\mathbb{E}_{S_1'}[\langle \mathrm{Sum}(S_1'), v\rangle^2] &= \sum_{u \in S_1} \frac{|S_1'|}{|S_1|}\langle u, v\rangle^2 + \sum_{u,u' \in S_1, u \ne u'} \frac{|S_1'|(|S_1'|-1)}{|S_1|(|S_1|-1)}\langle u, v\rangle\langle u', v\rangle \\
&\le \sum_{u \in S_1} \frac{|S_1'|}{|S_1|}\langle u, v\rangle^2 + \frac{|S_1'|(|S_1'|-1)}{|S_1|(|S_1|-1)} \sum_{u,u' \in S_1} \langle u, v\rangle\langle u', v\rangle \\
&\le |S_1'|\frac{\theta d}{10^5 Cm} + \frac{|S_1'|(|S_1'|-1)}{|S_1|(|S_1|-1)} \left(\langle v, \mathrm{Sum}(S_1)\rangle\right)^2 \\
&\le |S_1'|\frac{\theta d}{10^5 Cm} + (\mathbb{E}_{S_1'}[\langle \mathrm{Sum}(S_1'), v\rangle])^2.
\end{aligned}$$

Thus, the variance is at most $|S_1'|\theta d/(10^5 Cm)$. Now since $|S_1'| = |S_1||S_2|/m$ and $C \ge 1$, we have

$$\mathbb{E}_{S_1'}[\langle \mathrm{Sum}(S_1'), v\rangle] \ge 5\sqrt{\mathsf{Var}_{S_1'}(\langle \mathrm{Sum}(S_1'), v\rangle)}$$

and thus with probability at least 0.1,

$$\langle v, \mathrm{Sum}(S_1')\rangle \ge 0.5\mathbb{E}_{S_1'}[\langle \mathrm{Sum}(S_1'), v\rangle] \ge \frac{|S_1'|}{8}\sqrt{\frac{\theta d}{|S_1||S_2|}}\,.$$

Next note that by the constraints on the parameters, $|S_1'| \le \min(m, \sqrt{k})$ and thus by Claim 3.10,

$$\|\mathrm{Sum}(S_1')\| \le \sqrt{|S_1'|d + 2|S_1'|\sqrt{\lambda d}} \le 2\sqrt{|S_1'|d}$$

which implies that with 0.1 probability over the randomness of the choice of S_1'

$$\left\langle v, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|} \right\rangle \ge \frac{1}{16}\sqrt{\frac{S_1'\theta}{|S_1||S_2|}} \ge \frac{\sqrt{\theta}}{16\sqrt{Cm}}\,.$$

This holds for all $v \in S_2'$ where S_2' was constructed at the beginning of this proof according to Claim 3.6. By linearity of expectation, this means that there is some choice of S_1' such that there are at least

$$0.1|S_2'| \ge \frac{\theta|S_2|}{80\lambda}$$

vectors $v \in S_2$ such that

$$\left\langle v, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|} \right\rangle \geq \frac{\sqrt{\theta}}{16\sqrt{Cm}}$$

as desired. □

We now describe a single iteration of our algorithm (see Algorithm 1) where we take as input a parameter s and the goal is to eliminate all pairs of subsets S_1, S_2 with $s/2 \leq \min(|S_1|, |S_2|) \leq s$ that violate (λ, m, k)-balancedness. Repeating this algorithm over logarithmically many scales for s and then logarithmically many scales for k will give our full algorithm (see Algorithm 2).

Definition 3.13. Given a collection of (finite) sets of vectors in $\mathbb{R}^d$, say $F_1, \ldots, F_\ell$, and a parameter $\gamma \geq 0$, we define $\mathsf{Filter}_\gamma(F_1, \ldots, F_\ell) \subset \mathbb{R}^d$ to consist of all vectors $v \in \mathbb{R}^d$ such that

$$\max_{i \in [l]} |\langle v, \mathrm{Sum}(F_i)/\|\mathrm{Sum}(F_i)\| \rangle| \geq \gamma \,.$$

When we apply $\mathsf{Filter}_\gamma(F_1, \ldots, F_\ell)$ to a set $S \subset \mathbb{R}^d$, we remove from S all points that are in $\mathsf{Filter}_\gamma(F_1, \ldots, F_\ell)$.

Algorithm 1 Single Filtering Iteration

Input: Finite set of samples $S \subset \mathbb{R}^d$
Input: Parameters λ, m, k, s, p
Partition S into two sets A, B where each element is independently assigned to A with probability p
Set $\tau = \frac{sp^6}{10^{11} \log |S|}$
Set $\gamma = \sqrt{\frac{\lambda p^{50}}{10^{100} m}}$
Set $F_1, F_2, \ldots, F_k = \emptyset$
for All collections of subsets $T_1, \ldots T_\ell \subset A$ with $|T_1| + |T_2| + \cdots + |T_\ell| \leq \tau$ **do**
  Set $\mathsf{check} =$ **True**
  for All disjoint pairs $S_1, S_2 \subset A \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ with $|S_1| = p^2 s/(2 \cdot 10^6), |S_2| = p^2 k/(2 \cdot 10^6 s)$ **do**
    if $|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2) \rangle| \geq \frac{p^4 \sqrt{\lambda |S_1||S_2| d}}{10^{14}}$ **then**
      Set $\mathsf{check} =$ **False**
  if check **then**
    Set $F_1 = T_1, \ldots, F_\ell = T_\ell$
    Break
Set $B' = B \backslash \mathsf{Filter}_\gamma(F_1, \ldots, F_\ell)$
Output: B'

Lemma 3.14 (Analysis of Algorithm 1)**.** *Assume that the set S is $(\lambda, m, k/2)$-balanced and $\lambda p^{50}/(10^{100} m)$-bounded. Assume the parameters satisfy $\lambda \leq d, m \leq kp^{20}/10^{50}, p \leq 1/2$. Also, assume that there is a subset $R \subset S$ with $|R| \leq (p^{50} m)/(10^{100} \log |S|)$ such that for any subset $T \subset S$, we have*

$$\left| \left\{ v \in S \backslash R \;\middle|\; |\langle v, \mathrm{Sum}(T)/\|\mathrm{Sum}(T)\| \rangle| \geq \sqrt{\frac{\lambda p^{50}}{10^{100} m}} \right\} \right| \leq 2|T| \,.$$

Then with probability $1 - 2^{-sp^6/10^{11}}$*, the set* B' *output by Algorithm 1 has the property that for any disjoint sets* $S_1, S_2 \subset B'$ *with* $s/2 \le |S_1| \le s, |S_1| \le |S_2| \le m$ *and* $k/2 \le |S_1||S_2| \le k$*,*

$$|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle| \le \sqrt{\lambda |S_1||S_2| d}\,.$$

Proof. Throughout this proof we set $\gamma = \sqrt{\frac{\lambda p^{50}}{10^{100} m}}$ just as in Algorithm 1. We now introduce some terminology. We say that $S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ is unclean if there exist disjoint $S_1, S_2 \subset S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ such that $s/2 \le |S_1| \le s, |S_1| \le |S_2| \le m$ and $k/2 \le |S_1||S_2| \le k$ and

$$|\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle| \ge \sqrt{\lambda |S_1||S_2| d}$$

and otherwise we say that $S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ is clean. Similarly, if $A \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ contains two disjoint sets $S'_{1,A}, S'_{2,A}$ such that

$$\begin{aligned} |S'_{1,A}| &= \frac{p^2 s}{2 \cdot 10^6} \\ |S'_{2,A}| &= \frac{p^2 k}{2 \cdot 10^6 s} \\ |\langle \mathrm{Sum}(S'_{1,A}), \mathrm{Sum}(S'_{2,A})\rangle| &\ge \frac{p^4 \sqrt{\lambda |S_1||S_2| d}}{10^{14}} \end{aligned}$$

then we say $A \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ is unclean and otherwise we say that it is clean.

There are at most $|S|^\tau$ distinct filters considered in Algorithm 1. For each of these filters $\{T_1, \ldots, T_\ell\}$, we apply Lemma 3.7 to $S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$. If it is unclean, then, with probability at least $1 - 2^{-sp^6/10^{10}}$, A is unclean. This is because if S_1, S_2 witness $S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ being unclean, then since $s/2 \le |S_1| \le s$ and $k/(2s) \le |S_2| \le 2k/s$, we can choose random subsets $S'_{1,A}, S'_{2,A}$ of the appropriate size from the sets guaranteed by Lemma 3.7. (Note: we can apply Lemma 3.7 because $|S_1||S_2| \le k$ and $|S_1| \le |S_2| \le m$ by definition of unclean, and since $S \supset S_1 \cup S_2$ is assumed to be $(\lambda, m, k/2)$-balanced.) Now we can union bound this over all $|S|^\tau$ distinct filters and since by the definition of τ,

$$|S|^\tau \le 2^{\frac{sp^6}{5 \cdot 10^{10}}}$$

and thus with probability $1 - 2^{-sp^6/10^{11}}$, we have the following property:

For any $\{T_1, \ldots, T_\ell\}$ if $A \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ is clean then $S \backslash \mathsf{Filter}_\gamma(T_1, \ldots, T_\ell)$ is clean. If Algorithm 1 chooses $F_1, \ldots, F_\ell$ such that $S \backslash \mathsf{Filter}_\gamma(F_1, \ldots, F_\ell)$ is clean then we are done. Thus, it remains to show that there actually exists a filter $F_1, \ldots, F_\ell$ that cleans A.

We construct such a filter iteratively. Start with an empty filter. Now if we are not done, then there must exist a pair $S'_{1,A}, S'_{2,A}$ that witnesses A being unclean. We will apply Lemma 3.12 on this pair (with $\theta \leftarrow \lambda p^4/10^{20}, C = 10^{30}/p^6, k \leftarrow k/2$). First we verify that the conditions of Lemma 3.12 are met. We have

$$|S'_{1,A}||S'_{2,A}| = \frac{p^4 k}{4 \cdot 10^{12}} \ge \frac{10^{37} m}{p^{16}} \ge 10 C m$$

and also clearly $|S'_{1,A}||S'_{2,A}| \le k/2$. Also, $|S'_{1,A}|, |S'_{2,A}| \le m$ since even $|S_1|, |S_2| \le m$. Recall that we have

$$|\langle \mathrm{Sum}(S'_{1,A}), \mathrm{Sum}(S'_{2,A})\rangle| \ge \frac{p^4 \sqrt{\lambda |S_1||S_2| d}}{10^{14}} \ge \sqrt{\theta |S'_{1,A}||S'_{2,A}| d}$$

and also $S'_{1,A} \cup S'_{2,A}$ is $(\lambda, m, k/2)$-balanced by assumption. Finally,

$$\frac{\lambda p^{50}}{10^{100} m} \leq \frac{\theta}{10^5 C m},$$

so the boundedness condition is satisfied, and clearly $\theta \leq \lambda \leq d$ and $C \geq 1$. Thus, Lemma 3.12 tells us that we can find a subset F_1 with $|F_1| \leq p^6 k/(10^{40} m)$ such that

$$\left| \left\{ v \in A \middle| |\langle v, \mathrm{Sum}(F_1)/\left\|\mathrm{Sum}(F_1)\right\|\rangle| \geq \sqrt{\frac{\lambda p^{50}}{10^{100} m}} \right\} \right| \geq \frac{p^6 k}{10^{30} s}.$$

Thus, by our assumption on R, the number of elements in the above set that are in R is at least $\frac{p^6 \cdot k}{10^{30} s} - 2|F_1|$, which is at least $\frac{p^6 k}{2 \cdot 10^{30} s}$, since $s/2 \leq |S_1| \leq m$. In particular, we added at most $p^6 k/(10^{40} m)$ elements to our filter and eliminated at least $\frac{p^6 k}{2 \cdot 10^{30} s}$ elements of R. Now we can iterate the above argument on $A \backslash \mathsf{Filter}_\gamma(F_1)$. Overall, repeating this process, the total number of elements that we will add to our filter is at most

$$\frac{p^6 k}{10^{40} m} \left(\frac{|R|}{\frac{p^6 k}{2 \cdot 10^{30} s}} + 1 \right) \leq \tau.$$

This completes the proof. □

Algorithm 2 Full Sample Splitting

Input: Finite set of samples $S \subset \mathbb{R}^d$
Input: Parameters λ, m, δ
  Set $k = 10^{200} m \log^{100}(|S|m/\delta)$
  Set $S_{\mathsf{filt}} = S$
  while $k \leq m^2$ **do**
    Set $s = 10^{199} \log^{100}(|S|m/\delta)$
    while $s \leq m$ **do**
      Run Algorithm 1 on S with parameters $\lambda, m, k, s, p = 1/(5 \log^2 m)$
      Set $S_{\mathsf{filt}} \leftarrow B'$ where B' is the output of Algorithm 1
      $s \leftarrow 2s$
    $k \leftarrow 2k$
  Output: S_{filt}

Lemma 3.15 (Analysis of Algorithm 2)**.** *Let $S \subset \mathbb{R}^d$ be a finite set of vectors and λ, m, δ be some parameters with $\lambda \leq d$. Assume that S is γ^2-bounded where $\gamma = \sqrt{\frac{\lambda}{10^{200} m \log^{100}(|S|m/\delta)}}$. Also assume that there is a subset $R \subset S$ with $|R| \leq \frac{m}{10^{200} \log^{100}(|S|m/\delta)}$ such that for all subsets $T \subset S$,*

$$\left| \left\{ v \in S \backslash R \;\middle|\; |\langle v, \mathrm{Sum}(T)/\left\|\mathrm{Sum}(T)\right\|\rangle| \geq \gamma \right\} \right| \leq 2|T|.$$

Then if we run Algorithm 2 on S, with probability $1 - \delta$, the output S_{filt} will be (λ, m, m^2)-balanced.

Proof. First by Claim 3.11, we have that S is (λ, m', m')-balanced with $m' = m \cdot 10^{200} \log^{100}(|S|m/\delta)$ (and thus also (λ, m, m')-balanced). Now we prove that after every execution of the outer while loop (for a fixed value of k, before doubling k) in Algorithm 2, the set S_{filt} will be (λ, m, k)-balanced. We do this by induction, where the base case follows from the preceding statement. Now after doubling k, we know that S_{filt} is $(\lambda, m, k/2)$-balanced. Next we apply Lemma 3.14 for each execution of the inner while loop in Algorithm 2. Note that this is valid because $\lambda \le d$, k is initialized sufficiently large, and our upper bound on $|R|$ is sufficiently small. Also s is initialized sufficiently large so we can union bound the failure probability over all iterations and deduce that with probability $1 - \delta$, the conclusion of Lemma 3.14 every time we run Algorithm 1. If after the completion of the inner while loop, the set S_{filt} is not (λ, m, k)-balanced then there must be some disjoint $|S_1|, |S_2|$ with $|S_1|, |S_2| \le m$, $k/2 \le |S_1||S_2| \le k$, and $|\langle \text{Sum}(S_1), \text{Sum}(S_2)\rangle| \ge \sqrt{\lambda |S_1||S_2| d}$. WLOG $|S_1| \le |S_2|$. By the inductive hypothesis, we must have $|S_1||S_2| \ge k/2$ and since $|S_2| \le m$,

$$|S_1| = \frac{|S_1| \cdot |S_2|}{|S_2|} \ge \frac{k/2}{m} \ge 5 \cdot 10^{199} \log^{100}(|S|m/\delta)$$

and thus there was some value of s for which we executed the inner while loop and $s/2 \le |S_1| \le s$. However, applying the guarantee of Lemma 3.14 for this execution of Algorithm 1 implies that such S_1, S_2 cannot exist and this is a contradiction. Thus, actually S_{filt} must be (λ, m, k)-balanced and this completes the induction. Since we keep increasing k up to m^2, at the end we know that S_{filt} is (λ, m, m^2)-balanced and we are done. □

Now we can use Lemma 3.15 to prove Theorem 3.2.

Proof of Theorem 3.2. Consider starting with a set S of $n(10\log(nd))^{10}$ obliviously ε-contaminated samples. First, we remove all points $X_i \in S$ with $\|X_i\|^2 \ge d + \sqrt{(\log(nd))^{100} d}$ or $\|X_i\|^2 \le d - \sqrt{(\log(nd))^{100} d}$. Next, for all pairs of distinct points X_i, X_j with $|\langle X_i, X_j\rangle| \ge \sqrt{(\log(nd))^{200} d}$, we remove both of them. By Fact 2.8 and Claim 3.3, with probability 0.999, this only removes contaminated points. Furthermore, the remaining dataset is equivalent to an obliviously ε-contaminated one (since it is equivalent to first remove the subset of contaminated points that violate the previous conditions and then draw the uncontaminated points).

Case 1: $\varepsilon n \lesssim d$ We first consider the case where $(10\log(nd))^{1000}\varepsilon n \le d$. We run Algorithm 2 with $\delta = 0.001$ and

$$m = \varepsilon n (10\log(nd))^{200}$$
$$\lambda = \varepsilon n (10\log(nd))^{1000}.$$

Recall that Algorithm 2 runs $O(\log^2 m)$ iterations of Algorithm 1. For all executions of Algorithm 1, we have $\gamma \ge (10\log(nd))^{300}$. Also note that we can imagine drawing the uncontaminated points in B after drawing the points in A. On the other hand, the filters are constructed only from A. Thus, with $1 - 1/(nd)^{100}$ probability, none of the uncontaminated points are removed by the filters. We can union bound this failure probability over all executions of Algorithm 1 to get that with probability 0.999, no uncontaminated points are removed by any filters throughout the execution of Algorithm 2. By the construction of S_{filt}, we conclude that with 0.999 probability $|S_{\text{filt}}| \ge n$ and the number of contaminated points in S_{filt} is at most εn. Also, none of the filters constructed throughout Algorithm 2 depend on the points in S_{filt} so it is equivalent to an obliviously ε-contaminated dataset (since it is equivalent to simply apply these filters to the contaminated points before drawing the rest of the dataset). It remains to argue that with high probability, S_{filt} is κ-friendly

and then we can apply the tester that we assumed works under κ-friendly oblivious ε-contamination to complete the proof.

Note that $\lambda = \varepsilon n(10\log(nd))^{1000} \le d$ by assumption and after the initial filtering step (where we filter by norm and pairwise inner product), we know that the dataset is $(\log(nd))^{200}$-bounded. Also, by Claim 3.4, there is a set $R \subset S_{\mathsf{filt}}$ with $|R| \le \varepsilon n$ (consisting of exactly the contaminated points) such that for all subsets $T \subset S_{\mathsf{filt}}$,

$$\left|\left\{v \in S\backslash R \;\middle|\; |\langle v, \mathrm{Sum}(T)/\left\|\mathrm{Sum}(T)\right\|\rangle| \ge 100\log n\right\}\right| \le 2|T|\,.$$

Thus, we can apply Lemma 3.15 to get that S_{filt} is (λ, m, m^2)-balanced. This then implies that S_{filt} is equivalent to a κ-friendly obliviously ε-contaminated dataset and we are done in this case.

Case 2: $\varepsilon n \gtrsim d$ Now it remains to consider the case where $(10\log(nd))^{1000}\varepsilon n \ge d$. We can increase the dimension by adding dummy coordinates to all of the points. We can draw these coordinates independently from $N(0,1)$ and pad the dimension to $d' = (10\log(nd))^{1000}\varepsilon n$. We will use S' to denote the padded dataset and X_i' to denote points in S'. Recall that we filtered by norm and inner product at the beginning. Since all of the additional coordinates are i.i.d. standard Gaussians, with 0.999 probability, we have that after the padding, for all $X_i' \in S'$

$$d' - \sqrt{(\log(nd'))^{100}d'} \le \|X_i'\|^2 \le d' + \sqrt{(\log(nd'))^{100}d'}$$

and for all distinct $X_i', X_j' \in S'$,

$$|\langle X_i', X_j'\rangle| \le \sqrt{(\log nd'))^{200}d'}\,.$$

Now we can run Algorithm 2 as in the previous case on the padded points. By the same argument, we end up with an obliviously ε-contaminated dataset S'_{filt} such that S'_{filt} is $(\log(nd'))^{200}$-bounded and $(\varepsilon n(\log(nd'))^{1000}, \varepsilon n(\log(nd'))^{200}, \varepsilon^2 n^2(\log(nd'))^{400})$-balanced. WLOG say $S'_{\mathsf{filt}} = \{X_1', \ldots, X_n'\}$. Now we take S'_{filt} and remove the padding to get $S_{\mathsf{filt}} = \{X_1, \ldots, X_n\}$. Let Π_{pad} denote the operator that projects onto the padded coordinates. For a set $A \subset [n]$, $\sum_{i\in A}\Pi_{\mathsf{pad}}X_i'$ is just a vector in $\mathbb{R}^{d'-d}$ distributed according to $N(0, |A|I)$. Union bounding over all choices of disjoint sets $A, B \subset [n]$ with $|A|, |B| \le \varepsilon n$ using Fact 2.9, with probability 0.999 over the randomness in the padded coordinates,

$$\left|\left\langle \sum_{i\in A}\Pi_{\mathsf{pad}}X_i', \sum_{i\in B}\Pi_{\mathsf{pad}}X_i'\right\rangle\right| \le \sqrt{|A||B|\varepsilon nd'}(\log(nd'))^{100} \le \sqrt{|A||B|}\varepsilon n(10\log(nd))^{1000}\,.$$

Combining the above and the balancedness of S'_{filt}, we get that after removing the padding for any disjoint sets $A, B \subset [n]$,

$$\left|\left\langle \sum_{i\in A}X_i, \sum_{i\in B}X_i\right\rangle\right| \le \sqrt{|A||B|}\varepsilon n(10\log(nd))^{1000} + \sqrt{\varepsilon n(\log(nd'))^{1000}|A||B|d'} \le \kappa\sqrt{|A||B|}\varepsilon n\,.$$

Thus, S_{filt} is κ-friendly (recall the latter two conditions follow from the filtering by norm and inner product that we did at the beginning) and we are done. □

4 Mean Testing Robustly Against Oblivious Adversaries

In this section, we prove our main technical result, the upper bound against oblivious adversaries:

Theorem 4.1. *Suppose that* $n \geq \tilde{O}\left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4} + \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)$*, and that* $1 \geq \alpha \geq \varepsilon \cdot \log(nd)^{O(1)}$*. Then, there exists an* ε*-robust mean tester using* n *samples.*

4.1 Setup and Algorithm

From Section 3, we may assume we are dealing with the friendly oblivious contamination model. We restate the definition for convenience.

Definition 3.1. [(Friendly) Oblivious Contamination Model] We say $X_1, \ldots, X_n$ are obliviously ε-contaminated samples from a distribution $\mathcal{D}$ if they are drawn as follows: first $Y_1, \ldots, Y_{\varepsilon n}$ are chosen adversarially, then $Y_{\varepsilon n+1}, \ldots, Y_n \sim \mathcal{D}$ i.i.d., and finally $Y_1, \ldots, Y_n$ are randomly permuted to produce $X_1, \ldots, X_n$.

In the *friendly* oblivious contamination model, we additionally have the following assumption about the data:

Assumption 2. A dataset $X_1, \ldots, X_n \in \mathbb{R}^d$ is κ*-friendly* if the following all hold:

1. For any disjoint subsets $S, T \subset [n]$ of sizes $k_1, k_2 \leq \varepsilon \cdot n$,
$$\left|\left\langle \sum_{i \in S} X_i, \sum_{i \in T} X_i \right\rangle\right| \leq \kappa \cdot (\sqrt{k_1 k_2} \cdot \max(\sqrt{\varepsilon n d}, \varepsilon n)).$$
2. For every distinct $i \neq j \in [n]$, $|\langle X_i, X_j \rangle| \leq \kappa \cdot \sqrt{d}$.
3. For every $i \in [n]$, $\|X_i\|_2^2 = d \pm \kappa\sqrt{d}$.

We think of κ as a sufficiently large $\log(nd)^{O(1)}$ term.

Given this promise on the data, the algorithm, roughly speaking, checks the mean and the variance in the direction of the sum of the points. If both look reasonable for a set of samples from the null distribution, we accept, otherwise, we reject. Formally, we use the algorithm described in Algorithm 3.

Algorithm 3 Robust mean tester for obliviously-corrupted data satisfying Assumption 2. Input: $X_1, \ldots, X_n \in \mathbb{R}^d$, $\alpha, \varepsilon > 0$.

Let $\mathbf{S} := \sum_{i \in [n]} X_i$.
if $\left|\|\mathbf{S}\|_2^2 - nd\right| > 0.01\alpha^2 n^2$ **then**
    return REJECT.
else if $\frac{1}{n}\sum_{i \in [n]} \left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \geq 1 + 0.025\frac{\alpha^4}{\varepsilon} \cdot \frac{n}{d}$ **and** $n \leq O(\kappa^5) \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon}{\alpha^2}\right)$ **then**
    return REJECT.
else
    return ACCEPT.

We have the following results which lead to our main theorem.

Lemma 4.2. *Suppose that* $X_1, \ldots, X_n$ *are drawn from the friendly* ε*-oblivious contamination model. Moreover, assume that* $d \geq \varepsilon \cdot n$ *and* $n \leq \frac{d}{\alpha^2}$*. Then, Algorithm 3 can solve robust mean testing in* $n = O(\kappa^5) \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4} + \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$ *samples, whenever* $\alpha \geq \kappa^5 \cdot \varepsilon$*.*

Lemma 4.3. *Suppose that $X_1, \dots, X_n$ are drawn from the friendly ε-oblivious contamination model. Then, Algorithm 3 can solve robust mean testing in $n = O(\kappa^5) \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon}{\alpha^2}\right)$ samples, whenever $\alpha \geq \kappa^5 \cdot \varepsilon$.*

We can think of "solving robust mean testing" (as stated in Lemmas 4.2 and 4.3) as follows. If given n ε-obliviously contaminated samples from $\mathcal{N}(0, I)$, with high probability we either return ACCEPT or the samples were not κ-friendly. Likewise, if given n ε-obliviously contaminated samples from $\mathcal{N}(0, I)$, with high probability we either return REJECT or the samples were not κ-friendly. In this section, we always use the phrase ***with high probability*** to mean the failure probability is at most $\frac{1}{\text{poly}(n,d)}$.

Given Lemmas 4.2 and 4.3, we explain how to prove our main theorem, Theorem 4.1.

Proof of Theorem 4.1. First, we assume that the data was drawn from the κ-friendly ε-oblivious contamination model. Suppose that $n \geq \kappa^6 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4} + \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)$ and $\alpha \geq \kappa^5 \cdot \varepsilon$. Then, if $n \geq \frac{d}{\alpha^2}$, we have $n \geq \kappa^{10} \cdot \frac{d\varepsilon^2}{\alpha^4}$, so we can use Theorem 7.1. Alternatively, if $n \leq \frac{d}{\alpha^2}$ and $d \geq \varepsilon \cdot n$, we can use either Lemma 4.2 or 4.3. Finally, if $d \leq \varepsilon \cdot n$, but $n \geq \kappa^6 \cdot \frac{\sqrt{d}}{\alpha^2}$, then $\kappa^5 \cdot \frac{d\varepsilon}{\alpha^2} = \kappa^5 \cdot \frac{d}{\varepsilon} \cdot \left(\frac{\varepsilon}{\alpha}\right)^2 \leq \kappa^{-5} \cdot \frac{d}{\varepsilon}$. Since $n \geq \kappa^{-1} \cdot \frac{d}{\varepsilon}$, this means that $O(\kappa^5) \cdot \frac{d\varepsilon}{\alpha^2} \leq O(\kappa^{-4}) \cdot n$. Therefore, $n \geq O(\kappa^5) \cdot \frac{\sqrt{d}}{\alpha^2} + O(\kappa^5) \cdot \frac{d\varepsilon}{\alpha^2}$, so we can apply Lemma 4.3.

By Theorem 3.2, we may remove the assumption about the data being friendly. This completes the proof. □

The rest of the section is primarily devoted to Lemma 4.2, but we prove Lemma 4.3 in Section 4.6. By Proposition 2.17, we may assume that the alternative hypothesis is $\|\mu\|_2 \in [\alpha, 2\alpha]$. In fact, for simplicity we will pretend the alternative is $\|\mu\|_2 = \alpha$. Indeed, if $\|\mu\|_2 = \alpha' \in [\alpha, 2\alpha]$, then our proof will show that either $|\|\mathbf{S}\|_2^2 - nd| > 0.01(\alpha')^2 n^2 \geq 0.01\alpha^2 n^2$ or $\frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \geq 1 + 0.025\frac{(\alpha')^4}{\varepsilon} \cdot \frac{n}{d} \geq 1 + 0.025\frac{\alpha^4}{\varepsilon} \cdot \frac{n}{d}$.

