2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS) | 979-8-3503-1894-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/FOCS57990.2023.00133

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

Clément Canonne
School of Computer Science
University of Sydney
Sydney, Australia
clement.canonne@sydney.edu.au

Samuel B. Hopkins
Department of EECS
Massachusetts Institute of Technology
Cambridge, USA
samhop@mit.edu

Jerry Li
Machine Learning Foundations
Microsoft Research
Redmond, USA
jerrl@microsoft.com

Allen Liu
Department of EECS
Massachusetts Institute of Technology
Cambridge, USA
cliu568@mit.edu

Shyam Narayanan
Department of EECS
Massachusetts Institute of Technology
Cambridge, USA
shyamsn@mit.edu

***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.

CC was supported by an ARC DECRA (DE230101329) and an unrestricted gift from Google Research. SH was supported by NSF Award No. 2238080 and MLA@CSAIL. AL was supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship. SN was supported by an NSF Graduate Fellowship and a Google Fellowship.

***Index Terms*—Property testing, robust statistics, identity testing, corrupted data, data poisoning**

I. 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 $X_1, \ldots, X_n \in \mathbb{R}^d$, the goal is to decide between two hypotheses:

H$_0$: $X_1, \ldots, X_n$ were drawn from $\mathcal{N}(0, I)$, an origin-centered identity-covariance Gaussian.

H$_1$: $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ 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) [18], [20], and yet basic

979-8-3503-1894-4/23/$31.00 ©2023 IEEE

DOI 10.1109/FOCS57990.2023.00133

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

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 [5]. Second, the question is pertinent to *data poisoning attacks* in machine learning [11], [25], 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 [29]. 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 [18]) 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 [31] 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 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 I.1)! 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 I.4). (This complements a reduction from planted clique by Brennan and Bresler [6] 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.

A. 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 [16], [18] 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 (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 2 (Oblivious contamination model)**.** The adversary chooses εn indices $i_1, \dots, i_{\varepsilon n}$ and values $X''_{i_1}, \dots, X''_{i_{\varepsilon n}}$.

[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 [14]. See Section I-C 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.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

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.

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 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 $X_1, \dots, X_n$ 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.

B. 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 I.3), which is information-theoretically necessary, up to the factor $(\log n)^{O(1)}$.

Theorem I.1 (Obliviously-robust mean testing (Informal; see full paper for formal statement))**.** *In the* oblivious *contamination model, there is a mean tester which is robust to ε-contamination, which uses*

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

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.

To see that Theorem I.1 implies a strict separation between the oblivious and adaptive models, we recall:

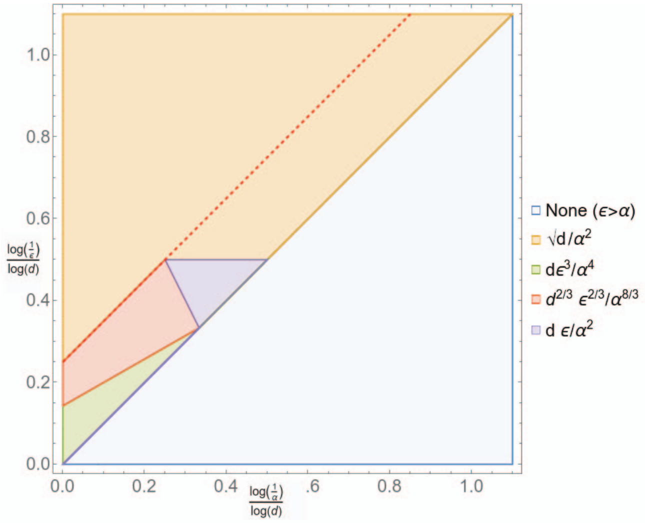

Fig. 1. The various phases of the sample complexity of robust mean testing in the oblivious contamination model, as stated in Theorem I.1: each area of this plot corresponds to which term of the sample complexity dominates, as a function of d, ε, α. The separation between adaptive and oblivious contamination 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.

Theorem I.2 ([31])**.** *In the* adaptive *contamination model, the optimal sample complexity of ε-robust mean testing is*

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

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 the full paper we show explicitly how to obtain Theorem I.2 by combining Narayanan's result on differentially-private mean testing with known robust-privacy equivalence results (as in e.g. [2], [24], [26]). 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 I.1 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

[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.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

samples, rather than from adaptivity itself.[4] We show in the full paper that the lower bound of Theorem I.2 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 [31].

