High-Dimensional Gaussian Mean Estimation under Realizable Contamination

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin-Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin-Madison

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study mean estimation for a Gaussian distribution with identity covariance in $\mathbb{R}^d$ under a missing data scheme termed *realizable ε-contamination* model. In this model an adversary can choose a function $r(x)$ between 0 and ε and each sample x goes missing with probability $r(x)$. Recent work (Ma et al., 2024) proposed this model as an intermediate-strength setting between Missing Completely At Random (MCAR)—where missingness is independent of the data—and Missing Not At Random (MNAR)—where missingness may depend arbitrarily on the sample values and can lead to non-identifiability issues. That work established information-theoretic upper and lower bounds for mean estimation in the realizable contamination model. Their proposed estimators incur runtime exponential in the dimension, leaving open the possibility of computationally efficient algorithms in high dimensions. In this work, we establish an information–computation gap in the Statistical Query model (and, as a corollary, for Low-Degree Polynomials and PTF tests), showing that algorithms must either use substantially more samples than information-theoretically necessary or incur exponential runtime. We complement our SQ lower bound with an algorithm whose sample–time tradeoff nearly matches our lower bound. Together, these results qualitatively characterize the complexity of Gaussian mean estimation under ε-realizable contamination.

Keywords: Robust Statistics, Missing Data, Mean Estimation, High-Dimensional Inference

1. Introduction

The fundamental assumption underlying much of classical statistics is that datasets consist of i.i.d. samples drawn independently from the distribution we aim to learn. In practice, however, this assumption is often violated. A common violation arises when the observations are incomplete. This can occur for a variety of reasons, for example due to data collection via crowdsourcing (Vuurens et al., 2011) or peer grading (Piech et al.; Kulkarni et al., 2013), and has led to the development of a broad literature studying estimators that are robust to missing data (see, e.g., Rubin (1976); Tsiatis; Little and Rubin (2019b) for standard references).

Different kinds of missingness patterns can be classified based on their nature. The simplest and most benign form of missingness is known as "Missing Completely at Random" (MCAR), meaning that the mechanism causing the missingness is independent of the data itself. A large body of work studies statistical estimation and inference under MCAR assumptions. Examples include sparse linear regression (Loh and Wainwright; Belloni et al., 2017), classification (Tony Cai and Zhang, 2019; Sell et al., 2024), PCA (Elsener and van de Geer, 2019; Zhu et al., 2022; Yan et al., 2024),

covariance and precision matrix estimation (Lounici, 2014; Loh and Tan, 2018) and changepoint estimation (Xie et al., 2012; Follain et al., 2022).

However, the MCAR assumption fails to capture many settings of interest where missingness is *systematic*. For example, individuals with depression are more likely to submit incomplete questionnaires (Carreras et al., 2021), and data may be missing for patients who discontinue treatment or go off protocol due to poor tolerability (Little et al., 2012). Other commonly studied missingness mechanisms include MAR, in which the missingness dependence on the data values is only through the observed portion of the data (Little and Rubin, 2019a; Seaman et al., 2013; Farewell et al., 2022), and MNAR, where the missingness may depend in any way on the data (Robins, 1997; Rotnitzky and Robins, 1997; Scharfstein et al., 1999; Shpitser et al., 2015; Adak et al., 2020; Diakonikolas et al., 2025a). While these models are more expressive, the resulting statistical guarantees are sometimes weak, for example failing to ensure identifiability of the target parameters (Ma et al., 2024).

Recent work (Ma et al., 2024), proposed and studied a different model—which they termed *realizable contamination*; see Definition 1. This missingness model is not MCAR, but it is milder than MAR and MNAR: the missingness depends on the data, but in a more structured manner. As we explain below, the realizable contamination model can also be viewed as an analogue of Massart noise (Massart and Nedelec, 2006)—a widely studied label-corruption model in supervised learning that lies between purely random and adversarial corruptions—in the context of unsupervised setting.

Definition 1 (Realizable ε-contamination model) *Let $\varepsilon \in (0,1)$ be a contamination parameter. Let P be a distribution on a domain $\mathcal{X}$ with probability density function (pdf) $p : \mathcal{X} \to \mathbb{R}_+$. An ε-corrupted version $\widetilde{P}$ of P is any distribution that can be obtained as follows: First, an adversary chooses a function $f : \mathcal{X} \to \mathbb{R}_+$ such that $(1-\varepsilon)p(x) \le f(x) \le p(x)$. $\widetilde{P}$ is then defined to be the distribution whose samples are generated as follows:*

- *With probability $\int_{\mathcal{X}} f(x)\mathrm{d}x$ the sample is drawn from the distribution with pdf $f(x)/\int_{\mathcal{X}} f(x)\mathrm{d}x$.*
- *With probability $1 - \int_{\mathcal{X}} f(x)\mathrm{d}x$ the sample is set to the special symbol $\perp$.*

Ma et al. (2024) define this model using a somewhat different formalism that is equivalent to Definition 1. The equivalence of the definitions is discussed in Section A.1. Below, we discuss connections between Definition 1 and preexisting concepts in the statistics and ML literature.

First, the realizable contamination model shares some similarities with Huber's contamination model from classical robust statistics (Tukey, 1960; Huber, 1964). In Huber's model, the observed samples are drawn i.i.d. from a mixture that with probability $(1-\varepsilon)$ outputs a sample from the inlier (clean) distribution and with probability ε ouputs a sample from an arbitrary outlier distribution. As explained in Section A.1, the realizable contamination model of Definition 1 can equivalently be described as a mixture where, with probability $1-\varepsilon$, a sample is drawn from P, and otherwise from an MNAR version of P. Thus, the inlier component is the same in both models. That said, Huber's model is more powerful due to its ability to introduce arbitrary outlier samples.

Second, the model of Definition 1 can be viewed as an unsupervised analogue of *Massart* noise (Massart and Nedelec, 2006). In supervised learning, Random Classification Noise (Angluin and Laird, 1988) and adversarial noise represent two extremes, and Massart noise was introduced as a realistic intermediate model. There, the label of x is flipped with probability $\eta(x)$, with $\int \eta(x)dx$ equal to the total corruption rate. As shown in Section A.1, Definition 1 admits an equivalent description: a sample $x \sim P$ is discarded with probability $1 - f(x)/p(x)$, closely mirroring the Massart noise model.

Lastly, as observed in Ma et al. (2024), realizable ε-contamination generalizes several previous approaches for studying restricted forms of MNAR, including certain biased sampling models (Vardi, 1985; Gill et al., 1988; Bickel and Ritov, 1991; Aronow and Lee, 2013; Sahoo et al., 2022), related restrictions known as sensitivity conditions in the causal inference literature (Rosenbaum, 1987; Zhao et al., 2019), and the missingness model used in the truncated statistics literature (Daskalakis et al., 2018; Kontonis et al., 2019; Diakonikolas et al., 2024a). The latter can be viewed, at a high level, as corresponding to Definition 1 with $\varepsilon = 1$, albeit with important differences between the two models. A more detailed discussion of related work appears in Section A.1.

In this paper, we study the complexity of arguably the most fundamental statistical task in the presence of realizable ε-contamination: estimating the mean of a multivariate Gaussian with identity covariance. Specifically, given access to ε-corrupted samples from $P = \mathcal{N}(\mu, I_d)$ on $\mathbb{R}^d$ where the mean vector μ is unknown, and a desired accuracy δ, the goal is to compute an estimate $\widehat{\mu}$ that lies within Euclidean distance δ of the true mean μ. The only known results for this problem are information-theoretic. Ma et al. (2024) established matching upper and lower bounds on the task's sample complexity. For high constant probability of success, it is

$$n = \Theta\left(\frac{d}{\delta^2(1-\varepsilon)}\right) + \exp\left(\Theta\left(\frac{\log(1+\frac{\varepsilon}{1-\varepsilon})}{\delta}\right)^2\right). \tag{1}$$

We note that the first term in (1) corresponds to the standard sample complexity for Gaussian mean estimation from the $n(1-\varepsilon)$ clean samples, while the second term captures the additional cost due to realizable ε-contamination. Remarkably, these results hold for all $\varepsilon, \delta \in (0,1)$, highlighting two important differences from Huber's contamination model (and other robust statistics/missing data models). First, $\delta \to 0$ implies consistent estimation is possible; that is, one can achieve arbitrarily small error, whereas under Huber's model, estimation is only possible for $\delta \gg \varepsilon$. Second, estimation remains possible even for $\varepsilon \geq 1/2$, i.e., when the majority of samples are corrupted.

For the sake of completeness, the Appendix provides a self-contained and more concise proof of the sample complexity bounds, albeit with slightly weaker guarantees: Appendix E establishes a lower bound in the one-dimensional case, while Appendix F presents a (computationally inefficient) algorithm using $\frac{d}{\varepsilon^2}\exp((\log(1+\varepsilon/(1-\varepsilon))/\delta)^2)$ samples.

Although the information-theoretic aspects of our problem are well-understood, much less is known about the computational aspects (the focus of this paper). Specifically, essentially the only known multivariate estimator is based on computing a cover of the unit sphere with $2^{\Theta(d)}$ directions and applying the one-dimensional estimator to the projections of the samples along each direction. This has runtime $2^{\Theta(d)}\,\mathrm{poly}(n,d)$ and therefore begs the following questions:

Is there a sample (near-)optimal and polynomial-time mean estimator in the presence of realizable contamination? More broadly, what is the computational complexity of this task?

In this paper we provide strong formal evidence (in the form of Statistical Query/Low-Degree Polynomial or PTF lower bounds) that the problem exhibits an *information–computation gap*—meaning that either the runtime or sample size required is inherently large. Second, we give an algorithm whose sample-time tradeoff nearly matches our lower bound. Together, these results qualitatively characterize the complexity of Gaussian mean estimation with realizable contamination.

1.1. Our Results

We begin with our first main result: an information-computation gap for Gaussian mean estimation in the ε-realizable contamination model. As is typical for such gaps, rather than proving them unconditionally, one usually establishes them under complexity-theoretic hardness assumptions or for restricted (yet natural) families of algorithms. We state the result for the family of Statistical Query (SQ) algorithms below (and discuss other models later on). Before stating the result, we provide a brief summary of SQ.

SQ Model Basics The model, introduced by Kearns (1998) and extensively studied since, see, e.g., Feldman et al. (2013), considers algorithms that, instead of drawing individual samples from the target distribution, have indirect access to the distribution using the following oracle:

Definition 2 (STAT Oracle) *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $f : \mathbb{R}^d \to [-1, 1]$. For $\tau > 0$, the $\mathrm{STAT}(\tau)$ oracle responds to the query f with a value v such that $|v - \mathbb{E}_{X\sim D}[f(X)]| \le \tau$. We call τ the tolerance of the statistical query.*

An *SQ lower bound* for a learning problem is an unconditional statement that any SQ algorithm for the problem either needs to perform a large number q of queries, or at least one query with very small tolerance τ. Note that, by Hoeffding-Chernoff bounds, a query of tolerance τ is implementable by non-SQ algorithms by drawing $1/\tau^2$ samples and averaging them. Thus, an SQ lower bound intuitively serves as a tradeoff between runtime of $\Omega(q)$ and sample complexity of $\Omega(1/\tau)$.

Our SQ lower bound for Gaussian mean estimation under ε-realizable contamination shows that any SQ algorithm must either make an exponential number of queries or make at least one query with tolerance $d^{-\widetilde{\Omega}\left(\frac{\log(1+\varepsilon/(1-\varepsilon))}{\delta}\right)^2}$ (corresponding to a sample complexity of $d^{\widetilde{\Omega}\left(\frac{\log(1+\varepsilon/(1-\varepsilon))}{\delta}\right)^2}$ for sample-based algorithms). The lower bound is established for a testing version of the problem: distinguishing between two means that differ by δ in ℓ_2 distance. Since any estimator with accuracy $\delta/2$ solves this testing problem, the SQ hardness directly carries over to the estimation setting.

Theorem 3 (SQ lower bound) *There exists a sufficiently small absolute constant $c > 0$ such that the following holds for all $\varepsilon \in (0, 1)$ and $\delta \in (0, c\varepsilon)$. Define $m = \lfloor c\gamma^2/\log\gamma \rfloor$ for $\gamma := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$. For any dimension $d \ge (m \log d)^2$ we have the following: Any SQ algorithm that distinguishes between $\mathcal{N}(0, I_d)$ and $\mathcal{N}(\delta u, I_d)$ for u being a unit vector, under the contamination model of Definition 1 needs either $2^{d^{\Omega(1)}}$ queries or at least one query with $d^{-m/16}$ tolerance.*

The above result is based on framing our problem as a special case of Non-Gaussian Component Analysis (NGCA), a general testing problem that is known to be hard in many restricted models of computation, beyond the SQ model, including low-degree polynomial methods (Brennan et al., 2021), PTFs (Diakonikolas et al., 2025c) and the Sum-of-Squares framework (Diakonikolas et al., 2024b). As such, we highlight that qualitatively similar hardness results to those in Theorem 3 also hold in these models. The formal statements are deferred to Appendix G.

Our second main result presents an algorithm whose leading term in the sample complexity is $d^{O\left(\log(1+\frac{\varepsilon}{1-\varepsilon})/\delta\right)^2}$, with a runtime that is nearly polynomial in n and d. In light of our SQ lower bound, this conceptually means that (up to certain factors that we explain below) our algorithm achieves the optimal tradeoff between information and computation.

Theorem 4 (Algorithmic result) *Let $\varepsilon \in (0,1)$ and $\delta \leq \varepsilon$ be parameters.*[1] *There exists an algorithm that takes as input samples from an ε-corrupted version of $\mathcal{N}(\mu, I_d)$, the parameters ε, δ and returns $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 \leq \delta$ with probability at least 0.9. Moreover, the sample complexity of the algorithm is $n = \frac{(kd)^{O(k)}}{\varepsilon^2}$ where $k := \left(\frac{1}{\delta}\log(1+\frac{\varepsilon}{1-\varepsilon})\right)^2$ and the runtime of the algorithm is $\exp(e^{\widetilde{O}(k)}/\varepsilon^2)\,\mathrm{poly}(n,d)$.*

Up to the $k^{O(k)}/\varepsilon^2$ factor (which is independent of the dimension), the sample complexity achieved by our algorithm matches our SQ lower bound. The runtime of our algorithm is polynomial in n and d, up to the term $\exp(e^{\widetilde{O}(k)}/\varepsilon^2)$ (which is also dimension-independent).