In the rest of this section, we use $\mathbf{S}$ to represent $\text{Sum}([n]) = \sum_{i\in[n]} X_i$. We also will split the data into good (uncorrupted) points G and bad (corrupted) points B. We will always use $\mathbf{R}$ to denote $\text{Sum}(B) = \sum_{i\in B} X_i$. If the good samples are drawn as $\mathcal{N}(\mu, I)$, we always use $\mathbf{T}$ to denote $\sum_{i\in G}(X_i - \mu)$, and $\mathbf{Q}$ to denote $|G| \cdot \mu$. Note that $\mathbf{S} = \mathbf{Q} + \mathbf{R} + \mathbf{T}$. Also, note that in the null case, $\mathbf{Q} = 0$ and $\mathbf{T} = \sum_{i\in G} X_i$, which means $\mathbf{S} = \mathbf{R} + \mathbf{T}$.

4.2 Consequences of Assumption 2

In this section, we prove a series of propositions that will be useful in bounding the mean and variance. In all of the following, let $X_1, \dots, X_n \in \mathbb{R}^d$ be any vectors satisfying Assumption 2.

First, we have the following bound on the norm of any sum of at most εn points.

Proposition 4.4. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ satisfy Assumption 2. Then, for any subset B' of size $k \leq \varepsilon n$, $\|\sum_{i\in B'} X_i\|_2^2 = kd \pm O(\kappa) \cdot k \cdot (\sqrt{\varepsilon n d} + \varepsilon n)$.*

Proof. Let B_1', B_2' be a random partition of B' into sets of equal size. For any distinct $i, j \in B'$, let $p = \Pr(i \in B_1', j \in B_2') = \Omega(1)$. Then

$$\left|\sum_{i\neq j\in B'} \langle X_i, X_j\rangle\right| = \left|\frac{1}{p} \cdot \mathbb{E}_{B_1', B_2'}\left\langle \sum_{i\in B_1'} X_i, \sum_{j\in B_2'} X_j\right\rangle\right| \leq \kappa \cdot k \cdot O(\sqrt{\varepsilon n d} + \varepsilon n),$$

by Item 1 of Assumption 2. Finally, $\sum_{i\in B'} \|X_i\|_2^2 = kd \pm k \cdot \kappa\sqrt{d}$. Overall, this means $\|\sum_{i\in B'} X_i\|_2^2 = kd \pm \kappa \cdot k \cdot O(\sqrt{\varepsilon n d} + \varepsilon n)$. □

Next, we have the following proposition.

Proposition 4.5. *Let $X_1, \dots, X_n$ satisfy Assumption 2, with $\varepsilon n \le d$. Let $B \subset [n]$ be any subset of size εn, and let $\mathbf{R} = \sum_{i \in B} X_i$. Let $B' \subset B$ be a subset of size $k \le \varepsilon \cdot n$. Then, $\sum_{i \in B'} \langle X_i, \mathbf{R} \rangle = kd \pm O(\kappa \cdot \varepsilon n \sqrt{kd})$.*

Proof. We have

$$\sum_{i \in B'} \langle X_i, \mathbf{R} \rangle = \left\langle \sum_{i \in B'} X_i, \sum_{i \in B} X_i \right\rangle = \left\| \sum_{i \in B'} X_i \right\|_2^2 + \left\langle \sum_{i \in B'} X_i, \sum_{i \in B \setminus B'} X_i \right\rangle.$$

By Proposition 4.4, we know that $\| \sum_{i \in B'} X_i \|_2^2 = kd \pm O(\kappa \cdot k\sqrt{\varepsilon nd})$. By Item 1 of Assumption 2, we know that $\left| \langle \sum_{i \in B'} X_i, \sum_{i \in B \setminus B'} X_i \rangle \right| \le \kappa \cdot \sqrt{k \cdot \varepsilon n} \cdot \sqrt{\varepsilon nd} = \kappa \cdot \varepsilon n \cdot \sqrt{kd}$. Since $k \le \varepsilon n$, $k\sqrt{\varepsilon nd} \le \varepsilon n \sqrt{kd}$, so overall, we have that $\sum_{i \in B'} \langle X_i, \mathbf{R} \rangle = kd \pm O(\kappa \cdot \varepsilon n \sqrt{kd})$. □

As a result, we have the following.

Proposition 4.6. *Let $X_1, \dots, X_n$ satisfy Assumption 2, with $\varepsilon n \le d$. Then, for any subset $B \subset [n]$ of size εn, we have that $\sum_{i \in B} (\langle X_i, \mathbf{R} \rangle - d)^2 \le O(\kappa^3 \cdot \varepsilon^2 n^2 d)$, where $\mathbf{R} := \sum_{i \in B} X_i$.*

Proof. For each $i \in B$, define $y_i := \langle X_i, \mathbf{R} \rangle - d$. Consider the kth largest y_i. It must be at most $O(\kappa \cdot \varepsilon n \sqrt{d/k})$, or else the sum of the k largest y_i would exceed $O(\kappa \cdot \varepsilon n \sqrt{kd})$, contradicting Proposition 4.5. Likewise, the kth smallest y_i must be greater than or equal to $-O(\kappa \varepsilon n \sqrt{d/k})$, so the kth largest $|y_i|$ is at most $O(\kappa \varepsilon n \sqrt{d/k})$.

This means that

$$\sum_{i \in B} (\langle X_i, \mathbf{R} \rangle - d)^2 = \sum_{i \in B} |y_i|^2 \le \sum_{k=1}^{\varepsilon n} O\left(\kappa \cdot \varepsilon n \sqrt{\frac{d}{k}} \right)^2 = \sum_{k=1}^{\varepsilon n} O\left(\frac{\kappa^2 \varepsilon^2 n^2 d}{k} \right) = O(\kappa^3 \varepsilon^2 n^2 d). \qquad \square$$

We also have the following bound.

Proposition 4.7. *Let $X_1, \dots, X_n$ satisfy Assumption 2, with $\varepsilon n \le d$. Let $B \subset [n]$ have size εn, and let $\mathbf{R} = \sum_{i \in B} X_i$. Then, $\| \sum_{i \in B} (\langle X_i, \mathbf{R} \rangle - d) X_i \|_2 \le O(\kappa^2 \cdot \varepsilon nd)$.*

Proof. Write $y_i := \langle X_i, \mathbf{R} \rangle - d$. We can then write

$$\begin{aligned}
\sum_{i \in B} (\langle X_i, \mathbf{R} \rangle - d) X_i &= \sum_{i \in B} y_i X_i \\
&= \sum_{i \in B \,:\, y_i \ge 0} \int_0^{y_i} X_i \, dt - \sum_{i \in B \,:\, y_i < 0} \int_0^{|y_i|} X_i \, dt \\
&= \underbrace{\int_0^{\infty} \left(\sum_{i \in B : y_i \ge t} X_i \right) dt}_{A_+} - \underbrace{\int_0^{\infty} \left(\sum_{i \in B : y_i \le -t} X_i \right) dt}_{A_-}.
\end{aligned}$$

For simplicity, we will just bound $\|A_+\|_2$, as the argument for bounding $\|A_-\|_2$ is identical. As in Proposition 4.6, for any real number $t \ge 0$, the number of indices $i \in B$ such that $\langle X_i, \mathbf{R} \rangle - d \ge t$ is some $k(t)$ for $t \le O(\kappa \cdot \varepsilon n \sqrt{d/k(t)})$, so $k(t) \le O\left(\frac{\kappa^2 \varepsilon^2 n^2 d}{t^2} \right)$ for all t. In addition, $k(t) \le \varepsilon n$ always, because we are only counting indices in B.

For any $t \geq 0$, define $A_+(t) := \left\|\sum_{i\in B: y_i \geq t} X_i\right\|_2$. By Proposition 4.4 (which we can apply since $k(t) \leq \varepsilon n$),

$$\|A_+(t)\|_2 = O\left(\sqrt{k(t)\cdot d} + \sqrt{\kappa \cdot k(t)}(\varepsilon n d)^{1/4}\right) = O\left(\frac{\kappa \varepsilon n d}{t} + \frac{\kappa^{3/2}(\varepsilon n)^{5/4} d^{3/4}}{t}\right) \leq O\left(\frac{\kappa^{1.5}\varepsilon n d}{t}\right),$$

since we are assuming $d \geq \varepsilon n$.

In addition, for any $t \geq 0$, as $k(t) \leq \varepsilon n$, we have $\|A_+(t)\|_2 \leq O(\sqrt{\varepsilon n d} + \sqrt{\kappa}(\varepsilon n)^{3/4} d^{1/4}) = O(\sqrt{\kappa} \cdot \sqrt{\varepsilon n d})$. Also, note that $A_+(t) = 0$ for $t \geq O(\kappa \varepsilon n \sqrt{d})$, as this will imply $k(t) < 1$, so $k(t) = 0$.

Overall, we have that

$$\begin{aligned}\|A_+\|_2 &\leq \int_0^\infty \|A_+(t)\|_2 dt \\ &\leq \int_0^{\kappa\varepsilon n} O(\sqrt{\kappa \cdot \varepsilon n d}) dt + \int_{\kappa\varepsilon n}^{\kappa \cdot \varepsilon n\sqrt{d}} O\left(\frac{\kappa^{1.5}\varepsilon n d}{t}\right) dt.\end{aligned}$$

The first integral is trivially bounded by $O(\kappa^{1.5} \cdot \varepsilon n \sqrt{\varepsilon n d}) \leq O(\kappa^{1.5} \cdot \varepsilon n d)$, since $\varepsilon n \leq d$. The second integral equals

$$\kappa^{1.5}\varepsilon n d \cdot \log \frac{\kappa \varepsilon n \sqrt{d}}{\kappa \varepsilon n} \leq \kappa^2 \cdot \varepsilon n d.$$

An identical calculation for A_-, combined with the triangle inequality, completes the proof. □

Finally, we will bound the Frobenius norm and operator norm of $\sum X_i X_i^\top$, over any subset of εn points.

Proposition 4.8. *Suppose $B \subset [n]$ has size εn. Then, under Assumption 2, and if $\varepsilon n \leq d$, we have that $\|\sum_{i\in B} X_i X_i^\top\|_F \leq O(\kappa \cdot d\sqrt{\varepsilon n})$.*

Proof. Note that

$$\left\|\sum_{i\in B} X_i X_i^\top\right\|_F^2 = \sum_{i,j\in B} \mathrm{Tr}(X_i X_i^\top X_j X_j^\top) = \sum_{i,j\in B} \langle X_i, X_j\rangle^2.$$

By Item 3 of Assumption 2, we know that for $i = j$, $\langle X_i, X_j\rangle^2 = \|X_i\|_2^4 \leq \kappa^2 d^2$, and by Item 2 of Assumption 2, for each $i \neq j$, $\langle X_i, X_j\rangle^2 \leq \kappa^2 \cdot d$. So, because $|B| = \varepsilon \cdot n \leq d$,

$$\left\|\sum_{i\in B} X_i X_i^\top\right\|_F^2 \leq O(\kappa^2 \cdot \varepsilon n \cdot d^2 + \kappa^2 \cdot (\varepsilon n)^2 \cdot d) = O(\kappa^2 \cdot \varepsilon n \cdot d^2).$$

We take the square root, and the result follows. □

Proposition 4.9. *Suppose $B \subset [n]$ has size εn. Then, under Assumption 2, and if $\varepsilon n \leq d$, we have that $\|\sum_{i\in B} X_i X_i^\top\|_{op} \leq O(\kappa^2 \cdot d)$.*

Proof. Choose a unit vector w. Since $\sum_{i\in B} X_i X_i^\top$ is PSD, it suffices to show that $w^\top \left(\sum_{i\in B} X_i X_i^\top\right) w = \sum_{i\in B}\langle X_i, w\rangle^2$ is at most $O(\kappa^2 \cdot d)$.

First, we consider the kth largest value of $\langle X_i, w\rangle$. For any subset $B' \subset B$ of $k \leq \varepsilon n$ elements, we have $\|\sum_{i\in B'} X_i\|_2 \leq \sqrt{kd + \kappa \cdot k\sqrt{\varepsilon n d}} + \kappa \cdot k \cdot \varepsilon n \leq O(\sqrt{\kappa \cdot kd})$, by Proposition 4.4 and since $\varepsilon n \leq d$. Therefore, $\sum_{i\in B'}\langle X_i, w\rangle = \langle \sum_{i\in B'} X_i, w\rangle \leq O(\sqrt{\kappa \cdot kd})$. This means that the kth largest value of $\langle X_i, w\rangle$ is at most $O(\sqrt{\kappa \cdot d/k})$, and the kth smallest value of $\langle X_i, w\rangle$ is at least $-O(\sqrt{\kappa \cdot d/k})$, which means the kth largest value of $\langle X_i, w\rangle^2$ is at most $O(\kappa \cdot d/k)$. Adding this over $1 \leq k \leq \varepsilon \cdot n$, we have that $\sum_{i\in B}\langle X_i, w\rangle^2 \leq \sum_{k=1}^{\varepsilon n} O\left(\kappa \cdot \frac{d}{k}\right) = O(\kappa^2 \cdot d)$. □

4.3 The Null Case: Mean

In this subsection, we verify that, under the null hypothesis and Assumption 2, with high probability Algorithm 3 not reject on the first step, assuming sufficiently many samples. In this subsection, we do not assume that $\varepsilon n \le d$.

Define $v = \left(\sum_{i\in[n]} X_i\right) / \|\sum_{i\in n} X_i\|_2$ to be the unit vector representing the *direction* of the sum of all points. Also, define $z_i := \langle X_i, v\rangle$ for all $i \le n$. Recall that $G \subset [n]$ represents the set of good (uncorrupted) data points, and $B = [n]\backslash G$ represents the set of bad (corrupted) data points. Note that $|G| = (1-\varepsilon)n$ and $|B| = \varepsilon n$. We now prove the main lemma for this subsection.

Lemma 4.10. *Assume the null hypothesis, meaning that each X_i for $i \in G$ is drawn i.i.d. as $\mathcal{N}(0, I)$. Also, assume $n \ge \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}\right)$ and $\alpha \ge \kappa^5 \cdot \varepsilon$. Then, under Assumption 2, with high probability, $\left\|\sum_{i\in[n]} X_i\right\|_2^2 = nd \pm 0.01\alpha^2 n^2$.*

Proof. We write

$$\left\|\sum_{i\in[n]} X_i\right\|_2^2 = \underbrace{\left\|\sum_{i\in G} X_i\right\|_2^2}_{a} + \underbrace{\left\|\sum_{i\in B} X_i\right\|_2^2}_{b} + 2\cdot \underbrace{\left\langle \sum_{i\in G} X_i, \sum_{i\in B} X_i\right\rangle}_{c}.$$

Using standard concentration, we can write $a = (1-\varepsilon)nd \pm O(\kappa n\sqrt{d})$ with high probability, and using Proposition 4.4, we can write $b = \varepsilon nd \pm \kappa\cdot\varepsilon n \cdot O(\sqrt{\varepsilon nd} + \varepsilon n)$. Finally, we know that the samples X_i for $i \in G$ are drawn independently, from the samples in B, which means that with very high probability, $|c| \le O\left(\frac{\kappa}{\sqrt{d}}\right) \cdot \|\sum_{i\in G} X_i\|_2 \cdot \|\sum_{i\in B} X_i\|_2$. We can bound this as $O\left(\frac{\kappa}{\sqrt{d}} \cdot \sqrt{\kappa nd} \cdot \sqrt{\kappa}(\sqrt{\varepsilon nd} + \varepsilon n)\right) = \kappa^2(\sqrt{\varepsilon n^2 d} + \varepsilon n^{3/2})$, by Proposition 4.4.

Overall, we have that

$$\left\|\sum_{i\in[n]} X_i\right\|_2^2 = nd \pm \kappa^2 \cdot O\left(n\sqrt{d} + (\varepsilon n)^{3/2}\sqrt{d} + \varepsilon^2 n^2 + \sqrt{\varepsilon n^2 d} + \varepsilon n^{3/2}\right).$$

But note that $\sqrt{\varepsilon n^2 d} < n\sqrt{d}$, so the only relevant error terms are the other ones, $n\sqrt{d}$, $(\varepsilon n)^{3/2}\sqrt{d}$, $\varepsilon^2 n^2$, and $\varepsilon n^{3/2}$. By assuming that $n \ge \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}\right)$, we have that the first two error terms $O(\kappa^2 n\sqrt{d})$ and $O(\kappa^2(\varepsilon n)^{3/2} d)$ are each bounded by $0.001\alpha^2 n^2$. The third term $O(\kappa^2 \cdot \varepsilon^2 n^2)$ is at most $0.001\alpha^2 n^2$, assuming that $\alpha \ge \kappa^5 \cdot \varepsilon$. The final term $O(\kappa^2 \cdot \varepsilon n^{3/2})$ is at most $0.001\alpha^2 n^2$ assuming that $n \ge \kappa^5 \cdot \frac{\varepsilon^2}{\alpha^4}$, which is true if $n \ge \kappa^5 \cdot \frac{\sqrt{d}}{\alpha^2}$ and $\varepsilon \le \alpha$. Hence, we have that $\left\|\sum_{i\in[n]} X_i\right\|_2^2 = nd \pm 0.01\alpha^2 n^2$. □

4.4 The Null Case: Variance

In this subsection, we verify that, under the null hypothesis and Assumption 2, with high probability Algorithm 3 does not reject on the second step, assuming sufficiently many samples. Hence, Algorithm 3 accepts. In this and the next subsection, we may additionally assume that $\varepsilon n \le d$ and $n \le \frac{d}{\alpha^2}$. More formally, we make the following assumption in this subsection.

Assumption 3. We make Assumption 2. In addition, we assume that $n \ge \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4} + \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$ and $\alpha \ge \kappa^5 \cdot \varepsilon$. Finally, we also assume that $\varepsilon n \le d$ and $n \le \frac{d}{\alpha^2}$.

We recall that $\mathbf{R} := \sum_{i\in B} X_i$ and $\mathbf{T} = \sum_{i\in G} X_i$. Also, recall that $\mathbf{S} = \sum_{i\in[n]} X_i = \mathbf{R} + \mathbf{T}$, and $v = \mathbf{S}/\|\mathbf{S}\|_2$. We wish to provide an upper bound for

$$\frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i,\mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 = \frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i,\mathbf{R}\rangle + \langle X_i,\mathbf{T}\rangle - d}{\|\mathbf{S}\|_2}\right)^2.$$

First, we consider the bad terms, i.e., we bound

$$\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i,\mathbf{R}\rangle + \langle X_i,\mathbf{T}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 = \frac{1}{n\cdot\|\mathbf{S}\|_2^2}\sum_{i\in B}\left[(\langle X_i,\mathbf{R}\rangle - d)^2 + 2\left(\langle X_i,\mathbf{R}\rangle - d\right)\cdot\langle X_i,\mathbf{T}\rangle + \langle X_i,\mathbf{T}\rangle^2\right].$$

Using the results from Section 4.2, we prove the following bound.

Lemma 4.11. *Assume the null hypothesis. Then, under Assumption 3, $\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i,\mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \le \varepsilon + 0.01\cdot\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d}$ with high probability.*

Proof. First, note that

$$\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)^2 = O(\kappa^3\cdot\varepsilon^2 n^2 d) \tag{4}$$

by Proposition 4.6. Next, we can write $\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)\langle X_i,\mathbf{T}\rangle = \langle \mathbf{T}, \sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i\rangle$. However, by Proposition 4.7, $\|\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i\|_2 \le O(\kappa^2\cdot\varepsilon n d)$ and $\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i$ is independent of $\mathbf{T}$, which is drawn as $\mathcal{N}(0,(1-\varepsilon)nI)$. Therefore, with high probability,

$$\left|\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)\langle X_i,\mathbf{T}\rangle\right| = \left|\left\langle \mathbf{T}, \sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i\right\rangle\right| \le O(\kappa\sqrt{n})\cdot O(\kappa^2\cdot\varepsilon n d) = O(\kappa^3\cdot\varepsilon n^{3/2} d). \tag{5}$$

Finally, we can write $\sum_{i\in B}\langle X_i,\mathbf{T}\rangle^2 = \mathbf{T}^\top\left(\sum_{i\in B}X_iX_i^\top\right)\mathbf{T} = (1-\varepsilon)n\cdot Z^\top\left(\sum_{i\in B}X_iX_i^\top\right)Z$, where Z is a standard Gaussian independent of $\{X_i\}_{i\in B}$. We can then use the Hanson-Wright inequality (Lemma 2.10) to say that with high probability, $\left|Z^\top\left(\sum_{i\in B}X_iX_i^\top\right)Z - \mathrm{Tr}(\sum_{i\in B}X_iX_i^\top)\right| \le O\left(\kappa\cdot\|\sum_{i\in B}X_iX_i^\top\|_F\right)$. We can write $\mathrm{Tr}(\sum_{i\in B}X_iX_i^\top) = \sum_{i\in B}\|X_i\|_2^2 = \varepsilon n d \pm O(\kappa\cdot\varepsilon n\sqrt{d})$, by Item 3 of Assumption 2. So, by using Proposition 4.8, we have that

$$\sum_{i\in B}\langle X_i,\mathbf{T}\rangle^2 = (1-\varepsilon)n\cdot(\varepsilon n d \pm O(\kappa\cdot\varepsilon n\sqrt{d} + \kappa^2\cdot d\sqrt{\varepsilon n})) \le \varepsilon n^2 d + O(\kappa^2\cdot\varepsilon^{1/2}n^{3/2}d), \tag{6}$$

since we are assuming $\varepsilon n \le d$. By combining Equations (4), (5), and (6), we have that

$$\begin{aligned}\sum_{i\in B}(\langle X_i,\mathbf{S}\rangle - d)^2 &= \sum_{i\in B}\left[(\langle X_i,\mathbf{R}\rangle - d)^2 + 2\left(\langle X_i,\mathbf{R}\rangle - d\right)\cdot\langle X_i,\mathbf{T}\rangle + \langle X_i,\mathbf{T}\rangle^2\right]\\ &\le \varepsilon n^2 d + \kappa^3\cdot O(\varepsilon^2 n^2 d + \varepsilon^{1/2}n^{3/2}d).\end{aligned}$$

As we are assuming that $n \ge \kappa^5\cdot\frac{d\varepsilon^3}{\alpha^4}$, this implies that $O(\kappa^3\cdot\varepsilon^2 n^2 d) \le 0.001\cdot\frac{\alpha^4 n^3}{\varepsilon}$. Moreover, we are assuming that $n \ge \kappa^5\cdot\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}} \ge \kappa^5\cdot\frac{d^{2/3}\varepsilon}{\alpha^{8/3}}$, which implies that $O(\kappa^3\cdot\varepsilon^{1/2}n^{3/2}d) \le 0.001\cdot\frac{\alpha^4 n^3}{\varepsilon}$. In summary,

$$\sum_{i\in B}(\langle X_i,\mathbf{S}\rangle - d)^2 \le \varepsilon n^2 d + 0.002\cdot\frac{\alpha^4}{\varepsilon}\cdot n^3. \tag{7}$$

Next, we note that $\|\mathbf{S}\|_2^2 = nd \pm 0.01\alpha^2 n^2 = nd \cdot \left(1 \pm 0.01\frac{\alpha^2 n}{d}\right)$, using Lemma 4.10. As a result, as we are assuming that $n \le \frac{d}{\alpha^2}$, then the reciprocal of $1 \pm 0.01\frac{\alpha^2 n}{d}$ is in the range $1 \pm 0.02\frac{\alpha^2 n}{d}$. Therefore, by (7),

$$\begin{aligned}
\frac{1}{n}\sum_{i\in B}\left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 &\le \frac{1}{n^2 d}\cdot\left(1 + 0.02\frac{\alpha^2 n}{d}\right)\cdot\left(\varepsilon n^2 d + 0.002\cdot\frac{\alpha^4}{\varepsilon}\cdot n^3\right) \\
&= \left(1 + 0.02\frac{\alpha^2 n}{d}\right)\cdot\varepsilon\cdot\left(1 + 0.002\cdot\frac{\alpha^4}{\varepsilon^2}\cdot\frac{n}{d}\right) \\
&\le \varepsilon\cdot\left(1 + 0.01\cdot\frac{\alpha^4}{\varepsilon^2}\cdot\frac{n}{d}\right) \\
&= \varepsilon + 0.01\cdot\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d},
\end{aligned}$$

where the penultimate line uses the fact that $\frac{\alpha^2 n}{d} < \min\left(1, 0.1\frac{\alpha^4}{\varepsilon^2}\cdot\frac{n}{d}\right)$ since we are assuming that $n \le \frac{d}{\alpha^2}$ and $\alpha \ge 10\varepsilon$. □

Next, we deal with the sum over good points.

Lemma 4.12. *Assume the null hypothesis. Then, under Assumption 3, with high probability,*

$$\frac{1}{n}\sum_{i\in G}\left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \le 1 - \varepsilon + 0.01\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d}.$$

Proof. Recall that $\mathbf{S} = \mathbf{R} + \mathbf{T}$, and suppose $\mathbf{R}, \mathbf{T}$ are fixed. Then, by Proposition 2.12, the posterior distribution of $\{X_i\}_{i\in G}$ conditioned on $\mathbf{R}$ and $\mathbf{T}$ is $\left\{\frac{\mathbf{T}}{(1-\varepsilon)n} + Y_i - \bar{Y}\right\}$, where $\{Y_i\}_{i\in G}$ are i.i.d. $\mathcal{N}(0, I)$, independent of $(\mathbf{R}, \mathbf{T})$, and $\bar{Y} = \frac{1}{(1-\varepsilon)n}\sum_{i\in G} Y_i$. As a result, the posterior distribution of $\{\langle X_i, \mathbf{S}\rangle - d\}_{i\in G}$ is $\left\{\frac{\langle \mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon)n} - d + \|\mathbf{S}\|_2\cdot(z_i - \bar{z})\right\}$, where $\{z_i\}_{i\in G}$ are i.i.d. univariate $\mathcal{N}(0,1)$, and $\bar{z} = \frac{1}{(1-\varepsilon)n}\sum_{i\in G} z_i$. Hence, we can rewrite the desired sum over good points as

$$\frac{1}{n}\sum_{i\in G}\left(\frac{\frac{\langle \mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon n)} - d}{\|\mathbf{S}\|_2} + (z_i - \bar{z})\right)^2.$$

Now, note that $\langle \mathbf{T}, \mathbf{S}\rangle = \|\mathbf{T}\|_2^2 + \langle \mathbf{T}, \mathbf{R}\rangle$. Since $\mathbf{T} \sim \mathcal{N}(0, (1-\varepsilon)nI)$ is independent of $\mathbf{R}$, and $\|\mathbf{R}\|_2 \le O(\kappa\sqrt{\varepsilon n d})$ by Proposition 4.4 and the assumption that $\varepsilon n \le d$, we have that $|\langle \mathbf{T}, \mathbf{R}\rangle| \le O(\kappa^2\sqrt{\varepsilon n^2 d})$ with high probability. In addition, $\|\mathbf{T}\|_2^2 = (1-\varepsilon)nd \pm O(\kappa\cdot n\sqrt{d})$ with high probability, as $\mathbf{T}$ is the sum of the uncorrupted samples. In sum, $\langle \mathbf{T}, \mathbf{S}\rangle = (1-\varepsilon)nd \pm O(\kappa^2\cdot n\sqrt{d})$. Therefore, $\left|\frac{\langle \mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon)n} - d\right| \le O(\kappa^2\sqrt{d})$. Since $\|\mathbf{S}\|_2^2 = nd\left(1 \pm 0.01\frac{\alpha^2 n}{d}\right) = \Theta(nd)$ as we are assuming $n \le \frac{d}{\alpha^2}$, this means $\left(\frac{\langle \mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon)n} - d\right)/\|\mathbf{S}\|_2 = \pm O\left(\frac{\kappa^2}{\sqrt{n}}\right)$.