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 [31] which works only when $\varepsilon \leq \alpha \cdot d^{-1/4}$.

Theorem I.3 (Adaptively-robust efficient mean testing (Informal; see full paper for formal statement))**.** *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 I.2 raises the question of whether a similar analogue of Theorem I.1 is possible. (The tester described in Theorem I.1 relies on a computationally inefficient "filtering step"; see Section I-D). 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 I.4 (Computational lower bound (Informal; see full paper for formal statement))**.** *In the oblivious contamination model, any ε-robust* low-degree *mean testing algorithm in the Huber contamination model has sample complexity*

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

Theorem I.4 complements a reduction from planted clique [6] 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 (see the full paper) 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.

C. *Related Work*

Gaussian Mean Testing. Gaussian mean testing is known in statistics as the Gaussian sequence model [4], [22], [27]; the understanding that it is possible to use fewer samples than dimensions appears relatively recent [33]. A recent influential work, [20], 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 [23], testing with unknown covariance [9], [19], testing subject to differential privacy [10], [31], robustly testing the covariance [17], or (distributed) testing giving partial observations from each sample [1], [34].

(Algorithmic) Robust Statistics. Algorithmic robust statistics, especially in high dimensions, has experienced a recent renaissance following a range of algorithmic breakthroughs; see the book [18]. 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 [13].

Connection to (Differential) Privacy. A recent line of work [2], [24], [26] established a (two-way) correspondence between adversarially robust and differentially private algorithms for a range of tasks, a connection we use to obtain Theorem I.2. 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 [3]. For instance, the statistical query model was invented to capture a class of PAC learning algorithms which tolerate *random classification noise* [28]. A full survey is out of scope, but some highlights include *nasty noise*, which is essentially the adaptive contamination model we consider here [8], [21], and Massart noise, which has led to exciting recent algorithmic advances [12], [15], [32]. 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, [5] shows some equivalences between adaptive and oblivious adversaries up to polynomial factors in sample complexity, for restricted classes of algorithms (SQ) or adversaries (additive). [14], [20] 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

[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.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

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.

D. Overview of Techniques

1) Exploiting Obliviousness to Robustly Test with Fewer Samples:

a) Our Approach.: We focus first on our main technical contribution, the mean tester from Theorem I.1. 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:

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

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 $\left\|\sum_{i\in B} X_i\right\|_2$ larger than it would be for a set of εn good samples. Building on this idea, we can design a tester using $\tilde{\Theta}\Big(\max\Big(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\Big(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\Big)\Big)\Big)$ 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 $\left\|\sum_{i\in S} X_i\right\|_2^2$. 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 4 (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 $|T| \le \varepsilon n$,

$$\|\mathrm{Sum}(T)\|_2^2 \le |T|d + \tilde{O}(\varepsilon^{1.5} n^{1.5}\sqrt{d} + \varepsilon^2 n^2)$$

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 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 5 (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 $X_1, \ldots, X_n$ 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 $\|\mathrm{Sum}(S)\|_2^2 - (1-\varepsilon)d \ge \Omega(\alpha^2 n^2)$ 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 (Main Lemma for Narayanan's Tester)**.** *For any two good-enough subsets S, S' of $X_1, \ldots, X_n \in \mathbb{R}^d$,*

[5] A similar tester can be extracted from [31]. While Narayanan's paper focuses on differentially private mean testing, the tester can be shown to be robust by virtue of its privacy guarantees. The tester we describe here is simpler than Narayanan's original tester, in part because we need only robustness, not privacy.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

$\left|\|\mathrm{Sum}(S)\|_2^2 - \|\mathrm{Sum}(S')\|_2^2\right| \ll \alpha^2 n^2$*, 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

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

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 $\varepsilon n d$ bad samples B will satisfy $\|\mathrm{Sum}(B)\|_2^2 \leq O(\varepsilon n d)$; 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,

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

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, $|\|\mathrm{Sum}(B)\|_2^2 - \varepsilon n d| \leq O(\varepsilon n d) \ll \alpha^2 n^2$ 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 in the null case

$$\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle \sim \mathcal{N}\Big(0, (1-\varepsilon)n\|\mathrm{Sum}(B)\|_2^2\Big)$$

and in the alternative case

$$\begin{aligned}&\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle\\ &\sim \mathcal{N}\Big((1-\varepsilon)n\langle \mu, \mathrm{Sum}(B)\rangle, (1-\varepsilon)n\|\mathrm{Sum}(B)\|_2^2\Big) .\end{aligned}$$