1.2. Our Techniques

SQ lower bound. Our SQ lower bound builds on the framework of Diakonikolas et al. (2017), which establishes the following. If A is a one-dimensional distribution whose first m moments match those of $\mathcal{N}(0,1)$, then distinguishing between $\mathcal{N}(0,I)$ and the d-dimensional distribution that agrees with A along an unknown direction and is standard Gaussian in all orthogonal directions requires either $q = 2^{d^{\Omega(1)}}$ queries or tolerance $\tau < d^{-\Omega(m)}$ in the SQ model. We show that the robust mean estimation problem we consider can be cast in this form. The main challenge is to design an adversary, as in Definition 1, that corrupts $\mathcal{N}(\delta,1)$ so that it matches $m = \widetilde{\Omega}(\log(1+\frac{\varepsilon}{1-\varepsilon})/\delta)^2$ moments of $\mathcal{N}(0,1)$. We accomplish this via a two-step approach.

First, we construct a function f for the adversary in Definition 1 that uses $\varepsilon/2$ of the corruption budget to match the standard Gaussian over a large interval $x \in [-B, B]$, where $B = \log(1+\frac{\varepsilon}{1-\varepsilon})/\delta$. This relies on the observation that the probability density functions of two unit-variance Gaussians whose means differ by δ are within a $(1 \pm \varepsilon)$ multiplicative factor of each other, except in the B-tails.

After this step, the moments do not exactly match those of $\mathcal{N}(0,1)$ due to discrepancies outside the interval $[-B, B]$. However, for a small constant c, the difference in the first cB^2 moments is exponentially small because the mass in the tails decays exponentially for Gaussians. Consequently, by adding a very small polynomial $p(x)$ to $f(x)$ on $[-1,1]$, we can eliminate these remaining discrepancies. By imposing the moment-matching constraints and expanding the polynomial $p(x)$ in the basis of Legendre polynomials, we show that such a polynomial indeed $p(\cdot)$ exists. Adding this polynomial correction corresponds to censoring each point x with probability $\varepsilon/4 + p(x)(1-\varepsilon/4)\mathbb{1}(|x|<1)/\phi(x)$, where ϕ denotes the standard Gaussian pdf.

Algorithmic result. As discussed in the introduction, from an information-theoretic perspective, a number of samples as shown in Equation (1) suffices to estimate the mean up to error δ. The algorithm achieving this guarantee relies on the fact that two Gaussians whose means differ by δ have probability density functions that are within a multiplicative factor of $1 \pm \varepsilon$ of each other, except in the $(\log(1+\frac{\varepsilon}{1-\varepsilon})/\delta)$-tails. Consequently, by accurately estimating the densities in these tails using $\exp((\log(1+\frac{\varepsilon}{1-\varepsilon})/\delta)^2)$ samples, one can rule out any candidate mean μ' that lies at distance more than δ from the true mean μ. This yields a one-dimensional estimator; (also see Appendix F for more details). A multivariate extension can be obtained by repeating this procedure along every direction in a fine cover of the unit sphere, thereby learning an ℓ_2-approximation of μ. However, such a cover must have $2^{\Theta(d)}$ size, rendering the resulting algorithm computationally infeasible.

1. If $\delta \gg \varepsilon$, pre-existing algorithms from robust statistics (see, e.g., Diakonikolas and Kane (2023)) obtain error $O(\delta)$.

For computational efficiency, rather than computing tails, we rely on moments as a substitute. Our approach is motivated by the following observation. Let $k > (\log(1 + \frac{\varepsilon}{1-\varepsilon})/\delta)^2$. Then, for two unit-variance Gaussians whose means differ by δ (e.g., $\mathcal{N}(0,1)$ and $\mathcal{N}(\delta,1)$), there must exist a moment of order $t \in [k]$ that differs by at least ε (Lemma 12). This structural result motivates the following algorithm: First, we compute a rough estimate $\widehat{\mu}$ of the true mean μ and translate all samples via the transformation $x \leftarrow x - \widehat{\mu}$. After this preprocessing step, the samples behave as if they were drawn from a distribution with mean $\mu - \widehat{\mu}$, and the remaining goal is to estimate this difference to error δ. For each $t \in [k]$, we compute the order-t moment tensor T_t. By the structural result of Lemma 12 we can certify that for all v such that $\langle v^{\otimes t}, T_t \rangle$ is small below a certain threshold η it should hold $|v^\top(\mu - \widehat{\mu})| \leq \delta/2$ (Claim 14). It therefore suffices to restrict attention to the subspace V corresponding to directions with large moments, and to apply the inefficient estimator described above only within this subspace to estimate $\mu - \widehat{\mu}$ up to error $\delta/2$. By the triangle inequality, the resulting estimate has error at most δ. The runtime of this approach is $2^{O(\dim(V))}$. Crucially, we show that $\dim(V)$ is small—bounded solely as a function of ε and δ, with no dependence on d. This follows from the fact that the ℓ_2-norm of the full moment tensor is bounded in terms of ε, δ, a property that continues to hold under our corruption model (cf. Lemma 15). A small norm implies that only few eigenvalues can be large, which in turn bounds $\dim(V)$.

There are two technical complications. First, identifying the subspace V is computationally hard: even finding a single direction v with a large projection of a degree-4 moment tensor is intractable. To circumvent this, instead of searching for directions v with small $\langle v^{\otimes t}, T_t \rangle$, we flatten T_t into a matrix $M \in \mathbb{R}^d \times \mathbb{R}^{d^{t-1}}$ and define V as the span of singular vectors of M with singular values exceeding η, which is computable in polynomial time. This relaxation suffices since, for any $v \in V$, $\langle v^{\otimes t}, T_t \rangle \leq \sup_{u \in \mathbb{R}^{d^{t-1}} : \|u\|_2 = 1} \langle v, Mu \rangle \leq \delta$.

The second challenge concerns the initial rough estimate $\widehat{\mu}$. While the robust statistics literature offers many black-box estimators, most assume stronger contamination models and require $\varepsilon < 1/2$, failing when a majority of samples are corrupted. To our knowledge, no polynomial-time rough estimator exists for Definition 1 when $\varepsilon > 1/2$. We therefore use a list-decodable estimator, which tolerates a majority of outliers but, due to non-identifiability, outputs a list of candidate means, one of which is close to μ, but all others could be arbitrarily inaccurate. Fortunately, there is a tournament-based procedure known in the literature that we can use to identify an element of the list that has error comparable to the best among all of them. This enables us to end up with a single warm start vector and completes the proof sketch of the algorithm.

2. Preliminaries

Basic Notation We use $\mathbb{Z}_+$ for the set of positive integers. We denote $[n] = \{1, \ldots, n\}$. For a vector x we denote by $\|x\|_2$ its Euclidean norm. Let I_d denote the $d \times d$ identity matrix (omitting the subscript when it is clear from the context). We use $\top$ for the transpose of matrices and vectors. For a tensor T, we define by $\|T\|_2 = \sqrt{\sum_i T_i^2}$ the ℓ_2 or (Frobenius) norm. We use $a \lesssim b$ to denote that there exists an absolute universal constant $C > 0$ (independent of the variables or parameters on which a and b depend) such that $a \leq Cb$. In our notation $a = O(b)$ has the same meaning as $a \lesssim b$ (similarly for $\Omega(\cdot)$ notation) We use $\widetilde{O}$ and $\widetilde{\Omega}$ to hide polylogarithmic factors in the argument.

2.1. Non-Gaussian Component Analysis (NGCA)

We give a brief background on the SQ hardness of the Non-Gaussian Component Analysis problem (NGCA). First, the testing version of NGCA is defined as distinguishing between a standard Gaussian and a Gaussian that has a non-Gaussian component planted in an unknown direction, defined below.

Definition 5 (Hidden direction distribution) *Let A be a distribution on $\mathbb{R}$. For a unit vector v, we denote by $P_{A,v}$ the distribution with the density $P_{A,v}(x) := A(v^\top x)\phi_{\perp v}(x)$, where $\phi_{\perp v}(x) = \exp\left(-\|x-(v^\top x)v\|_2^2/2\right)/(2\pi)^{(d-1)/2}$, i.e., the distribution that coincides with A on the direction v and is standard Gaussian in every orthogonal direction.*

Problem 6 (Non-Gaussian Component Analysis (NGCA)) *Let A be a distribution on $\mathbb{R}$ and $P_{A,v}$ the distribution from Definition 5. We define the following hypothesis testing problem:*

- *H_0: The data distribution is $\mathcal{N}(0, I_d)$.*
- *H_1: The data distribution is $P_{A,v}$, for some vector $v \in \mathcal{S}^{d-1}$ in the unit sphere.*

A known result is that NGCA is hard in the SQ model if A matches a lot of moments with the standard Gaussian. Here we use the statement from Theorem 1.5 in Diakonikolas et al. (2023) using $\lambda = 1/2$ and $c = (1-\lambda)/8 = 1/16$ and $\nu = 0$. Hardness results for other models, beyond SQ are summarized in Appendix G.

Condition 7 (Moment matching condition) $\mathbb{E}_{x\sim A}[x^i] - \mathbb{E}_{x\sim\mathcal{N}(0,1)}[x^i] = 0$ *for all* $i \in [m]$.

Proposition 8 (Theorem 1.5 in Diakonikolas et al. (2023)) *Let d, m be positive integers with $d \geq (m\log d)^2$. Any SQ algorithm that solves Problem 6 for a distribution A satisfying Condition 7 requires either $2^{d^{\Omega(1)}}$ many queries or at least one query with accuracy $d^{-m/16}$.*

3. Statistical Query Lower Bound: Proof of Theorem 3

In order to show Theorem 3, we will prove the following moment-matching proposition.

Proposition 9 (Moment Matching) *There exists a sufficiently small absolute constant $c > 0$ such that the following holds. For every $\varepsilon \in (0,1)$ and $\delta \in (0, c\varepsilon)$, there exists a distribution A on $\mathbb{R}$ such that the following statements are satisfied for $m = \lfloor c\gamma^2/\log\gamma \rfloor$ where $\gamma := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$:*

- *A is the conditional distribution on the visible (non-deleted) samples of an ε-corrupted version of $\mathcal{N}(\delta, 1)$ according to Definition 1.*
- *It holds $\mathbb{E}_{x\sim A}[x^i] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[x^i]$ for $i = 1, \ldots, m$.*

We briefly explain how Proposition 9 yields Theorem 3. First note that the conclusion of Proposition 9 is phrased in terms of the conditional distribution on the visible samples. However, in Theorem 3 we are interested in the hypothesis testing between the corrupted versions of $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu, \delta)$, i.e., the entire distributions on both visible and non-visible samples. Although this might seem like a discrepancy, it suffices to focus on the conditional distributions on the visible samples since one may also randomly delete an ε-fraction of points from the null hypothesis, so that the number of deleted points matches the corresponding number in the alternative hypothesis.

For v being the unit vector in the direction of μ, let $P_{A,v}$ be the distribution as in Problem 6. Then, $P_{A,v}$ is an ε-corrupted version of $\mathcal{N}(\delta v, I)$, and the hypothesis testing problem of Theorem 3

is an instance of non-Gaussian component analysis testing with A as the hidden direction distribution. An application of Proposition 8 with $m = \lfloor c\gamma^2/\log\gamma \rfloor$ then yields Theorem 3. Finally, note that Theorem 3 shows hardness of distinguishing between the ground truth $\mu = 0$ and $\|\mu\|_2 \geq \delta$. This immediately implies hardness of the problem of estimating μ up to error $\delta/2$. This is because if one has an estimator it can run it to obtain $\widehat{\mu}$ with $\|\widehat{\mu} - \mu\|_2 \leq \delta/2$ and reject the null hypothesis iff $\|\widehat{\mu}\|_2 > \delta/2$.

3.1. Proof of Proposition 9

Denote by ϕ the pdf of $\mathcal{N}(0,1)$. Following the Definition 1 we need to find a function $f : \mathbb{R} \to \mathbb{R}_+$ such that $(1-\varepsilon)\phi(x-\delta) \leq f(x) \leq \phi(x-\delta)$ and the distribution with pdf $f(x)/\int_{\mathbb{R}} f(x)\mathrm{d}x$ matches the first m moments with $\mathcal{N}(0,1)$. The argument to do so consists of two parts:

1. We will show that there exists $g : \mathbb{R} \to \mathbb{R}_+$ such that $(1-\varepsilon)\phi(x-\delta) \leq g(x) \leq \phi(x-\delta)$, $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$ and $\frac{g(x)}{1-\varepsilon/2} = \phi(x)$ for all $x \in [-B, B]$ for $B := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$. This means that g defines an ε-corrupted version of $\mathcal{N}(\delta, 1)$ that matches exactly $\mathcal{N}(0,1)$ in the entire range $[-B, B]$. Due to the mismatch outside this interval, the first m moments will not match exactly, but will differ by only a small amount.
2. In order to correct the moments, we will find an appropriate polynomial $p(x)$ to add to the function g from the previous step in $[-1, 1]$ so that (i) the distribution with pdf proportional to $f(x) := g(x) + p(x)\mathbb{1}(|x| \leq 1)$ now matches the m first moments with $\mathcal{N}(0,1)$ exactly and (ii) f satisfies $(1-\varepsilon)\phi(x-\delta) \leq f(x) \leq \phi(x-\delta)$, i.e., it is still a valid ε-corruption of $\mathcal{N}(\delta, 1)$.

We now present the proofs of the two steps in Lemma 10 and Lemma 11, respectively.

Lemma 10 *Denote by $\phi(x)$ the pdf of $\mathcal{N}(0,1)$. For any $\varepsilon, \delta \in (0,1)$ there exists a function $g : \mathbb{R} \to \mathbb{R}_+$ such that $(1-\varepsilon)\phi(x-\delta) \leq g(x) \leq \phi(x-\delta)$, $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$ and $\frac{g(x)}{1-\varepsilon/2} = \phi(x)$ for all $x \in [-B + \delta/2, B + \delta/2]$ where $B := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$.*

Proof Sketch The full proof is provided in Appendix C. For convenience, we will prove the claim with everything shifted by $\delta/2$ to the left, i.e., we will show that,

$$(1-\varepsilon)\phi(x-\delta/2) \leq g(x) \leq \phi(x-\delta/2) \tag{2}$$

as well as $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$ and $\frac{g(x)}{1-\varepsilon/2} = \phi(x+\delta/2)$ for all $x \in [-B, B]$. For simplicity of notation, we will use $p_+(x) := \phi(x-\delta/2)$ and $p_-(x) := \phi(x+\delta/2)$ to denote the two Gaussian densities for the rest of the proof.

The main idea is to let $g(x) = (1-\varepsilon/2)p_-(x)$ for as all x in an interval around zero which is as large as possible without violating the condition (2). Once we find which is the biggest possible such interval, we will need to correct $g(x)$ outside of it so that it still respects condition (2).