Next, defining $\tilde{z}_i := z_i - \bar{z}$, we have $\sum_{i\in G}\tilde{z}_i^2 = \sum_{i\in G} z_i^2 - (1-\varepsilon)n\cdot\bar{z}^2$. Clearly, $\bar{z} \sim \mathcal{N}(0, \frac{1}{(1-\varepsilon)n})$, so $|\bar{z}| \le \kappa/\sqrt{n}$ with high probability. Thus, $\sum_{i\in G}\tilde{z}_i^2 = \sum_{i\in G} z_i^2 - (1-\varepsilon)n\cdot\bar{z}^2 = (1-\varepsilon)n \pm O(\kappa\sqrt{n} + \kappa^2) =$

$(1-\varepsilon)n \pm O(\kappa^2\sqrt{n})$, by Proposition 2.11. Hence, because the average of $\tilde{z}_i$ over $i \in G$ is 0,

$$\begin{aligned}\frac{1}{n}\sum_{i\in G}\left(\frac{\frac{\langle \mathbf{T},\mathbf{S}\rangle}{(1-\varepsilon n)}-d}{\|\mathbf{S}\|_2}+\tilde{z}_i\right)^2 &= \frac{1}{n}\cdot\sum_{i\in G}\tilde{z}_i^2+(1-\varepsilon)\cdot\left(\frac{\frac{\langle \mathbf{T},\mathbf{S}\rangle}{(1-\varepsilon n)}-d}{\|\mathbf{S}\|_2}\right)^2\\ &= \frac{1}{n}\cdot\sum_{i\in G}\tilde{z}_i^2+(1-\varepsilon)\cdot O\left(\frac{\kappa^2}{\sqrt{n}}\right)^2\\ &= (1-\varepsilon)\pm O\left(\frac{\kappa^4}{\sqrt{n}}\right).\end{aligned}$$

As long as $n \geq \kappa^5\cdot\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$ the error term is at most $0.01\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d}$, which completes the proof. □

By combining Lemmas 4.11 and 4.12, we have the following.

Lemma 4.13. *Assume the null hypothesis. Then, under Assumption 3,* $\frac{1}{n}\sum_{i=1}^n\left(\frac{\langle X_i,\mathbf{S}\rangle-d}{\|\mathbf{S}\|_2}\right)^2 \leq 1+0.02\cdot\frac{\alpha^4 n}{\varepsilon d}$, *with high probability.*

4.5 The Alternative Case: Variance

Let $\mu = \alpha\cdot v$, where v is a unit vector. Recall that $\mathbf{Q} = (1-\varepsilon)n\cdot\alpha v$, $\mathbf{R} = \sum_{i\in B}X_i$, and $\mathbf{T} = \sum_{i\in G}(X_i-\alpha v) = (\sum_{i\in G}X_i)-\mathbf{Q}$. Let $\mathbf{S} = \mathbf{Q}+\mathbf{R}+\mathbf{T} = \sum_{i\in[n]}X_i$. We wish to bound

$$\frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i,\mathbf{S}\rangle-d}{\|\mathbf{S}\|_2}\right)^2 = \frac{1}{n}\sum_{i\in[n]}\left(\frac{\langle X_i,\mathbf{Q}+\mathbf{R}\rangle+\langle X_i,\mathbf{T}\rangle-d}{\|\mathbf{S}\|_2}\right)^2.$$

We can again split $[n]$ into bad and good points. For the bad points B, our goal is to bound

$$\frac{1}{n}\cdot\sum_{i\in B}\left(\frac{\langle X_i,\mathbf{S}\rangle-d}{\|\mathbf{S}\|_2}\right)^2 = \frac{1}{n\cdot\|\mathbf{S}\|_2^2}\cdot\sum_{i\in B}\left[\langle X_i,\mathbf{T}\rangle^2+2\langle X_i,\mathbf{T}\rangle\cdot(\langle X_i,\mathbf{Q}+\mathbf{R}\rangle-d)+(\langle X_i,\mathbf{Q}+\mathbf{R}\rangle-d)^2\right].$$

Before doing so, we will consider the relationship between the values $\mathbf{R},\mathbf{Q},\mathbf{T}$. Note that $\mathbf{T}$ is independent of both $\mathbf{R}$ and $\mathbf{Q}$, whereas $\mathbf{R}$ may depend on $\mathbf{Q}$.

Proposition 4.14. *Suppose that $X_1,\dots,X_n$ satisfy Assumption 2, that $\varepsilon n \leq d$, and Algorithm 3 does not reject in Line 3. Then, $\|\mathbf{R}\|_2^2 = \varepsilon nd \pm O(\kappa\cdot(\varepsilon n)^{3/2}\sqrt{d})$, and $\|\mathbf{Q}+\mathbf{R}\|_2^2 = \varepsilon nd \pm 0.01\alpha^2n^2 \pm \kappa^2\cdot O\left(n\sqrt{d}+\alpha n^{3/2}\right)$.*

Proof. By Proposition 4.4, we know that $\|\mathbf{R}\|_2^2 = \|\sum_{i\in B}X_i\|_2^2 = \varepsilon nd\pm O(\kappa\cdot(\varepsilon n)^{3/2}\sqrt{d})$. Next, because Algorithm 3 did not reject in Line 3, we have that $\|\mathbf{Q}+\mathbf{R}+\mathbf{T}\|_2^2 = \|\sum_{i\in[n]}X_i\|_2^2 = nd\pm 0.01\alpha^2n^2$. However, we can write $\|\mathbf{Q}+\mathbf{R}+\mathbf{T}\|_2^2 = \|\mathbf{Q}+\mathbf{R}\|_2^2+\|\mathbf{T}\|_2^2+2\langle\mathbf{Q}+\mathbf{R},\mathbf{T}\rangle$. Let $A = \|\mathbf{Q}+\mathbf{R}\|_2$. Then, with high probability, $\|\mathbf{T}\|_2^2 = (1-\varepsilon)nd\pm O(\kappa n\sqrt{d})$ and $\langle\mathbf{Q}+\mathbf{R},\mathbf{T}\rangle = \pm O(\kappa A\sqrt{n})$, since $\mathbf{T}\sim\mathcal{N}(0,(1-\varepsilon)nI)$ is independent of $\mathbf{Q}+\mathbf{R}$. This means $nd\pm 0.01\alpha^2n^2 = A^2+(1-\varepsilon)nd\pm O(\kappa n\sqrt{d})\pm O(\kappa\cdot A\sqrt{n})$, so $A^2 = \varepsilon nd\pm 0.01\alpha^2n^2\pm O(\kappa n\sqrt{d})\pm O(\kappa\cdot A\sqrt{n})$. Finally, we know that $A\leq\|\mathbf{Q}\|_2+\|\mathbf{R}\|_2\leq\alpha n+\sqrt{\kappa\cdot\varepsilon nd}$, which means

$$\|\mathbf{Q}+\mathbf{R}\|_2^2 = A^2 = \varepsilon nd\pm 0.01\alpha^2n^2\pm O(\kappa^2)\cdot(n\sqrt{d}+\alpha n^{3/2}).$$ □

Hence, we have the following corollary.

Proposition 4.15. *Suppose that $X_1, \dots, X_n$ satisfy Assumption 2, $\varepsilon \le 0.5$, , and Algorithm 3 does not reject in Line 3. Then, if $n \ge \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}\right)$, $\left|\|\mathbf{R}\|_2^2 - \|\mathbf{Q}+\mathbf{R}\|_2^2\right| \le 0.1 \cdot \|\mathbf{Q}\|_2^2$ with high probability.*

Proof. As a direct consequence of Proposition 4.14,

$$\left|\|\mathbf{R}\|_2^2 - \|\mathbf{Q}+\mathbf{R}\|_2^2\right| \le 0.01\alpha^2 n^2 + \kappa^2 \cdot O\left((\varepsilon n)^{3/2}\sqrt{d} + n\sqrt{d} + \alpha n^{3/2}\right). \tag{8}$$

Assuming that $n \ge \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}\right)$, each error term is at most $0.01\alpha^2 n^2$, so (8) is at most $0.05\alpha^2 n^2$. Since $\|\mathbf{Q}\|_2 = \alpha(1-\varepsilon)n \ge 0.5\alpha n$, this means that $\left|\|\mathbf{R}\|_2^2 - \|\mathbf{Q}+\mathbf{R}\|_2^2\right| \le 0.1 \cdot \|\mathbf{Q}\|_2^2$. □

We now turn to bounding the sum for the bad points B.

Lemma 4.16. *Suppose that Algorithm 3 does not reject in Line 3, and $\varepsilon \le 0.1$. Then, with high probability under the alternative hypothesis and Assumption 3,*

$$\sum_{i\in B} (\langle X_i, \mathbf{S}\rangle - d)^2 \ge \varepsilon n^2 d \left(1 + \frac{0.05\alpha^4 n}{\varepsilon^2 d}\right).$$

Proof. We can rewrite the left-hand side of the above expression as

$$\sum_{i\in B} \left((\langle X_i, \mathbf{Q}+\mathbf{R}\rangle - d)^2 + \langle X_i, \mathbf{T}\rangle^2 + 2(\langle X_i, \mathbf{Q}+\mathbf{R}\rangle - d)\cdot\langle X_i, \mathbf{T}\rangle\right)$$

First, we consider $\sum_{i\in B}(\langle X_i, \mathbf{Q}+\mathbf{R}\rangle - d)^2$. Since $\sum_{i\in B} X_i = \mathbf{R}$, by Jensen's inequality this is at least $\varepsilon n \cdot \left(\langle \frac{\mathbf{R}}{\varepsilon n}, \mathbf{Q}+\mathbf{R}\rangle - d\right)^2 = \frac{1}{\varepsilon n}\cdot(\langle \mathbf{R}, \mathbf{Q}+\mathbf{R}\rangle - \varepsilon n d)^2$. However, we can write $\langle \mathbf{R}, \mathbf{Q}+\mathbf{R}\rangle = \frac{\|\mathbf{Q}+\mathbf{R}\|_2^2 + \|\mathbf{R}\|_2^2 - \|\mathbf{Q}\|_2^2}{2}$, and since $\|\mathbf{R}\|_2^2 \le \|\mathbf{Q}+\mathbf{R}\|_2^2 + 0.1\|\mathbf{Q}\|_2^2$ by Proposition 4.15, this means that $\langle \mathbf{R}, \mathbf{Q}+\mathbf{R}\rangle \le \|\mathbf{Q}+\mathbf{R}\|_2^2 - 0.45\|\mathbf{Q}\|_2^2$. By Proposition 4.14, we have $\|\mathbf{Q}+\mathbf{R}\|_2^2 = \varepsilon n d \pm 0.01\alpha^2 n^2 \pm O(\kappa^2)\cdot\left(n\sqrt{d} + \alpha n^{3/2}\right)$. Therefore, since $\|\mathbf{Q}\|_2^2 = \alpha^2(1-\varepsilon)^2 n^2 \ge 0.8\alpha^2 n^2$ as $\varepsilon \le 0.1$, this means that

$$\langle \mathbf{R}, \mathbf{Q}+\mathbf{R}\rangle - \varepsilon n d \le 0.01\alpha^2 n^2 + O(\kappa^2)\cdot\left(n\sqrt{d} + \alpha n^{3/2}\right) - 0.36\alpha^2 n^2 = -0.35\alpha^2 n^2 + O(\kappa^2)\cdot\left(n\sqrt{d} + \alpha n^{3/2}\right).$$

Therefore,

$$\begin{aligned}
\sum_{i\in B}(\langle X_i, \mathbf{Q}+\mathbf{R}\rangle - d)^2 &\ge \frac{1}{\varepsilon n}\cdot\left(0.35\alpha^2 n^2 - O(\kappa^2)\cdot(n\sqrt{d} + \alpha n^{3/2})\right)^2 \\
&\ge 0.1\frac{\alpha^4}{\varepsilon} n^3 - O(\kappa^2)\cdot\left(\frac{\alpha^2}{\varepsilon}\cdot n^2\sqrt{d} + \frac{\alpha^3}{\varepsilon} n^{5/2}\right) \\
&\ge 0.08\cdot\frac{\alpha^4}{\varepsilon} n^3.
\end{aligned} \tag{9}$$

Above, the second inequality follows because $(A-B)^2 \ge A^2 - 2AB$ for any real A, B, and the last inequality follows because the two error terms are each at most $0.01\frac{\alpha^4}{\varepsilon} n^3$ if $n \ge \kappa^5 \cdot \frac{\sqrt{d}}{\alpha^2}$.

To bound $\sum_{i\in B}\langle X_i,\mathbf{T}\rangle^2$, we can write this as $\mathbf{T}^\top\left(\sum_{i\in B}X_iX_i^\top\right)\mathbf{T} = (1-\varepsilon)n\cdot Z^\top\left(\sum_{i\in B}X_iX_i^\top\right)Z$, where $Z\sim\mathcal{N}(0,I)$ is independent of $\{X_i\}_{i\in B}$. We apply Hanson-Wright (Lemma 2.10) along with Proposition 4.8, to say that with high probability, $\left|Z^\top\left(\sum_{i\in B}X_iX_i^\top\right)Z - \mathrm{Tr}(\sum_{i\in B}X_iX_i^\top)\right| \le O\left(\kappa\cdot\|\sum_{i\in B}X_iX_i^\top\|_F\right) \le O(\kappa^2\cdot d\sqrt{\varepsilon n})$. In addition, $\mathrm{Tr}(\sum_{i\in B}X_iX_i^\top) = \sum_{i\in B}\|X_i\|_2^2 = \varepsilon nd \pm \kappa\cdot\varepsilon n\sqrt{d}$. Therefore, since $\varepsilon n\le d$,

$$\sum_{i\in B}\langle X_i,\mathbf{T}\rangle^2 = (1-\varepsilon)n\cdot\left(\varepsilon nd \pm O(\kappa^2\cdot d\sqrt{\varepsilon n} + \kappa\cdot\varepsilon n\sqrt{d})\right) = \varepsilon n^2 d - \varepsilon^2 n^2 d \pm O\left(\kappa^2\cdot\varepsilon^{1/2}n^{3/2}d\right). \quad (10)$$

To bound the final term $\sum_{i\in B}(\langle X_i,\mathbf{Q}+\mathbf{R}\rangle - d)\langle X_i,\mathbf{T}\rangle$, we first bound $\|\sum_{i\in B}(\langle X_i,\mathbf{Q}+\mathbf{R}\rangle - d)X_i\|_2$. We can use Proposition 4.7 to obtain that $\|\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i\|_2 \le O(\kappa^2\varepsilon nd) \le O(\kappa^2\alpha nd)$, as we assumed that $\varepsilon\le\alpha$. Next, to bound $\|\sum_{i\in B}\langle X_i,\mathbf{Q}\rangle X_i\|_2$, we can write $\sum_{i\in B}\langle X_i,\mathbf{Q}\rangle X_i = \left(\sum_{i\in B}X_iX_i^\top\right)\cdot\mathbf{Q}$, which has norm at most $\left\|\sum_{i\in B}X_iX_i^\top\right\|_{op}\cdot\|\mathbf{Q}\|_2 \le O(\kappa^2\cdot d\cdot\alpha n)$, using Proposition 4.9. Therefore, since $\{X_i\}_{i\in B}$ and $\mathbf{Q}$ are independent of $\mathbf{T}\sim\mathcal{N}(0,(1-\varepsilon)nI)$, with high probability we have that

$$\begin{aligned}\left|\sum_{i\in B}(\langle X_i,\mathbf{Q}+\mathbf{R}\rangle - d)\cdot\langle X_i,\mathbf{T}\rangle\right| &= \left|\left\langle\mathbf{T},\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i + \sum_{i\in B}\langle X_i,\mathbf{Q}\rangle X_i\right\rangle\right| \\ &\le O(\kappa\sqrt{n})\cdot\left(\left\|\sum_{i\in B}(\langle X_i,\mathbf{R}\rangle - d)X_i\right\|_2 + \left\|\sum_{i\in B}\langle X_i,\mathbf{Q}\rangle X_i\right\|_2\right) \\ &\le O(\kappa^3\cdot\alpha n^{3/2}d). \end{aligned} \quad (11)$$

In summary, by combining Equations (9), (10), and (11), we have

$$\sum_{i\in B}(\langle X_i,\mathbf{S}\rangle - d)^2 \ge \frac{0.08\alpha^4n^3}{\varepsilon} + \varepsilon n^2d - O(\kappa^3)\cdot\left(\varepsilon^2n^2d + \varepsilon^{1/2}n^{3/2}d + \alpha n^{3/2}d\right).$$

Now, assuming that $n\ge\kappa^5\cdot\left(\frac{d\varepsilon^3}{\alpha^4} + \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right) \ge \kappa^5\cdot\left(\frac{d\varepsilon^3}{\alpha^4} + \frac{d^{2/3}\varepsilon}{\alpha^{8/3}} + \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^2}\right)$, each of the three error terms is at most $0.01\frac{\alpha^4}{\varepsilon}n^3$. So overall,

$$\sum_{i\in B}(\langle X_i,\mathbf{S}\rangle - d)^2 \ge \frac{0.05\alpha^4n^3}{\varepsilon} + \varepsilon n^2d = \varepsilon n^2d\left(1 + \frac{0.05\alpha^4n}{\varepsilon^2d}\right). \qquad \square$$

Corollary 4.17. *Suppose that Algorithm 3 does not reject in Line 3, and $\varepsilon\le 0.1$. Then, under the alternative hypothesis and Assumption 3, with high probability $\frac{1}{n}\cdot\sum_{i\in B}\left(\frac{\langle X_i,\mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \ge \varepsilon + 0.04\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d}$.*

Proof. Since Algorithm 3 did not reject on Line 3, this means $\|\mathbf{S}\|_2^2 = nd\pm 0.01\alpha^2n^2 = nd\cdot\left(1\pm 0.01\frac{\alpha^2n}{d}\right)$. So, $\frac{1}{n\cdot\|\mathbf{S}\|_2^2} = \frac{1}{n^2d}\cdot\left(1\pm 0.02\frac{\alpha^2n}{d}\right)$, because we assumed $n\le\frac{d}{\alpha^2}$. Hence, by Lemma 4.16, we have

$$\begin{aligned}\frac{1}{n}\cdot\sum_{i\in B}\left(\frac{\langle X_i,\mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 &\ge \varepsilon\cdot\left(1 + 0.05\frac{\alpha^4}{\varepsilon^2}\cdot\frac{n}{d}\right)\cdot\left(1 - 0.02\alpha^2\cdot\frac{n}{d}\right) \\ &\ge \varepsilon\cdot\left(1 + 0.04\frac{\alpha^4}{\varepsilon^2}\cdot\frac{n}{d}\right) \\ &= \varepsilon + 0.04\frac{\alpha^4}{\varepsilon}\cdot\frac{n}{d}. \end{aligned} \qquad \square$$

Finally, we bound the good samples.

Lemma 4.18. *Assume that* $n \geq \kappa^5 \cdot \left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$. *Then, with high probability*

$$\frac{1}{n}\sum_{i\in G}\left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \geq 1 - \varepsilon - \frac{0.01\alpha^4 n}{\varepsilon d}.$$

Proof. Let's fix the vectors $\mathbf{Q}, \mathbf{R}, \mathbf{T}$, and consider the posterior distribution of the good samples $\{X_i\}_{i\in G}$. By Proposition 2.12, we can write $X_i = \frac{\mathbf{Q}+\mathbf{T}}{(1-\varepsilon)n} + Y_i - \bar{Y}$, where $\{Y_i\}_{i\in G}$ are distributed as i.i.d. $\mathcal{N}(0, I)$ and $\bar{Y} = \frac{1}{(1-\varepsilon)n}\sum_{i\in G} Y_i$. Hence, $\{\langle X_i, \mathbf{S}\rangle\}_{i\in G}$ is distributed as $\frac{\langle \mathbf{Q}+\mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon)n} + \|\mathbf{S}\|_2 \cdot z_i - \bar{z}$, where z_i are distributed as i.i.d. $\mathcal{N}(0,1)$ and $\bar{z} = \frac{1}{(1-\varepsilon)n}\sum_{i\in G} z_i$. Hence, defining $\tilde{z}_i = z_i - \bar{z}$, since $\tilde{z}_i$ have mean 0, we can rewrite our expression as

$$\frac{1}{n}\sum_{i\in G}\left(\frac{\frac{\langle \mathbf{Q}+\mathbf{T}, \mathbf{S}\rangle}{(1-\varepsilon)n} - d}{\|\mathbf{S}\|_2} + \tilde{z}_i\right)^2 \geq \frac{1}{n}\sum_{i\in G}(\tilde{z}_i)^2 \geq (1-\varepsilon) - \frac{\kappa^2}{\sqrt{n}}.$$

The final inequality above combines the facts that $\sum_{i\in G}\tilde{z}_i^2 = \sum_{i\in G} z_i^2 - (1-\varepsilon)n\bar{z}^2$, that $\sum_{i\in G} z_i^2 = (1-\varepsilon)n \pm \kappa\sqrt{n}$ by Proposition 2.11, and that $|\bar{z}| \leq \kappa/\sqrt{n}$. Finally, because we are assuming $n \geq \kappa^5 \cdot \left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$, we have that $\frac{\kappa^2}{\sqrt{n}} \leq 0.01\frac{\alpha^4 n}{\varepsilon d}$. This completes the proof. □

By combining Corollary 4.17 and Lemma 4.18, the following lemma is immediate.

Lemma 4.19. *Suppose that Algorithm 3 does not reject in Line 3, and* $\varepsilon \leq 0.1$. *Then, under the alternative hypothesis and Assumption 3, with high probability*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\langle X_i, \mathbf{S}\rangle - d}{\|\mathbf{S}\|_2}\right)^2 \geq 1 + 0.03 \cdot \frac{\alpha^4 n}{\varepsilon d}.$$

As a direct consequence of Lemmas 4.10, 4.13, and 4.19, Lemma 4.2 is immediate.

4.6 Proof of Lemma 4.3

In this section, we finish the proof of Theorem 4.1, by proving Lemma 4.3.

It suffices to prove the following lemma.

Lemma 4.20. *Assume the alternative hypothesis, and that* $n \geq \kappa^5 \cdot \left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon}{\alpha^2}\right)$ *and that* $\varepsilon \leq 0.1$ *and* $\alpha \geq \kappa^5 \cdot \varepsilon$. *Then, under Assumption 2, with high probability* $\left\|\sum_{i\in[n]} X_i\right\|_2^2 \geq nd + 0.1\alpha^2 n^2$.

Proof. As usual, we write $\sum_{i\in[n]} X_i = \mathbf{Q} + \mathbf{R} + \mathbf{T}$, so $\|\sum_{i\in[n]} X_i\|_2^2 = \|\mathbf{Q}+\mathbf{R}+\mathbf{T}\|_2^2 = \|\mathbf{Q}+\mathbf{R}\|_2^2 + \|\mathbf{T}\|_2^2 + 2\langle \mathbf{Q}+\mathbf{R}, \mathbf{T}\rangle$.

Let $A = \|\mathbf{Q}+\mathbf{R}\|_2$. Note that $A \geq \|\mathbf{Q}\|_2 - \|\mathbf{R}\|_2 = 0.9\alpha n - \|\mathbf{R}\|_2$, assuming $\varepsilon \leq 0.1$. In addition, by Proposition 4.4, we have that $\|\mathbf{R}\|_2^2 = \|\sum_{i\in B} X_i\|_2^2 \leq \varepsilon nd + O(\kappa)\cdot((\varepsilon n)^{3/2}\sqrt{d} + (\varepsilon n)^2)$. So, $\|\mathbf{R}\|_2 \leq O(\sqrt{\varepsilon nd} + \kappa\cdot\varepsilon n)$. Assuming that $n \geq \kappa^5 \cdot \frac{d\varepsilon}{\alpha^2}$ and $\alpha \geq \kappa^5 \cdot \varepsilon$, both $O(\sqrt{\varepsilon nd})$ and $O(\kappa\cdot\varepsilon n)$ are at most $0.1\alpha n$. Thus, $A \geq 0.7\alpha n$.

Since $\mathbf{T} \sim \mathcal{N}(0, (1-\varepsilon)nI)$ is independent of $\mathbf{Q}, \mathbf{R}$, this means that $\|\mathbf{T}\|_2^2 \geq (1-\varepsilon)nd - \kappa n\sqrt{d}$ and $|\langle \mathbf{Q}+\mathbf{R}, \mathbf{T}\rangle| \leq (\kappa\sqrt{n}) \cdot A$ with high probability. Thus,

$$\|\mathbf{Q}+\mathbf{R}+\mathbf{T}\|_2^2 \geq A^2 + (1-\varepsilon)nd - \kappa n\sqrt{d} - 2\kappa\sqrt{n} \cdot A \geq (1-\varepsilon)nd + (A - \kappa\sqrt{n})^2 - \kappa^2 n\sqrt{d}.$$

Since $n \geq \kappa^5 \cdot \frac{\sqrt{d}}{\alpha^2} \geq \frac{\kappa^5}{\alpha^2}$, this means that $\kappa\sqrt{n} \leq 0.2\alpha n$, so $A - \kappa\sqrt{n} \geq 0.5\alpha n$. Moreover, $\kappa^2 n\sqrt{d} \leq 0.05\alpha^2 n^2$. Thus,

$$\|\mathbf{Q}+\mathbf{R}+\mathbf{T}\|_2^2 \geq (1-\varepsilon)nd + (0.5\alpha n)^2 - 0.05\alpha^2 n^2 = (1-\varepsilon)nd + 0.2\alpha^2 n^2.$$

Assuming that $n \geq \kappa^5 \cdot \frac{d\varepsilon}{\alpha^2}$, $\varepsilon nd \leq 0.1\alpha^2 n^2$, which means this is at least $nd + 0.1\alpha^2 n^2$. □

By combining Lemmas 4.10 and 4.20, Lemma 4.3 is immediate (since $\frac{d\varepsilon^3}{\alpha^4} < \frac{d\varepsilon}{\alpha^2}$). Note that we never assumed $\varepsilon n \leq d$ in either of these lemmas.

5 Lower bound in the Huber model

In this section, we prove that under the Huber model, one needs $n = \Omega(d\varepsilon^3/\alpha^4)$ samples to solve robust mean testing. Our lower bound even holds in the restricted setting where under the null hypothesis, the distribution must be uncorrupted.

5.1 Main Lower Bound

We are now ready to prove our main lower bound.

Theorem 5.1. *Let $\mathcal{D}_0$ represent the distribution of $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, and let $\mathcal{D}_1$ represent the distribution of $(X_1, \dots, X_n)$ where we choose a random vector $v \sim \mathcal{N}(0, \frac{1}{d} \cdot I)$ and conditional on v, each X_i is drawn i.i.d. from the mixture $(1-\varepsilon) \cdot \mathcal{N}(\alpha \cdot v, I) + \varepsilon \cdot \mathcal{N}(-\frac{1-\varepsilon}{\varepsilon} \cdot \alpha \cdot v, I)$. Then, there exists a small absolute constant $c > 0$ such that if $n = c \cdot \frac{d\varepsilon^3}{\alpha^4}$, $\alpha \geq \varepsilon$, and $c \cdot \frac{d\varepsilon^3}{\alpha^4} \geq \frac{\sqrt{d}}{\alpha^2}$, then $\mathrm{d_{TV}}(\mathcal{D}_0, \mathcal{D}_1) \leq 0.1$.*

Since the total variation distance is at most 0.1, no algorithm can successfully distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$ with probability more than 0.55. Moreover, $\|v\|_2 \leq 1 + o(1)$, and therefore $\|\alpha v\|_2 \leq \alpha(1 + o(1))$, with very high probability. Hence, this proves the desired lower bound when $c \cdot \frac{d\varepsilon^3}{\alpha^4} \geq \frac{\sqrt{d}}{\alpha^2}$. Alternatively, if $c \cdot \frac{d\varepsilon^3}{\alpha^4} < \frac{\sqrt{d}}{\alpha^2}$, the lower bound is immediate from the non-robust lower bound [DKS17]. When $\alpha \geq \varepsilon$, it is well-known that this problem is impossible, since the null and alternative distributions have total variation distance $\leq \varepsilon$.

We will bound the $\mathrm{d_{TV}}(\mathcal{D}_0, \mathcal{D}_1)$ through χ^2 divergence. As $\mathrm{D}_{\chi^2}(\mathcal{D}_1 || \mathcal{D}_0)$ is actually too large and thus does not suffice, we instead bound $\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0)$ for some $\mathcal{D}_1'$ that is close in total variation distance to $\mathcal{D}_1$.

For a sample $X = (X_1, \dots, X_n) \sim \mathcal{D}_1$, we will let a set $S \subset [n]$ correspond to X where $i \in S$ iff X_i was drawn from the mixture component $\mathcal{N}(\alpha \cdot v, I)$. Note that S is not determined by X. We will choose $\mathcal{D}_1'$ to be $\mathcal{D}_1$ restricted to having S with size $(1-\varepsilon)n \pm K\sqrt{\varepsilon n}$ for some large constant K. Call such sets S good, and let $\mathcal{S}$ be the set of all good sets. By standard properties of Binomial distributions, if $K \geq 100$, with probability at least $1 - 10^{-4}$, a random subset S obtained by including each element $i \in [n]$ with probability $1 - \varepsilon$ is good. Hence, $\mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_1') \leq 2 \cdot 10^{-4}$. Thus, it now suffices to upper bound $\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0)$ (which then upper bounds $\mathrm{d_{TV}}(\mathcal{D}_1', \mathcal{D}_0)$ by Fact 2.3).