So, $|\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle| \leq O(n\alpha \cdot \sqrt{\varepsilon n d} + n\sqrt{\varepsilon d}) \ll \alpha^2 n^2$, as $\|\mathrm{Sum}(B)\|_2^2 \leq O(\varepsilon n d)$ 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 6 (Informal, see full paper for formal statement)**.** A *friendly* oblivious adversary introduces $\{X_i\}_{i \in B}$ such that

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

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

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 $\|\mathrm{Sum}(B)\|_2^2 \le O(\varepsilon n d)$ from above, observe that subject to friendliness, for any $S \subseteq B$,

$$\begin{aligned}\|\mathrm{Sum}(S)\|_2^2 &= |S| \cdot (d \pm \tilde{O}(\sqrt{d})) \\ &\quad + O(\mathbb{E}_{S_1,S_2} \langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle) \\ &= d|S| + \tilde{O}(|S|\sqrt{\varepsilon n d} + |S|\sqrt{d})\end{aligned}$$

where S_1, S_2 is a random partition of S. In particular, $\|\mathrm{Sum}(B)\|_2^2 = \varepsilon n d \pm o(\alpha^2 n^2)$ 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: Given $X_1, \ldots, X_n \in \mathbb{R}^d$, if $\|\mathrm{Sum}([n])\|_2^2 - nd \ge \Omega(\alpha^2 n^2)$, or if

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

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 $|\langle \mathrm{Sum}(G), \mathrm{Sum}(B)\rangle| \le O(n\sqrt{\varepsilon d})$ using standard concentration tools and obliviousness, and $\|\mathrm{Sum}(B)\|_2^2 = \varepsilon n d + \tilde{O}(\varepsilon^{1.5} n^{1.5}\sqrt{d} + \varepsilon n\sqrt{d})$ by friendliness. All together,

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

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,

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

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 the full paper.

Of course, outputting $\mathbf{H}_1$ when

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

only makes sense if the adversary cannot make this happen in the null model. We show 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 $X_1, \ldots, X_n$ 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 $\|\mathrm{Sum}(S)\|_2^2 \ge 100\varepsilon n d$, 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 $X_1, \ldots, X_n$ into U and V.
2) Using only U, identify a set of unit vectors $v_1, \ldots, v_\ell \in \mathbb{R}^d$.
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 $T_1, \ldots, T_\ell \subseteq U$ such that, for each $i \in [\ell]$,

[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.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

- $|T_i| \ll \varepsilon n/(\log n)^{O(1)}$ (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 $T_1, \ldots, T_m$ "cleans" U.

We need to establish two things: first, that such a family $T_1, \ldots, T_\ell$ which cleans U exists, and second, with high probability over the random split U, V, any $T_1, \ldots, T_\ell \subseteq \{X_1, \ldots, X_n\}$ which cleans U also cleans V. However, these are in tension. For the first, we would like to be able to choose the sets $T_1, \ldots, T_\ell$ 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 $T_1, \ldots, T_\ell$ the cleaning algorithm could make, because we will need to make a union bound over all such choices; the smaller the sets $T_1, \ldots, T_\ell$ 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 2 (Small Witness Lemma). *Let $S_1, S_2 \subseteq \mathbb{R}^d$ have $|S_1|, |S_2| = \varepsilon n$ and $\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle \geq \varepsilon n \cdot \sqrt{\theta d}$. Suppose $S_1 \cup S_2$ is λ-friendly, for some $\lambda \gg \theta$, and that there is some parameter $C > 0$ such that $|\langle X, X'\rangle| \leq \theta\sqrt{d}/C$ 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 $|S_1'| \leq \varepsilon n/C$ and $\Omega(\varepsilon n)$ vectors $X \in S_2$ such that $\left\langle X_i, \frac{\mathrm{Sum}(S_1')}{\|\mathrm{Sum}(S_1')\|_2}\right\rangle \geq \Omega\left(\sqrt{\frac{\theta}{C\varepsilon n}}\right)$.*

In Lemma 2, we think of $\theta \approx \varepsilon n(\log n)^{O(1)}$, so that $\langle \mathrm{Sum}(S_1), \mathrm{Sum}(S_2)\rangle \geq \varepsilon n\sqrt{\theta d}$ is a violation of friendliness, and $C \approx (\log n)^{O(1)}$ so that S_1' is significantly smaller than S_1. Proving Lemma 2 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 $\mathrm{Sum}(S_1)$, and then show this is preserved when we replace S_1 with a random subset $S_1' \subset S_1$. Lemma 2 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 $T_1, \ldots, T_\ell$ which cleans U, then with high probability it also cleans V. Consider the set $\mathcal{T}$ of all possible $(T_1, \ldots, T_\ell) \in \binom{n}{\varepsilon n/(\log n)^{O(1)}}^\ell$; note that $|\mathcal{T}| \leq 2^{\varepsilon n/(\log n)^{O(1)}}$ because $\ell = O(1)$.