For the first part of our proof argument (finding the largest interval for which setting $g(x) = (1-\varepsilon/2)p_-(x)$ inside it satisfies condition (2)) we solve the equations $(1-\varepsilon/2)p_-(x) = p_+(x)$ and $(1-\varepsilon/2)p_-(x) = (1-\varepsilon)p_+(x)$. The two solutions are

$$x_+ = \frac{1}{\delta}\log\left(\frac{1-\varepsilon/2}{1-\varepsilon}\right) \quad \text{and} \quad x_- = \frac{1}{\delta}\log\left(1 - \frac{\varepsilon}{2}\right).$$

This means that if we define $B := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$ we have that the function defined as $g(x) := p_-(x)(1-\varepsilon/2)$ satisfies condition (2) for all $x \in [-B, B]$.

We now need to show how to define $g(x)$ outside of $[-B, B]$. We show in the full proof provided in Appendix C that it is possible to extend the definition outside of $[-B, B]$ in a way that condition (2) continues to hold and $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$. A first, unsuccessful approach would be to set $g(x) = (1-\varepsilon)p_+(x)$ for $x > B$ and $g(x) = p_+(x)$ for $x < -B$. Although this ensures condition (2), we can show that $\int_{\mathbb{R}} g(x)\mathrm{d}x < 1 - \varepsilon/2$ is not satisfied. However, this is not the only choice for defining $g(x)$ outside of $[-B, B]$. As we show in Appendix C, choosing the other extreme of $g(x) = p_-(x)$ in $x > B$ results in $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x > 1 - \varepsilon/2$. By continuity this would mean that there exists a way of defining $g(x)$ in $x > B$ that achieves $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$. ■

Lemma 11 *Let $\varepsilon \in (0, 1), \delta \ll \varepsilon$ and $g(x)$ be as in Lemma 10. There exists a polynomial $p(x)$ such that the function $f(x) := g(x) + p(x)\mathbb{1}(|x| \le 1)$ satisfies $(1-\varepsilon)\phi(x-\delta) \le f(x) \le \phi(x-\delta)$ and the distribution with pdf $f(x)/\int_{\mathbb{R}} f(x)\mathrm{d}x$ matches the first m moments with $\mathcal{N}(0, 1)$ for some $m = \Omega(\gamma^2/\log\gamma)$, where $\gamma := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$.*

Proof Sketch Recall that $g(x)$ from Lemma 10 satisfies $(1-\varepsilon)\phi(x-\delta) \le g(x) \le \phi(x-\delta)$ for all $x \in \mathbb{R}$, integrates to $1 - \varepsilon/2$ and is exactly $g(x) = \phi(x)(1-\varepsilon/2)$ for $x \in [-B + \delta/2, B + \delta/2]$ where $B := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$. We will show the existence of a polynomial p such that

1. $|p(x)| \le c\varepsilon$, where c is a sufficiently small absolute constant,
2. $\int_{-1}^{1} p(x)\mathrm{d}x = 0$,
3. $\int_{\mathbb{R}} x^i \frac{g(x)+p(x)\mathbb{1}(|x|\le 1)}{1-\varepsilon/2}\mathrm{d}x = \int_{\mathbb{R}} x^i\phi(x)\mathrm{d}x$ for all $i \in [m]$.

This suffices because of the following. First, the moment matching property in the conclusion of Lemma 11 directly follows by the third item in the above list. We now show how the part that $(1-\varepsilon)\phi(x-\delta) \le f(x) \le \phi(x-\delta)$ for all $x \in \mathbb{R}$ follows from the above. This can be seen by verifying that (i) $\phi(x-\delta) - g(x) = \Omega(\varepsilon)$ for all $x \in [-1, 1]$ and (ii) $g(x) - (1-\varepsilon)\phi(x-\delta) = \Omega(\varepsilon)$ for all $x \in [-1, 1]$. We show the part (i) since the other part can be seen with identical arguments. The smallest value of $\phi(x-\delta) - g(x)$ happens at $x = -1$. On that point:

$$\phi(x-\delta) - (1-\varepsilon/2)\phi(x) = \phi(x-\delta) - \phi(x) + \frac{\varepsilon}{2}\phi(x) \ge -O(\delta) + \Omega(\varepsilon) \ge \Omega(\varepsilon),$$

where the first inequality uses the fact that $\phi(x-\delta) - \phi(x) = x\phi(x)\delta + \frac{\xi^2-1}{2}\phi(\xi)\delta^2$ for some $x - \delta \le \xi \le x$, by Taylor's theorem), and we also used that $\phi(x) = \Omega(1)$ for $x \in [-1, 1]$. The last inequality above used that $\delta \ll \varepsilon$.

We now turn to showing the existence of the polynomial p for which we draw some analytic ideas from Diakonikolas et al. (2017). Recall the moment matching condition that we want to ensure: $\int_{-1}^{1} x^i p(x)\mathrm{d}x = \int_{-\infty}^{\infty} x^i\phi(x)\mathrm{d}x - \int_{-\infty}^{\infty} x^i\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x$. Using the fact that $g(x)/(1-\varepsilon/2) = \phi(x)$ in the interval $[-B + \delta, B + \delta]$ the moment matching condition becomes:

$$\int_{-1}^{1} x^i p(x)\mathrm{d}x = \int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} x^i\phi(x)\mathrm{d}x - \int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} x^i\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x, \tag{3}$$

for $i = 1, \ldots, m$. First, we can express $p(x)$ as a linear combination of Legendre polynomials P_k: $p(x) = \sum_{k=0}^{m} a_k P_k(x)$, where $a_k = \frac{2k+1}{2}\int_{-1}^{1} P_k(x)p(x)\mathrm{d}x$. By properties of Legendre polynomials (see Fact 26 and relevant background in Section B.1) we have that $|p(x)| \le \sum_{k=0}^{m} |a_k|$ for all $x \in [-1, 1]$, thus it suffices to bound the coefficients a_k. The idea towards that end is to use Equation (3)

and focus on upper bounding $\left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\phi(x)\mathrm{d}x\right|$ and $\left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x\right|$. As we show in Appendix C, $\left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\phi(x)\mathrm{d}x\right| \lesssim 4^k e^{-(B-\delta)^2/4}k^k$ which is enough to yield the conclusion in Lemma 11. The details can be found in Appendix C. ■

4. Efficient Mean Estimation with Realizable Contamination: Proof of Theorem 4

The algorithm will be based on the lemma below establishing that if the ground-truth Gaussian has a mean that deviates from zero, then some sufficiently high-order moment must necessarily deviate from the corresponding standard Gaussian moment. The proof follows by explicit computations and is provided in Section D.2. Background on Hermite polynomials can be found in Section B.2.

Lemma 12 (Structural Lemma) *Let $\varepsilon \in (0,1)$ and $\delta \in (0, \log(1+\frac{2\varepsilon}{1-\varepsilon}))$ be parameters. Let k be an even integer which satisfies $k \geq 3(\frac{1}{\delta}\log(1+\frac{2\varepsilon}{1-\varepsilon}))^2$ and $P = \mathcal{N}(\delta, 1)$ be a Gaussian distribution. Then for any ε-corrupted version $\widetilde{P}$ of P under the model of Definition 1, if P' denotes the conditional distribution of $\widetilde{P}$ on the non-missing samples, it holds $|\mathbb{E}_{x\sim P'}[h_k(x)]| > \frac{\epsilon}{(k+1)^{k/2}}$, where h_k is the normalized probabilist's Hermite polynomial.*

The pseudocode for the algorithm is given in Algorithm 1. It begins by invoking a list-decodable mean estimation procedure as a black box, i.e., a procedure that returns a poly-sized list L of candidate means such that at least one element $\widetilde{\mu}_0$ is $O(\sqrt{\log(1/\varepsilon)})$ from the true mean. As explained in Section 1.2 this vector would be a good warm start for the subsequent steps. However, since the identity of this good candidate is unknown we use a standard tournament-style selection procedure from robust statistics (Diakonikolas et al., 2025b) to select a vector $\widehat{\mu}_0$ from the list L, with error approximately the best among the errors of L's candidates.

Having this $\widehat{\mu}_0$ warm start vector in hand, the algorithm draws a dataset and shifts all samples by $\widehat{\mu}_0$. After this translation, the data may be viewed as having ground truth mean $\mu - \widehat{\mu}_0$, and the task reduces to estimating this offset. We then have the main, dimension-reduction part of the algorithm. For many values t, it computes the t-th order moment tensor and considers the subspace V_t spanned by eigenvectors whose corresponding eigenvalues exceed a carefully chosen threshold. The algorithm then sets V to be the span of all the V_t's. By the structural result in Lemma 12, we show that $\mu - \widehat{\mu}_0$ is largely contained inside V. Finally, the algorithm applies a computationally inefficient estimator restricted to V to recover the component of the ground truth mean that lies in this subspace. In the proof of correctness, we will show that the dimension of V is small, thus the complexity of this step is as stated in the theorem.

Due to the lack of space, we defer formal statements for the guaranties of the two subroutines (estimation of mean estimated on the list and selection based on tournaments) to Section D.1. In the remainder of this section we prove Theorem 4, with some details deferred to the Appendices. We first argue that the final output has error δ. We then bound the complexity of the algorithm.

Error analysis In the second line, we know that the list L contains one estimate $\widehat{\mu}_0$ with $\|\widehat{\mu}_0 - \mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. In what follows we will analyze the for loop and show that the estimator produced in its last line has ℓ_2-error at most δ.

Consider a $\widehat{\mu}_0$ for which $\|\widehat{\mu}_0 - \mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. Because of the centering transformation, we can equivalently analyze things as if there was no re-centering transformation but instead all samples come from the model with mean μ of bounded norm $\|\mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$.

Algorithm 1 Spectral algorithm for multivariate mean estimation

Input: Sample access to the distribution of Definition 1, parameters $\varepsilon, \delta \in (0,1)$.
Output: Vector $\widehat{\mu} \in \mathbb{R}^d$.

Let C be a sufficiently large absolute constant, $k := \left\lceil C\left(\frac{1}{\delta}\log(1+\frac{\varepsilon}{1-\varepsilon})\right)^2\right\rceil$ and $\eta := \frac{1}{C}\frac{\varepsilon}{(k+1)^{k/2}}$.
Use a list-decoding algorithm to compute a list L of candidate mean estimates in $\mathbb{R}^d$ such that L contains at least one $\widetilde{\mu}_0$ with $\|\widetilde{\mu}_0 - \mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. // Using Fact 34
$\widehat{\mu}_0 \leftarrow \text{TOURNAMENTIMPROVE}(L, \varepsilon, O(\sqrt{\log(1/(1-\varepsilon))}))$. // Using Fact 35
Draw samples $S = \{x_1, \ldots, x_n\}$ from Definition 1 for $n = (kd)^{Ck}/\varepsilon^2$.
Recenter the dataset $S' \leftarrow \{x - \widehat{\mu}_0 : x \in S\}$.
for $t = 0, 1, \ldots, k$ **do**
  Compute the empirical Hermite tensor $\widehat{T}_t \leftarrow \frac{1}{n}\sum_{x\in S'} H_t(x)$. // cf. Definition 27
  Let $M(\widehat{T}_t)$ be the flattened $d \times d^{t-1}$ matrix.
  Compute the right singular vectors $v_1, \ldots, v_d$ of $M(\widehat{T}_t)$ with singular values $\sigma_1, \ldots, \sigma_d$.
  Define $\mathcal{I}_t = \{i \in [d] : \sigma_i > \eta\}$ and $V_t = \mathrm{span}(\{v_i : i \in \mathcal{I}_t\})$.
Let $V = \mathrm{span}(V_1, \ldots, V_k)$.
Compute $\widehat{\mu} \leftarrow \mathrm{BruteForce}(\mathrm{Proj}_V(S'), \varepsilon, \delta/2)$. // Using Theorem 45
return $\widehat{\mu} + \widehat{\mu}_0$.

First we will require the following concentration lemma, which we show in Section B.2.

Lemma 13 (Hermite tensor concentration) *Let $\eta, \varepsilon \in (0,1)$ be parameters, C be a sufficiently large absolute constant and $\mu \in \mathbb{R}^d$ be a vector with $\|\mu\|_2 = O(\sqrt{\log(\frac{1}{1-\varepsilon})})$. Let $\widetilde{P}$ be an ε-corrupted version of $\mathcal{N}(\mu, I)$ (cf. Definition 1) and let P' denote the conditional distribution of $\widetilde{P}$ on the non-missing samples. Let $x_1, \ldots, x_n \sim P'$ be i.i.d. samples and define $\widehat{T} := \frac{1}{n}\sum_{i=1}^n H_k(x_i)$, and $T := \mathbb{E}_{x\sim P'}[H_k(x)]$, where $H_k(x)$ denotes the Hermite tensor from Definition 27. If $n > C\frac{d^{3k}2^{O(k)}(k\log(\frac{1}{1-\varepsilon}))^{k/2}}{(1-\varepsilon)\eta^2\tau}$, then with probability at least $1-\tau$ we have that $\left\|\widehat{T} - T\right\|_2 \le \eta$.*

We will use $\eta := \frac{1}{C}\frac{\varepsilon}{(k+1)^{k/2}}$ as in the pseudocode. The number of samples in Theorem 4 has been chosen to allow a union bound over $t = 0, \ldots, k$ and all iterations of the for loop of the algorithm so that we have

$$\left\|\widehat{T}_t - T_t\right\|_2 \le \eta \quad \forall t = 0, \ldots, k \tag{4}$$

with probability at least 0.99. Using the formula for the sample complexity of Lemma 13 and simplifying it bit, it can be seen that this concentration event can be achieved with $(kd)^{O(k)}\varepsilon^{-2}$ samples.

We now focus on showing that the estimate $\widehat{\mu}$ towards the end of the for loop has error δ. We will first argue that the mean has a small component of size at most $\delta/2$ in the subspace $V^\perp$. Since we estimate the mean up to error $\delta/2$ on the subspace V, this will immediately mean that the total error is at most δ.