It will be convenient to use the following notation throughout this section: let Z be the probability that a random subset of $[n]$ obtained by including each element $i \in [n]$ with probability $1 - \varepsilon$ is good. We begin by computing the likelihood ratio.

Claim 5.2. *Let $X = (X_1, \ldots, X_n)$ be a set of samples. Let $p_{\mathcal{D}_0}(X), p_{\mathcal{D}_1'}(X)$ be the PDFs of seeing that sample from $\mathcal{D}_0$ and $\mathcal{D}_1'$ respectively. Then*

$$\frac{p_{\mathcal{D}_1'}(X)}{p_{\mathcal{D}_0}(X)} = \frac{1}{Z}\sum_{S\in\mathcal{S}}(1-\varepsilon)^{|S|}\varepsilon^{n-|S|}\left(\frac{d+t_S}{d}\right)^{-d/2}\exp\left(\frac{\alpha_S(X)}{2(t_S+d)}\right)$$

where we define for subsets $S \subset [n]$,

$$\begin{aligned} X_S &:= \frac{\alpha}{\varepsilon}\cdot\left(\varepsilon\cdot\sum_{i\in S}X_i - (1-\varepsilon)\cdot\sum_{i\notin S}X_i\right) \\ \alpha_S(X) &:= \|X_S\|^2 \\ t_S &:= \frac{\alpha^2}{\varepsilon^2}\cdot\left(\varepsilon^2\cdot|S| + (1-\varepsilon)^2\cdot(n-|S|)\right) \end{aligned}$$

Proof. For any sample $X = (X_1, \ldots, X_n)$, the PDF of seeing that sample from $\mathcal{D}_0$ is

$$p_{\mathcal{D}_0}(X) = \prod_{i=1}^{n} e^{-\|X_i\|^2/2} \tag{12}$$

The probability of seeing that sample from $\mathcal{D}_1$ is

$$\begin{aligned} p_{\mathcal{D}_1}(X) &= \mathbb{E}_v \prod_{i=1}^{n}\left((1-\varepsilon)\cdot e^{-\|X_i-\alpha v\|^2/2} + \varepsilon\cdot e^{-\|X_i+(1-\varepsilon)\alpha/\varepsilon v\|^2/2}\right) \\ &= \sum_{S\subset[n]}\mathbb{E}_v\left((1-\varepsilon)^{|S|}\varepsilon^{(n-|S|)}\cdot\prod_{i\in S}e^{-\|X_i-\alpha v\|^2/2}\cdot\prod_{i\notin S}e^{-\|X_i+(1-\varepsilon)\alpha/\varepsilon v\|^2/2}\right). \end{aligned} \tag{13}$$

By restricting ourselves to good sets $S \in \mathcal{S}$, the probability of seeing X drawn from $\mathcal{D}_1'$ is

$$p_{\mathcal{D}_1'}(X) = \frac{1}{Z}\sum_{S\in\mathcal{S}}\mathbb{E}_v\left((1-\varepsilon)^{|S|}\varepsilon^{(n-|S|)}\cdot\prod_{i\in S}e^{-\|X_i-\alpha v\|^2/2}\cdot\prod_{i\notin S}e^{-\|X_i+(1-\varepsilon)\alpha/\varepsilon v\|^2/2}\right), \tag{14}$$

where Z is the probability of a random set S being good if each $i \in [n]$ is included in S independently with probability $1-\varepsilon$.

From (12) and (14), it is simple to compute the ratio

$$\frac{p_{\mathcal{D}_1'}(X)}{p_{\mathcal{D}_0}(X)} = \frac{1}{Z}\cdot\sum_{S\in\mathcal{S}}\left((1-\varepsilon)^{|S|}\varepsilon^{(n-|S|)}\cdot\underbrace{\mathbb{E}_v\left[\prod_{i\in S}e^{-\alpha\langle X_i,v\rangle-\alpha^2\|v\|^2/2}\cdot\prod_{i\notin S}e^{(1-\varepsilon)\alpha/\varepsilon\cdot\langle X_i,v\rangle-(1-\varepsilon)^2\alpha^2/\varepsilon^2\cdot\|v\|^2/2}\right]}_{A_S(X)}\right).$$

We use $A_S(X)$ as a shorthand in simplifying the expression above. Now we can explicitly compute $A_S(X)$. With X_S, t_S as defined above, we can write

$$A_S(X) = \mathbb{E}_v\left[e^{-\langle X_S,v\rangle - t_S/2\cdot\|v\|^2}\right].$$

Since $v \sim \mathcal{N}(0, \frac{1}{d} \cdot I) = \frac{1}{\sqrt{d}} \cdot \mathcal{N}(0, I)$, we can use the rotational symmetry of v and Fact 2.4 to rewrite

$$\begin{aligned} A_S(X) &= \mathbb{E}_{x \sim \mathcal{N}(0,1)} \left[e^{-\sqrt{\alpha_S(X)/d} \cdot x - (t_S/2d) \cdot x^2} \right] \cdot \left(\mathbb{E}_{x \sim \mathcal{N}(0,1)} \left[e^{-(t_S/2d) \cdot x^2} \right] \right)^{d-1} \\ &= \frac{\exp\left(\frac{\alpha_S(X)/d}{2 + 2t_S/d} \right)}{\sqrt{1 + t_S/d}} \cdot \left(\frac{1}{\sqrt{1 + t_S/d}} \right)^{d-1} \\ &= \exp\left(\frac{\alpha_S(X)}{2(t_S + d)} \right) \cdot \left(\frac{d + t_S}{d} \right)^{-d/2}. \end{aligned}$$

□

Using expression for the likelihood ratio in Claim 5.2, we can explicitly compute the χ^2 divergence $\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0)$.

Lemma 5.3. *We have*

$$\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0) = \frac{1}{Z^2} \cdot \sum_{S,T \subset \mathcal{S}} (1-\varepsilon)^{|S|+|T|} \varepsilon^{(n-|S|)+(n-|T|)} \cdot \left(1 - \left(\frac{t_{S,T}}{d} \right)^2 \right)^{-d/2}$$

where $t_{S,T} = \frac{\alpha^2}{\varepsilon^2} \cdot (\varepsilon^2 |S \cap T| - \varepsilon(1-\varepsilon)|S \triangle T| + (1-\varepsilon)^2 |(S \cup T)^c|)$ *and* $\triangle$ *denotes symmetric difference.*

Proof. Using Claim 5.2, we can write

$$\begin{aligned} \mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0) = \frac{1}{Z^2} \cdot \sum_{S,T \subset \mathcal{S}} (1-\varepsilon)^{|S|+|T|} \varepsilon^{(n-|S|)+(n-|T|)} \cdot \left(\frac{d + t_S}{d} \right)^{-d/2} \left(\frac{d + t_T}{d} \right)^{-d/2} \\ \cdot \underbrace{\mathbb{E}_{X \sim \mathcal{D}_0} \left[\exp\left(\frac{\alpha_S(X)}{2(t_S + d)} + \frac{\alpha_T(X)}{2(t_T + d)} \right) \right]}_{B_{S,T}}. \end{aligned} \tag{15}$$

where $\alpha_S(X) = \|X_S\|^2$, $\alpha_T(X) = \|X_T\|^2$ and

$$\begin{aligned} X_S &= \frac{\alpha}{\varepsilon} \left(\varepsilon \cdot \sum_{i \in S} X_i - (1-\varepsilon) \cdot \sum_{i \notin S} X_i \right) \\ X_T &= \frac{\alpha}{\varepsilon} \left(\varepsilon \cdot \sum_{i \in T} X_i - (1-\varepsilon) \cdot \sum_{i \notin T} X_i \right) \end{aligned}$$

are as defined in Claim 5.2. Now we explicitly compute the expression above labelled $B_{S,T}$. In each coordinate $j \in [d]$, $((X_S)_j, (X_T)_j)$ forms a bivariate Gaussian, and $((X_S)_j, (X_T)_j)$ over all $j \in [d]$ are independent and identically distributed. Through direct computation, we get that $((X_S)_j, (X_T)_j) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} t_S & t_{S,T} \\ t_{S,T} & t_T \end{pmatrix}$$

and

$$t_S = \frac{\alpha^2}{\varepsilon^2} \cdot \left(\varepsilon^2 \cdot |S| + (1-\varepsilon)^2 \cdot (n - |S|)\right)$$
$$t_T = \frac{\alpha^2}{\varepsilon^2} \cdot \left(\varepsilon^2 \cdot |T| + (1-\varepsilon)^2 \cdot (n - |T|)\right)$$
$$t_{S,T} = \frac{\alpha^2}{\varepsilon^2} \cdot \left(\varepsilon^2 |S \cap T| - \varepsilon(1-\varepsilon)|S \triangle T| + (1-\varepsilon)^2 |(S \cup T)^c|\right) .$$

Therefore, $\left(\frac{(X_S)_j}{\sqrt{2(t_S+d)}}, \frac{(X_T)_j}{\sqrt{2(t_T+d)}}\right) \sim \mathcal{N}(\mathbf{0}, \Sigma')$, where

$$\Sigma' = \begin{pmatrix} \frac{t_S}{2(t_S+d)} & \frac{t_{S,T}}{2\sqrt{(t_S+d)(t_T+d)}} \\ \frac{t_{S,T}}{2\sqrt{(t_S+d)(t_T+d)}} & \frac{t_T}{2(t_T+d)} \end{pmatrix} .$$

By Corollary 2.7, this implies that

$$\mathbb{E}\left[\exp\left(\frac{(X_S)_j^2}{2(t_S+d)} + \frac{(X_T)_j^2}{2(t_T+d)}\right)\right] = \frac{1}{\sqrt{\left(1 - \frac{t_S}{t_S+d}\right) \cdot \left(1 - \frac{t_T}{t_T+d}\right) - \frac{t_{S,T}^2}{(t_S+d)(t_T+d)}}} .$$

Therefore, by multiplying this over all j (since $((X_S)_j, (X_T)_j)$ are i.i.d. across all $j \in [d]$), we have that

$$B_{S,T} = \left(\left(1 - \frac{t_S}{t_S+d}\right) \cdot \left(1 - \frac{t_T}{t_T+d}\right) - \frac{t_{S,T}^2}{(t_S+d)(t_T+d)}\right)^{-d/2}$$
$$= \left(\frac{d^2 - t_{S,T}^2}{(t_S+d)(t_T+d)}\right)^{-d/2}$$

So, (15) can be rewritten as

$$\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0) = \frac{1}{Z^2} \cdot \sum_{S,T \subset \mathcal{S}} (1-\varepsilon)^{|S|+|T|} \varepsilon^{(n-|S|)+(n-|T|)} \cdot \left(1 - \left(\frac{t_{S,T}}{d}\right)^2\right)^{-d/2} .$$

□

Now we can complete the proof of Theorem 5.1 by upper bounding the RHS of Lemma 5.3.

Proof of Theorem 5.1. Recall that it suffices to prove that $\mathrm{D}_{\chi^2}(\mathcal{D}_1' || \mathcal{D}_0) = \mathbb{E}_{X \sim \mathcal{D}_0}\left(\frac{p_{\mathcal{D}_1'}(X)}{p_{\mathcal{D}_0}(X)}\right)^2 \leq 1.01$ as, by Fact 2.3, this implies that the TV distance between $\mathcal{D}_1'$ and $\mathcal{D}_0$ is at most 0.1. By Lemma 5.3 it now suffices to bound the expression

$$\frac{1}{Z^2} \cdot \sum_{S,T \subset \mathcal{S}} (1-\varepsilon)^{|S|+|T|} \varepsilon^{(n-|S|)+(n-|T|)} \cdot \left(1 - \left(\frac{t_{S,T}}{d}\right)^2\right)^{-d/2} .$$

We can think of S, T as random subsets of $[n]$ where each element i is chosen to be in S (and likewise T) with probability $1-\varepsilon$, and then conditioning on S, T having size $(1-\varepsilon)n \pm K\sqrt{\varepsilon n}$ for some sufficiently

large constant K. In this case, if we use $S, T \sim \mathcal{S}$ to denote this distribution, the above expression is equivalent to

$$\mathbb{E}_{S,T\sim\mathcal{S}}\left(1-\left(\frac{\alpha^2}{\varepsilon^2}\cdot\frac{\varepsilon^2|S\cap T|-\varepsilon(1-\varepsilon)|S\triangle T|+(1-\varepsilon)^2|(S\cup T)^c|}{d}\right)^2\right)^{-d/2}.$$

So, now we just need to show that if $n = c\cdot\frac{d\varepsilon^3}{\alpha^4}$ for some small constant c, that the above expectation is at most 1.01.

Now, recall that we assumed $\frac{\sqrt{d}}{\alpha^2}\le c\cdot\frac{d\varepsilon^3}{\alpha^4}$. This means that $c\cdot\frac{\sqrt{d}\varepsilon^3}{\alpha^2}\ge 1$, or equivalently $\frac{d\varepsilon^3}{\alpha^4}\ge c^{-2}\varepsilon^{-3}$. Hence, we may assume that $n\ge c^{-1}\cdot\varepsilon^{-3}$.

If $|S| = a$ and $|T| = b$, and we let $Y := |(S\cup T)^c|$,

$$\begin{aligned}
&\varepsilon^2|S\cap T|-\varepsilon(1-\varepsilon)|S\triangle T|+(1-\varepsilon)^2|(S\cup T)^c|\\
&=(1-\varepsilon)^2\cdot Y-\varepsilon(1-\varepsilon)(n-a-Y+n-b-Y)+\varepsilon^2(a+b-n+Y)\\
&=Y-\varepsilon(1-\varepsilon)(2n-a-b)+\varepsilon^2(a+b-n)\\
&=Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n\,.
\end{aligned}$$

Recall that we may always assume $a, b = (1-\varepsilon)n\pm K\sqrt{\varepsilon n}$. Also, note that $Y\sim\mathrm{HGeom}(n, n-a, n-b)$. Therefore, if we condition on fixed $a, b\in[(1-\varepsilon)n-K\sqrt{\varepsilon n},(1-\varepsilon)n+K\sqrt{\varepsilon n}]$, we have that $\mathbb{E}[Y|a,b] = \frac{(n-a)(n-b)}{n} = \varepsilon^2 n\pm 2K\varepsilon\sqrt{\varepsilon n}\pm K^2\varepsilon$. By our assumption that $n\ge c^{-1}\cdot\varepsilon^{-3}$ and choosing c sufficiently small in terms of K, this can be bounded as $\varepsilon^2 n\pm 3K\varepsilon\sqrt{\varepsilon n}$.

Moreover, by Proposition 2.16, since $n-a, n-b\le\varepsilon n+K\sqrt{\varepsilon n}\le 2\varepsilon n$,

$$\mathbb{P}\left(|Y-\mathbb{E}[Y|a,b]|\ge t\sqrt{\varepsilon n}|a,b\right)\le 2e^{-2t^2(\varepsilon n)/(n-a)}\le 2e^{-t^2}\,.$$

Because $|\mathbb{E}[Y|a,b]-\varepsilon^2 n|\le 3K\varepsilon\sqrt{\varepsilon n}$, this means $\mathbb{P}(|Y-\varepsilon^2 n|\ge(3K\varepsilon+t)\sqrt{\varepsilon n})\le 2e^{-t^2}$. Hence, because $\varepsilon(a+b)-\varepsilon(2-\varepsilon)n = -\varepsilon^2 n^2\pm 2K\sqrt{\varepsilon n}$, this means $\mathbb{P}(|Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n|\ge(5K+t)\sqrt{\varepsilon n})\le 2e^{-t^2}$. In addition, we know that Y is bounded by $\min(n-a, n-b)\le 2\varepsilon n$, so overall $|Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n|$ is also bounded by $4\varepsilon n$ with probability 1.

We can rewrite our goal as bounding

$$\mathbb{E}_{S,T\sim\mathcal{S}}\left(1-\left(\frac{\alpha^2}{\varepsilon^2}\cdot\frac{Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n}{d}\right)^2\right)^{-d/2}.$$

Note that if $|x|\le 0.2$, then $1-x^2\ge e^{-2x^2}$, so $(1-x^2)^{-d/2}\le e^{-2x^2\cdot-d/2} = e^{dx^2}$. We know that $|Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n|\le 4\varepsilon n$ with probability 1, so as long as $\frac{\alpha^2}{\varepsilon^2}\cdot\frac{4\varepsilon n}{d}\le 0.4$, which holds when $n\le 0.1\cdot\frac{d\varepsilon^3}{\alpha^4}\le 0.1\cdot\frac{d\varepsilon}{\alpha^2}$, we just need to bound

$$\mathbb{E}_{S,T\sim\mathcal{S}}\left[\exp\left(\frac{\alpha^4}{\varepsilon^4}\cdot\frac{1}{d}\cdot(Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n)^2\right)\right]. \tag{16}$$

Defining C such that $Y+\varepsilon(a+b)-\varepsilon(2-\varepsilon)n = C\sqrt{\varepsilon n}$, then $\mathbb{P}(|C|\ge 5K+t)\le 2e^{-t^2}$. So, (16) equals

$$\mathbb{E}_{S,T\sim\mathcal{S}}\left[\exp\left(\frac{\alpha^4}{\varepsilon^4}\cdot\frac{1}{d}\cdot C^2\varepsilon n\right)\right] = \mathbb{E}_{S,T\sim\mathcal{S}}\left[\exp\left(C^2\cdot\frac{\alpha^4}{\varepsilon^3}\cdot\frac{n}{d}\right)\right] = \mathbb{E}_{S,T\sim\mathcal{S}}\left[e^{C^2\cdot c}\right],$$

since $n\le cd\cdot\frac{\varepsilon^3}{\alpha^4}$. By our bounds on C, if we assume c is sufficiently small in terms of K, this is at most 1.01, which means $\mathrm{D}_{\chi^2}(\mathcal{D}_1'||\mathcal{D}_0)\le 1.01$. This concludes the proof, since Fact 2.3 implies $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_1',\mathcal{D}_0)\le 0.05$, and we already know that $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_1',\mathcal{D}_1)\le 2\cdot 10^{-4}$. □

6 Improved Lower Bound against Oblivious Adversaries

In this section, we further improve our lower bound from Section 5 against an oblivious adversary.

6.1 Lower bound instance

We first construct the distributions for the lower bound instance. Fix parameters $\varepsilon < \alpha \le 1$ and dimension d, and consider drawing n samples for some choice of n. We will also set an auxiliary parameter β, which will depend on $\varepsilon, \alpha, d, n$.

The null distribution $\mathcal{D}_0$ will simply be n i.i.d. samples from $\mathcal{N}(0, I)$. To generate the alternative distribution $\mathcal{D}_1$, we perform the following steps:

1. Select a subset $A \subset [n]$ of size εn randomly. Let $A^c = [n] \backslash A$
2. Draw $\varepsilon \cdot n$ points $\{X_i\}_{i \in A}$ i.i.d. from $\mathcal{N}(0, I)$. Set $\mathbf{R}_A := \mathrm{Sum}(A) = \sum_{i \in A} X_i$.
3. Draw the vector $z \in \mathbb{R}^d$ from $\mathcal{N}\left(0, \frac{\alpha^2}{d} \cdot I\right)$.
4. Define $\mu := -\beta \cdot \mathbf{R}_A - z$, and draw $(1-\varepsilon)n$ points $\{X_i\}_{i \in A^c}$ from the distribution $\mathcal{N}(\mu, I)$.

For simplicity, we may write $X = (X_1, \dots, X_n)$, both in the null and alternative settings.

Note that with very high probability, $\|z\|_2 \le 2\alpha$. We will also ensure that β is chosen so that with very high probability, $\beta \cdot \|\mathbf{R}_A\|_2 \le 2\alpha$. As a result, this alternative construction indeed has $\|\mu\|_2 \le O(\alpha)$.

In the rest of this section, we prove that it is statistically hard to distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$, for an appropriate choice of β.

Theorem 6.1. *Suppose that $n \le c \cdot \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)$ for some sufficiently small constant $c > 0$, and that $\varepsilon \le \alpha \le 1$ and $n \ge \frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}$. Then $\mathrm{d_{TV}}(\mathcal{D}_0, \mathcal{D}_1) \le 0.1$.*

This implies that no algorithm can successfully distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$ with probability more than 0.55, which proves the desired lower bound when $c \cdot \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right) \ge \frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}$. Alternatively, we may either use the non-robust lower bound [DKS17] or Theorem 5.1. Finally, when $\alpha \ge \varepsilon$, it is well-known that this problem is impossible, since the null and alternative distributions have total variation distance $\le \varepsilon$.

We will prove the lower bound via a chi-square computation. For this, we must compute likelihood ratios, which we will do in the next subsection.

6.2 Likelihood Ratio Computation

First, we will compute a formula for the likelihood ratio between $\mathcal{D}_1$ and $\mathcal{D}_0$, if we condition on the set $A \subset [n]$ in the alternative hypothesis.

Definition 6.2. Recall that $p_{\mathcal{D}_0}(X), p_{\mathcal{D}_1}(X)$ denote the joint PDF of the points $X = (X_1, \dots, X_n)$ drawn according to $\mathcal{D}_0$ and $\mathcal{D}_1$, respectively. We also define $p_A(X)$ to denote the PDF of $X_1, \dots, X_n$ drawn according to $\mathcal{D}_1$, conditioned on the first step selecting A.

In addition, we will define $\mathbf{R}_A := \mathrm{Sum}(A) = \sum_{i \in A} X_i$, and $\mathbf{T}_A(X) := \sum_{i \in A^c} X_i$. Usually, the choice of $X_1, \dots, X_n$ will be clear, in which case we will drop the argument X.

Lemma 6.3. *Conditioned on A, the likelihood ratio is*

$$\frac{p_A(X)}{p_{\mathcal{D}_0}(X)} = \left(1 + \frac{(1-\varepsilon)n \cdot \alpha^2}{d}\right)^{-d/2} \cdot \exp\left(-\frac{(1-\varepsilon)n\beta^2 d \cdot \|\mathbf{R}_A\|_2^2 + 2\beta d \cdot \langle \mathbf{R}_A, \mathbf{T}_A \rangle - \alpha^2 \cdot \|\mathbf{T}_A\|_2^2}{2((1-\varepsilon)\alpha^2 n + d)}\right).$$

Proof. Suppose we additionally condition on the value $z \sim \mathcal{N}(0, \frac{\alpha^2}{d} \cdot I)$. Then,

$$\begin{aligned}
\log \frac{p_A(X|z)}{p_{\mathcal{D}_0}(X)} &= \sum_{i \in A^c} \left(-\frac{1}{2} \|X_i + \beta \cdot \mathbf{R}_A + z\|_2^2 + \frac{1}{2} \|X_i\|_2^2 \right) \\
&= \sum_{i \in A^c} \left(-\langle X_i, \beta \cdot \mathbf{R}_A + z \rangle - \frac{1}{2} \|\beta \cdot \mathbf{R}_A + z\|_2^2 \right) \\
&= -\langle \mathbf{T}_A, \beta \cdot \mathbf{R}_A + z \rangle - \frac{(1-\varepsilon)n}{2} \cdot \|\beta \cdot \mathbf{R}_A + z\|_2^2 \\
&= -\beta \cdot \langle \mathbf{T}_A, \mathbf{R}_A \rangle - \frac{(1-\varepsilon)n}{2} \cdot \beta^2 \cdot \|\mathbf{R}_A\|_2^2 - \langle \mathbf{T}_A + (1-\varepsilon)n \cdot \beta \cdot \mathbf{R}_A, z \rangle - \frac{(1-\varepsilon)n}{2} \cdot \|z\|_2^2.
\end{aligned}$$

So,

$$\frac{p_A(X|z)}{p_{\mathcal{D}_0}(X)} = \exp\left(-\beta \cdot \langle \mathbf{T}_A, \mathbf{R}_A \rangle - \frac{(1-\varepsilon)n}{2} \cdot \beta^2 \cdot \|\mathbf{R}_A\|_2^2 - \langle \mathbf{T}_A + (1-\varepsilon)n \cdot \beta \cdot \mathbf{R}_A, z \rangle - \frac{(1-\varepsilon)n}{2} \cdot \|z\|_2^2 \right).$$

Next, we remove the conditioning on z. Indeed, by using the above equation followed by Proposition 2.5, we have

$$\begin{aligned}
&\frac{p_A(X)}{p_{\mathcal{D}_0}(X)} \\
&= \mathbb{E}_{z \sim \mathcal{N}(0, \frac{\alpha^2}{d} \cdot I)} \exp\left(-\beta \cdot \langle \mathbf{T}_A, \mathbf{R}_A \rangle - \frac{(1-\varepsilon)n}{2} \cdot \beta^2 \cdot \|\mathbf{R}_A\|_2^2 - \langle \mathbf{T}_A + (1-\varepsilon)n \cdot \beta \cdot \mathbf{R}_A, z \rangle - \frac{(1-\varepsilon)n}{2} \cdot \|z\|_2^2 \right) \\
&= \exp\left(-\beta \langle \mathbf{T}_A, \mathbf{R}_A \rangle - \frac{(1-\varepsilon)n}{2} \beta^2 \|\mathbf{R}_A\|_2^2 \right) \mathbb{E}_{z \sim \mathcal{N}(0, \frac{\alpha^2}{d} \cdot I)} \exp\left(-\frac{(1-\varepsilon)n}{2} \|z\|_2^2 + \langle \mathbf{T}_A + (1-\varepsilon)n\beta \mathbf{R}_A, z \rangle \right) \\
&= \exp\left(-\beta \langle \mathbf{T}_A, \mathbf{R}_A \rangle - \frac{(1-\varepsilon)n}{2} \beta^2 \|\mathbf{R}_A\|_2^2 \right) \cdot \left(1 + \frac{(1-\varepsilon)n \cdot \alpha^2}{d} \right)^{-d/2} \cdot \exp\left(\frac{\|(1-\varepsilon)n\beta \mathbf{R}_A + \mathbf{T}_A\|_2^2}{2((1-\varepsilon)n + d/\alpha^2)} \right).
\end{aligned}$$

We can combine the terms that are in terms of $\|\mathbf{R}_A\|_2^2$, $\|\mathbf{T}_A\|_2^2$, and $\langle \mathbf{R}_A, \mathbf{T}_A \rangle$, to simplify this as

$$\left(1 + \frac{(1-\varepsilon)n \cdot \alpha^2}{d} \right)^{-d/2} \cdot \exp\left(-\frac{(1-\varepsilon)n\beta^2 d \cdot \|\mathbf{R}_A\|_2^2 + 2\beta d \cdot \langle \mathbf{R}_A, \mathbf{T}_A \rangle - \alpha^2 \cdot \|\mathbf{T}_A\|_2^2}{2((1-\varepsilon)\alpha^2 n + d)} \right). \qquad \square$$

Now, recall that the χ^2 divergence $\mathrm{D}_{\chi^2}(\mathcal{D}_1 || \mathcal{D}_0)$ equals

$$\mathop{\mathbb{E}}_{X \sim \mathcal{D}_0} \left(\frac{p_{\mathcal{D}_1}(X)}{p_{\mathcal{D}_0}(X)} \right)^2 = \mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim \mathcal{N}(0,I)} \mathop{\mathbb{E}}_{A, B \subset [n]} \left(\frac{p_A(X) p_B(X)}{p_{\mathcal{D}_0}(X)^2} \right),$$

where A, B will always denote random subsets of size εn in $[n]$. Using Lemma 6.3, we can write this as

$$\begin{aligned}
&\left(1 + \frac{(1-\varepsilon)n\alpha^2}{d} \right)^{-d} \cdot \mathop{\mathbb{E}}_{A,B} \mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim \mathcal{N}(0,I)} \exp\Bigg[\\
&-\frac{1}{2} \cdot \frac{(1-\varepsilon)n\beta^2 d \cdot (\|\mathbf{R}_A\|_2^2 + \|\mathbf{R}_B\|_2^2) + 2\beta d \cdot (\langle \mathbf{R}_A, \mathbf{T}_A \rangle + \langle \mathbf{R}_B, \mathbf{T}_B \rangle) - \alpha^2 \cdot (\|\mathbf{T}_A\|_2^2 + \|\mathbf{T}_B\|_2^2)}{(1-\varepsilon)\alpha^2 n + d} \Bigg].
\end{aligned} \tag{17}$$

Now, the exponential term can be decomposed coordinate-wise, and since each coordinate of $X_1, \dots, X_n$ is independent if we condition on A, B, we can therefore write (17) after removing the expectation on A, B as

$$\left(1 + \frac{(1-\varepsilon)n\alpha^2}{d}\right)^{-d} \cdot \left(\mathop{\mathbb{E}}_{x_1,\dots,x_n \sim \mathcal{N}(0,1)} \exp\left(-\frac{1}{2} \cdot \frac{x^\top (M_A + M_B)x}{(1-\varepsilon)\alpha^2 n + d}\right)\right)^d \tag{18}$$

Above, each x_i is a standard univariate Gaussian, and M_A is the $n \times n$ matrix with blocks

$$\begin{array}{c} \\ A \\ A^c \end{array} \begin{array}{c} \overbrace{}^{A} \quad \overbrace{}^{A^c} \\ \begin{pmatrix} (1-\varepsilon)n\beta^2 d & \beta d \\ \beta d & -\alpha^2 \end{pmatrix} \end{array}$$

and M_B is defined similarly. Here, each block is dependent on whether the row/column indices are in A or A^c, and all entries in the same block are the same. Note that M_A has rank at most 2. Moreover, by projecting onto the space of vectors v where v_i is constant for all $i \in A$, and constant for all $i \in A^c$, we have that M_A has the same nonzero eigenvalues as $\sqrt{D_A}\Sigma_A\sqrt{D_A}$, where

$$D_A = \begin{pmatrix} \varepsilon n & 0 \\ 0 & (1-\varepsilon)n \end{pmatrix}, \quad \Sigma_A = \begin{pmatrix} (1-\varepsilon)n\beta^2 d & \beta d \\ \beta d & -\alpha^2 \end{pmatrix}.$$

If we define $M_{A,B} = M_A + M_B$, we can write $M_{A,B}$ in a similar block-diagonal format, where the rows/-columns are split based on the index being in $A \cap B$, $A \cap B^c$, $A^c \cap B$, or $A^c \cap B^c$. Therefore, if $|A \cap B| = \gamma \cdot n$ for some $0 \le \gamma \le \varepsilon$, $M_{A,B}$ has the same nonzero eigenvalues as $\sqrt{D_{A,B}}\Sigma_{A,B}\sqrt{D_{A,B}}$, where

$$D_{A,B} = \begin{pmatrix} \gamma n & 0 & 0 & 0 \\ 0 & (\varepsilon-\gamma)n & 0 & 0 \\ 0 & 0 & (\varepsilon-\gamma)n & 0 \\ 0 & 0 & 0 & (1-2\varepsilon+\gamma)n \end{pmatrix}$$

and

$$\Sigma_{A,B} := \begin{pmatrix} 2(1-\varepsilon)n\beta^2 d & (1-\varepsilon)n\beta^2 d + \beta d & (1-\varepsilon)n\beta^2 d + \beta d & 2\beta d \\ (1-\varepsilon)n\beta^2 d + \beta d & (1-\varepsilon)n\beta^2 d - \alpha^2 & 2\beta d & \beta d - \alpha^2 \\ (1-\varepsilon)n\beta^2 d + \beta d & 2\beta d & (1-\varepsilon)n\beta^2 d - \alpha^2 & \beta d - \alpha^2 \\ 2\beta d & \beta d - \alpha^2 & \beta d - \alpha^2 & -2\alpha^2 \end{pmatrix}.$$

Now, for any subsets $A, B \subset [n]$ of size $\varepsilon \cdot n$, we define $G_A = \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot M_A$ and $G_{A,B} = G_A + G_B = \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot M_{A,B}$. We note the following basic proposition.