Fixing some $(T_1, \ldots, T_\ell) \in \mathcal{T}$, our goal is to show that with probability at least $1 - 2^{-\Omega(\varepsilon n)}$ over the random split U, V, if $T_1, \ldots, T_\ell$ 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 $X_1, \ldots, X_n$ 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.

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 $\tilde{\Omega}\left(\min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)$ 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 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 $n \ll \tilde{\Omega}\left(\min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)$. 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 [20] – 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 [20] by using a *conditional* second moment (a.k.a. conditional χ^2 divergence) approach. ([20] 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.)

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

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 [7]) using the same instance from [20]; this is a direct computation using now-standard techniques from [30].

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 4). 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 $n \le d$. 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 $|T| \le \varepsilon n$, then $1_T^\top M 1_T$ certifies the first part of good-enough-ness:

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

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 $n \le d$; $n > d$ 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 the full paper 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.

Acknowledgments

The authors would like to thank Guy Blanc and Gautam Kamath for some helpful suggestions.

References

[1] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. volume 125 of *Proceedings of Machine Learning Research*, pages 41–63. PMLR, 09–12 Jul 2020.

[2] Hilal Asi, Jonathan R. Ullman, and Lydia Zakynthinou. From robustness to privacy and back. *CoRR*, abs/2302.01855, 2023.

[3] Maria-Florina Balcan and Nika Haghtalab. Noise in classification., 2020.

[4] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, pages 577–606, 2002.

[5] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. On the power of adaptivity in statistical adversaries. In *Conference on Learning Theory*, pages 5030–5061. PMLR, 2022.

[6] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

[7] Matthew Brennan, Guy Bresler, Samuel B Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *arXiv preprint arXiv:2009.06107*, 2020.

[8] Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[9] 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*, pages 321–336. SIAM, 2021.

[10] Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan R. Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. In *NeurIPS*, 2020.

[11] 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. *Advances in Neural Information Processing Systems*, 34:10862–10875, 2021.

[12] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[14] 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*, pages 2683–2702. SIAM, 2018.

[15] Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 35:3624–3636, 2022.

[16] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

[17] Ilias Diakonikolas and Daniel M. Kane. The sample complexity of robust covariance testing. In *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 1511–1521. PMLR, 2021.

[18] Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023. To appear. Draft available at https://sites.google.com/view/ars-book/.

[19] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Gaussian mean testing made simple. In *SOSA*, pages 348–352. SIAM, 2023.

[20] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS*, pages 73–84. IEEE Computer Society, 2017.

[21] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the*

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.

50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1061–1073, 2018.

[22] Michael Sergeevich Ermakov. Minimax detection of a signal in a gaussian white noise. *Theory of Probability & Its Applications*, 35(4):667–679, 1991.

[23] Anand Jerry George and Clément L. Canonne. Robust testing in high-dimensional sparse models. In *NeurIPS*, 2022.

[24] Kristian Georgiev and Samuel B. Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In *NeurIPS*, 2022.

[25] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

[26] Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. *CoRR*, abs/2212.05015, 2022.

[27] Yuri Ingster, Jurij I Ingster, and IA Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

[28] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[29] 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)*, pages 69–75. IEEE, 2020.

[30] 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*, pages 1–50. Springer, 2022.

[31] Shyam Narayanan. Private high-dimensional hypothesis testing. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 3979–4027. PMLR, 2022.

[32] Rajai Nasser and Stefan Tiegel. Optimal sq lower bounds for learning halfspaces with massart noise. In *Conference on Learning Theory*, pages 1047–1074. PMLR, 2022.

[33] Muni S Srivastava and Meng Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402, 2008.

[34] Botond Szabó, Lasse Vuursteen, and Harry van Zanten. Optimal high-dimensional and nonparametric distributed testing under communication constraints. 2022.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 24,2024 at 18:34:42 UTC from IEEE Xplore. Restrictions apply.