Claim 14 (Mean certification) *Consider the notation of Algorithm 1 and that the event of (4) holds. The subspace V mentioned in Algorithm 1 satisfies $v \in V^\perp \Rightarrow |v^\top \mu| \le \delta/2$.*

Proof Let $\widetilde{P}$ denote the ε-corrupted version of $\mathcal{N}(\mu, I)$ according to Definition 1 and let P' denote the conditional distribution on the non-missing samples. We prove the claim by contradiction. Suppose that $|v^\top \mu| > \delta/2$. Then by Lemma 12 there exists a $t \in \{0, \dots, k\}$ with $|\mathbb{E}_{x\sim P'}[h_t(v^\top x)]| > \eta$. Using the Hermite tensor property (see Section B.2 for background on Hermite tensors) $\mathbb{E}_{x\sim P'}[h_t(v^\top x)] = \langle v^{\otimes t}, \mathbb{E}_{x\sim P'}[H_t(x)]\rangle$ this means that $|\langle v^{\otimes t}, T_t\rangle| > \eta$. By the concentration of the event in (4), we have that $|\langle v^{\otimes t}, \widehat{T}_t\rangle| > \eta$. Therefore, v belongs to V_t, which contradicts our assumption $v \in V^\perp$. ■

Claim 14 shows that $\|\mathrm{Proj}_{V^\perp}(\mu)\|_2 \le \delta/2$. For the orthogonal subspace our algorithm computes a $\widehat{\mu} \in V$ with $\|\widehat{\mu} + \widehat{\mu}_0 - \mathrm{Proj}_V(\mu)\|_2 \le \delta/2$ (because of the guarantee stated in Theorem 45). Thus by the Pythagorean theorem, our estimator $\widehat{\mu}$ towards the end of the for loop has error at most δ. This concludes the error analysis part of the proof. In the remainder of this section we analyze the complexity of the algorithm.

Complexity of the list-decoding and tournament subroutines The sample complexity of this part is given by Fact 34 and it can be checked that it is smaller than the $(kd)^{O(k)}$ that is mentioned in the statement of Theorem 4. The runtime of list-decoding is $\mathrm{poly}(n)$. It can also be checked that the sample complexity and runtime of the tournament subroutine step, stated in Fact 35 are smaller than what is stated in Theorem 4.

Complexity of a single iteration of the outer loop We will bound the runtime of the main for loop, conditioned on the event that $\widehat{\mu}_0$ is within $O(\sqrt{\log(1/(1-\varepsilon))})$ of the true mean. The runtime of every step except the application of the brute force algorithm is polynomial on all parameters n, d and $1/(1-\varepsilon)$, so it remains to analyze the runtime of the brute force algorithm. By Theorem 45, the runtime is $2^{O(\dim(V))}\,\mathrm{poly}(n, d)$ thus we need a bound on $\dim(V)$.

Suppose that we have shown a bound $\|\widehat{T}_t\|_2 \le \gamma$. Having that bound will allow us to argue as follows: If σ_i denote the singular values of $M(T_t)$, then $\sqrt{\sum_{i=1}^d \sigma_i^2} = \|M(T_t)\|_{\mathrm{F}} = \|T_t\|_2 \le \gamma$ which means that for the set $\mathcal{I}_t = \{i : \sigma_i \ge \eta\}$ it holds $|\mathcal{I}_t| \le \gamma^2/\eta^2$. This means that $\dim(V_k) \le \gamma^2/\eta^2$. Thus $\dim(V) \le \sum_{t=0}^k \dim(V_t) \le (k+1)\gamma^2/\eta^2$.

We will show in Lemma 15 the ℓ_2-norm bound $\|\mathbb{E}[\widehat{T}_t]\|_2 \le e^{\widetilde{O}(k)}$. Under the event (4) that we conditioned on in the beginning, this will also imply that $\|\widehat{T}_t\|_2 \le e^{\widetilde{O}(k)}$. Having this and plugging $\gamma = e^{\widetilde{O}(k)}$, $\eta = O(\varepsilon/(k+1)^{k/2})$ to the bounds of the previous paragraph we will finally have the bounds below which conclude that the runtime of the algorithm is $\exp(e^{\widetilde{O}(k)}/\varepsilon^2)\,\mathrm{poly}(n, d)$.

$$\dim(V) \le (k+1)\frac{\gamma^2}{\eta^2} \le k\frac{e^{\widetilde{O}(k)}}{\eta^2} \le \frac{e^{\widetilde{O}(k)}}{\eta^2} \le \frac{e^{\widetilde{O}(k)}(k+1)^k}{\varepsilon^2} \le \frac{k^{\widetilde{O}(k)}}{\varepsilon^2} \le \frac{e^{\widetilde{O}(k)}}{\varepsilon^2},$$

Lemma 15 (Moment tensor norm bound) *Let $\widetilde{P}$ be the ε-corrupted version of $\mathcal{N}(\mu, I)$ mentioned in the statement of Theorem 4 and P' denote the conditional distribution on the non-missing samples. Assume that $\|\mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. Let $T_t = \mathbb{E}_{x\sim P'}[H_t(x)]$ denote the tensors used in Algorithm 1. We have that*

$$\|T_t\|_2 \le \frac{1}{1-\varepsilon} O\left(\log(1/(1-\varepsilon))\right)^{t/2} + \frac{1}{1-\varepsilon}\exp\left(O\left(t\log(1/(1-\varepsilon))\right)\right). \tag{5}$$

We defer the proof to Section D.2. Note that the RHS in inequality (5) can be further bounded from above by the simpler expression $e^{\widetilde{O}(k)}$ by using $t \le k$, $k = \left(\frac{1}{\delta}\log(1+\varepsilon/(1-\varepsilon))\right)^2$, $\delta \le \varepsilon$.

5. Conclusions

We studied mean estimation under the realizable ε-realizable contamination model. Our results suggest several natural directions for future work. One direction is to extend this framework to other statistical tasks, such as linear regression and covariance estimation. A second direction is to investigate broader distributional assumptions. While the Gaussian setting serves as the canonical starting point, a natural next step is to consider subgaussian distributions. However, in full generality, consistency is known to be impossible in this model (Ma et al., 2024). This raises the question of identifying structured subclasses of subgaussian distributions, or alternative distribution families, for which consistent estimation remains achievable.

Acknowledgments

Ilias Diakonikolas is supported by NSF Medium Award CCF-2107079, ONR award number N00014-25-1-2268, and an H.I. Romnes Faculty Fellowship. Daniel M. Kane is supported by NSF Medium Award CCF-2107547. Thanasis Pittas is supported by NSF Medium Award CCF-2107079.

References

M. F. Adak, P. Lieberzeit, P. Jarujamrus, and N. Yumusak. Classification of alcohols obtained by QCM sensors with different characteristics using ABC based neural network. 23(3):463–469, 2020. ISSN 2215-0986. doi: 10.1016/j.jestch.2019.06.011.

G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. 1999.

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

P. M. Aronow and D. KK Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.

M-F Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680, 2008.

A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):939–956, 2017.

P. J. Bickel and J. Ritov. Large sample theory of estimation in biased sampling regression models. i. *The Annals of Statistics*, 19(2):797–816, 1991.

M. S. Brennan, G. Bresler, S. Hopkins, J. Li, and T. Schramm. Statistical Query Algorithms and Low Degree Tests Are Almost Equivalent. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 774–774. PMLR, 2021. URL https://proceedings.mlr.press/v134/brennan21a.html.

G. Carreras, G. Miccinesi, A. Wilcock, N. Preston, D. Nieboer, L. Deliens, M. Groenvold, U. Lunder, A. van der Heide, M. Baccini, et al. Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the action study. *BMC medical research methodology*, 21(1):13, 2021.

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, pages 47–60, 2017.

Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett. Optimal Robust Linear Regression in Nearly Linear Time, 2020.

C. Daskalakis, D. Rohatgi, and E. Zampetakis. Truncated Linear Regression in High Dimensions. In *Advances in Neural Information Processing Systems*, volume 33, pages 10338–10347. Curran Associates, Inc., a. URL https://proceedings.neurips.cc/paper/2020/hash/751f6b6b02bf39c41025f3bcfd9948ad-Abstract.html.

C. Daskalakis, P. Stefanou, R. Yao, and E. Zampetakis. Efficient Truncated Linear Regression with Unknown Noise Variance. In *Advances in Neural Information Processing Systems*, volume 34, pages 1952–1963. Curran Associates, Inc., b. URL https://proceedings.neurips.cc/paper/2021/hash/0ed8861dc36bee580d100f91283d0559-Abstract.html.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient Statistics, in High Dimensions, from Truncated Samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018. doi: 10.1109/focs.2018.00067.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Computationally and Statistically Efficient Truncated Regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 955–960. PMLR, 2019. URL https://proceedings.mlr.press/v99/daskalakis19a.html.

I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 1 edition, 2023. doi: 10.1017/9781108943161.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. 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, 2018a.

I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1047–1060, 2018b.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 2019a.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient Algorithms and Lower Bounds for Robust Linear Regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2745–2754. SIAM, 2019b. doi: 10.1137/1.9781611975482.170.

I. Diakonikolas, D. Kane, S. Karmalkar, A. Pensia, and T. Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. *Advances in Neural Information Processing Systems*, 35:13947–13960, 2022a.

I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, pages 5061–5117. PMLR, 2022b.

I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. Sq lower bounds for non-gaussian component analysis with weaker assumptions. *Advances in Neural Information Processing Systems*, 2023.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. Statistical Query Lower Bounds for Learning Truncated Gaussians. In *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 1336–1363. PMLR, 2024a. URL https://proceedings.mlr.press/v247/diakonikolas24b.html.

I. Diakonikolas, S. Karmalkar, S. Pang, and A. Potechin. Sum-of-squares lower bounds for non-gaussian component analysis. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 949–958. IEEE, 2024b.

I. Diakonikolas, J. Diakonikolas, D. M. Kane, J. C. H. Lee, and T. Pittas. Linear regression under missing or corrupted coordinates. *arXiv preprint arXiv:2509.19242*, 2025a.

I. Diakonikolas, D. M. Kane, S. Liu, and T. Pittas. Entangled mean estimation in high dimensions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1680–1688, 2025b.

I. Diakonikolas, D. M. Kane, S. Liu, and T. Pittas. Ptf testing lower bounds for non-gaussian component analysis. In *Proceedings of the 66th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2025c.

A. Elsener and S. van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019.

DM Farewell, RM Daniel, and SR Seaman. Missing at random: a stochastic process perspective. *Biometrika*, 109(1):227–241, 2022.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.

B. Follain, T. Wang, and R. J. Samworth. High-dimensional changepoint estimation with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84 (3):1023–1055, 2022.

F. Galton. An Examination into the Registered Speeds of American Trotting Horses, with Remarks on Their Value as Hereditary Data. 62:310–315, 1897. ISSN 0370-1662. URL https://www.jstor.org/stable/115734.

R. D. Gill, Y. Vardi, and J. A. Wellner. Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, pages 1069–1112, 1988.

R. D. Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

S. Karmalkar, A. Klivans, and P. Kothari. List-decodable linear regression. *Advances in neural information processing systems*, 32, 2019.

M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

A. R. Klivans, P. K. Kothari, and R. Meka. Efficient Algorithms for Outlier-Robust Regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018. URL http://proceedings.mlr.press/v75/klivans18a.html.

V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient Truncated Statistics with Unknown Truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019. doi: 10.1109/focs.2019.00093.

P. K. Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.

C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. 20(6):33:1–33:31, 2013. ISSN 1073-0516. doi: 10.1145/2505057.

D. Kunisky, A. S. Wein, and A. S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2022.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic Estimation of Mean and Covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016. doi: 10.1109/focs.2016.76.

R. J. Little, R. D'agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019a. ISBN 978-0-470-52679-8.

R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019b.

P.-L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under epsilon-contamination. 2018.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. 40(3):1637–1664. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS1018.

K. Lounici. High-dimensional covariance matrix estimation with missing observations. 2014.

T. Ma, K. A. Verchand, T. B. Berrett, T. Wang, and R. J. Samworth. Estimation beyond Missing (Completely) at Random, 2024.

P. Massart and E. Nedelec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

K. Pearson. On the Systematic Fitting of Curves to Observations and Measurements. 1(3):265–303, 1902. ISSN 0006-3444. doi: 10.2307/2331540.

K. Pearson and A. Lee. On the Generalised Probable Error in Multiple Normal Correlation. 6(1): 59–68, 1908. ISSN 0006-3444. doi: 10.2307/2331556.

A. Pensia, V. Jog, and P.-L. Loh. Robust Regression with Covariate Filtering: Heavy Tails and Adversarial Contamination. pages 1–12, 2024. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2024.2392906.

C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, pages 153–160. International Educational Data Mining Society. URL http://www.educationaldatamining.org/EDM2013/papers/rn_paper_23.pdf.

P. Raghavendra and M. Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.

J. M. Robins. Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data. 16(1):21–37, 1997. ISSN 1097-0258. doi: 10.1002/(SICI)1097-0258(19970115)16:1<21::AID-SIM470>3.0.CO;2-F.

S. Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge University Press, 1 edition, 2024. doi: 10.1017/9781009305129.

P. R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

A. Rotnitzky and J. Robins. Analysis of Semi-Parametric Regression Models with Non-Ignorable Non-Response. 16(1):81–102, 1997. ISSN 1097-0258. doi: 10.1002/(SICI)1097-0258(19970115)16:1<81::AID-SIM473>3.0.CO;2-0.

D. B. Rubin. Inference and Missing Data. 63(3):581–592, 1976. ISSN 0006-3444. doi: 10.2307/2335739.

R. Sahoo, L. Lei, and S. Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448): 1096–1120, 1999.

S. Seaman, J. Galati, D. Jackson, and J. Carlin. What is meant by "missing at random"? 2013.

T. Sell, T. B. Berrett, and T. I. Cannings. Nonparametric classification with missing data, 2024.

I. Shpitser, K. Mohan, and J. Pearl. Missing data as a causal and probabilistic problem. Technical report, 2015.

G. Szegö. *Orthogonal Polynomials*. Number τ. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967. ISBN 9780821889527. URL https://books.google.com/books?id=3hcW8HBh7gsC.

T Tony Cai and Linjun Zhang. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):675–705, 2019.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York. ISBN 978-0-387-32448-7. doi: 10.1007/0-387-37345-4.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

Y. Vardi. Empirical distributions in selection bias models. *The Annals of Statistics*, pages 178–203, 1985.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

J. Vuurens, A. P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, 2011.

Y. Xie, J. Huang, and R. Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2012.

Y. Yan, Y. Chen, and J. Fan. Inference for heteroskedastic pca with missing data. *The Annals of Statistics*, 52(2):729–756, 2024.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.