Proposition 6.4. *Suppose that $n \le \frac{0.1d}{\alpha^2}$ and $0 \le \beta \le \frac{0.1}{n}$. Then, all eigenvalues of G_A are strictly greater than $-\frac{1}{2}$.*

As a direct corollary, all eigenvalues of $G_{A,B}$, for any A, B, are strictly greater than -1.

Proof. It suffices to prove the claim for $\hat{G}_A := \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot \sqrt{D_A}\Sigma_A\sqrt{D_A}$. Note that $\hat{G}_A$ is a 2×2 symmetric matrix. If $\hat{G}_A$ has eigenvalues λ_1, λ_2, then we need to show that $\lambda_1 + \frac{1}{2}, \lambda_2 + \frac{1}{2} > 0$. It therefore suffices to show that $(\lambda_1 + \frac{1}{2}) + (\lambda_2 + \frac{1}{2}) = \operatorname{Tr}(\hat{G}_A) + 1$ and $(\lambda_1 + \frac{1}{2}) \cdot (\lambda_2 + \frac{1}{2}) = \det(\hat{G}_A) + \frac{1}{2}\operatorname{Tr}(\hat{G}_A) + \frac{1}{4}$ are both strictly greater than 0.

Note that

$$\mathrm{Tr}(\hat{G}_A) = \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot \mathrm{Tr}(\Sigma_A \cdot D_A) = \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot \Big[(1-\varepsilon)n\beta^2 d \cdot \varepsilon n - \alpha^2 \cdot (1-\varepsilon)n\Big] \geq \frac{-\alpha^2(1-\varepsilon)n}{(1-\varepsilon)\alpha^2 n + d}.$$

We are assuming that $n \leq \frac{0.1d}{\alpha^2}$, which means that $(1-\varepsilon)\alpha^2 n \leq 0.1d$. So in fact, $\mathrm{Tr}(\hat{G}_A) \geq -0.1$, so $\mathrm{Tr}(\hat{G}_A) + 1 \geq 0.9 > 0$.

Next,

$$\begin{aligned}
\det(\hat{G}_A) &= \frac{1}{((1-\varepsilon)\alpha^2 n + d)^2} \cdot \det(D_A) \cdot \det(\Sigma_A) \\
&= \frac{\varepsilon n \cdot (1-\varepsilon)n}{((1-\varepsilon)\alpha^2 n + d)^2} \cdot \Big((1-\varepsilon)n\beta^2 d \cdot (-\alpha^2) - (\beta d)^2\Big) \\
&= -\frac{\varepsilon(1-\varepsilon)n^2 \cdot \beta^2 d \cdot ((1-\varepsilon)\alpha^2 n + d)}{((1-\varepsilon)\alpha^2 n + d)^2} \\
&= -\frac{\varepsilon(1-\varepsilon)n^2 \cdot \beta^2 d}{(1-\varepsilon)\alpha^2 n + d}.
\end{aligned}$$

Since $0 \leq \beta \leq 0.1/n$, this means $\det(\hat{G}_A) \geq -\frac{0.01\varepsilon(1-\varepsilon)d}{(1-\varepsilon)\alpha^2 n + d} \geq -\frac{0.01d}{d} = -0.01$. So, $\det(\hat{G}_A) + \frac{1}{2}\mathrm{Tr}(\hat{G}_A) + \frac{1}{4} \geq -0.01 - 0.05 + 0.25 > 0$. □

As a result of Proposition 6.4, we can apply Proposition 2.6 to obtain the following.

Lemma 6.5. *Assuming that $n \leq \frac{0.1d}{\alpha^2}$ and $\beta \leq \frac{0.1}{n}$, the χ^2 divergence $\mathrm{D}_{\chi^2}(\mathcal{D}_1||\mathcal{D}_0)$ equals*

$$\mathop{\mathbb{E}}_{A,B}\left[\left(\left(\frac{d + (1-\varepsilon)n\alpha^2}{d}\right)^2 \cdot \det(I + G_{A,B})\right)^{-d/2}\right].$$

Proof. We have the following chain of equalities. The first equality follows by combining (17) and (18), the second follows by the definition of $G_{A,B}$, the third follows by Proposition 2.6, and the final follows by basic manipulation.

$$\begin{aligned}
\mathrm{D}_{\chi^2}(\mathcal{D}_1||\mathcal{D}_0) &= \mathop{\mathbb{E}}_{A,B}\left[\left(1 + \frac{(1-\varepsilon)n\alpha^2}{d}\right)^{-d} \cdot \left(\mathop{\mathbb{E}}_{x_1,\ldots,x_n \sim \mathcal{N}(0,1)} \exp\left(-\frac{1}{2} \cdot \frac{x^\top (M_A + M_B)x}{(1-\varepsilon)\alpha^2 n + d}\right)\right)^d\right] \\
&= \mathop{\mathbb{E}}_{A,B}\left[\left(1 + \frac{(1-\varepsilon)n\alpha^2}{d}\right)^{-d} \cdot \left(\mathop{\mathbb{E}}_{x_1,\ldots,x_n \sim \mathcal{N}(0,1)} \exp\left(-\frac{1}{2} \cdot x^\top \cdot G_{A,B} \cdot x\right)\right)^d\right] \\
&= \mathop{\mathbb{E}}_{A,B}\left[\left(1 + \frac{(1-\varepsilon)n\alpha^2}{d}\right)^{-d} \cdot \det(I + G_{A,B})^{-d/2}\right] \\
&= \mathop{\mathbb{E}}_{A,B}\left[\left(\left(\frac{d + (1-\varepsilon)n\alpha^2}{d}\right)^2 \cdot \det(I + G_{A,B})\right)^{-d/2}\right].
\end{aligned}$$

□

6.3 Final Computation

Through some tedious computations, one can show the following:

Lemma 6.6. *Suppose that* $|A \cap B| = \gamma \cdot n$*, for some* $0 \leq \gamma \leq \varepsilon$*. Then,*

$$\det(I + G_{A,B}) = \frac{1}{(d + (1-\varepsilon)\alpha^2 n)^2} \cdot \left[(d + \beta^2 d(\varepsilon^2 - \gamma)n^2)^2 - (\alpha^2 n - 2\beta d\varepsilon n + \beta^2 d\varepsilon n^2 - 2\alpha^2 \varepsilon n + \alpha^2 \gamma n + 2\beta d\gamma n - 2\beta^2 d\varepsilon^2 n^2 + \beta^2 d\varepsilon\gamma n^2)^2\right] \tag{19}$$

Proof. Note that $G_{A,B}$ has the same eigenvalues as $\hat{G}_{A,B} := \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot \sqrt{D_{A,B}}\Sigma_{A,B}\sqrt{D_{A,B}}$. Hence,

$$\det(I + G_{A,B}) = \det\left(I + \frac{1}{(1-\varepsilon)\alpha^2 n + d} \cdot \sqrt{D_{A,B}}\Sigma_{A,B}\sqrt{D_{A,B}}\right) = \det\left(I + \frac{1}{(1-\varepsilon)\alpha^2 n + d} D_{A,B} \cdot \Sigma_{A,B}\right).$$

We can then can compute and factor the determinant as an expression of $\alpha, \varepsilon, \beta, \gamma, d$, and n. Writing the output as a difference of squares, one then obtains (19).[9] □

Now, we will set β to satisfy the quadratic equation $\alpha^2 n - 2\beta d\varepsilon n + \beta^2 d\varepsilon n^2 = 0$. This is equivalent to $\beta^2(d\varepsilon n) - (2d\varepsilon)\beta + \alpha^2 = 0$, for which we will set β to be the solution

$$\beta = \frac{d\varepsilon - \sqrt{d^2\varepsilon^2 - \alpha^2 \cdot d\varepsilon n}}{d\varepsilon n} = \frac{1}{n} \cdot \left(1 - \sqrt{1 - \frac{\alpha^2 n}{d\varepsilon}}\right).$$

Note that this is only possible if $\alpha^2 n < d\varepsilon$, so $n < \frac{d\varepsilon}{\alpha^2}$. In this case, we can simplify our expression as

$$\det(I + G_{A,B}) = \frac{(d + \beta^2 d(\varepsilon^2 - \gamma)n^2)^2 - O(\alpha^2\varepsilon n + \alpha^2\gamma n + \beta d\gamma n + \beta^2 d\varepsilon^2 n^2 + \beta^2 d\varepsilon\gamma n^2)^2}{(d + (1-\varepsilon)\alpha^2 n)^2}.$$

Using the fact that $\gamma \leq \varepsilon$, we can ignore the terms $\alpha^2\gamma n$ (smaller than $\alpha^2\varepsilon n$) and $\beta^2 d\varepsilon\gamma n^2$ (smaller than $\beta^2 d\varepsilon^2 n^2$). So, this simplifies to

$$\det(I + G_{A,B}) = \frac{(d + \beta^2 d(\varepsilon^2 - \gamma)n^2)^2 - O(\alpha^2\varepsilon n + \beta d\gamma n + \beta^2 d\varepsilon^2 n^2)^2}{(d + (1-\varepsilon)n\alpha^2)^2}.$$

In addition, note that if $n < \frac{d\varepsilon}{\alpha^2}$, then $\sqrt{1 - \frac{\alpha^2 n}{d\varepsilon}} \geq 1 - \frac{\alpha^2 n}{d\varepsilon}$, which means $\beta \leq \frac{1}{n} \cdot \frac{\alpha^2 n}{d\varepsilon} = \frac{\alpha^2}{d\varepsilon}$. Hence, if $n < \frac{d\varepsilon}{\alpha^2}$,

$$\det(I + G_{A,B}) \geq \frac{(d - \frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2)^2 - O(\alpha^2\varepsilon n + \frac{\alpha^2}{\varepsilon} \cdot \gamma n + \frac{\alpha^4}{d} \cdot n^2)^2}{(d + (1-\varepsilon)n\alpha^2)^2}. \tag{20}$$

Note that if $n \leq \frac{0.1 d\varepsilon}{\alpha^2}$, then $\beta \leq \frac{\alpha^2}{d\varepsilon} \leq \frac{0.1}{n}$. So, by combining (20) with Lemma 6.5, we have the following lemma.

[9] Some Mathematica code to verify the computation is provided in Appendix A.

Lemma 6.7. *Suppose that* $n \le \frac{0.1d\varepsilon}{\alpha^2}$ *and* $\beta = \frac{1}{n} \cdot (1 - \sqrt{1 - (\alpha^2 n)/(d\varepsilon)})$. *Then, for* $\gamma := \frac{|A \cap B|}{n}$, *we have*

$$\mathrm{D}_{\chi^2}(\mathcal{D}_1 || \mathcal{D}_0) \le \mathop{\mathbb{E}}_{A,B} \left[\left(\frac{(d - \frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2)^2 - O(\alpha^2 \varepsilon n + \frac{\alpha^2}{\varepsilon} \cdot \gamma n + \frac{\alpha^4}{d} \cdot n^2)^2}{d^2} \right)^{-d/2} \right]. \quad (21)$$

Recall that γ is the fraction of $[n]$ in both A and B, so the distribution of γ is $\frac{1}{n} \cdot \mathrm{HGeom}(n, \varepsilon n, \varepsilon n)$. Hence, $\gamma \in [0, \varepsilon]$ with probability 1, and by Corollary 2.15, $\mathbb{P}(\max(\gamma - \varepsilon^2, 0) > t) \le \exp\left(-\min\left(\frac{t^2 \cdot n}{4\varepsilon^2}, \frac{t \cdot n}{4}\right)\right)$.

We note the following simple proposition.

Proposition 6.8. *Suppose that* $n \le c \cdot \frac{d\varepsilon}{\alpha^2}$ *for some small constant* c. *Then, for any* A *and* B, *each of* $\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2$, $\alpha^2 \varepsilon n$, $\frac{\alpha^2}{\varepsilon} \cdot \gamma \cdot n$, *and* $\frac{\alpha^4}{d} \cdot n^2$ *is smaller than* $c \cdot d$.

Proof. Since $\gamma \le \varepsilon$, $\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 \le \frac{\alpha^4 n^2}{d\varepsilon}$. If $\frac{\alpha^4 n^2}{d\varepsilon} \ge c \cdot d$, then $n \ge \sqrt{c} \cdot \frac{d\sqrt{\varepsilon}}{\alpha^2} > c \cdot \frac{d\varepsilon}{\alpha^2}$. Next, $\alpha^2 \varepsilon n, \frac{\alpha^2}{\varepsilon} \cdot \gamma n \le \alpha^2 n$. If $\alpha^2 n \ge c \cdot d$, then $n \ge c \cdot \frac{d}{\alpha^2} > c \cdot \frac{d\cdot\varepsilon}{\alpha^2}$. Finally, if $\frac{\alpha^4}{d} \cdot n^2 \le c \cdot d$, then $n \ge \sqrt{c} \cdot \frac{d}{\alpha^2} > c \cdot \frac{d\cdot\varepsilon}{\alpha^2}$. □

The importance of Proposition 6.8 is that if $0 < x \le c$ for a sufficiently small constant c, $1 - x \ge e^{-2x}$. Therefore, we can rewrite the right-hand side of (21) as at most

$$\begin{aligned} &\mathop{\mathbb{E}}_{A,B} \left[\exp\left(O\left(\frac{\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2}{d} + \left(\frac{\alpha^2 \varepsilon n + \frac{\alpha^2}{\varepsilon} \cdot \gamma n + \frac{\alpha^4}{d} \cdot n^2}{d} \right)^2 \right) \cdot \frac{d}{2} \right) \right] \\ = &\mathop{\mathbb{E}}_{A,B} \left[\exp\left(O\left(\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 + \left(\frac{\alpha^2 \varepsilon n + \frac{\alpha^2}{\varepsilon} \cdot \gamma n + \frac{\alpha^4}{d} \cdot n^2}{\sqrt{d}} \right)^2 \right) \right) \right]. \quad (22) \end{aligned}$$

First, note that if we additionally have $n \le c \cdot \min\left(\frac{\sqrt{d}}{\alpha^2\varepsilon}, \frac{d^{3/4}}{\alpha^2}\right)$ for a small constant c, then $\alpha^2 \varepsilon n \le c\sqrt{d}$ and $\frac{\alpha^4}{d} \cdot n^2 \le c^2\sqrt{d}$. Also, note that $\frac{\alpha^2}{\varepsilon} \cdot \gamma n \le \alpha^2 \varepsilon n + \frac{\alpha^2}{\varepsilon} \cdot \max(0, \gamma - \varepsilon^2) n$, and that $\left[(\frac{\alpha^2}{\varepsilon} \cdot \max(0, \gamma - \varepsilon^2) n)/\sqrt{d} \right]^2 = \frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2)^2 \cdot n^2 \le \frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2$, since $\max(0, \gamma - \varepsilon^2) \le 1$. As a result, we can bound (22) as at most

$$\mathop{\mathbb{E}}_{A,B} \left[\exp\left(O\left(\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 + c^2 \right) \right) \right],$$

assuming $n \le c \cdot \min\left(\frac{d\varepsilon}{\alpha^2}, \frac{\sqrt{d}}{\alpha^2\varepsilon}, \frac{d^{3/4}}{\alpha^2}\right)$.

Now, for any value $t > 0$, we have that by Corollary 2.15,

$$\begin{aligned} \mathbb{P}_{A,B}\left(\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 > t \right) &= \mathbb{P}_{A,B}\left(\gamma - \varepsilon^2 > t \cdot \frac{d\varepsilon^2}{\alpha^4 n^2} \right) \\ &\le \exp\left(-\min\left(t^2 \cdot \frac{d^2\varepsilon^4}{\alpha^8 n^4} \cdot \frac{n}{4\varepsilon^2}, t \cdot \frac{d\varepsilon^2}{\alpha^4 n^2} \cdot \frac{n}{4} \right) \right) \\ &= \exp\left(-\min\left(t^2 \cdot \frac{d^2\varepsilon^2}{4\alpha^8 n^3}, t \cdot \frac{d\varepsilon^2}{4\alpha^4 n} \right) \right). \end{aligned}$$

If we additionally assume that $n \le c \cdot \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$ and $n \le c \cdot \frac{d\varepsilon^2}{\alpha^4}$, this is at most $\exp\left(-\min(t^2/4c^3, t/4c)\right)$. So, for $t \ge c$, the probability that $\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 \ge t$ is at most $e^{-t/4c}$, which means that

$$\mathop{\mathbb{E}}_{A,B}\left[\exp\left(O\left(\frac{\alpha^4}{d\varepsilon^2} \cdot \max(0, \gamma - \varepsilon^2) \cdot n^2 + c^2\right)\right)\right] \le e^{O(c)} \le 1.01.$$

To summarize what we have proved, in combination with Lemma 6.7, we have the following.

Lemma 6.9. *Assuming that*

$$n \le c \cdot \min\left(\frac{d\varepsilon}{\alpha^2}, \frac{\sqrt{d}}{\alpha^2\varepsilon}, \frac{d^{3/4}}{\alpha^2}, \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon^2}{\alpha^4}\right), \tag{23}$$

for some sufficiently small constant c, we have that

$$\mathrm{D}_{\chi^2}(\mathcal{D}_1||\mathcal{D}_0) \le 1.01.$$

However, we note that we can remove several of the terms in (23). More precisely, we have the following proposition.

Proposition 6.10. *Suppose that $n \le c \cdot \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)$ for some sufficiently small constant $c > 0$, and that $\varepsilon \le 1$ and $n \ge \frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^3}{\alpha^4}$. Then, $n \le c \cdot \min\left(\frac{d\varepsilon}{\alpha^2}, \frac{\sqrt{d}}{\alpha^2\varepsilon}, \frac{d^{3/4}}{\alpha^2}, \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon^2}{\alpha^4}\right)$.*

Proof. First, note that $\left(\frac{\sqrt{d}}{\alpha^2\varepsilon}\right)^{2/3} \cdot \left(\frac{d\varepsilon^3}{\alpha^4}\right)^{1/3} = \frac{d^{2/3}\varepsilon^{1/3}}{\alpha^{8/3}} \ge \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$, since $\varepsilon \le 1$. Therefore, if $\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}} > \frac{d\varepsilon^3}{\alpha^4}$, then $\frac{\sqrt{d}}{\alpha^2\varepsilon} > \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$. Thus, if $n \le c \cdot \min\left(\frac{d\varepsilon}{\alpha^2}, \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}\right)$ and $n \ge \frac{d\varepsilon^3}{\alpha^4} + \frac{\sqrt{d}}{\alpha^2}$, then also $n \le c \cdot \frac{\sqrt{d}}{\alpha^2\varepsilon}$. In addition, $\frac{d^{3/4}}{\alpha^2} = \sqrt{\frac{\sqrt{d}}{\alpha^2\varepsilon} \cdot \frac{d\varepsilon}{\alpha^2}}$, which means we also obtain $n \le c \cdot \frac{d^{3/4}}{\alpha^2}$, because we just showed that $n \le c \cdot \frac{\sqrt{d}}{\alpha^2\varepsilon}$ and we are assuming that $n \le c \cdot \frac{d\varepsilon}{\alpha^2}$. Finally, $\left(\frac{\sqrt{d}}{\alpha^2}\right)^{2/3} \cdot \left(\frac{d\varepsilon^2}{\alpha^4}\right)^{1/3} = \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$, which means that if $\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}} > \frac{\sqrt{d}}{\alpha^2}$, then $\frac{d\varepsilon^2}{\alpha^4} > \frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}$. So, by our assumptions, $n \le c \cdot \frac{d\varepsilon^2}{\alpha^4}$ as well. □

From here, the proof of Theorem 6.1 is straightforward.

Proof of Theorem 6.1. By Lemma 6.9 and Proposition 6.10, we have that under the assumptions of Theorem 6.1, $\mathrm{D}_{\chi^2}(\mathcal{D}_1||\mathcal{D}_0) \le 1.01$. By Fact 2.3, we have that $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_1||\mathcal{D}_0) \le 0.1$.

Finally, note that we created the adversarial samples and the mean vector μ first, and then generated the uncorrupted data, so the adversary is oblivious. Finally, $\|\mu\|_2 \le \|z\|_2 + \beta \cdot \|\mathbf{R}_A\|_2$. However, $z \sim \mathcal{N}(0, \frac{\alpha^2}{d} \cdot I)$ means $\|z\|_2 \le 2\alpha$ with very high probability. Moreover, $\beta \le \frac{\alpha^2}{\varepsilon d}$ and $\mathbf{R}_A$ is the sum of $\varepsilon \cdot n$ i.i.d. $\mathcal{N}(0, 1)$, so $\|\mathbf{R}_A\|_2 \le 2\sqrt{\varepsilon n d}$ with very high probability. So, $\beta \cdot \|\mathbf{R}_A\|_2 \le \frac{\alpha^2 \cdot 2\sqrt{\varepsilon n d}}{\varepsilon d} = 2\alpha^2\sqrt{\frac{n}{\varepsilon d}}$. Assuming that $n \le \frac{d\varepsilon}{\alpha^2}$, this is at most 2α. So overall, $\|\mu\|_2 \le 4\alpha$. We can replace α with $\alpha/4$ in the construction to finish the proof. □

7 The Sample Complexity under Strong Contamination

In this section, we leverage the tight sample complexity bounds for differentially private mean testing [Nar22], along with the robust-private equivalence of [GH22; HKMN22; AUZ23], to obtain the optimal sample complexity of robust mean testing under the strong contamination model:

Theorem 7.1. *For $\alpha \geq \varepsilon \cdot \mathrm{polylog}(d, \frac{1}{\varepsilon}, \frac{1}{\alpha})$,[10] the sample complexity of Gaussian mean testing in the adaptive contamination model is*

$$\tilde{\Theta}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d\varepsilon^2}{\alpha^4}\right).$$

The rest of this section is dedicated to the proof of this theorem. First, we recall the definition of differential privacy: for simplicity, and as it suffices for our purposes, we focus on "fully approximate" differentially private decision algorithms.

Definition 7.2. A randomized algorithm $\mathcal{A}\colon \mathcal{X}^n \to \{0,1\}$ is $(0,\delta)$-*differentially private* (DP) if for all datasets $X, X' \in \mathcal{X}^n$ that only differ in a single data point $X_i \neq X'_i$,

$$\left|\mathbb{P}(\mathcal{A}(X) = 1) - \mathbb{P}(\mathcal{A}(X') = 1)\right| \leq \delta.$$

Upper bound. To prove our upper bound, we will require the tight upper bound for DP mean testing:

Theorem 7.3 ([Nar22])**.** *For any parameters $0 < \alpha, \delta \leq \frac{1}{2}$, there exists a $(0,\delta)$-DP algorithm $\mathcal{A}$ that on*

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

samples $X_1, \ldots, X_n$, satisfies:

- *If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, then with probability at least 0.99 (over both the randomness of the samples and the algorithm), $\mathcal{A}(X) = 0$.*[11]
- *For any vector μ with $\|\mu\|_2 \geq \alpha$, if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$, then with probability at least 0.99, $\mathcal{A}(X) = 1$.*

Moreover, this is tight: any $(0,\delta)$-DP algorithm with these guarantees must take $\tilde{\Omega}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d^{1/3}}{\alpha^{4/3}\delta^{2/3}} + \frac{1}{\alpha\delta}\right)$ samples.

It is essentially folklore (see also [GH22]) that any $(0,\delta)$-DP decision algorithm that succeeds with 0.99 probability given n samples is automatically ε-robust in the strong (adaptive) corruption model for $\varepsilon := \frac{1}{10\delta n}$, and succeeds with at least $2/3$ probability over the input. For completeness, we briefly reproduce the argument here for the algorithm $\mathcal{A}$. If $X_1, \ldots, X_n$ are i.i.d., then with at least 0.9 probability, $\mathbb{P}(\mathcal{A}(X_1, \ldots, X_n) = 0) \geq 0.9$ over the randomness of the algorithm $\mathcal{A}$. Hence, for any ε-corruption of the data X' (i.e., $\varepsilon n = \frac{1}{10\delta}$ individual data points are possibly adaptively changed from X to X'), by the definition of privacy,

$$\left|\mathbb{P}(\mathcal{A}(X_1, \ldots, X_n) = 0) - \mathbb{P}(\mathcal{A}(X'_1, \ldots, X'_n) = 0)\right| \leq \delta \cdot \frac{1}{10\delta} = \frac{1}{10}$$

Hence, for any such corruption X', $\mathbb{P}(\mathcal{A}(X') = 0) \geq 0.89 > 2/3$. The same argument can be used to show that if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$ where $\|\mu\|_2 \geq \alpha$, with probability at least 0.9 over $X_1, \ldots, X_n$, $\mathbb{P}(\mathcal{A}(X') = 1) > 2/3$ for any ε-corruption of X.

[10] While this condition may seem somewhat restrictive, it is in fact inconsequential. Indeed, for $\varepsilon \leq \alpha \leq \varepsilon \cdot \mathrm{polylog}(d, 1/\varepsilon, 1/\alpha)$, one can use a robust *learning* algorithm with sample complexity $O\left(d/\alpha^2\right)$, which in this parameter regime becomes $\tilde{O}\left(d\varepsilon^2/\alpha^4\right)$.