Z. Zhu, T. Wang, and R. J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022.

Appendix A. Omitted Details from Section 1

A.1. Additional Discussion on Related Work

Comparison of Definition 1 with the definition in Ma et al. (2024) As described in Section 1, the contamination model (Definition 1) studied in this paper was proposed in the recent work of Ma et al. (2024) as an attempt to formalize missingness mechanisms that are non-MCAR yet milder than MAR or MNAR. At first glance, the model appears to be defined slightly differently in eq. (6) of Ma et al. (2024). The ε-corrupted version $\widetilde{P}$ of the original distribution P there is defined as any mixture of the form (as a note regarding notation in that paper, we are using $q = 1$ in the notation of that paper, i.e., there is no MCAR component in eq. (6))

$$(1-\varepsilon)P + \varepsilon Q, \tag{6}$$

where Q is any MNAR version of P that the adversary can choose.

However, Ma et al. (2024, Proposition 2) provides a characterization that establishes its equivalence with the definition used in this paper. As explained following Proposition 2, if samples are interpreted as being generated by first drawing a value and then applying a missingness mechanism, and if $h(x)$ denotes the probability that a sample is missing conditional on the original value being x, then Proposition 2 shows that the realizable ε-contamination model is equivalent to the condition $1-\varepsilon \le h(x) \le 1$. In the language of our paper, this characterization leads to the alternative definition of the realizable ε-contamination model stated in Definition 16 below.

Definition 16 (Contamination model; Alternate definition) *Let P be a distribution on a domain $\mathcal{X}$ with pdf $p : \mathcal{X} \to \mathbb{R}_+$. An ε-corrupted version $\widetilde{P}$ of P is any distribution that can be obtained as follows: First, an adversary chooses a function $f : \mathcal{X} \to \mathbb{R}_+$ with $(1-\varepsilon)p(x) \le f(x) \le p(x)$. $\widetilde{P}$ is then defined to be the distribution whose samples are generated as follows:*

1. *Draw x from P.*
2. *With probability $1 - f(x)/p(x)$, replace x by a special symbol $\perp$.*

As shown below, it is then straightforward to check that Definition 16 is equivalent to Definition 1.

Claim 17 *Definition 1 and Definition 16 are equivalent.*

Proof Let us denote by X the initial value of the sample before the missingness pattern is applied and by Z the sample after its application. In Definition 16 we have

$$\mathbb{P}[Z \neq \perp] = \int_{\mathcal{X}} \mathbb{P}[Z \neq \perp \,|X = x]\, \mathbb{P}[X = x]\mathrm{d}x = \int_{\mathcal{X}} \frac{f(x)}{p(x)} p(x)\mathrm{d}x = \int_{\mathcal{X}} f(x)\mathrm{d}x,$$

and we also have

$$\mathbb{P}[X = x | Z \neq \perp] = \frac{\mathbb{P}[Z \neq \perp \,|X = x]\, \mathbb{P}[X = x]}{\mathbb{P}[Z \neq \perp]} = \frac{(1 - \mathbb{P}[Z = \perp \,|X = x])\, p(x)}{\int_{\mathcal{X}} f(x)\mathrm{d}x}$$
$$= \frac{\frac{f(x)}{p(x)} p(x)}{\int_{\mathcal{X}} f(x)\mathrm{d}x} = \frac{f(x)}{\int_{\mathcal{X}} f(x)\mathrm{d}x},$$

which agrees with Definition 1.

■

Robust Statistics The field of robust statistics was initiated in the 1960s through the seminal works of Huber and Tukey (Tukey, 1960; Huber, 1964), with the goal of developing estimators that are robust to data contamination. In this setting, data corruptions are formalized by allowing a small fraction $\varepsilon < 1/2$ of the samples to come from an arbitrary distribution. While early work in the 1960s studied the information-theoretic aspects of one-dimensional inference in this model (including optimal error rates and sample complexity), extensions of these algorithms to higher dimensions required exponential time. It was not until 2016 that the first polynomial-time algorithms were obtained (Lai et al., 2016; Diakonikolas et al., 2019a). This led to a revitalization of the field, with improved algorithms for a variety of problems such as mean estimation (Kothari and Steurer, 2017; Diakonikolas et al., 2018a, 2022b) and linear regression (Klivans et al., 2018; Diakonikolas et al., 2019b; Pensia et al., 2024; Cherapanamjeri et al., 2020).

Apart from the fact that robust statistics considers data corruptions rather than missingness, the differences with the current work are as follows:

(i) Due to the arbitrary nature of outliers, identifiability is only possible when the corruption fraction satisfies $\varepsilon < 1/2$. If a majority of the samples are corrupted, the dataset may consist of two equally sized subsets corresponding to different underlying distributions, in which case it is impossible to determine which subset contains the inliers. The robust statistics literature has nevertheless considered the regime $\varepsilon > 1/2$. Since identifiability is impossible in this case, the goal shifts to outputting a list of candidate solutions with the guarantee that at least one is close to the ground truth. Algorithms of this type are known as *list-decoding* algorithms (Balcan et al., 2008; Charikar et al., 2017), and such algorithms have been developed for several tasks, including mean estimation (Charikar et al., 2017; Diakonikolas et al., 2018b, 2022a) and linear regression (Karmalkar et al., 2019; Raghavendra and Yau, 2020).

(ii) Even in the regime $\varepsilon < 1/2$, where identifiability is possible, consistency—i.e., the property that the estimation error vanishes as the number of samples tends to infinity—is still unattainable in robust statistics. For example, for Gaussian mean estimation with an ε fraction of arbitrary corruptions, the optimal error is $\Theta(\varepsilon)$, regardless of the sample size (and analogous lower bounds hold for other distributions, with different dependencies on ε). In contrast, Definition 1 allows for consistency due to the additional structure imposed on the missingness pattern.

Truncated Statistics Although it is a form of Missing Not At Random, this large subfield of statistics—tracing back to Galton, Pearson, and Lee (Galton, 1897; Pearson, 1902; Pearson and Lee, 1908)—developed largely orthogonally to the rest of the missing data literature. Truncated statistics concerns scenarios in which there is a truncation set (which may be known or unknown to the learning algorithm), and only samples that fall within this set are observed. Despite early work on this problem, efficient algorithms for fundamental tasks were obtained only in the last decade, including Gaussian mean estimation (Daskalakis et al., 2018; Kontonis et al., 2019) and linear regression (Daskalakis et al., 2019, a,b).

More concretely, the notion of missingness used in truncated statistics is defined as follows (we state it for identity-covariance Gaussians to match the inlier distribution considered in this paper). Samples are drawn from $\mathcal{N}(\mu, I)$ but are revealed only if they fall in some subset $S \subseteq \mathbb{R}^d$ whose probability mass is assumed to be lower bounded by a parameter $\alpha > 0$; otherwise, the samples are hidden (e.g., represented by a special symbol $\perp$). More generally, the literature also considers a setting in which hidden samples are completely unobserved, so the algorithm does not know the ratio of missing to visible samples.

At a high level, this can be viewed as a special case of the ε-realizable contamination model considered in this paper (Definition 1) with $\varepsilon = 1$, but there are important differences. First, there is no analogue in Definition 1 of the requirement that the truncation set have probability mass bounded away from zero. As a result, the adversary in Definition 1 could make all samples missing, which is the main reason the problem is unsolvable when $\varepsilon = 1$ (the sample complexity in Equation (1) diverges as $\varepsilon \to 1$).

Second, even when such a lower-bound condition is imposed, additional nuances arise in the truncated statistics literature depending on whether the truncation set S is known to the algorithm. If the set is unknown, mean estimation remains information-theoretically impossible to arbitrary accuracy (Daskalakis et al., 2018). Estimation becomes possible only if (i) the algorithm has oracle access to S (Daskalakis et al., 2018), or (ii) the set has bounded complexity, for example bounded VC dimension or bounded Gaussian surface area (Kontonis et al., 2019). In the latter case, there is an information–computation trade-off, providing evidence that polynomial-time algorithms often require more samples than the information-theoretic optimum (Diakonikolas et al., 2024a).

Appendix B. Omitted Details from Section 2

B.1. Useful Facts

Fact 18 (Gaussian tail bound) *Let $Z \sim N(0,1)$. Then for all $x > 0$,*

$$\mathbb{P}[Z \geq x] \leq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}.$$

Fact 19 (Mills ratio inequality Gordon (1941)) *Let ϕ, Φ denote the pdf and cdf of $\mathcal{N}(0,1)$ respectively. The following holds for all $x > 0$:*

$$x < \frac{\phi(x)}{1-\Phi(x)} < x + \frac{1}{x}.$$

Fact 20 (Dvoretzky–Kiefer–Wolfowitz (DKW) inequality) *Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with cumulative distribution function F, and let the empirical cumulative distribution function be $F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right| > \varepsilon\right] \leq 2\, e^{-2n\varepsilon^2}.$$

Fact 21 (Maximal coupling (see, e.g., Roch (2024))) *Let P and Q be distributions on some domain $\mathcal{X}$. It holds that $D_{\mathrm{TV}}(P,Q) = \inf_{\Pi} \mathbb{P}_{(X,Y)\sim\Pi}[X \neq Y]$ where the infimum is over all possible couplings between P and Q. Moreover, there exists Π such that $\mathbb{P}_{(X,Y)\sim\Pi}[X \neq Y] = D_{\mathrm{TV}}(P,Q)$.*

Fact 22 (see, e.g., Corollary 4.2.13 in Vershynin (2018)) *Let $\xi > 0$. There exists a set $\mathcal{C}$ of unit vectors of $\mathbb{R}^d$ such that $|\mathcal{C}| < (1 + 2/\xi)^d$ and for every $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$ it holds $\min_{y\in\mathcal{C}} \|y - u\|_2 \leq \xi$.*

Corollary 23 (see, e.g., Exercise 4.4.3 (b) in Vershynin (2018)) *There exists a subset $\mathcal{C}$ of the d-dimensional unit ball with $|\mathcal{C}| \leq 7^d$ such that $\|x\|_2 \leq 2\max_{v\in\mathcal{C}} |v^\top x|$ for all $x \in \mathbb{R}^d$ and $\|A\|_{\mathrm{op}} \leq 3\max_{x\in\mathcal{C}} x^\top A x$ for every symmetric $A \in \mathbb{R}^{d\times d}$.*

Fact 24 *Let $x \sim \mathcal{N}(0, I_d)$ and let $j \geq 1$. Then $\mathbb{E}[\|x\|^j] \leq (2\sqrt{jd})^j$.*

Proof For a standard normal $z \sim N(0,1)$ it is well known that $\mathbb{E}\,|z|^j \leq (2\sqrt{j})^j$. Since $\|x\|^j = \left(\sum_{i=1}^d x_i^2\right)^{j/2}$ and for $r \geq 1$ we have $(\sum_{i=1}^d a_i)^r \leq d^{r-1}\sum_{i=1}^d a_i^r$, applying this with $a_i = x_i^2$ and $r = j/2$ gives

$$\|x\|^j \leq d^{j/2-1}\sum_{i=1}^d |x_i|^j.$$

Taking expectations and using identical marginals,

$$\mathbb{E}\,\|x\|^j \leq d^{j/2}\,\mathbb{E}\,|z|^j \leq d^{j/2}\,(2\sqrt{j})^j = (2\sqrt{jd})^j.$$

■

Proposition 25 (Le Cam's lemma) *For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_{\Psi}\left\{\mathop{\mathbb{P}}_{X\sim P_1}(\Psi(X)\neq 1) + \mathop{\mathbb{P}}_{X\sim P_2}(\Psi(X)\neq 2)\right\} = 1 - D_{\mathrm{TV}}(P_1, P_2),$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Legendre Polynomials In this work, we make use of the Legendre Polynomials which are orthogonal polynomials over $[-1, 1]$. Some of their properties are:

Fact 26 (Szegö (1967)) *The Legendre polynomials P_k for $k \in \mathbb{Z}$, satisfy the following properties:*

1. *P_k is a k-degree polynomial and $P_0(x) = 1$ and $P_1(x) = x$.*
2. *$\int_{-1}^1 P_i(x)P_j(x)\mathrm{d}x = 2/(2i+1)\mathbb{1}\{i = j\}$, for all $i, j \in \mathbb{Z}$.*
3. *$|P_k(x)| \leq 1$ for all $|x| \leq 1$.*
4. *$|P_k(x)| \leq (4|x|)^k$ for $|x| \geq 1$.*
5. *$P_k(x) = (-1)^k P_k(-x)$.*
6. *$P_k(x) = 2^{-k}\sum_{i=1}^{\lceil k/2\rceil}\binom{k}{i}\binom{2k-2i}{k}x^{k-2i}$.*

B.2. Hermite Analysis

Definition 27 (Hermite tensor) *For $k \in \mathbb{N}$ and $x \in \mathbb{R}^n$, we define the k-th Hermite tensor as*

$$(H_k(x))_{i_1,i_2,\ldots,i_k} = \frac{1}{\sqrt{k!}}\sum_{\substack{\text{Partitions } P \text{ of } [k]\\ \text{into sets of size 1 and 2}}}\bigotimes_{\{a,b\}\in P}(-I_{i_a,i_b})\bigotimes_{\{c\}\in P}x_{i_c}\,.$$

Fact 28 *If $v \in \mathbb{R}^d$ is a unit vector it holds $H_k(v^\top x) = \langle v^{\otimes k}, H_k(x)\rangle$.*

Fact 29 $\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[H_k(x)] = \mu^{\otimes k}/\sqrt{k!}$.

Hermite polynomials form a complete orthogonal basis of the vector space $L_2(\mathbb{R}, \mathcal{N}(0,1))$ of all functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[f^2(x)] < \infty$. There are two commonly used types of Hermite polynomials. The *physicist's* Hermite polynomials, denoted by H_k for $k \in \mathbb{Z}$ satisfy the following orthogonality property with respect to the weight function e^{-x^2}: for all $k, m \in \mathbb{Z}$, $\int_{\mathbb{R}} H_k(x)H_m(x)e^{-x^2}\mathrm{d}x = \sqrt{\pi}2^k k!\mathbf{1}(k = m)$. The *probabilist's* Hermite polynomials H_{e_k} for $k \in \mathbb{Z}$ satisfy $\int_{\mathbb{R}} H_{e_k}(x)H_{e_m}(x)e^{-x^2/2}\mathrm{d}x = k!\sqrt{2\pi}\mathbf{1}(k = m)$ and are related to the physicist's polynomials through $H_{e_k}(x) = 2^{-k/2}H_k(x/\sqrt{2})$. We will mostly use the *normalized probabilist's* Hermite polynomials $h_k(x) = H_{e_k}(x)/\sqrt{k!}$, $k \in \mathbb{Z}$ for which $\int_{\mathbb{R}} h_k(x)h_m(x)e^{-x^2/2}\mathrm{d}x = \sqrt{2\pi}\mathbf{1}(k = m)$. These polynomials are the ones obtained by Gram-Schmidt orthonormalization of the basis $\{1, x, x^2, \ldots\}$ with respect to the inner product $\langle f, g\rangle_{\mathcal{N}(0,1)} = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[f(x)g(x)]$. Every function $f \in L_2(\mathbb{R}, \mathcal{N}(0,1))$ can be uniquely written as $f(x) = \sum_{i\in\mathbb{Z}} a_i h_i(x)$ and we have $\lim_{n\to\infty} \mathbb{E}_{x\sim\mathcal{N}(0,1)}[(f(x) - \sum_{i=0}^n a_i h_i(x))^2] = 0$ (see, e.g., Andrews et al. (1999)). Extending the normalized probabilist's Hermite polynomials to higher dimensions, an orthonormal basis of $L_2(\mathbb{R}^d, \mathcal{N}(0, I_d))$ (with respect to the inner product $\langle f, g\rangle = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f(x)g(x)]$) can be formed by all the products of one-dimensional Hermite polynomials, i.e., $h_a(x) = \prod_{i=1}^d h_{a_i}(x_i)$, for all multi-indices $a \in \mathbb{Z}^d$ (we are now slightly overloading notation by using multi-indices as subscripts). The total degree of h_a is $|a| = \sum_{i=1}^d a_i$.

Claim 30 (Univariate Gaussian shift bound) *Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree at most d satisfying $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[p(x)^2] = 1$. Then for every $\mu \in \mathbb{R}$, $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[p(x+\mu)^2\right] \le e^{d\mu^2}$.*

Proof Let $h_k(x) = He_k(x)/\sqrt{k!}$ denote the normalized probabilists' Hermite polynomials. Expand p in the orthonormal Hermite basis:

$$p(x) = \sum_{k=0}^{d} c_k\, h_k(x), \qquad \sum_{k=0}^{d} c_k^2 = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[p(x)^2] = 1.$$

The following identity holds for the normalized probabilists' Hermite polynomials:

$$h_k(x+\mu) = \sum_{r=0}^{k} \binom{k}{r} \mu^{k-r} \sqrt{\frac{r!}{k!}}\, h_r(x).$$

Using orthonormality of $\{h_r\}$ and independence of the coefficients in the expansion, we have that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_k(x+\mu)^2] = \sum_{r=0}^{k} \binom{k}{r}^2 \mu^{2(k-r)}\, \frac{r!}{k!} = \sum_{s=0}^{k} \binom{k}{s} \frac{\mu^{2s}}{s!},$$

where $s = k - r$. Bounding $\binom{k}{s} \le k^s/s!$ yields

$$\binom{k}{s} \frac{\mu^{2s}}{s!} \le \frac{(k\mu^2)^s}{s!}.$$

Therefore

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_k(x+\mu)^2] \le \sum_{s=0}^{\infty} \frac{(k\mu^2)^s}{s!} = e^{k\mu^2}.$$

Now expand $p(x+\mu)$:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[p(x+\mu)^2] = \sum_{k=0}^{d} c_k^2 \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_k(x+\mu)^2] \le \sum_{k=0}^{d} c_k^2 e^{k\mu^2} \le e^{d\mu^2}\sum_{k=0}^{d} c_k^2 = e^{d\mu^2}.$$

■

Claim 31 (Multivariate Gaussian shift bound) *Let $p:\mathbb{R}^n\to\mathbb{R}$ be a polynomial of total degree at most D satisfying $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[p(x)^2]=1$. For a multi-index $\alpha\in\mathbb{N}^n$, define $h_\alpha(x)=\prod_{i=1}^n h_{\alpha_i}(x_i)$ and $|\alpha|=\sum_{i=1}^n\alpha_i$. Then for every $\mu\in\mathbb{R}^n$, it holds $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[p(x+\mu)^2]\le e^{D\|\mu\|_2^2}$.*

Proof Expand p in the multivariate orthonormal Hermite basis:

$$p(x)=\sum_{|\alpha|\le D} c_\alpha\, h_\alpha(x), \qquad \sum_{|\alpha|\le D} c_\alpha^2 = \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[p(x)^2]=1.$$

Because $h_\alpha(x)=\prod_i h_{\alpha_i}(x_i)$, and the coordinates of x are independent, we have

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_\alpha(x+\mu)^2] = \prod_{i=1}^n \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_{\alpha_i}(x_i+\mu_i)^2].$$

Applying the univariate bound of Claim 30 to each coordinate,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_{\alpha_i}(x_i+\mu_i)^2] \le e^{\alpha_i\mu_i^2},$$

so

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[h_\alpha(x+\mu)^2] \le \exp\left(\sum_{i=1}^n \alpha_i\mu_i^2\right) \le \exp\big(|\alpha|\,\|\mu\|_2^2\big) \le e^{D\|\mu\|_2^2}.$$

Finally,

$$\mathbb{E}[p(X+\mu)^2] = \sum_{|\alpha|\le D} c_\alpha^2\,\mathbb{E}[h_\alpha(X+\mu)^2] \le \sum_{\alpha} c_\alpha^2\, e^{D\|\mu\|_2^2} = e^{D\|\mu\|_2^2}.$$

■

Claim 32 *Let H_k denote the k-th Hermite tensor for d dimensions. Then, the following bound holds: $\|H_k(x)\|_2 \le d^{k/2}(1+\|x\|^k)2^{O(k)}$.*

Proof

For a degree-k tensor A, we use A^π to denote the matrix that $A^\pi_{i_1,\dots,i_k} = A_{\pi(i_1,\dots,i_k)}$. Note that $\|A\|_2=\|A^\pi\|_2$. Then from the definition of Hermite tensor, we have that

$$H_k(x) = \frac{1}{\sqrt{k!}}\sum_{t=1}^{\lfloor k/2\rfloor}\sum_{\text{Permutation }\pi\text{ of }[k]} \frac{1}{2^t t!(k-2t)!}\left(I^{\otimes t}x^{\otimes(k-2t)}\right)^\pi.$$

Thus the norm is

$$\begin{aligned}
\|H_k(x)\|_2 &= \left\| \frac{1}{\sqrt{k!}} \sum_{t=1}^{\lfloor k/2 \rfloor} \sum_{\text{Permutation } \pi \text{ of } [k]} \frac{1}{2^t t!(k-2t)!} \left(I^{\otimes t} x^{\otimes (k-2t)} \right)^{\pi} \right\|_2 \\
&\leq \sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\sqrt{k!}}{2^t t!(k-2t)!} \max\left(\|I^{\otimes t}\|_2 \|x\|_2^{k-2t}, 1 \right) \\
&\leq \sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\sqrt{k!}}{2^t t!(k-2t)!} \max(d^{t/2} \|x\|_2^{k-2t}, 1) \\
&\leq \sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\sqrt{k!}}{2^t t!(k-2t)!} \left(d^{t/2} \|x\|_2^{k-2t} + 1 \right) \\
&\leq \sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\sqrt{k!}}{2^t t!(k-2t)!} \left(d^{t/2} \max(\|x\|_2^k, 1) + 1 \right) \\
&\leq 2d^{k/2}(1 + \|x\|^k) \sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\sqrt{k!}}{2^t t!(k-2t)!}.
\end{aligned}$$

One can see that the denominator is minimized when $t = k/2 - O(\sqrt{k})$. Using that, we have that the right hand side above is at most $d^{k/2}(1 + \|x\|^k)2^{O(k)}$.

■

We restate and prove the following concentration of Hermite moments.

Lemma 13 (Hermite tensor concentration) *Let $\eta, \varepsilon \in (0,1)$ be parameters, C be a sufficiently large absolute constant and $\mu \in \mathbb{R}^d$ be a vector with $\|\mu\|_2 = O(\sqrt{\log(\frac{1}{1-\varepsilon})})$. Let $\widetilde{P}$ be an ε-corrupted version of $\mathcal{N}(\mu, I)$ (cf. Definition 1) and let P' denote the conditional distribution of $\widetilde{P}$ on the non-missing samples. Let $x_1, \ldots, x_n \sim P'$ be i.i.d. samples and define $\widehat{T} := \frac{1}{n}\sum_{i=1}^n H_k(x_i)$, and $T := \mathbb{E}_{x \sim P'}[H_k(x)]$, where $H_k(x)$ denotes the Hermite tensor from Definition 27. If $n > C\frac{d^{3k}2^{O(k)}(k\log(\frac{1}{1-\varepsilon}))^{k/2}}{(1-\varepsilon)\eta^2\tau}$, then with probability at least $1-\tau$ we have that $\left\|\widehat{T} - T\right\|_2 \leq \eta$.*

Proof Consider one entry $\widehat{T}_{i_1 i_2 \cdots i_k}$ of the estimator. It holds

$$\begin{aligned}
\mathrm{Var}(\widehat{T}_{i_1 i_2 \cdots i_k}) &= \frac{1}{n} \operatorname*{Var}_{x \sim P'}(H_k(x)_{i_1 i_2 \cdots i_k}) \\
&\leq \frac{1}{n} \mathop{\mathbb{E}}_{x \sim P'}[(H_k(x)_{i_1 i_2 \cdots i_k})^2] \\
&\leq \frac{1}{n(1-\varepsilon)} \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\mu, I)} \left[(H_k(x)_{i_1 i_2 \cdots i_k})^2 \right] && \text{(by Definition 1)} \\
&\leq \frac{1}{n(1-\varepsilon)} \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\mu, I)} \left[\|H_k(x)\|_2^2 \right] \\
&\leq \frac{d^{k/2}2^{O(k)}}{n(1-\varepsilon)} \left(1 + \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\mu, I)} \left[\|x\|_2^k \right] \right) && \text{(by Claim 32)}
\end{aligned}$$

$$\leq \frac{d^k 2^{O(k)} k^{k/2}}{n(1-\varepsilon)}, \tag{7}$$

where the last step can be shown as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{N}(\mu,I)}\left[\|x\|_2^k\right] &= \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,I)}\left[\|z+\mu\|_2^k\right] \\
&\leq \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,I)}\left[(\|z\|_2+\|\mu\|_2)^k\right] \\
&\leq 2^{k-1}\left(\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,I)}\left[\|z\|_2^k\right]+\|\mu\|_2^k\right) \\
&\leq 2^{O(k)}\left((kd)^{k/2}+2^{O(k)}\log(1/(1-\varepsilon))^{k/2}\right) \qquad \text{(using Fact 24)} \\
&\leq 2^{O(k)}k^{k/2}d^{k/2}\log(1/(1-\varepsilon))^{k/2}.
\end{aligned}$$

Having the variance bound of inequality (7), an application of Chebyshev's inequality yields that if the number of samples is $n > C\frac{d^{2k}2^{O(k)}k^{k/2}\log(1/(1-\varepsilon))^{k/2}}{(1-\varepsilon)\eta^2\tau'}$ then

$$\mathbb{P}\left[|\widehat{T}_{i_1i_2\cdots i_k} - T_{i_1i_2\cdots i_k}| > \frac{\eta}{d^{k/2}}\right] \leq \frac{d^{2k}2^{O(k)}k^{k/2}\log(1/(1-\varepsilon))^{k/2}}{n(1-\varepsilon)\eta^2} \leq \tau'.$$

We will use $\tau' = \tau d^{-k}$. By union bound the probability of having $|\widehat{T}_{i_1i_2\cdots i_k} - T_{i_1i_2\cdots i_k}| \leq \frac{\eta}{d^{k/2}}$ for all entries simultaneously is at least $1-\tau$. In that event we have that $\|\widehat{T}-T\|_2 = \sqrt{\sum_{i_1\cdots i_k}|\widehat{T}_{i_1i_2\cdots i_k} - T_{i_1i_2\cdots i_k}|^2} \leq \eta$ which completes the proof. ■

Appendix C. Omitted Details from Section 3

We restate and prove the following two lemmata.

Lemma 10 *Denote by $\phi(x)$ the pdf of $\mathcal{N}(0,1)$. For any $\varepsilon,\delta \in (0,1)$ there exists a function $g:\mathbb{R}\to\mathbb{R}_+$ such that $(1-\varepsilon)\phi(x-\delta) \leq g(x) \leq \phi(x-\delta)$, $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1-\varepsilon/2$ and $\frac{g(x)}{1-\varepsilon/2} = \phi(x)$ for all $x \in [-B+\delta/2, B+\delta/2]$ where $B := \frac{1}{\delta}\log\left(1+\frac{\varepsilon/2}{1-\varepsilon/2}\right)$.*

Proof For convenience we will prove the claim with everything shifted by $\delta/2$ to the left, i.e., we will show that,

$$(1-\varepsilon)\phi(x-\delta/2) \leq g(x) \leq \phi(x-\delta/2) \tag{8}$$

as well as $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1-\varepsilon/2$ and $\frac{g(x)}{1-\varepsilon/2} = \phi(x+\delta/2)$ for all $x\in[-B,B]$. For simplicity of notation, we will use $p_+(x) := \phi(x-\delta/2)$ and $p_-(x) := \phi(x+\delta/2)$ to denote the two Gaussian densities for the rest of the proof.

The main idea is to let $g(x) = (1-\varepsilon/2)p_-(x)$ for all x in an interval around zero which is as large as possible without violating the condition (8). Once we find which is the biggest possible such interval, we will need to correct $g(x)$ outside of it so that it still respects condition (8).

For the first part of our proof argument (finding the largest interval for which setting $g(x) = (1-\varepsilon/2)p_-(x)$ inside it satisfies condition (8)) we solve the equations $(1-\varepsilon/2)p_-(x) = p_+(x)$ and $(1-\varepsilon/2)p_-(x) = (1-\varepsilon)p_+(x)$. The two solutions are

$$x_+ = \frac{1}{\delta}\log\left(\frac{1-\varepsilon/2}{1-\varepsilon}\right) \quad \text{and} \quad x_- = \frac{1}{\delta}\log\left(1-\frac{\varepsilon}{2}\right).$$

This means that if we define $B := \frac{1}{\delta}\log\left(1+\frac{\varepsilon/2}{1-\varepsilon/2}\right)$ we have that the function defined as $g(x) := p_-(x)(1-\varepsilon/2)$ satisfies condition (8) for all $x \in [-B, B]$.