[11] While [Nar22] did not state a 0.99 success probability, one can amplify the success probability by running several independent copies and using the majority output.

Thus, the algorithm of Theorem 7.3 readily implies an ε-robust one for robust mean testing, for $\varepsilon = \frac{1}{10\delta n}$. Plugging $\delta = \frac{1}{10\varepsilon n}$ in its sample complexity, it suffices for n to satisfy

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

This is equivalent to requiring

$$\alpha \geq \tilde{O}(\varepsilon) \quad \text{and} \quad n \geq \tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d\varepsilon^2}{\alpha^4}\right),$$

where $\tilde{O}$ may hide polylogarithmic factors in $d, \alpha^{-1}, \varepsilon^{-1}$. Hence, there exists a robust algorithm against strong contamination with sample complexity $\tilde{O}\left(\frac{d^{1/2}}{\alpha^2} + \frac{d\varepsilon^2}{\alpha^4}\right)$.

Lower bound. We next show this sample complexity is optimal, again by a reduction between robust and private algorithms. Suppose there exists a robust mean testing algorithm $\mathcal{A}$ that uses n samples. We set $\delta := \frac{1}{\varepsilon n}$, and construct a $(0, \delta)$-differentially private algorithm for mean testing using a black-box robustness-to-privacy transformation [HKMN22; AD20]. We will then use a lower bound from [Nar22], which will create a contradiction if n is too small.

To explain this transformation, first, for any two datasets X, X' of size n, we define the Hamming distance $d_{\mathrm{H}}(X, X')$ to be the number of indices i such that $X_i \neq X'_i$. Now, for any dataset $X = (X_1, \ldots, X_n)$, define the *score* $\mathcal{S}(X; \mathcal{A})$ of X (for $\mathcal{A}$) to be the smallest nonnegative integer k such that there exists a dataset X' of size n with $d_{\mathrm{H}}(X, X') = k$ and $\mathcal{A}(X') = 1$. Equivalently, $\mathcal{S}(X; \mathcal{A})$ represents the smallest number of points we need to alter from X to obtain some X' on which the robust algorithm would reject. (Note that if $\mathcal{A}(X) = 1$, then the score of X is simply 0.)

The differentially private algorithm $\mathcal{A}'$ on X computes $\mathcal{S}(X; \mathcal{A})$, and then outputs 1 with probability $\min(0, 1 - \delta \cdot \mathcal{S}(X; \mathcal{A}))$.

Note that $\mathcal{S}(X; \mathcal{A})$ changes by at most 1 between adjacent datasets X, X', because if $\mathcal{S}(X; \mathcal{A}) = k$, there exists X'' with $d_{\mathrm{H}}(X, X'') = k$ and $\mathcal{A}(X'') = 1$. But then, $d_{\mathrm{H}}(X', X'') \leq k+1$, so $\mathcal{S}(X'; \mathcal{A}) \leq k+1$. Likewise, we can show $\mathcal{S}(X'; \mathcal{A}) \geq k - 1$. This proves that the algorithm is $(0, \delta)$-differentially private, since the probability of outputting 1 changes by at most δ if the score changes by at most 1.

Next, if $X = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, then by the property of the robust algorithm, with probability at least $2/3$, every dataset X' of Hamming distance at most εn from X satisfies $\mathcal{A}(X') = 0$. Whenever this happens, $\mathcal{S}(X; \mathcal{A}) \geq \varepsilon n$, and thus conditioned on this the algorithm $\mathcal{A}'$ outputs 1 with probability 0, and hence always outputs 0.

Finally, if $X = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$, then with probability at least $2/3$, $\mathcal{A}(X) = 1$. Hence, with probability at least $2/3$ we have $\mathcal{S}(X; \mathcal{A}) = 0$, and conditioned on this the algorithm $\mathcal{A}'$ outputs 1 with probability 1.

Overall, this means that if $\mathcal{A}$ is robust against strong contamination, then there exists an algorithm $\mathcal{A}'$ that is $(0, \frac{1}{\varepsilon n})$-differentially private for the Gaussian mean testing problem, with the same number of samples n.

However, we can now invoke the lower bound part of Theorem 7.3 for DP Gaussian mean testing. From the above reduction, a robust algorithm using n samples yields an $(0, \frac{1}{\varepsilon n})$-DP algorithm with the same sample complexity, which by Theorem 7.3 means that one must have

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

This implies that

$$n = \tilde{\Omega}\left(\frac{d^{1/2}}{\alpha^2}\right) \qquad \text{and} \qquad n = \tilde{\Omega}\left(\frac{d\varepsilon^2}{\alpha^4}\right).$$

This concludes the proof of Theorem 7.1. □

Lower Bound against Additive Adversaries. Finally, we note that the same lower bound holds even if we restrict ourselves to adaptive adversaries that only can *add* points, and can never remove points. This again follows readily from the results of [Nar22], but is a consequence of an intermediate result proven in the paper rather than a direct black-box application of their private sample complexity lower bound. The lemma that we require is the following.

Lemma 7.4 ([Nar22, Theorem D.6, restated])**.** *Fix any $\alpha, \delta \le 1$ and any dimension d. There exists a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with support only on $\{\mu \in \mathbb{R}^d : \|\mu\|_2 \ge \alpha\}$, with the following property. Suppose $\mathcal{U}$ is the distribution over $(X_1, \dots, X_n) \in (\mathbb{R}^d)^n$ where each $X_i \sim \mathcal{N}(0, I)$, and $\mathcal{V}$ is the distribution over $(X_1, \dots, X_n) \in (\mathbb{R}^d)^n$ where we first draw $\mu \sim \mathcal{D}$ and then draw each $X_i \sim \mathcal{N}(\mu, I)$.*

Then, for some universal constants $c_1, c_2 > 0$, if $n \le c_1 \cdot \frac{d^{1/3}}{\alpha^{4/3}\cdot\delta^{2/3}}$ there exist distributions $\mathcal{U}', \mathcal{V}'$ over $(\mathbb{R}^d)^n$ such that $\mathrm{d_{TV}}(\mathcal{U},\mathcal{U}') \le 1/4$, $\mathrm{d_{TV}}(\mathcal{V},\mathcal{V}') \le 1/4$, and there is a coupling of $(\mathcal{U}', \mathcal{V}')$ such that $\mathbb{E}_{(X,Y)\sim(\mathcal{U}',\mathcal{V}')}[d_{\mathrm{H}}(X, Y)] \le c_2/\delta$, where d_{H} denotes the Hamming distance, i.e., the number of points that differ between X and Y.

Now, we show why Lemma 7.4 implies that any robust algorithm cannot distinguish between i.i.d. samples from $\mathcal{N}(0, I)$ and $\mathcal{N}(\mu, I)$, where μ is drawn from the distribution $\mathcal{D}$ in Lemma 7.4 under adaptive ε-additive contamination, unless the number of samples is at least $\Omega\left(\frac{d\varepsilon^2}{\alpha^4}\right)$. This would conclude the claim.

Fix $\alpha, \varepsilon \le 1$, and define $\delta = \frac{10c_2}{\varepsilon n}$, so that $\frac{10c_2}{\delta} = \varepsilon n$. Suppose that $n \le c_1 \cdot \frac{d^{1/3}}{\alpha^{4/3}\delta^{2/3}}$, which for $\delta = \frac{10c_2}{\varepsilon n}$ is equivalent to $n \le \frac{c_1^3}{100c_2^2} \cdot \frac{d\varepsilon^2}{\alpha^4}$. By Lemma 7.4 and Markov's inequality, $\mathbb{P}_{(X,Y)\sim(\mathcal{U}',\mathcal{V}')}\left[d_{\mathrm{H}}(X, Y) \ge 10c_2/\delta\right] \le 1/10$, which means by the coupling between $\mathcal{U}$ and $\mathcal{U}'$ and between $\mathcal{V}$ and $\mathcal{V}'$, there exists a coupling between $\mathcal{U}$ and $\mathcal{V}$ with $\mathbb{P}_{(X,Y)\sim(\mathcal{U},\mathcal{V})}\left[d_{\mathrm{H}}(X, Y) \ge 10c_2/\delta\right] \le 1/4 + 1/4 + 1/10 = 3/5$, i.e., such that $\mathbb{P}_{(X,Y)\sim(\mathcal{U},\mathcal{V})}\left[d_{\mathrm{H}}(X, Y) \le 10c_2/\delta\right] \ge 2/5$.

Consider such a coupling between $\mathcal{U}$ and $\mathcal{V}$. Suppose we generate $(X, Y) \sim (\mathcal{U}, \mathcal{V})$, and in the $2/5$ probability event $\{d_{\mathrm{H}}(X, Y) \le 10c_2/\delta\}$, we let $\hat{X} = \hat{Y} = X \cup Y$. Note that $\hat{X}$ can be created by adding at most $10c_2/\delta$ points to X and at most $10c_2/\delta$ points to Y. Otherwise, we let $\hat{X} = X$ and $\hat{Y} = Y$. Importantly, this means there exists a distribution over $\hat{X}$ and $\hat{Y}$ (which are generated only by additive adaptive contamination of $\frac{10c_2}{\delta} = \varepsilon n$ points) such that with $2/5$ probability, $\hat{X}$ and $\hat{Y}$ are the same. So, the total variation distance between the distributions is at most $3/5$, which means no algorithm can successfully distinguish between the two distributions with more than 80% probability.

In summary, there cannot exist an algorithm that uses $n \le \frac{c_1^3}{100c_2^2} \cdot \frac{d\varepsilon^2}{\alpha^4}$ samples and distinguishes between samples from $\mathcal{N}(0, I)$ and $\mathcal{N}(\mu, I)$ where $\mu \sim \mathcal{D}$, under ε-additive adaptive contamination. Finally, because there exists an $\Omega(\sqrt{d}/\alpha^2)$-lower bound even against uncorrupted samples [SD08; DKS17], we conclude that the sample complexity of robust Gaussian mean testing against additive adaptive adversaries is

$$\Omega\left(\frac{\sqrt{d}}{\alpha^2} + \frac{d\varepsilon^2}{\alpha^4}\right),$$

as claimed.

8 Polynomial-Time Algorithm

Theorem 8.1. *Let $d \in \mathbb{N}, \delta > 0$. Let $\alpha = O(1)$ and assume that $C\varepsilon\sqrt{\log 1/\varepsilon} \le \alpha$ and*

$$n \ge \Omega\left(\frac{\sqrt{d}\log 1/\delta}{\alpha^2} + \frac{d\varepsilon^2 \log 1/\delta}{\alpha^4}\operatorname{poly}\log(d, 1/\varepsilon, 1/\alpha, \log 1/\delta)\right).$$

Then, there is an algorithm which runs in time $O(\varepsilon n^2 d \min(n,d) + nd)$ with the following guarantees:

- *For every $\mu \in \mathbb{R}^d$ with $\|\mu\| = \alpha$, with probability $1-\delta$ over n independent samples $X_1, \dots, X_n$ from $\mathcal{N}(\mu, I)$, given any adaptive ε-corruption of $X_1, \dots, X_n$, the algorithm outputs* YES.
- *With probability $1-\delta$ over n independent samples $X_1, \dots, X_n$ from $\mathcal{N}(0, I)$, given any adaptive ε-corruption of $X_1, \dots, X_n$, the algorithm outputs* NO.

We briefly mention that the assumption that $\|\mu\|$ is exactly α is largely for notational convenience this section. It is straightforward to verify that the same arguments also extend to testing when the mean of the alternative hypothesis satisfies $\alpha \le \|\mu\| \le O(1)$, which implies the same results for $\alpha \le \|\mu\|$ (see Subsection 2.3).

8.1 Regularity conditions

We will seek to algorithmically enforce a set of regularity conditions which are guaranteed to be satisfied by any set of uncorrupted points, from either the null or alternate hypothesis. We demonstrate that if these regularity conditions are satisfied, then the norm of the sum of the samples will suffice to distinguish between the two cases, with high probability. Concretely, the regularity condition we will require is the following:

Definition 8.2. Let $S = \{X_1, \dots, X_n\}$ be a set of points in $\mathbb{R}^d$. We say that S is $(\varepsilon, \beta_1, \beta_2)$-regular if for all sets $T \subset S$ with $|T| \le \varepsilon n$, we have:

(i) $\sum_{i \in T} \|X_i\|^2 = |T|d \pm O(\beta_1)$,

(ii) $\|\mathrm{Sum}(T)\|^2 = |T|d \pm O(\beta_2)$, and

(iii) $|\langle \mathrm{Sum}(T), \mathrm{Sum}(S)\rangle| = |T|d \pm O(\sqrt{n}\beta_1)$.

We first note the following bound:

Lemma 8.3. *Let $\alpha = O(1)$, let $\varepsilon, \delta > 0$ be at most a sufficiently small constant, and let $S = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$ be a set of n independent draws from $\mathcal{N}(\mu, I)$, where $\|\mu\|_2 \le \alpha$, and suppose that $n \ge \log(1/\delta)/\varepsilon$. Then, with probability $1-\delta$, S is*

$$\left(\varepsilon, \varepsilon n\sqrt{d}\log(n/\delta), (\varepsilon n)^2 \log 1/\varepsilon + \varepsilon n\sqrt{\varepsilon n d \log 1/\varepsilon}\right)\text{-regular}.$$

Proof. We prove that S satisfies each bullet point in sequence. To prove the first bullet point, by Fact 2.8, with probability $1-\delta/3$, for all $i \in [n]$ $|\|X_i\|^2 - d| \le 10\left(\sqrt{\log(3n/\delta)d} + \log(3n/\delta)\right) \le 30\sqrt{d}\cdot\log(n/\delta)$. Assuming this holds for all i, then for any subset $T \subset S$, $\sum_{i\in T}\|X_i\|^2 = |T|d \pm |T|\cdot 30\sqrt{d}\log(n/\delta)$. Hence, if $|T| \le \varepsilon n$, this equals $|T|d \pm O(\varepsilon n\sqrt{d}\log(n/\delta))$, as desired.

We now prove the second bullet point. Fix any T satisfying $|T| \le \varepsilon n$. Then, $\mathrm{Sum}(T) \sim \mathcal{N}(|T|\mu, |T|I)$, so if we let $Z = |T|^{-1/2}(\mathrm{Sum}(T) - |T|\mu)$, we have that $Z \sim \mathcal{N}(0, I)$, and

$$\|\mathrm{Sum}(T)\|^2 = \|\mu\|^2 \cdot |T|^2 + |T|^{3/2}\langle \mu, Z\rangle + |T|\,\|Z\|^2 .$$

Hence, we have that, for any $C > 0$, there exists $C' > 0$ so that

$$\Pr\left[|\langle \mu, Z\rangle| \geq C'\alpha\sqrt{\varepsilon n \log 1/\varepsilon}\right] \leq \exp(-C\varepsilon n \log 1/\varepsilon) \ , \text{ and}$$
$$\Pr\left[\left|\|Z\|^2 - d\right| > C'\sqrt{\varepsilon n d \log 1/\varepsilon} + C'\varepsilon n \log 1/\varepsilon\right] \leq \exp(-C\varepsilon n \log 1/\varepsilon) \ .$$

The number of subsets $T \subset S$ of size at most εn is at most $\exp\left(O(\varepsilon n \log 1/\varepsilon)\right)$. Therefore, by a union bound over all choices of T, with probability $1 - \exp(-\Omega(\varepsilon n \log 1/\varepsilon)) = 1 - \delta/3$, we have that

$$\left|\|\mathrm{Sum}(T)\|^2 - (\|\mu\|^2 \cdot |T|^2 + |T|d)\right| \leq O(\varepsilon n\sqrt{\varepsilon n d \log 1/\varepsilon} + (\varepsilon n)^2 \log 1/\varepsilon) \ , \tag{24}$$

for all T with $|T| \leq \varepsilon n$. Since $\|\mu\|^2 \cdot |T|^2 \leq \alpha^2 |T|^2 = O(\varepsilon n)^2$, this implies that

$$\left|\|\mathrm{Sum}(T)\|^2 - |T|d\right| \leq O(\varepsilon n\sqrt{\varepsilon n d \log 1/\varepsilon} + (\varepsilon n)^2 \log 1/\varepsilon) \ , \tag{25}$$

as claimed. Condition on this event holding for the rest of the proof.

Finally, we prove the third bullet point. For any $i = 1, \ldots, n$, note that

$$\langle X_i, \mathrm{Sum}(S)\rangle = \|X_i\|^2 + \langle X_i, \mathrm{Sum}(S \setminus \{i\})\rangle \ .$$

The second term on the RHS is the inner product of two independent Gaussians, and hence is subexponential, with variance proxy $(n-1)d$. Therefore, with probability $1 - \delta/3$, we have that $|\langle X_i, \mathrm{Sum}(S \setminus \{i\})\rangle| \leq \sqrt{nd}\log(n/\delta)$ for all $i = 1, \ldots, n$. Condition on this holding. Then, for any fixed T satisfying $|T| \leq \varepsilon n$, we have that

$$\begin{aligned}
\langle \mathrm{Sum}(T), \mathrm{Sum}(S)\rangle &= \sum_{i \in T} \langle X_i, \mathrm{Sum}(S)\rangle \\
&= \sum_{i \in T} \|X_i\|_2^2 + \sum_{i \in T} \langle X_i, \mathrm{Sum}(S \setminus \{i\})\rangle \\
&= d|T| \pm O\left(\varepsilon n\sqrt{d}\log(n/\delta) + \varepsilon n\sqrt{nd}\log(n/\delta)\right) \\
&= d|T| \pm O\left(\varepsilon n\sqrt{nd}\log(n/\delta)\right) \ ,
\end{aligned}$$

as claimed. Combining these bounds immediately yields the desired claim. □

Next, we note some simple consequences of regularity. For any subset $T \subset S$, we let $\mathbf{1}_T$ denote the indicator vector of T, and $\mathbf{1} = \mathbf{1}_S$ denote the vector which is all 1's. We also define $X_T := \sum_{i \in T} X_i$.

Proposition 8.4. *Suppose S is $(2\varepsilon, \beta_1, \beta_2)$-regular. Then, for any sets T, T' of size at most εn,*

$$\langle X_T, X_{T'}\rangle = d \cdot |T \cap T'| \pm O(\beta_2).$$

Proof. It is straightfoward to verify that

$$\langle X_T, X_{T'}\rangle = \frac{1}{2}\left[\|X_{T \cup T'}\|^2 + \|X_{T \cap T'}\|^2 - \|X_{T \setminus T'}\|^2 - \|X_{T' \setminus T}\|^2\right].$$

Each of $T \cup T', T \cap T', T \setminus T', T' \setminus T$ have size at most $2\varepsilon n$. So, by regularity (Part ii of Definition 8.2),

$$\langle X_T, X_{T'}\rangle = \frac{1}{2} \cdot d \cdot \left(|T \cup T'| + |T \cap T'| - |T \setminus T'| - |T' \setminus T|\right) \pm O(\beta_2) = d \cdot |T \cap T'| \pm O(\beta_2). \quad \square$$

We note the following useful convexity lemma.

Fact 8.5. *Let $k \le n$ be a nonnegative integer, and $w \in [0,1]^n$ be an n-dimensional vector with $\|w\|_1 \le k$. Then, w is a convex combination of the points $\mathbf{1}_T$ over $T \subseteq S, |T| \le k$.*

Proof. Without loss of generality assume that the coordinates of w are sorted in increasing order, i.e. $0 \le w_1 \le \ldots, w_n \le 1$, and additionally define $w_0 = 0$ and $w_{n+k} = 1$ for all $k \ge 0$. For all integers $i \ge 0$, let $S_i = \{i, \ldots, i+k\} \cap [n]$, so that in particular $S_{n+j} = \emptyset$ for all $j > 0$. Recursively define weights $a_1, \ldots, a_n$ by $a_1 = w_1$, and $a_i = w_i - \sum_{j=i-k}^{i-1} a_j$, where we set $a_j = 0$ for $j < 0$. Then by construction, we have that $w = \sum_{i=1}^n a_i \mathbf{1}_{S_i}$. We will show that $a_i \ge 0$ for all i, and that $\sum_{i=1}^n a_i \le 1$, from which the claim immediately follows. To prove the first claim, we proceed by induction. Note that the base case is trivial, and moreover, if the claim is true for some $i < n$, then

$$\sum_{j=i-k+1}^{i} a_i = a_i + \sum_{j=i-k+1}^{i-1} a_i = w_i - a_{i-k} \le w_i \le w_{i+1} \ ,$$

so in particular $a_{i+1} \ge 0$, which proves the induction.

To prove the second claim, we simply observe that by nonnegativity, we have that

$$k \ge \sum_{i=1}^n w_i = \sum_{i=1}^n \sum_{j=i-k}^{i} a_i \ge k \sum_{i=1}^n \ ,$$

where the last inequality follows from the fact that each a_i appears at most k times in the sum. Simplifying then immediately yields the claim. □

We use Fact 8.5 to generalize Proposition 8.4 as follows.

Proposition 8.6. *Suppose S is $(2\varepsilon, \beta_1, \beta_2)$-regular, and $a, b \in [0,1]^n$ are n-dimensional vectors such that $\sum a_i, \sum b_i \le \varepsilon \cdot n$. Then, we have*

$$\left\langle \sum_{i \in S} a_i X_i, \sum_{i \in S} b_j X_j \right\rangle = d \cdot \left(\sum_{i \in S} a_i b_i \right) \pm O(\beta_2) \quad \text{and} \quad \left\langle \sum_{i \in S} a_i X_i, X_S \right\rangle = d \cdot \left(\sum_{i \in S} a_i \right) \pm O(\sqrt{n} \cdot \beta_1).$$

Proof. By Fact 8.5, we can write a as a convex combination of $\mathbf{1}_T$ over $T \subset S, |T| \le \varepsilon n$. In other words, there exists a distribution $\mathcal{T}_1$ over $T_1 \subset S, |T_1| \le \varepsilon n$ such that $a_i = \mathbb{P}_{T_1 \sim \mathcal{T}_1}(i \in T_1)$. Likewise, there exists an (independent) distribution $\mathcal{T}_2$ over $T_2 \subset S, |T_2| \le \varepsilon n$ such that $b_j = \mathbb{P}_{T_2 \sim \mathcal{T}_2}(j \in T_2)$. Now,

$$\left\langle \sum a_i X_i, \sum b_j X_j \right\rangle = \mathbb{E}_{T_1 \sim \mathcal{T}_1, T_2 \sim \mathcal{T}_2} \langle X_{T_1}, X_{T_2} \rangle.$$

By Proposition 8.4,

$$\mathbb{E}_{T_1 \sim \mathcal{T}_1, T_2 \sim \mathcal{T}_2} \langle X_{T_1}, X_{T_2} \rangle = \mathbb{E}_{T_1 \sim \mathcal{T}_1, T_2 \sim \mathcal{T}_2} \left[d \cdot |T_1 \cap T_2| \right] \pm O(\beta_2).$$

Next, the expectation of $|T_1 \cap T_2|$, using linearity of expectation and independence of $\mathcal{T}_1, \mathcal{T}_2$, equals

$$\sum_{i \in S} \mathbb{P}_{T_1 \sim \mathcal{T}_1, T_2 \sim \mathcal{T}_2}(i \in T_1 \cap T_2) = \sum_{i \in S} \mathbb{P}_{T_1 \sim \mathcal{T}_1}(i \in T_1) \cdot \mathbb{P}_{T_2 \sim \mathcal{T}_2}(i \in T_2) = \sum_{i \in S} a_i b_i.$$

Overall, this implies that

$$\left\langle \sum a_i X_i, \sum b_j X_j \right\rangle = d \cdot \left(\sum_{i \in S} a_i b_i \right) \pm O(\beta_2).$$

Next, by regularity (part iii of Definition 8.2), we have that

$$\left\langle \sum a_i X_i, X_S \right\rangle = \mathbb{E}_{T_1 \sim \mathcal{T}_1} \langle x_{T_1}, x_S \rangle = \mathbb{E}_{T_1 \sim \mathcal{T}_1} \left(d \cdot |T_1| \right) \pm O(\sqrt{n} \cdot \beta_1) = d \cdot \left(\sum_{i \in S} a_i \right) \pm O(\sqrt{n} \cdot \beta_1). \quad \square$$

8.2 Filtering preliminaries

Our algorithm for doing so will be based on the (soft) filter framework developed for robust estimation in other contexts. Here we establish the notation and preliminaries we will require to design and analyze our algorithm. We note that it will be convenient for us to use slightly nonstandard versions of the notation compared to the literature.

Our algorithm will assign weights to each point, that we will monotonically decrease over time. For any n, let Γ_n denote the set of valid weights:

$$\Gamma_n = \{ w \in \mathbb{R}^n : w_i \in [0,1] \text{ for all } i = 1, \dots, n \} .$$

Recall that for any set $T \subseteq S$, $\mathbf{1}_T \in \Gamma_n$ denotes the indicator vector for T, and $\mathbf{1} = \mathbf{1}_S$.

Let K be some value to be specified later. Given a set of points $S = \{X_1, \dots, X_n\}$, we associate it weight vectors $w^{(t)} \in \Gamma_n$, for $i = 1, \dots, n$ and $t = 1, \dots, K$ where initially we set $w^{(1)} = \mathbf{1}$.[12] For any such weight vector w, we let

$$\mathrm{Sum}(w, S) = \sum_{i \in S} \sqrt{w_i} X_i \text{ , and } M(w, S) = \sum_{i \in S} w_i X_i X_i^\top .$$

When the context is clear, we will drop the S from the notation for simplicity, i.e. we will let $\mathrm{Sum}(w) = \mathrm{Sum}(w, S)$. For any set T, and for any set of weights w on S, we let w_T denote the set of weights restricted to the indices in $T \cap S$. We also let $\mathrm{Gram}(w, S) = \mathrm{Gram}(w)$ be the $n \times n$ matrix given by

$$\mathrm{Gram}(w)_{ij} = \sqrt{w_i w_j} \langle X_i, X_j \rangle .$$

Note that by design the nontrivial eigenvalues of $\mathrm{Gram}(w)$ and $M(w)$ are identical.

Recall that when the samples S are an ε-corruption of G, this means that there are sets B, R so that $|B| = |R| = \varepsilon n$ so that $R \subset G$ and so that $S = (G \setminus R) \cup B$. For the remainder of this section, S, G, B, R will always refer to these sets. We define the following important set:

$$\mathfrak{S}_n = \{ w \in \Gamma_n : \left\| \mathbf{1}_{G \setminus R} - w_{G \setminus R} \right\|_1 \leq 5 \left\| \mathbf{1}_B - w_B \right\|_1 \} ,$$

that is, $\mathfrak{S}_n$ is the set of weights where we have removed at most five times as much weight from the good samples as we have removed from the the bad samples.

Finally, we will also seek to enforce regularity conditions on weighted subsets of points. We will require the following natural generalization of Definition 8.2:

[12] As is common in this literature, for simplicity of notation we will conflate S with the set of indices in S.

Definition 8.7. Let $S = \{X_1, \dots, X_n\}$ be a set of points in $\mathbb{R}^d$, and let $w \in \Gamma_n$. We say that w is $(\varepsilon, \beta_1, \beta_2)$-regular if for all sets $T \subset S$ with $|T| \le \varepsilon n$, we have:

(i) $\sum_{i \in T} \|X_i\|^2 = |T| d \pm O(\beta_1)$,

(ii) $\|\mathrm{Sum}(w, T)\|^2 = \|w_T\|_1 \, d \pm O(\beta_2)$, and

(iii) $|\langle \mathrm{Sum}(w, T), \mathrm{Sum}(w, S) \rangle| = \|w_T\|_1 \, d \pm O(\sqrt{n}\beta_1)$.