We now need to show how to define $g(x)$ outside of $[-B, B]$. We show that it is possible to extend the definition outside of $[-B, B]$ in a way that condition (8) continues to hold and $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1-\varepsilon/2$. A first, unsuccessful approach would be to set $g(x) = (1-\varepsilon)p_+(x)$ for $x > B$ and $g(x) = p_+(x)$ for $x < -B$. Although this ensures condition (8), the other desideratum $\int_{\mathbb{R}} g(x)\mathrm{d}x = 1-\varepsilon/2$ is not satisfied. To see this, let us define A_1 and A_2 be the following areas:

$$A_1 := \int_B^{+\infty} ((1-\varepsilon)p_+(x) - (1-\varepsilon/2)p_-(x))\,\mathrm{d}x,$$
$$A_2 := \int_{-\infty}^{-B} ((1-\varepsilon/2)p_-(x) - p_+(x))\,\mathrm{d}x.$$

Then the integral of g is

$$\begin{aligned}\int_{x\in\mathbb{R}} g(x)\mathrm{d}x &= \int_{-\infty}^{+\infty}(1-\varepsilon/2)p_-(x)\mathrm{d}x + \int_{x_+}^{B}((1-\varepsilon/2)p_(x) - (1-\varepsilon)p_+(x)))\mathrm{d}x + A_1 - A_2 \\ &< 1 - \frac{\varepsilon}{2} + A_1 - A_2,\end{aligned} \tag{9}$$

where the first integral is simply $1-\varepsilon/2$ and the integral from x_+ to B is negative (by definition of x_+). We can finally check that $A_2 > A_1$ to conclude the proof of $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x < 1-\varepsilon/2$. To see this, first note that we can rewrite $A_2 = \int_{-\infty}^{-B} p_+(-x) - (1-\varepsilon/2)p_-(-x)\mathrm{d}x = \int_B^{\infty}(1-\varepsilon/2)p_+(x) - p_-(x)\mathrm{d}x$ where we used that $p_+(-x) = p_-(x)$ and a change of variable. Then,

$$A_2 - A_1 = \int_B^{\infty} \frac{\varepsilon}{2}\left(p_+(x) - p_-(x)\right)\mathrm{d}x > 0\,.$$

However, the above choice of $g(x)$ for $x > B$ is not the only one allowed by condition (8). We could alternatively choose any $g(x) \in [(1-\varepsilon)p_+(x), p_+(x)]$ for $x > B$. We just saw that the first extreme choice $g(x) = (1-\varepsilon)p_+(x)$ results in $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x < 1-\varepsilon/2$. We will now show that the other extreme choice of setting $g(x) = p_-(x)$ in $x > B$ results in $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x > 1-\varepsilon/2$. By continuity this would mean that there exists a way of defining $g(x)$ in $x > B$ that achieves $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = 1-\varepsilon/2$.

We now show the remaining claim above, that the choice $g(x) = p_+(x)$ for all $x > B$ (and $g(x) = p_+(x)$ for $x < -B$, $g(x) = (1-\varepsilon/2)p_+(x)$ for $x \in [-B, B]$ as before) results in $\int_{x\in\mathbb{R}} g(x)\mathrm{d}x > 1-\varepsilon/2$. In this case, similarly to inequality (9), we have

$$\int_{x\in\mathbb{R}} g(x)\mathrm{d}x = \int_{-\infty}^{+\infty}(1-\varepsilon/2)p_-(x)\mathrm{d}x + \widetilde{A}_1 - A_2$$

where A_2 is the same as before, but $\widetilde{A}_1 = \int_B^\infty (p_+(x)\mathrm{d}x - (1-\varepsilon/2)p_-(x))\,\mathrm{d}x$. Now, this gives

$$\widetilde{A}_1 - A_2 = \int_B^\infty \frac{\varepsilon}{2}\,(p_+(x) + p_-(x))\,\mathrm{d}x > 0.$$

■

Lemma 11 *Let $\varepsilon \in (0,1), \delta \ll \varepsilon$ and $g(x)$ be as in Lemma 10. There exists a polynomial $p(x)$ such that the function $f(x) := g(x) + p(x)\mathbb{1}(|x| \le 1)$ satisfies $(1-\varepsilon)\phi(x-\delta) \le f(x) \le \phi(x-\delta)$ and the distribution with pdf $f(x)/\int_{\mathbb{R}} f(x)\mathrm{d}x$ matches the first m moments with $\mathcal{N}(0,1)$ for some $m = \Omega(\gamma^2/\log\gamma)$, where $\gamma := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$.*

Proof The function $g(x)$ from Lemma 10 satisfies the following:

1. $(1-\varepsilon)\phi(x-\delta) \le g(x) \le \phi(x-\delta)$ for all $x \in \mathbb{R}$
2. $\int_{\mathbb{R}} g(x)\mathrm{d}x = 1 - \varepsilon/2$
3. $g(x) = \phi(x)(1-\varepsilon/2)$ for $x \in [-B+\delta/2, B+\delta/2]$ where $B := \frac{1}{\delta}\log\left(1 + \frac{\varepsilon/2}{1-\varepsilon/2}\right)$.

We will show the existence of a polynomial p such that

1. $|p(x)| \le c\varepsilon$, where c is a sufficiently small absolute constant,
2. $\int_{-1}^1 p(x)\mathrm{d}x = 0$,
3. $\int_{\mathbb{R}} x^i \frac{g(x)+p(x)\mathbb{1}(|x|\le 1)}{1-\varepsilon/2}\mathrm{d}x = \int_{\mathbb{R}} x^i\phi(x)\mathrm{d}x$ for all $i \in [m]$.

Before showing that such a polynomial exists, we first show how Lemma 11 follows given the above points. First, the moment matching property in the conclusion of Lemma 11 directly follows by the third item in the above list. We now show how the part that $(1-\varepsilon)\phi(x-\delta) \le f(x) \le \phi(x-\delta)$ for all $x \in \mathbb{R}$ follows from the above. This can be seen by verifying that (i) $\phi(x-\delta) - g(x) = \Omega(\varepsilon)$ for all $x \in [-1,1]$ and (ii) $g(x) - (1-\varepsilon)\phi(x-\delta) = \Omega(\varepsilon)$ for all $x \in [-1,1]$. We show the part (i) since the other part can be seen with identical arguments. The smallest value of $\phi(x-\delta) - g(x)$ happens at $x = -1$. On that point:

$$\phi(x-\delta) - (1-\varepsilon/2)\phi(x) = \phi(x-\delta) - \phi(x) + \frac{\varepsilon}{2}\phi(x) \ge -O(\delta) + \Omega(\varepsilon) \ge \Omega(\varepsilon),$$

where the first inequality uses the fact that $\phi(x-\delta) - \phi(x) = x\phi(x)\delta + \frac{\xi^2-1}{2}\phi(\xi)\delta^2$ for some $x - \delta \le \xi \le x$, by Taylor's theorem), and we also used that $\phi(x) = \Omega(1)$ for $x \in [-1,1]$. The last inequality above used that $\delta \ll \varepsilon$.

We now turn to showing the existence of the polynomial p. This part of the proof follows an argument similar to the one in Diakonikolas et al. (2017). Recall the moment matching condition that we want to ensure: $\int_{-1}^1 x^i p(x)\mathrm{d}x = \int_{-\infty}^\infty x^i\phi(x)\mathrm{d}x - \int_{-\infty}^\infty x^i\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x$. Using the fact that $g(x)/(1-\varepsilon/2) = \phi(x)$ in the interval $[-B+\delta, B+\delta]$ the moment matching condition becomes:

$$\int_{-1}^1 x^i p(x)\mathrm{d}x = \int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} x^i\phi(x)\mathrm{d}x - \int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} x^i\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x, \tag{10}$$

for $i = 1,\ldots,m$. First, we express $p(x)$ as a linear combination of Legendre polynomials P_k:

Fact 33 *We can write* $p(x) = \sum_{k=0}^{m} a_k P_k(x)$*, where* $a_k = \frac{2k+1}{2}\int_{-1}^{1} P_k(x)p(x)\mathrm{d}x$.

By properties of Legendre polynomials (Fact 26) we have that $|p(x)| \le \sum_{k=0}^{m} |a_k|$ for all $x \in [-1,1]$, thus it suffices to bound the coefficients a_k. Towards that end,

$$\left|\int_{-1}^{1} P_k(x)p(x)\mathrm{d}x\right| = \left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\phi(x)\mathrm{d}x - \int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x\right|$$

(due to Equation (10))

$$\le \left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\phi(x)\mathrm{d}x\right| + \left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\frac{g(x)}{1-\varepsilon/2}\mathrm{d}x\right| \quad (11)$$

We will show how to bound the first term (the proof for the first term is almost identical). First,

$$\begin{aligned}
\left|\int_{\mathbb{R}\setminus[-B+\delta,B+\delta]} P_k(x)\phi(x)\mathrm{d}x\right| &\le \left|\int_{\mathbb{R}\setminus[-(B-\delta),(B-\delta)]} 4^k|x|^k\phi(x)\mathrm{d}x\right| && \text{(by Fact 26)}\\
&\le \int_{-\infty}^{-(B-\delta)} 4^k|x|^k\phi(x)\mathrm{d}x + \int_{B-\delta}^{\infty} 4^k|x|^k\phi(x)\mathrm{d}x\\
&= 2\int_{B-\delta}^{\infty} 4^k x^k\phi(x)\mathrm{d}x\\
&\lesssim 4^k\int_{\beta}^{\infty} x^k e^{-x^2/2}\mathrm{d}x && \text{(denote } \beta := B-\delta)\\
&\le 4^k\int_{\beta}^{\infty} x^k e^{-\beta^2/2+\beta}e^{-x}\mathrm{d}x\\
& && (-x^2/2+x \text{ is decreasing for } x>1)\\
&\le 4^k\int_{0}^{\infty} x^k e^{-\beta^2/2+\beta}e^{-x}\mathrm{d}x\\
&\le 4^k e^{-\beta^2/2+\beta}\int_{0}^{\infty} x^k e^{-x}\mathrm{d}x\\
&= 4^k e^{-\beta^2/2+\beta}\Gamma(k+1) && (\Gamma(\cdot) \text{ is the Gamma function})\\
&\le 4^k e^{-\beta^2/4}k^k\ .
\end{aligned}$$

Combining the inequality $|p(x)| \le \sum_{k=0}^{m}|a_k|$ with the formula for a_k and the above bound, we have that

$$|p(x)| \le \sum_{k=0}^{m}|a_k| \lesssim \sum_{k=0}^{m}\frac{2k+1}{2}4^k e^{-\beta^2/4}k^k \lesssim 4^m e^{-\beta^2/4}m^m\sum_{k=0}^{m}\frac{2k+1}{2} \lesssim 4^m e^{-\beta^2/4}m^{m+2}.$$

If m is a sufficiently small multiple of $\frac{\beta^2-\log(1/\varepsilon)}{\log(\beta^2-\log(1/\varepsilon))}$ then the RHS above is at most ε. Note that by our assumption $\delta \ll \varepsilon$ we have $\beta := B-\delta = \frac{1}{\delta}\log(1+\frac{\varepsilon/2}{1-\varepsilon/2}) - \delta = \Theta(\frac{1}{\delta}\log(1+\frac{\varepsilon/2}{1-\varepsilon/2}))$ and $\beta - \log(1/\varepsilon) = \Theta\left(\frac{1}{\delta}\log(1+\frac{\varepsilon/2}{1-\varepsilon/2})\right)$. Thus we can further simplify $\frac{\beta^2-\log(1/\varepsilon)}{\log(\beta^2-\log(1/\varepsilon))} = \Omega(\gamma^2\log\gamma)$ where $\gamma := \frac{1}{\delta}\log\left(1+\frac{\varepsilon/2}{1-\varepsilon/2}\right)$. Thus the moment matching can be achieved for m as high as a sufficiently small multiple of $\gamma^2/\log\gamma$.

■

Appendix D. Omitted Details from Section 4

D.1. Useful Subroutines from Robust Statistics

We require two subroutines used in Algorithm 1. The first handles mean estimation when more than half the samples are corrupted and, since exact recovery is impossible, outputs a list of candidate means containing one close to the truth.

Fact 34 (Mean list-decoding algorithm (see, e.g., Diakonikolas et al. (2022a))) *Let $\varepsilon \in (0,1)$ be a corruption rate parameter and $\tau \in (0,1)$ be a probability of failure parameter. There exists an algorithm that uses ε-corrupted samples from $\mathcal{N}(\mu, I)$ in the strong contamination model[2] and finds a list L of candidate means such that, with probability at least 0.99, there is at least one $\widehat{\mu}_0 \in L$ with $\|\widehat{\mu}_0 - \mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. The sample complexity of the algorithm is $n = (\log(1/(1-\varepsilon))d)^{O(\log(1/(1-\varepsilon)))}$, the size of the returned list is $|L| = O(\frac{1}{1-\varepsilon})$ and the runtime of the algorithm is $\mathrm{poly}(n)$.*

The second component is a pruning procedure that selects a near-optimal estimate from the list. The procedure runs a one-dimensional robust mean estimator along the line connecting each pair of vectors in the list. For each pair, it disqualifies the element that lies farther from the estimated mean along that line. At the end, any remaining element can be returned.

The proof is identical to that of Diakonikolas et al. (2025b), with the only difference being the choice of the one-dimensional robust mean estimator. Here, we may use an estimator designed for Gaussian mean estimation under an ε-fraction of arbitrary corruptions (unlike Definition 1, where arbitrary corruptions allow the adversary to modify an ε-fraction of the samples arbitrarily). In particular, the median or trimmed mean suffices, with sample complexity $\log(1/\tau)/\delta^2$ in one dimension. We set $\tau = 1/k^2$ to allow a union bound over all pairs in a list of size k.

Fact 35 (Tournament pruning (see Lemma 4.1 in Diakonikolas et al. (2025b))) *Let C be a sufficiently large absolute constant. Let $\varepsilon, \delta \in (0,1)$ be parameters. Let $L = \{\mu_1, \cdots, \mu_k\} \subset \mathbb{R}^d$ be a set of candidate estimates of $\mu \in \mathbb{R}^d$. There exists an algorithm* TOURNAMENTIMPROVE *that takes as input the list L, the parameters ε, δ and draws $n = O\left(\frac{\log k}{\delta^2}\right)$ samples according to the data generation model of Definition 1 with mean $\mu \in \mathbb{R}^d$ and corruption rate ε, and outputs some estimate $\mu_j \in L$ such that $\|\mu_j - \mu\|_2 \le 2\min_{i\in[k]} \|\mu_i - \mu\|_2 + \delta/2$ with probability at least 0.99. The runtime of the algorithm is $\mathrm{poly}(n, k, d)$.*

D.2. Omitted Details from the Proof of Correctness

We restate and prove the following structural result.