The key fact we will use is the following:

Lemma 8.8. *Let $\alpha = O(1)$, let $\varepsilon, \delta \in [0, 1)$, and let G be $(4\varepsilon, \beta_1, \beta_2)$-regular, where*

$$\beta_1 \ge \varepsilon n \sqrt{d} \log(n/\delta), \beta_2 \ge \varepsilon n \sqrt{\varepsilon n d \log 1/\varepsilon} + (\varepsilon n)^2 \log 1/\varepsilon \; .$$

Further, assume that

$$\|\mathrm{Sum}(G)\|^2 = dn + \|\mu\|^2 n^2 \pm O(\alpha n^{3/2} \sqrt{\log 1/\delta} + n\sqrt{d} \log(1/\delta)) \; .$$

Let S be an ε-contamination of G, and let $w \in \Gamma_n$ be a set of weights on w that satisfy $\|w\|_1 \ge (1 - \varepsilon)n$, and w is $(\varepsilon, \beta_1, \beta_2)$-regular. Then, we have that

$$\|\mathrm{Sum}(w, S)\|^2 = d \|w\|_1 + \|\mu\|^2 n^2 \pm O(\alpha n^{3/2} \sqrt{\log 1/\delta} + n\sqrt{d} \log(1/\delta) + \sqrt{n}\beta_1 + \beta_2) \; .$$

In particular, if $n \ge C \cdot \frac{\sqrt{d} \cdot \log 1/\delta}{\alpha^2}$ and $\sqrt{n}\beta_1, \beta_2 < \frac{1}{C} \cdot \alpha^2 n^2$ for C sufficiently large, then:

- *if $\mu = 0$, then $\left| \|\mathrm{Sum}(w, S)\|^2 - d \|w\|_1 \right| \le 0.4\alpha^2 n^2$, and*
- *if $\|\mu\| = \alpha$, then $\left| \|\mathrm{Sum}(w, S)\|^2 - d \|w\|_1 \right| \ge 0.7\alpha^2 n^2$.*

In other words, the norm of the sum of the set of points distinguishes between the null and alternative hypotheses.

Proof. First, we bound $\| \mathrm{Sum}(w, G \setminus R)\|^2$. Let a be the vector that equals 1 on the indices in R and $1 - \sqrt{w_i}$ on other indices, so that $\sum_{i \in G} a_i G_i + \mathrm{Sum}(w, G \setminus R) = \mathrm{Sum}(G)$. Then,

$$\| \mathrm{Sum}(w, G \setminus R)\|^2 = \| \mathrm{Sum}(G)\|^2 - 2 \left\langle \mathrm{Sum}(G), \sum_{i \in G} a_i G_i \right\rangle + \left\| \sum_{i \in G} a_i G_i \right\|^2 .$$

Let $\beta_3 := \alpha n^{3/2} \sqrt{\log 1/\delta} + n\sqrt{d} \log(1/\delta)$. Because $\|a\|_1 \le 2\varepsilon$ and G is $(4\varepsilon, \beta_1, \beta_2)$ regular, Proposition 8.6 and our assumption on $\| \mathrm{Sum}(G)\|^2$ imply that

$$\begin{aligned}
\|\mathrm{Sum}(w, G \setminus R)\|^2 &= dn + \|\mu\|^2 n^2 \pm O(\beta_3) - 2d \sum a_i \pm O(\sqrt{n}\beta_1) + d \sum a_i^2 \pm O(\beta_2) \\
&= d \cdot \left(\sum (1 - a_i)^2 \right) + \|\mu\|^2 n^2 \pm O(\sqrt{n}\beta_1 + \beta_2 + \beta_3) \\
&= d \left\| w_{G \setminus R} \right\|_1 + \|\mu\|^2 n^2 \pm O(\alpha n^{3/2} \sqrt{\log 1/\delta} + n\sqrt{d} \log(1/\delta) + \sqrt{n}\beta_1 + \beta_2) \; ,
\end{aligned}$$

since $(1 - a_i)^2 = w_i$ for $i \in G \setminus R$ and 0 otherwise.

Then, the regularity of w implies that

$$\begin{aligned}
\|\mathrm{Sum}(w, S)\|^2 &= \|\mathrm{Sum}(w, G \setminus R)\|^2 + 2\langle \mathrm{Sum}(w, G \setminus R), \mathrm{Sum}(w, B) \rangle + \|\mathrm{Sum}(w, B)\|^2 \\
&= \|\mathrm{Sum}(w, G \setminus R)\|^2 + 2\langle \mathrm{Sum}(w, S), \mathrm{Sum}(w, B) \rangle - \|\mathrm{Sum}(w, B)\|^2 \\
&= \|w\|_1 \, d + \|\mu\|^2 n^2 \pm O(\alpha n^{3/2} \sqrt{\log 1/\delta} + n\sqrt{d} \log(1/\delta) + \sqrt{n}\beta_1 + \beta_2) \; .
\end{aligned}$$

The second half of the claim then follows from straightforward calculations. □

8.3 Additional preliminaries

We first prove bounds on the eigenvalues of the Gram matrix of random Gaussian samples. To do so, we require properties about *Wishart matrices*, which we now define.

Definition 8.9. A Wishart matrix $W = \mathrm{W}_d(n)$ has distribution $W = H^\top H$, where $H \in \mathbb{R}^{n\times d}$ has every entry drawn as an i.i.d. standard Gaussian $\mathcal{N}(0,1)$. Note that $W \in \mathbb{R}^{d\times d}$, and W is positive semidefinite.

We will need the following concentration bound for the eigenvalues of a Wishart matrix.

Lemma 8.10. *(Follows from [DS03, Theorem II.13]) If $W \sim \mathrm{W}_d(n)$, then with probability $1-\delta$,*

$$\|S - n\cdot I\| \le O\left(\sqrt{nd} + \sqrt{n\log(1/\delta)} + d + \log(1/\delta)\right).$$

We now provide eigenvalue bounds for samples drawn from $\mathcal{N}(\mu, I)$.

Fact 8.11. *Let $\mu \in \mathbb{R}^d$ have $\|\mu\| \le \alpha$ and let $\delta > 0$. If $n \le d$ and $X_1,\dots,X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$, then with probability at least $1-\delta$, we have*

$$\|\operatorname{Gram}(\{X_1,\dots,X_n\}) - d\cdot I\| \le O(\max(\sqrt{nd}, \sqrt{d\log(1/\delta)}, \log(1/\delta), \alpha^2 n)).$$

If $n \ge d$ and $X_1,\dots,X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$, then with probability at least $1-\delta$, we have

$$\left\|\sum_{i\in[n]} X_iX_i^\top - n\cdot I\right\| \le O(\max(\sqrt{nd}, \sqrt{n\log(1/\delta)}, \log(1/\delta), \alpha^2 n)).$$

Proof. First, note that when $n \le d$, the nonzero (top n) eigenvalues of $\sum X_iX_i^\top$ match the eigenvalues of $\operatorname{Gram}(\{X_1,\dots,X_n\})$. So, in the $n \le d$ case we can focus on the top n eigenvalues of $\sum_{i\in[n]} X_iX_i^\top$. This will allow us to consolidate calculations for both the $n \le d$ and $n \ge d$ case.

Let $Y_i := X_i - \mu$. We can write

$$\sum_{i\le n} X_iX_i^\top = \sum_{i\le n} Y_iY_i^\top + \sum_{i\le n} Y_i\mu^\top + \sum_{i\le n} \mu Y_i^\top + n\cdot\mu\mu^\top.$$

Note that $\sum_{i\le n} Y_iY_i^\top \sim \mathrm{W}_d(n)$, so with probability at least $1-\delta$,

$$\left\|\sum_{i\in[n]} Y_iY_i^\top - n\cdot I\right\| \le O\left(\sqrt{nd} + \sqrt{n\log(1/\delta)} + d + \log(1/\delta)\right).$$

In the $n \le d$ case, we note that $\sum Y_iY_i^\top$ has the same nonzero eigenvalues as $\operatorname{Gram}(\{Y_1,\dots,Y_n\}) \sim W_n(d)$. So, with probability at least $1-\delta$, the top n eigenvalues of $\sum Y_iY_i^\top$ are in the range

$$d \pm O\left(\sqrt{nd} + \sqrt{d\log(1/\delta)} + n + \log(1/\delta)\right).$$

Next, $\sum_{i\le n} Y_i\mu^\top$ is a rank-1 matrix with operator norm $\|\sum_{i\le n} Y_i\|\cdot\|\mu\| \le \alpha\cdot\|\sum_{i\le n} Y_i\|$. Since $\sum_{i\le n} Y_i \sim \mathcal{N}(0, n\cdot I)$, with probability at least $1-\delta$ it has norm at most $O(\sqrt{nd + n\log(1/\delta)})$, which means $\|\sum_{i\le n} Y_i\mu^\top\| \le O(\alpha\sqrt{nd + n\log(1/\delta)}) \le O(\sqrt{nd} + \sqrt{n\log(1/\delta)})$. The same bound holds for $\sum_{i\le n} \mu Y_i^\top$. Finally, $\|n\mu\mu^\top\| = n\cdot\|\mu\|\cdot\|\mu^\top\| \le n\alpha^2$.

In the $n \ge d$ case, adding the bounds together completes the proof. In the $n \le d$ case, adding the bounds together tells us the top n eigenvalues of $\sum X_iX_i^\top$ are in the desired range, which completes the proof. □

The next fact we will need is a direct corollary of Lemma 4.1 in [DHL19].

Fact 8.12. *Let $\alpha = O(1)$, let $\mu \in \mathbb{R}^d$ have $\|\mu\| \le \alpha$, and let $\varepsilon, \delta > 0$. Suppose that $n \ge \Omega(1/\varepsilon)$. Then, there is some universal constant $c > 0$ so that with probability $1 - \delta$, we have that for all v with $\|v\| = 1$, and all w supported on G with $\|w\|_1 \le 10\varepsilon n$, it holds that*

$$\sum_{i \in G} w_i \langle v, X_i \rangle^2 \le c \cdot (\varepsilon n \log 1/\varepsilon + d + \log 1/\delta) \; .$$

Finally, we will require the following downweighting scheme:

Fact 8.13. *Let $w \in \mathfrak{S}_n$, and let $\tau_1, \dots, \tau_n$ be a set of nonnegative scores satisfying $\sum_{i \in G \setminus R} w_i \tau_i < 5 \sum_{i \in B} w_i \tau_i$. Let $w' \in \Gamma_n$ be defined by*

$$w'_i = \left(1 - \frac{\tau_i}{\max_{i \in S} \tau_i}\right) w_i \; .$$

Then $\mathrm{supp}(w') \subset \mathrm{supp}(w)$, and moreover $w' \in \mathfrak{S}_n$.

8.4 The filtering algorithm for $n \le d$

In this case, the filtering algorithm proceeds as follows. Let $\delta > 0$, and let

$$\gamma_2 := C \left(\sqrt{nd} + \alpha^2 n + \sqrt{(n+d)\log(1/\delta)} + \log(1/\delta) + \varepsilon n \log 1/\varepsilon \right) \; , \tag{26}$$

for some constant C sufficiently large. Initialize weights $w^{(1)} = \mathbf{1}$. Then, for $t = 1$ until termination, we proceed as follows. For any $w \in \Gamma_n$, let $D(w) = d \cdot \mathrm{diag}(w)$. Let λ denote the top singular value of $\mathrm{Gram}(w, S) - D(w)$, and let v be its associated singular unit vector (if there are multiple, choose any). If $\lambda < 5\gamma_2$, then terminate. Otherwise, for all $i \in S$, let $\tau_i = \frac{v_i^2}{w_i^{(t)}} \mathbb{I}[w_i^{(t)} > 0]$ (where τ_i defaults to 0 when $w_i^{(t)} = 0$), and proceed to sort the samples in decreasing order of τ_i. Then, define $w^{(t+1)}$ by

$$w_i^{(t+1)} = \left(1 - \frac{\tau_i}{\max_i \tau_i}\right) w_i^{(t)} \; .$$

The formal pseudocode for this algorithm appears in Algorithm 4.

Algorithm 4 Spectral filtering for $n \le d$. Input: $X_1, \dots, X_n \in \mathbb{R}^d$, $\gamma_2 > 0$.

Let $w^{(1)} = \mathbf{1}$, and let $t = 1$
while $\left\| \mathrm{Gram}(w^{(t)}, S) - D(w^{(t)}) \right\| \ge 5\gamma_2$ **do**
  Let v be the top singular vector of $\mathrm{Gram}(w^{(t)}, S) - D(w^{(t)})$
  For all i, let $\tau_i = \frac{v_i^2}{w_i^{(t)}} \mathbb{I}[w_i^{(t)} > 0]$ ▷ If $w_i^{(t)} = 0$, we set $\tau_i = 0$.
  Let

$$w_i^{(t+1)} = \left(1 - \frac{\tau_i}{\max_i \tau_i}\right) w_i^{(t)} \; .$$

  Let $t \leftarrow t + 1$
Return $w^{(t)}$

For the rest of the section, let us assume that G is $(\varepsilon, \varepsilon n\sqrt{d}\log(n/\delta), (\varepsilon n)^2 \log 1/\varepsilon + \varepsilon n\sqrt{\varepsilon n d \log 1/\varepsilon})$-regular, and additionally assume that

$$\|\mathrm{Gram}(G) - dI\| \leq \frac{\gamma_2}{10} \,. \tag{27}$$

By Lemma 8.3 and Fact 8.11, these two conditions hold together with probability $1-\delta$. Then, our main claim for this algorithm is the following:

Lemma 8.14. *Under the above assumptions, Algorithm 4 terminates in K iterations for some $K \leq 6\varepsilon n$, runs in time $O(dn^2)$ per iteration, and moreover, at termination, we have that*

- *$w^{(K)} \in \mathfrak{S}_n$, and*
- *for all $T \subset S$ with $|T| \leq \varepsilon n$, we have that*

$$\left\|\mathrm{Sum}(w^{(K)}, T)\right\|^2 = \left\|w^{(K)}(T)\right\|_1 d \pm O(\varepsilon n \cdot \gamma_2).$$

Proof. The runtime per iteration is clearly dominated by the time it takes to find the top singular vector of the centered gram matrix, which can be done in time $O(dn^2)$.

We will show that for all $t = 1, \ldots, K$, we have that $w^{(t)} \in \mathfrak{S}_n$. First, we demonstrate how this proves the overall lemma. First, note that after each iteration, some new w_i (with the maximum τ_i) becomes 0, so after $6\varepsilon n$ iterations, we have removed at least $6\varepsilon n$ mass from w. By definition of $\mathfrak{S}_n$, this means we have removed at least εn mass from the bad coordinates w_B, at which point no further updates can maintain the invariant that $w^{(i)} \in \mathfrak{S}_n$.

Next, we observe that if $w^{(K)} \in \mathfrak{S}_n$, then since we terminated, we must have that

$$\left\|\mathrm{Gram}(w^{(K)}) - D(w^{(K)})\right\| \leq 5\gamma_2 \,.$$

But then, for all T with $|T| \leq \varepsilon n$, let $\mathbf{1}_T \in \mathbb{R}^n$ be the indicator vector T. Then, we have that

$$\left\|\mathrm{Sum}(w^{(K)}, T)\right\|^2 = \mathbf{1}_T^\top \mathrm{Gram}(w^{(K)})\mathbf{1}_T = \left\|w_T^{(K)}\right\|_1 d \pm O(\varepsilon n \cdot \gamma_2) \,,$$

as claimed.

Thus, it suffices to prove the invariant that $w^{(t)} \in \mathfrak{S}_n$ for all $t = 1, \ldots, K$. We proceed by induction. Clearly $w^{(1)} \in \mathfrak{S}_n$. Now, suppose $w^{(t)} \in \mathfrak{S}_n$ for some $t < K$. Since we have not yet terminated, this implies that

$$\lambda = \left\|\mathrm{Gram}(w^{(t)}) - D(w^{(t)})\right\| \geq 5\gamma_2 \,.$$

But by (27) and the Cauchy interlacing theorem, we have that

$$\left\|\mathrm{Gram}(w^{(t)}, G \setminus R) - D(w^{(t)}_{G\setminus R})\right\| \leq \frac{\gamma_2}{10} \leq \frac{\lambda}{50} \,.$$

We claim that this implies that $5\sum_{i\in B} v_i^2 > \sum_{i \in G\setminus R} v_i^2$. Indeed, suppose not, and let v_G denote the restriction of v onto the coordinates in $G \setminus R$, and let v_B denote the restriction of v onto the coordinates in B. This means that

$$\begin{aligned}\left|v^\top \left(\mathrm{Gram}(w^{(t)}) - D(w)\right) v\right| &= \left|v_G^\top \left(\mathrm{Gram}(w^{(t)}, G\setminus R) - D(w^{(t)}_{G\setminus R})\right) v_G \right. \\ &\quad \left. + 2v_G^\top \left(\mathrm{Gram}(w^{(t)}) - D(w^{(t)})\right) v_B + v_B^\top \left(\mathrm{Gram}(w^{(t)}, B) - D(w^{(t)}_B)\right) v_B\right| \\ &\leq \|v_G\|^2 \cdot \frac{\lambda}{50} + 2\|v_G\|\|v_B\|\lambda + \|v_B\|^2 \lambda \leq 0.96\lambda \,,\end{aligned}$$

where the last inequality holds because $\|v_B\|^2 + \|v_G\|^2 = \|v\|^2 = 1$ and $\|v_B\|^2 \le \frac{1}{6}$. But this is a contradiction since v is the top singular vector of the centered Gram matrix. Therefore, by Fact 8.13 and the definition of τ_i, we obtain that $w^{(t+1)} \in \mathfrak{S}_n$, as claimed. (Note that if $w_i^{(t)} = 0$, the ith row and column of both $\mathrm{Gram}(w^{(t)}, S)$ and $D(w^{(t)})$ are 0 so $v_i = 0$, which means $w_i^{(t)} \cdot \tau_i = v_i^2$ even if $w_i^{(t)} = 0$.) This completes the proof. □

8.5 The filtering algorithm for $n > d$

The filtering algorithm proceeds similarly to above. Let γ_2 be as in (26). Initialize weights $w^{(1)} = \mathbf{1}$. Then, for $t = 1$ until termination, we proceed as follows. Let λ be the top singular value of $M(w^{(t)}) - nI$, and let v be its associated singular value (if there are multiple, again choose one arbitrarily). If $\lambda < 5\gamma_2$, then terminate. Otherwise, for all $i \in G$, let $\tau_i = \langle v, X_i \rangle^2 \mathbb{I}[w_i > 0]$. Proceed to sort the samples in decreasing order of τ_t. As before, by relabeling indices, assume that $\tau_1 \ge \tau_2 \ge \cdots \ge \tau_n$. Let I be the smallest index so that $\sum_{i \le I} w_i^{(t)} \ge 2\varepsilon n$, and define $w^{(t+1)}$ by

$$w_i^{(t+1)} = \begin{cases} \left(1 - \frac{\tau_i}{\tau_1}\right) w_i^{(t)} & \text{if } i \le I; \\ w_i^{(t)} & \text{if } i > I. \end{cases} \tag{28}$$

The formal pseudocode for the algorithm appears in Algorithm 5.

Algorithm 5 Spectral filtering for $n > d$. Input: $X_1, \ldots, X_n \in \mathbb{R}^d$, $\gamma_2 > 0$.

Let $w^{(1)} = \mathbf{1}$, and let $t = 1$
while $\|M(w, S) - nI\| \ge 5\gamma_2$ **do**
    Let v be the top singular vector of $M(w, S) - nI$
    For all i, let $\tau_i = \langle v, X_i \rangle^2 \mathbb{I}[w_i^{(t)} > 0]$
    Let $w^{(t+1)}$ be given by (28)
    Let $t \leftarrow t + 1$
Return $w^{(t)}$

As before, for the rest of this section, let us assume that G is $(\varepsilon, \varepsilon n\sqrt{d}\log(n/\delta), (\varepsilon n)^2 \log 1/\varepsilon + \varepsilon n\sqrt{\varepsilon n d \log 1/\varepsilon})$-regular, and additionally assume that

$$\|M(G) - nI\| \le \frac{\gamma_2}{10}, \tag{29}$$

and that Fact 8.12 holds. As before, direct applications of Lemma 8.3 and Facts 8.11 and 8.12 immediately imply that these conditions hold together with probability at least $1 - \delta$. Then, our main claim for this algorithm is the following:

Lemma 8.15. *Under the above regularity conditions, Algorithm 5 terminates in K iterations for some $K \le 6\varepsilon n$, runs in time $O(nd^2)$ per iteration, and moreover, at termination, we have that*

- *$w^{(K)} \in \mathfrak{S}_n$, and*
- *for all $T \subset S$ with $|T| \le \varepsilon n$, we have that*

$$\left\|\mathrm{Sum}(w^{(K)}, T)\right\|^2 \le 10\gamma_2 \cdot \varepsilon n.$$

Proof. As before, the per-iteration runtime is dominated by the runtime of PCA, which is $O(nd^2)$.

We will again inductively show that for all iterations t, we have that $w^{(t)} \in \mathfrak{S}_n$. We first show how to prove the lemma, assuming this claim. In this case, the number of iterations K can be bounded identically as in Lemma 8.14. Moreover, by construction, at termination we have that $\left\|M(w^{(K)}) - nI\right\| \leq 5\gamma_2$. Now suppose that there was some subset T with $|T| \leq \varepsilon n$ that had

$$\left\|\mathrm{Sum}(w^{(K)}, T)\right\|^2 > 10\gamma_2 \cdot \varepsilon n.$$

Then, there is a unit vector $v \in \mathbb{R}^d$ so that

$$\sum_{i \in T} (w_i^{(K)})^{1/2} \langle v, X_i \rangle > \sqrt{10\gamma_2 \cdot \varepsilon n}\ . \tag{30}$$

Thus, we have that

$$\sum_{i \in T \cup B} w_i^{(K)} \langle v, X_i \rangle^2 \geq \sum_{i \in T} w_i^{(K)} \langle v, X_i \rangle^2 \geq 10\gamma_2\ . \tag{31}$$

Above, the first inequality holds because every $w_i^{(K)}$ is nonnegative, and the second inequality follows from (30) and the Cauchy-Schwarz inequality. However, as $|R \cup T| \leq 2\varepsilon$ and $\|\mathbf{1} - w^{(K)}\|_1 \leq 6\varepsilon n$, (29) and Fact 8.12 together imply

$$\left\|M\left(w^{(K)}, G \setminus (R \cup T)\right) - nI\right\| \leq \frac{\gamma_2}{5}\ . \tag{32}$$

Since $G \setminus (R \cup T) = S \setminus (B \cup T)$, the inequalities (31) and (32) together imply that

$$\left\|M\left(w^{(K)}, S\right) - nI\right\| > 5\gamma_2\ ,$$

which is a contradiction.

Thus, as before, it suffices to prove that $w^{(t)} \in \mathfrak{S}_n$ for all iterations until termination. We will do so inductively. As before, the base case $t = 1$ is trivial. Now suppose that $w^{(t)} \in \mathfrak{S}_n$ for some $t < K$. Since we have not yet terminated, this means that $\left\|M(w^{(t)}, S) - nI\right\| > 5\gamma_2$. Then, (29) and Fact 8.12 together immediately imply that

$$\sum_{i \in B} w_i^{(t)} \langle v, X_i \rangle^2 \geq 3\gamma_2\ , \tag{33}$$

for v the top eigenvector of $M(w^{(t)}, S) - n \cdot I$. On the other hand, since $\sum_{i \leq I} w_i^{(t)} \leq 2\varepsilon n + 1$ and since $w^{(t)} \in \mathfrak{S}_n$ means we have removed at most $6\varepsilon n$ mass from all samples, this means $I \leq 8\varepsilon n + 1 \leq 10\varepsilon n$. So, Fact 8.12 implies that

$$\sum_{i \leq I, i \in G} w_i^{(t)} \langle v, X_i \rangle^2 < \gamma_2\ . \tag{34}$$

By definition of I, every $\langle v, X_i \rangle^2$ for $i \leq I$ is larger than every $\langle v, X_i \rangle^2$ for $i \in B \setminus [I]$. Therefore, since $\sum_{i \in B \setminus [I]} w_i^{(t)} \leq |B \setminus [I]| \leq \varepsilon n$ but $\sum_{i \leq I} w_i^{(t)} \geq 2\varepsilon n$, we have

$$2 \cdot \left(\sum_{i \in B} w_i^{(t)} \langle v, X_i \rangle^2 - \sum_{i \leq I} w_i^{(t)} \langle v, X_i \rangle^2\right) \leq 2 \cdot \sum_{i \in B \setminus [I]} w_i^{(t)} \langle v, X_i \rangle^2 \leq \sum_{i \in I} w_i^{(t)} \langle v, X_i \rangle^2. \tag{35}$$

Along with (33), (35) implies that

$$\sum_{i \leq I} w_i^{(t)} \langle v, X_i \rangle^2 \geq 2\gamma_2. \tag{36}$$

Hence, by combining (36) with (34), we have

$$\sum_{i \le I, i \in B} w_i^{(t)} \langle v, X_i \rangle^2 \ge \gamma_2 \ge \sum_{i \le I, i \in G} w_i^{(t)} \langle v, X_i \rangle^2 ,$$

and so the result for $w^{(t+1)}$ immediately follows from Fact 8.13. □

8.6 Bounding row sums

We now have a way to ensure that small subsets of points have means with bounded norm. We also need to enforce that row sums are bounded. To do so, we will simply remove the set of $O(\varepsilon n)$ points whose row sums have largest deviation from what we expect. More formally, given a set of weights $w \in \mathfrak{S}_n$, we will let

$$\tau_i = \left| \langle \sqrt{w_i} X_i, \sum_{j \in S} \sqrt{w_j} X_j \rangle - w_i d \right| \cdot \mathbb{I}[w_i > 0] . \tag{37}$$

We then sort the indices in decreasing order by τ_i. Again for simplicity of notation, assume that after some suitable reindexing we have that $\tau_1 \ge \tau_2 \ge \ldots \ge \tau_n$. Then, we replace w_i with 0 for all $i \le \varepsilon n$. We give the formal pseudocode for this algorithm in Algorithm 6.

Algorithm 6 Bounding row sums. Input: $X_1, \ldots, X_n \in \mathbb{R}^d$

1: For all i, let τ_i be as in (37).
2: Sort the indices in decreasing order by τ_i. ▷ By relabeling indices, for simplicity of notation assume that the i's are initally sorted
3: Set $w_i = 0$ for all $i \le \varepsilon n$.
4: **return** w

Lemma 8.16. *Assume G is $(12\varepsilon, \beta_1, \beta_2)$-regular, that Fact 8.12 holds for G, and $S = (G \backslash R) \cup B$, where $|R| = |B| = \varepsilon n$. Let $w \in \mathfrak{S}_n$, and assume that for all $T \subset S$ with $|T| \le 2\varepsilon n$, we have that $\|\mathrm{Sum}(w, T)\|^2 = \|w_T\|_1 \, d \pm O(\beta_2)$. Then for all $T \subset S \backslash B$ with $|T| \le \varepsilon n$, we have that*

$$\sum_{i \in T, j \in S} \sqrt{w_i w_j} \langle X_i, X_j \rangle = d \cdot \|w_T\|_1 \pm O(\sqrt{n}\beta_1 + \beta_2) .$$

Proof. Fix $T \in S \backslash B$ with $|T| \le \varepsilon n$. Since $S = (G \backslash R) \cup B$, we can write

$$\sum_{i \in T, j \in S} \sqrt{w_i w_j} \langle X_i, X_j \rangle = \underbrace{\sum_{i \in T, j \in G} \sqrt{w_i w_j} \langle X_i, X_j \rangle}_{A_1} - \underbrace{\sum_{i \in T, j \in R} \sqrt{w_i w_j} \langle X_i, X_j \rangle}_{A_2} + \underbrace{\sum_{i \in T, j \in B} \sqrt{w_i w_j} \langle X_i, X_j \rangle}_{A_3} .$$

Above, for $j \in R$, w_j is defined to equal $w_{j'}$ for the $j' \in B$ that replaces j.

Let $b_i := \sqrt{w_i}$ and $a_i := 1 - \sqrt{w_i}$. Since $|T| \le \varepsilon n$, we know that $\sum_{i \in T} b_i \le \varepsilon n$. Moreover, $\sum_{j \in G} w_j \ge (1 - 6\varepsilon)n$, so $\sum_{j \in G} a_i \le 6\varepsilon n$. Because G is $(12\varepsilon, \beta_1, \beta_2)$-regular and $T \subset G$, Proposition 8.6

implies that

$$\begin{aligned} A_1 &= \left\langle \sum_{i\in T} b_i X_i, \sum_{j\in G} X_j \right\rangle - \left\langle \sum_{i\in T} b_i X_i, \sum_{j\in G} a_j X_j \right\rangle \\ &= d \cdot \sum_{i\in T} b_i - d \sum_{i\in T} a_i b_i \pm O(\sqrt{n}\beta_1 + \beta_2) \\ &= d \cdot \sum_{i\in T} w_i \pm O(\sqrt{n}\beta_1 + \beta_2). \end{aligned} \tag{38}$$

Next, we bound A_2. Since $T, R \subset G$ are disjoint and $|T|, |R| \le \varepsilon n$, Proposition 8.6 implies that

$$A_2 = \left\langle \sum_{i\in T} b_i X_i, \sum_{j\in R} b_j X_j \right\rangle = \pm O(\sqrt{n}\beta_1 + \beta_2), \tag{39}$$

since the b_i terms in T and the b_j terms in R are from disjoint sets.

Finally, we bound A_3. We will only use the fact that T, B are disjoint sets in S of size at most εn and our assumption on w in the lemma statement. We can write

$$\begin{aligned} A_3 &= \frac{1}{2}\left(\|\mathrm{Sum}(w, T\cup B)\|^2 - \|\mathrm{Sum}(w,T)\|^2 - \|\mathrm{Sum}(w,B)\|^2 \right) \\ &= \frac{1}{2}\left(\|w_{T\cup B}\|_1 \cdot d - \|w_T\|_1 \cdot d - \|w_B\|_1 \cdot d \right) \pm O(\beta_2) \\ &= \pm O(\beta_2). \end{aligned} \tag{40}$$

Adding (38), (39), and (40) completes the proof. □

Lemma 8.17. *Assume Lemma 8.16 holds, and let w' be the output of Algorithm 6. Then, for all $T \subset S$ with $|T| \le \varepsilon n$, we have that*

$$\sum_{i\in T, j\in[n]} \sqrt{w'_i w'_j}\langle X_i, X_j\rangle = d \cdot \|w_T\|_1 \pm O(\sqrt{n}\beta_1 + \beta_2)\ .$$

Proof. Recall the definition of τ_i from (37). First, we note that for any $T \subset S\backslash B$ with $|T| \le \varepsilon n$, $\sum_{i\in T} \tau_i \le O(\sqrt{n}\beta_1 + \beta_2)$. To see why, we can split T into T^+ and T^-, where $i \in T^+$ if $\langle \sqrt{w_i}X_i, \sum_{j\in S}\sqrt{w_j}X_j\rangle \ge w_i d$ and $i \in T^-$ otherwise. Then, since $|T^+|, |T^-| \le \varepsilon n$, Lemma 8.16 implies that both $\sum_{i\in T^+}\tau_i$ and $\sum_{i\in T^-}\tau_i$ are at most $O(\sqrt{n}\beta_1 + \beta_2)$.

Since $\sum_{i\in T}\tau_i \le O(\sqrt{n}\beta_1 + \beta_2)$ for any subset T of $S\backslash B$ of size at most εn, and since we sorted the τ_i's in decreasing order, this implies $\sum_{i\in T}\tau_i \le O(\sqrt{n}\beta_1 + \beta_2)$ for any subset T of $S\backslash[\varepsilon n]$ of size at most εn. If w_i represents the values of w before setting the top εn indices to 0, and w'_i represents the values of w afterwards (i.e., $w'_i = 0$ for $i \le \varepsilon n$ and $w'_i = w_i$ for $i > \varepsilon n$), then

$$\begin{aligned} \left| \sum_{i\in T, j\in S} \sqrt{w'_i w_j}\langle X_i, X_j\rangle - d\cdot\|w'_T\|_1 \right| &\le \sum_{i\in T} \left| \left\langle \sqrt{w'_i}X_i, \sum_{j\in S}\sqrt{w_j}X_j \right\rangle - w'_i d \right| \\ &= \sum_{i\in T\backslash[\varepsilon n]} \tau_i \le O(\sqrt{n}\beta_1 + \beta_2). \end{aligned} \tag{41}$$

Next, we have

$$\sum_{i\in T, j\in[\varepsilon n]} \sqrt{w_i' w_j}\langle X_i, X_j\rangle = \sum_{i\in T\setminus[\varepsilon n], j\in[\varepsilon n]} \sqrt{w_i w_j}\langle X_i, X_j\rangle = \pm O(\beta_2), \tag{42}$$

by the same argument as in (40), since $T\setminus[\varepsilon n]$ and $[\varepsilon n]$ are disjoint sets in S and have size at most εn.

By subtracting (42) from (41), we obtain the desired bound

$$\left|\sum_{i\in T, j\in S} \sqrt{w_i' w_j'}\langle X_i, X_j\rangle - d\|w_T'\|_1\right| = \left|\sum_{i\in T, j\in S\setminus[\varepsilon n]} \sqrt{w_i' w_j}\langle X_i, X_j\rangle - d\|w_T'\|_1\right| \le O(\sqrt{n}\beta_1 + \beta_2). \quad \square$$

8.7 Putting it all together

We now have all the necessary pieces for the full algorithm. It will proceed as follows. First, remove any points whose norms differ from d by too much. Then, run the appropriate spectral filtering algorithm depending on whether or not $n \le d$. Next, use Algorithm 6 to bound row sums. Finally, check the sum of all entries in the (centered) Gram matrix, and if it is too large, then output YES, otherwise output NO. The full pseudocode is given in Algorithm 7.

Algorithm 7 Robust mean testing $X_1, \dots, X_n \in \mathbb{R}^d$

1: Remove any i satisfying $\left|\|X_i\|^2 - d\right| \ge O(\sqrt{d}\log n/\delta)$
2: Let γ_2 be as in (26).
3: **if** $n \le d$ **then**
4: Let w be the output of Algorithm 4 with parameter γ_2.
5: **else**
6: Let w be the output of Algorithm 5 with parameter γ_2.
7: Let w' be the output of Algorithm 6 with input w
8: **if** $\left|\|\mathrm{Sum}(w', S)\|^2 - d\,\|w\|_1\right| \ge 0.7\alpha^2 n^2$ **then**
9: **return** YES
10: **else**
11: **return** NO

Proof of Theorem 8.1. The runtime of the algorithm is clearly dominated by the runtime of the spectral filters, which both run in time $O(\varepsilon n^2 d \min(n, d))$, and the runtime of computing $\mathrm{Sum}(w, S)$, which is $O(nd)$. We now prove correctness.

By standard arguments, the first step (Line 1) of Algorithm 7 removes no uncorrupted points with probability $1 - \delta/3$.

Assume that G is $(12\varepsilon, \beta_1, \beta_2)$-regular for $\beta_1 = \varepsilon n\sqrt{d}\log(n/\delta)$ and $\beta_2 = \varepsilon n\sqrt{nd}\log(n/\delta) + (\varepsilon n)^2 \log 1/\varepsilon$. By Lemma 8.3, this occurs with probability $1 - \delta/3$. As argued above, for our choice of n, the conditions for Lemma 8.14 and Lemma 8.15 are also satisfied with probability $1 - \delta/3$ (and will even hold for all $|T| \le 2\varepsilon n$). Thus, we obtain that $w \in \mathfrak{S}_n$. Finally, regularity and standard sub-exponential moment bounds imply that the conditions for Lemma 8.17 are satisfied, as long as $\beta_1 = \varepsilon n\sqrt{d}\log(n/\delta)$ and $\beta_2 = O(\varepsilon n \cdot \gamma_2) = O(\varepsilon n\sqrt{nd} + \alpha^2\varepsilon n^2 + \varepsilon n\sqrt{(n+d)\log(1/\delta)} + \varepsilon n \log(1/\delta) + (\varepsilon n)^2 \log(1/\varepsilon)$. Therefore, the resulting set of weights w' is $(O(\varepsilon), \beta_1 + \frac{\beta_2}{\sqrt{n}}, \beta_2)$-regular. Plugging this into Lemma 8.8 yields the claim, since by our choice of n, one can verify that $\sqrt{n}\beta_1, \beta_2 \le \frac{1}{C} \cdot \alpha^2 n^2$. $\square$

9 Computational Lower Bound

We refer to [KWB22] for definitions concerning the low-degree method and the low-degree likelihood ratio, reviewing here only a little background. Originating in [BHK+19; Hop18; HS17], the *low-degree method* is a heuristic for understanding computational complexity of average-case problems, in this case a hypothesis testing problem. The heuristic unconditionally rules out a class of algorithms based on low-degree evaluating low-degree polynomials in the input; in this case, low-degree polynomials in the nd-length vector $(X_1, \dots, X_n)$. This low-degree model captures a surprisingly broad range of algorithms, including spectral methods, making it a powerful heuristic for detecting computational hardness.

Theorem 9.1. *For $n, d \in \mathbb{N}$ and $\varepsilon, \alpha > 0$, consider the following probability distributions on D_0, D_1 on $(\mathbb{R}^n)^d$.*

- *$D_0 = \mathcal{N}(0,1)^{\otimes nd}$*
- *D_1: first, sample a random unit vector $v \in \mathbb{R}^d$. Then, draw n i.i.d. vectors $X_1, \dots, X_n$ from the distribution $(1-\varepsilon)\mathcal{N}(\alpha v, I) + \varepsilon(-\alpha(1-\varepsilon)\varepsilon^{-1}v, I)$.*

For $D \in \mathbb{N}$, $D > 1$, let $L^{\le D}$ be the degree-D truncated likelihood ratio for D_1 with respect to D_0. Then

$$\|L^{\le D} - 1\| \le D^{O(D)} \cdot \frac{\sqrt{n}\alpha^2}{\sqrt{d}\varepsilon} \cdot \exp\left(\frac{\sqrt{n}\alpha^2}{\sqrt{d}\varepsilon} + \frac{\alpha^2}{\sqrt{d}\varepsilon^2}\right).$$

Consequently, if $n \le o(D^{-O(D)} \cdot \frac{d\varepsilon^2}{\alpha^4})$ and $n \ge \sqrt{d}/\alpha^2$ (since otherwise testing is information-theoretically impossible), we have $\|L^{\le D} - 1\| \le o(1)$.

Proof. We define the following auxiliary distribution P over $n \times d$ matrices – to draw a sample from P, first draw a random unit vector $v \in \mathbb{R}^d$, then sample each column of P independently to be equal to $\alpha \cdot v$ with probability $(1-\varepsilon)$ and otherwise equal to $-\alpha(1-\varepsilon)\varepsilon^{-1}$. Note that an equivalent way to sample from D_1 is to first draw $X \sim P$ and output $X + G$, where $G \in \mathbb{R}^{n \times d}$ has independent entries distributed as $\mathcal{N}(0,1)$. Thus, D_1 fits into the *Gaussian additive model.*

Using Theorem 2.6 of [KWB22], concerning low-degree likelihood ratio for Gaussian additive models, we have

$$\|L^{\le D} - 1\|^2 = \sum_{t=1}^{D} \frac{1}{t!} \cdot \mathbb{E}_{X, X' \sim P} \langle X, X' \rangle^t,$$

where X, X' are independent draws from P and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in nd dimensions.

Let v, w be the independent random unit vectors associated to separate draws $X, X' \sim P$, let $S_v \subseteq [n]$ be the columns of X equal to αv, and similarly for S_w. Then

$$\langle X, X' \rangle^t = \langle v, w \rangle^t \cdot (\alpha^2 |S_v \cap S_w| - \alpha^2(1-\varepsilon)\varepsilon^{-1}(|S_v \cap \overline{S_w}| + |S_w \cap \overline{S_v}) + \alpha^2(1-\varepsilon)^2\varepsilon^{-2}|\overline{S_v} \cap \overline{S_w}|)^t.$$

Furthermore, v, w are independent from S_v, S_w, so

$$\mathbb{E}\langle X, X' \rangle^t = \mathbb{E}\langle v, w \rangle^t \cdot \mathbb{E}(\alpha^2 |S_v \cap S_w| - \alpha^2(1-\varepsilon)\varepsilon^{-1}(|S_v \cap \overline{S_w}| + |S_w \cap \overline{S_v}|) + \alpha^2(1-\varepsilon)^2\varepsilon^{-2}|\overline{S_v} \cap \overline{S_w}|)^t.$$

The whole quantity is equal to zero for odd t, and for even t it's at most

$$\frac{O(t)^{t/2}}{d^{t/2}} \cdot \alpha^{2t} \cdot \mathbb{E}(|S_v \cap S_w| - (1-\varepsilon)\varepsilon^{-1}(|S_v \cap \overline{S_w}| + |S_w \cap \overline{S_v}|) + (1-\varepsilon)^2\varepsilon^{-2}|\overline{S_v} \cap \overline{S_w}|)^t,$$

using Fact 9.2. We have

$$\begin{aligned}
\mathbb{E}|S_v \cap S_w| &= n(1-\varepsilon)^2 \\
\mathbb{E}|S_v \cap \overline{S_w}| &= n\varepsilon(1-\varepsilon) \\
\mathbb{E}|\overline{S_v} \cap S_w| &= n\varepsilon(1-\varepsilon) \\
\mathbb{E}|\overline{S_v} \cap \overline{S_w}| &= n\varepsilon^2
\end{aligned}$$

and hence in particular

$$\mathbb{E}|S_v \cap S_w| - (1-\varepsilon)\varepsilon^{-1}(\mathbb{E}|S_v \cap \overline{S_w}| + \mathbb{E}|S_w \cap \overline{S_v}|) + (1-\varepsilon)^2\varepsilon^{-2}\mathbb{E}|\overline{S_v} \cap \overline{S_w}| = 0\,.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\langle X, X'\rangle^t \leq \frac{O(t)^{t/2}\alpha^{2t}}{d^{t/2}} \cdot \\
\sqrt{\mathbb{E}\left(|S_v \cap S_w| - \mathbb{E}|S_v \cap S_w|\right)^{2t}} \\
+ \sqrt{(1-\varepsilon)^{2t}\varepsilon^{-2t}\mathbb{E}\left(|S_v \cap \overline{S_w}| - \mathbb{E}|S_v \cap \overline{S_w}|\right)^{2t}} \\
+ \sqrt{(1-\varepsilon)^{4t}\varepsilon^{-4t}\mathbb{E}\left(|\overline{S_v} \cap \overline{S_w}| - \mathbb{E}|\overline{S_w} \cap \overline{S_v}|\right)^{2t}}
\end{aligned}$$

Observe that:

- $|S_v \cap S_w|$ follows a Binomial distribution $\mathrm{Bin}(n, (1-\varepsilon)^2)$, so

  $$\mathbb{E}(|S_v \cap S_w| - \mathbb{E}|S_v \cap S_w|)^{2t} \leq O(t)^{2t}n(1-\varepsilon)^2 + O(t)^t(n(1-\varepsilon)^2(2\varepsilon - \varepsilon^2))^t$$

  using Fact 9.3.
- $|S_v \cap \overline{S_w}|$ follows a Binomial distribution $\mathrm{Bin}(n, \varepsilon(1-\varepsilon))$, so

  $$\mathbb{E}(|S_v \cap \overline{S_w}| - \mathbb{E}|S_v \cap \overline{S_w}|)^{2t} \leq O(t)^{2t}n\varepsilon(1-\varepsilon) + O(t)^t(n\varepsilon(1-\varepsilon)(1-\varepsilon(1-\varepsilon)))^t$$

  using Fact 9.3.
- $|\overline{S_v} \cap \overline{S_w}|$ follows a Binomial distribution $\mathrm{Bin}(n, \varepsilon^2)$, so

  $$\mathbb{E}(|\overline{S_v} \cap \overline{S_w}| - \mathbb{E}|\overline{S_v} \cap \overline{S_w}|)^{2t} \leq O(t)^{2t}n\varepsilon^2 + O(t)^t(n\varepsilon^2(1-\varepsilon^2))^t$$

  using Fact 9.3.

Substituting these moment bounds and simplifying using $t \geq 1$, we get

$$\mathbb{E}\langle X, X'\rangle^t \leq \frac{O(t)^{O(t)}\alpha^{2t}}{d^{t/2}}\left(\sqrt{n}\cdot\eta^{-2t+1} + n^{t/2}\cdot\varepsilon^{-t}\right)\,.$$

Summing across $t \in [1, D]$ gives the result. □

Fact 9.2. *Let u, v be independent random unit vectors in d dimensions and let $t \in \mathbb{N}$. Then $\mathbb{E}\langle u, v\rangle^t \leq O(t)^{t/2}\cdot O(d)^{-t/2}$.*

Proof. The random variable $\langle u, v\rangle$ has the same distribution as $\langle x, g\rangle/\|g\|$ where x is any fixed unit vector and $g \sim \mathcal{N}(0, I)$. We have

$$\mathbb{E}\frac{\langle x, g\rangle^t}{\|g\|^t} \leq \left(\mathbb{E}\langle x, g\rangle^{2t}\right)^{1/2} \cdot \left(\mathbb{E}\|g\|^{-2t}\right)^{1/2} .$$

Since $\langle x, g\rangle$ is distributed as $\mathcal{N}(0, 1)$, the first term is at most $\sqrt{(2t)^t} = O(t)^{t/2}$. For the second term, $\|g\| \geq \Omega(\sqrt{d})$ with probability at least 0.9, so $\mathbb{E}\|g\|^{-2t} \leq O(1)^t \cdot d^{-t}$. □

The following is a special case of Rosenthal's inequality; see e.g. [Pin94].

Fact 9.3 (Moments of binomial distribution)**.** *There is a constant $C > 0$ such that for all $n, t \in \mathbb{N}$ and $p \in [0, 1]$, if $Y \sim \mathrm{Bin}(n, p)$ be a binomial random variable, $\mathbb{E}(Y - \mathbb{E}Y)^t \leq (Ct)^t \cdot np + (Ct)^{t/2} \cdot (np(1-p))^{t/2}$.*

Acknowledgments

The authors would like to thank Guy Blanc and Gautam Kamath for some helpful suggestions.

References

[ACT20] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. "Distributed Signal Detection under Communication Constraints". In: ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 41–63 (cit. on p. 5).

[AD20] Hilal Asi and John C Duchi. "Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms". In: *Advances in neural information processing systems*. Vol. 33. 2020, pp. 14106–14117 (cit. on p. 53).

[AUZ23] Hilal Asi, Jonathan R. Ullman, and Lydia Zakynthinou. "From Robustness to Privacy and Back". In: *CoRR* abs/2302.01855 (2023) (cit. on pp. 4, 5, 51).

[Bar02] Yannick Baraud. "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* (2002), pp. 577–606 (cit. on p. 5).

[BB20] Matthew Brennan and Guy Bresler. "Reducibility and statistical-computational gaps from secret leakage". In: *Conference on Learning Theory*. PMLR. 2020, pp. 648–847 (cit. on pp. 2, 5).

[BBH+20] Matthew Brennan, Guy Bresler, Samuel B Hopkins, Jerry Li, and Tselil Schramm. "Statistical query algorithms and low-degree tests are almost equivalent". In: *arXiv preprint arXiv:2009.06107* (2020) (cit. on p. 11).

[BEK02] Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. "PAC learning with nasty noise". In: *Theoretical Computer Science* 288.2 (2002), pp. 255–275 (cit. on p. 5).

[BH20] Maria-Florina Balcan and Nika Haghtalab. *Noise in Classification.* 2020 (cit. on p. 5).

[BHK+19] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. "A nearly tight sum-of-squares lower bound for the planted clique problem". In: *SIAM Journal on Computing* 48.2 (2019), pp. 687–735 (cit. on p. 68).

[BLMT22] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. "On the power of adaptivity in statistical adversaries". In: *Conference on Learning Theory*. PMLR. 2022, pp. 5030–5061 (cit. on pp. 1, 5).

[CCK+21] Clément L. Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. "Random Restrictions of High Dimensional Distributions and Uniformity Testing with Subcube Conditioning". In: *SODA*. SIAM, 2021, pp. 321–336 (cit. on p. 5).

[CKM+20] Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan R. Ullman, and Lydia Zakynthinou. "Private Identity Testing for High-Dimensional Distributions". In: *NeurIPS*. 2020 (cit. on p. 5).

[DGJ+21] Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Abhradeep Guha Thakurta. "A separation result between data-oblivious and data-aware poisoning attacks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10862–10875 (cit. on p. 1).

[DGT19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. "Distribution-independent pac learning of halfspaces with massart noise". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 5).

[DHL19] Yihe Dong, Samuel Hopkins, and Jerry Li. "Quantum entropy scoring for fast robust mean estimation and improved outlier detection". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 61).

[DK19] Ilias Diakonikolas and Daniel M. Kane. "Recent Advances in Algorithmic High-Dimensional Robust Statistics". In: *CoRR* abs/1911.05911 (2019) (cit. on p. 2).

[DK21] Ilias Diakonikolas and Daniel M. Kane. "The Sample Complexity of Robust Covariance Testing". In: *COLT*. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1511–1521 (cit. on p. 5).

[DK23] Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. To appear. Draft available at https://sites.google.com/view/ars-book/. Cambridge University Press, 2023 (cit. on pp. 1, 2, 5).

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robustly learning a gaussian: Getting optimal error, efficiently". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2018, pp. 2683–2702 (cit. on pp. 1, 5).

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robust estimators in high-dimensions without the computational intractability". In: *SIAM Journal on Computing* 48.2 (2019), pp. 742–864 (cit. on p. 5).

[DKMR22] Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. "Cryptographic hardness of learning halfspaces with massart noise". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3624–3636 (cit. on p. 5).

[DKP23] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. "Gaussian Mean Testing Made Simple". In: *SOSA*. SIAM, 2023, pp. 348–352 (cit. on p. 5).

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures". In: *FOCS*. IEEE Computer Society, 2017, pp. 73–84 (cit. on pp. 1, 5, 11, 40, 45, 54).

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. “Learning geometric concepts with nasty noise”. In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1061–1073 (cit. on p. 5).

[DS03] Kenneth R. Davidson and Slanislaw J. Szarek. “Local operator theory, random matrices and Banach spaces”. In: *Handbook of the geometry of Banach spaces* 2 (2003), pp. 317–366 (cit. on p. 60).

[Erm91] Michael Sergeevich Ermakov. “Minimax detection of a signal in a Gaussian white noise”. In: *Theory of Probability & Its Applications* 35.4 (1991), pp. 667–679 (cit. on p. 5).

[GC22] Anand Jerry George and Clément L. Canonne. “Robust Testing in High-Dimensional Sparse Models”. In: *NeurIPS*. 2022 (cit. on p. 5).

[GH22] Kristian Georgiev and Samuel B. Hopkins. “Privacy Induces Robustness: Information-Computation Gaps and Sparse Mean Estimation”. In: *NeurIPS*. 2022 (cit. on pp. 4, 5, 51, 52).

[GTX+22] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. “Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.2 (2022), pp. 1563–1580 (cit. on p. 1).

[GW17] Evan Greene and Jon A. Wellner. “Exponential bounds for the hypergeometric distribution”. In: *Bernoulli* 23.3 (2017), pp. 1911–1950 (cit. on p. 15).

[HKMN22] Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. “Robustness Implies Privacy in Statistical Estimation”. In: *CoRR* abs/2212.05015 (2022) (cit. on pp. 4, 5, 51, 53).

[Hop18] Samuel Hopkins. “Statistical inference and the sum of squares method”. PhD thesis. Cornell University, 2018 (cit. on p. 68).

[HS17] Samuel B Hopkins and David Steurer. “Efficient bayesian estimation from few samples: community detection and related problems”. In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 379–390 (cit. on p. 68).

[IIS03] Yuri Ingster, Jurij I Ingster, and IA Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2003 (cit. on p. 5).

[Kea98] Michael Kearns. “Efficient noise-tolerant learning from statistical queries”. In: *Journal of the ACM (JACM)* 45.6 (1998), pp. 983–1006 (cit. on p. 5).

[KNL+20] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. “Adversarial machine learning-industry perspectives”. In: *2020 IEEE security and privacy workshops (SPW)*. IEEE. 2020, pp. 69–75 (cit. on p. 1).

[KWB22] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. “Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio”. In: *Mathematical Analysis, its Applications and Computation: ISAAC 2019, Aveiro, Portugal, July 29–August 2*. Springer, 2022, pp. 1–50 (cit. on pp. 11, 68).

[Nar22] Shyam Narayanan. “Private High-Dimensional Hypothesis Testing”. In: *COLT*. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 3979–4027 (cit. on pp. 1, 4–6, 14, 15, 51–54).

[NT22] Rajai Nasser and Stefan Tiegel. “Optimal SQ lower bounds for learning halfspaces with massart noise”. In: *Conference on Learning Theory*. PMLR. 2022, pp. 1047–1074 (cit. on p. 5).

[Pin94] Iosif Pinelis. “Optimum bounds for the distributions of martingales in Banach spaces”. In: *The Annals of Probability* (1994), pp. 1679–1706 (cit. on p. 70).

[SD08] Muni S Srivastava and Meng Du. “A test for the mean vector with fewer observations than the dimension”. In: *Journal of Multivariate Analysis* 99.3 (2008), pp. 386–402 (cit. on pp. 5, 54).

[Ska13] Matthew Skala. “Hypergeometric tail inequalities: ending the insanity”. In: *CoRR* abs/1311.5939 (2013) (cit. on p. 15).

[SVZ22] Botond Szabó, Lasse Vuursteen, and Harry van Zanten. “Optimal high-dimensional and nonparametric distributed testing under communication constraints”. In: (2022). arXiv: 2202.00968 [math.ST] (cit. on p. 5).

A Mathematica code to verify the computation from Section 6.3

For completeness, we here provide some Mathematica code which can be used to verify the computations from the proof of Lemma 6.6:

```
s11[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] =
  2*(1 - \[Epsilon])*n*\[Beta]^2*d;
s12[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d + \[Beta]*d;
s13[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d + \[Beta]*d;
s14[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = 2*\[Beta]*d;
s21[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d + \[Beta]*d;
s22[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d - \[Alpha]^2;
s23[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = 2*\[Beta]*d;
s24[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = \[Beta]*d - \[Alpha]^2;
s31[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d + \[Beta]*d;
s32[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = 2*\[Beta]*d;
s33[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = (1 - \[Epsilon])*n*\[Beta]^2*d - \[Alpha]^2;
s34[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = \[Beta]*d - \[Alpha]^2;
s41[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = 2*\[Beta]*d;
s42[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = \[Beta]*d - \[Alpha]^2;
s43[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = \[Beta]*d - \[Alpha]^2;
s44[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_] = -2*\[Alpha]^2;

diag[\[Epsilon]_, \[Gamma]_, n_] =
  DiagonalMatrix[{\[Gamma]*n, (\[Epsilon] - \[Gamma])*
     n, (\[Epsilon] - \[Gamma])*n, (1 - 2*\[Epsilon] + \[Gamma])*n}];
I4 = DiagonalMatrix[{1, 1, 1, 1}];
sigma[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_,
   d_] = {{s11[\[Alpha], \[Beta], \[Epsilon], n, d],
    s12[\[Alpha], \[Beta], \[Epsilon], n, d],
    s13[\[Alpha], \[Beta], \[Epsilon], n, d],
    s14[\[Alpha], \[Beta], \[Epsilon], n,
     d]}, {s21[\[Alpha], \[Beta], \[Epsilon], n, d],
```

```
    s22[\[Alpha], \[Beta], \[Epsilon], n, d],
    s23[\[Alpha], \[Beta], \[Epsilon], n, d],
    s24[\[Alpha], \[Beta], \[Epsilon], n,
     d]}, {s31[\[Alpha], \[Beta], \[Epsilon], n, d],
    s32[\[Alpha], \[Beta], \[Epsilon], n, d],
    s33[\[Alpha], \[Beta], \[Epsilon], n, d],
    s34[\[Alpha], \[Beta], \[Epsilon], n,
     d]}, {s41[\[Alpha], \[Beta], \[Epsilon], n, d],
    s42[\[Alpha], \[Beta], \[Epsilon], n, d],
    s43[\[Alpha], \[Beta], \[Epsilon], n, d],
    s44[\[Alpha], \[Beta], \[Epsilon], n, d]}};
final[\[Alpha]_, \[Beta]_, \[Epsilon]_, n_, d_, \[Gamma]_] =
  I4 + Dot[diag[\[Epsilon], \[Gamma], n],
     sigma[\[Alpha], \[Beta], \[Epsilon], n,
      d]]/((1 - \[Epsilon])*\[Alpha]^2*n + d);

FullSimplify[( (d + \[Beta]^2 d (\[Epsilon]^2 \[Minus] \[Gamma]) \
n^2)^2 \[Minus] (\[Alpha]^2 n \[Minus]
       2 \[Beta] d \[Epsilon] n + \[Beta]^2 d \[Epsilon] n^2 \[Minus]
       2 \[Alpha]^2 \[Epsilon] n + \[Alpha]^2 \[Gamma] n +
       2 \[Beta] d \[Gamma] n \[Minus]
       2 \[Beta]^2 d \[Epsilon]^2 n^2 + \[Beta]^2 d \[Epsilon]\
  \[Gamma]  n^2)^2 ) ==
  Factor[Det[
     final[\[Alpha], \[Beta], \[Epsilon], n,
      d, \[Gamma]]]*(d + \[Alpha]^2*n - \[Alpha]^2 *\[Epsilon]*
        n)^2] ]
```