Proposition 36 *Let $\varepsilon \in (0,1)$ and $\delta \in (0, \log(1 + \frac{2\varepsilon}{1-\varepsilon}))$ be parameters. Let k be an even integer which satisfies $k \ge 3\left(\frac{1}{\delta}\log(1 + \frac{2\varepsilon}{1-\varepsilon})\right)^2$ and $P = \mathcal{N}(\delta, 1)$ be a Gaussian distribution. Then for any ε-corrupted version $\widetilde{P}$ of P under the model of Definition 1, if P' denotes the conditional distribution of $\widetilde{P}$ on the non-missing samples, it holds $\mathbb{E}_{x\sim P'}[x^k] - \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k] > \varepsilon$.*

2. Unlike Definition 1, in the *strong contamination model* $(1-\varepsilon)n$ samples are drawn from $\mathcal{N}(\mu, I)$ and an adversary can add the remaining εn points *arbitrarily*.

Proof We will prove this by showing the following two claims:

1. $\mathbb{E}_{y\sim P}[y^k] > (1+2\varepsilon/(1-\varepsilon))\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k]$.
2. $\mathbb{E}_{x\sim P'}[x^k] \geq (1-\varepsilon)\,\mathbb{E}_{y\sim P}[y^k]$.

These two claims suffice because if we combine them, we obtain

$$\mathop{\mathbb{E}}_{x\sim P'}[x^k] \geq (1-\varepsilon)\mathop{\mathbb{E}}_{y\sim P}[y^k] \geq (1-\varepsilon)(1+2\varepsilon/(1-\varepsilon))\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[z^k] \geq (1+\varepsilon)\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[z^k]\,.$$

Rearranging, this means that $\mathbb{E}_{x\sim P'}[x^k] - \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k] \geq \varepsilon\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k] \geq \varepsilon$. We now show the two claims. The second claim follows directly by the definition of the contamination model: if f denotes the function used in Definition 1, then $f(x) \geq (1-\varepsilon)p(x)$ (where p is the pdf of P). Thus

$$\mathop{\mathbb{E}}_{x\sim P'}[x^k] = \int_{\mathbb{R}} x^k \frac{f(x)}{\int_{\mathbb{R}} f(z)\mathrm{d}z}\mathrm{d}x \geq (1-\varepsilon)\int_{\mathbb{R}} x^k \frac{p(x)}{\int_{\mathbb{R}} f(z)\mathrm{d}z}\mathrm{d}x \geq (1-\varepsilon)\int_{\mathbb{R}} x^k p(x) = (1-\varepsilon)\mathop{\mathbb{E}}_{y\sim P}[y^k].$$

We now move to the first claim, i.e., that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[(z+\delta)^k] \geq \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k](1+2\varepsilon/(1-\varepsilon))$ when $k \geq 3\left(\frac{1}{\delta}\log(1+\frac{2\varepsilon}{1-\varepsilon})\right)^2$. First, we recall that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k] = (k-1)!!$. Using the binomial theorem, we can write the other non-centered moment as follows:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[(z+\delta)^k] = \sum_{\substack{j=0\\ j\text{ even}}}^{k}\binom{k}{j}\delta^k \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[z^{k-j}] = \sum_{\substack{j=0\\ j\text{ even}}}^{k}\binom{k}{j}\delta^k\,(k-j-1)!!\,. \tag{12}$$

We will now rewrite the right hand side above. First, for the $(k-j-1)!!$ we have the following. We can rewrite this as follows by taking the first $j/2$ odd factors of $(k-1)!!$:

$$(k-j-1)!! = \frac{(k-1)!!}{(k-1)(k-3)\cdots(k-j+1)} \geq \frac{(k-1)!!}{k^{j/2}}, \tag{13}$$

where we used that every factor in the denominator is at most k. We also have the following for the binomial coefficient:

$$\binom{k}{j} = \frac{k(k-1)\cdots(k-(j-1))}{j!} = \frac{k^j}{j!}\prod_{i=0}^{j-1}\left(1-\frac{i}{k}\right) \geq \frac{k^j}{j!}\prod_{i=0}^{j-1}\left(1-\frac{i}{j}\right)\frac{k^j}{j!}\frac{j!}{j^j} = \frac{k^j}{j^j}. \tag{14}$$

Combining equation (12) with inequalities (13) and (14), we have $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[(z+\delta)^k] \geq (k-1)!!\sum_{\substack{j=0\\ j\text{ even}}}^{k}\left(\frac{\sqrt{k}\delta}{j}\right)^j$. The sum can be lower bounded by a specific term of the sum. For this we will chose j to be a sufficiently small multiple of $\sqrt{k}\delta$ to obtain

$$\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[(z+\delta)^k] \geq (k-1)!!e^{\sqrt{k}\delta/3} = \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[z^k]e^{\sqrt{k}\delta/3}.$$

Thus, if $k \geq 3\left(\frac{1}{\delta}\log(1+\frac{2\varepsilon}{1-\varepsilon})\right)^2$ then $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[(z+\delta)^k] \geq (1+2\varepsilon/(1-\varepsilon))\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k]$. ■

We will need a Hermite-polynomial version of Proposition 36, obtained by expansion in the Hermite basis and an application of Cauchy–Schwarz.

Lemma 12 (Structural Lemma) *Let $\varepsilon \in (0,1)$ and $\delta \in (0, \log(1+\frac{2\varepsilon}{1-\varepsilon}))$ be parameters. Let k be an even integer which satisfies $k \geq 3(\frac{1}{\delta}\log(1+\frac{2\varepsilon}{1-\varepsilon}))^2$ and $P = \mathcal{N}(\delta, 1)$ be a Gaussian distribution. Then for any ε-corrupted version $\widetilde{P}$ of P under the model of Definition 1, if P' denotes the conditional distribution of $\widetilde{P}$ on the non-missing samples, it holds $|\mathbb{E}_{x\sim P'}[h_k(x)]| > \frac{\epsilon}{(k+1)^{k/2}}$, where h_k is the normalized probabilist's Hermite polynomial.*

Proof Proposition 36 states that $\mathbb{E}_{x\sim P'}[x^k] - \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^k] > \varepsilon$. We now expand the function x^k in the Hermite basis, i.e., $x^k = \sum_{t=0}^k a_t h_t(x)$ where $a_t := \mathbb{E}_{x\sim\mathcal{N}(0,1)}[x^k h_t(x)]$. Combining the result from Proposition 36 with Cauchy-Schwarz gives the following

$$\begin{aligned}
&\sqrt{\sum_{t=0}^k a_t^2}\sqrt{(k+1)\max_{t=0,\cdot,k}\left|\mathop{\mathbb{E}}_{x\sim P'}[h_t(x)] - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[h_t(z)]\right|^2} \\
&\quad\geq \sqrt{\sum_{t=0}^k a_t^2}\sqrt{\sum_{t=0}^k\left(\mathop{\mathbb{E}}_{x\sim P'}[h_t(x)] - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[h_t(z)]\right)^2} \\
&\quad\geq \sum_{t=0}^k a_t\left(\mathop{\mathbb{E}}_{x\sim P'}[h_t(x)] - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[h_t(z)]\right) \\
&\quad= \mathop{\mathbb{E}}_{x\sim P'}[x^k] - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[z^k] > \epsilon.
\end{aligned}$$

Rearranging we have

$$\begin{aligned}
\max_{t=0,\cdot,k}\left|\mathop{\mathbb{E}}_{x\sim P'}[h_t(x)] - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[h_t(z)]\right| &\geq \frac{\epsilon}{\sqrt{k+1}\sqrt{\sum_{t=0}^k a_t^2}} = \frac{\epsilon}{\sqrt{k+1}\sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^{2k}]}} \\
&= \frac{\epsilon}{\sqrt{(k+1)\frac{(2k)!}{2^k k!}}} \geq \frac{\epsilon}{(k+1)^{k/2}}.
\end{aligned}$$

Finally, noting that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[h_t(z)] = 0$ concludes the proof. ∎

We restate and prove the following lemma.

Lemma 15 (Moment tensor norm bound) *Let $\widetilde{P}$ be the ε-corrupted version of $\mathcal{N}(\mu, I)$ mentioned in the statement of Theorem 4 and P' denote the conditional distribution on the non-missing samples. Assume that $\|\mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. Let $T_t = \mathbb{E}_{x\sim P'}[H_t(x)]$ denote the tensors used in Algorithm 1. We have that*

$$\|T_t\|_2 \leq \frac{1}{1-\varepsilon}O\left(\log(1/(1-\varepsilon))\right)^{t/2} + \frac{1}{1-\varepsilon}\exp\left(O\left(t\log(1/(1-\varepsilon))\right)\right). \tag{5}$$

Proof Using the variational characterization of the ℓ_2-norm, we have

$$\|T_t\|_2 = \sup_{\|A\|_2=1}\langle A, T_t\rangle = \sup_{\|A\|_2=1}\mathop{\mathbb{E}}_{x\sim P'}[\langle A, H_t(x)\rangle].$$

Since $H_t(x)$ is a symmetric t-tensor and depends only on the symmetrization of A, we may restrict without loss of generality to symmetric A. For such A, the function $p_A(x) := \langle A, H_t(x)\rangle$ is a degree-t polynomial satisfying $\mathbb{E}_{z\sim\mathcal{N}(0,I)}[p_A(z)^2] = \|A\|_2^2 = 1$. Moreover, every degree-t polynomial that is orthonormal with respect to the Gaussian measure can be written in this way from a unique symmetric tensor A. Therefore,

$$\|T_t\|_2 = \sup_{\substack{p \text{ degree } t \\ \mathbb{E}_{z\sim\mathcal{N}(0,I)}[p(z)^2]=1}} \mathbb{E}_{x\sim\widetilde{P}}[p(x)]. \tag{15}$$

Thus we need to show that the $\mathbb{E}_{x\sim P'}[p(x)]$ is bounded for every unit-norm polynomial p of degree t. To this end, recall the definition of the distribution P' from Definition 16

$$\begin{aligned}
\mathbb{E}_{x\sim P'}[p(x)] &= \mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x) \mid x \text{ not missing}] = \frac{\mathbb{E}[p(x)\mathbb{1}(x \text{ not missing})]}{\mathbb{P}_{x\sim\mathcal{N}(\mu,I)}[x \text{ not missing}]} \\
&\le \frac{1}{1-\varepsilon}\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)\mathbb{1}(x \text{ not missing})] \\
&\le \frac{1}{1-\varepsilon}\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)] - \frac{1}{1-\varepsilon}\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)\mathbb{1}(x \text{ missing})] \\
&\le \frac{1}{1-\varepsilon}\left(\left|\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)]\right| + \left|\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)\mathbb{1}(x \text{ missing})]\right|\right).
\end{aligned}$$

We will analyze each term separately. For the first term we have the following:

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)] &\le \left\|\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[H_t(x)]\right\|_2 && \text{(by the variational characterization of } \ell_2\text{-norm)} \\
&= \left\|\mu^{\otimes t}\right\|_2/\sqrt{t!} \le \|\mu\|_2^t/\sqrt{t!} \\
&= O\left(\log(1/(1-\varepsilon))\right)^{t/2} && \text{(by Fact 29 and } \|\mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})\text{)}
\end{aligned}$$

For the second term, we have the following:

$$\begin{aligned}
\left|\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p(x)\mathbb{1}(x \text{ missing})]\right| &\le \sqrt{\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p^2(x)]}\sqrt{\mathbb{P}[x \text{ missing}]} \\
&\le \sqrt{\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p^2(x)]\varepsilon}
\end{aligned}$$

It remains to bound $\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p^2(x)]$. We know that $\mathbb{E}_{x\sim\mathcal{N}(0,I)}[p^2(x)] = 1$ however we need to bound the expectation over a translated Gaussian. By Claim 31 we have that $\mathbb{E}_{x\sim\mathcal{N}(\mu,I)}[p^2(x)] \le e^{t\|\mu\|^2} \le e^{O(t\log(1/(1-\varepsilon)))}$ (where we used $\|\mu\|_2 = O(\sqrt{\log(1/(1-\varepsilon))})$. Overall, by putting everything together, we have that

$$\mathbb{E}_{x\sim\widetilde{P}}[p(x)] \le \frac{1}{1-\varepsilon}O\left(\log(1/(1-\varepsilon))\right)^{t/2} + \frac{1}{1-\varepsilon}\exp\left(O\left(t\log(1/(1-\varepsilon))\right)\right).$$

■

Appendix E. Sample Complexity Lower Bound

We restate and prove the following result.

Theorem 37 (Sample complexity lower bound) *For every $\varepsilon \in (0,1), \delta \in (0, \log^{-1/2}(1+\frac{\varepsilon}{1-\varepsilon}))$ and $n \in \mathbb{Z}_+$ the following holds. If $\mathcal{A}$ is an algorithm that uses samples from an ε-corrupted version of a Gaussian $\mathcal{N}(\mu, 1)$ and outputs $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_2 \leq \delta$ with probability at least 0.9 then the sample complexity of $\mathcal{A}$ is*

$$n \geq \frac{1}{1-\varepsilon} \exp\left(\Omega\left(\frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{\delta}\right)^2\right). \tag{16}$$

Remark 38 *Some remarks follow:*

- *The bound in Equation (16) agrees with the sample complexity upper and lower bound shown in Ma et al. (2024) (Theorems 5 and 6 therein): Our model coincides with theirs when $\sigma = 1$ and $q = 1$. Solving for the second term (which is the dominant term) in Theorem 5 or 6 to be equal to δ^2 yields (up to constant factors) the same expression as the right-hand side of Equation (16).*
- *(Small ε regime) When $\varepsilon \to 0$ the bound becomes $\exp(\Omega(\varepsilon/\delta)^2)$.*
- *(Large ε regime) When $\varepsilon \to 1$ the bound behaves like $\exp\left(\Omega(\frac{1}{\delta}\log(\frac{1}{1-\varepsilon}))^2\right)$.*

The argument for showing the theorem consists of showing that there exist two distributions in this contamination model that are close in total variation distance.

Lemma 39 *For every $\varepsilon \in (0,1), \delta > 0, n \in \mathbb{Z}_+$ the following holds. Consider the two Gaussians $P_1 = \mathcal{N}(-\delta/2, 1)$ and $P_2 = \mathcal{N}(\delta/2, 1)$. There exist distributions Q_1, Q_2 on $\mathbb{R}$ such that*

- *Q_1 is an ε-corrupted version of P_1 according to Definition 1 and Q_2 is ε-corrupted version of P_2.*
- *$D_{\mathrm{TV}}(Q_1^{\otimes n}, Q_2^{\otimes n}) \leq \frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$.*

We first show how Theorem 37 follows given the lemma

Proof [Proof of Theorem 37] Define the following hypothesis testing problem: With probability $1/2$ all samples come from Q_1 and with probability $1/2$ all samples come from Q_2. If a mean estimator existed that had accuracy $\delta/2$ with probability 0.9 then we would be able to solve the testing problem with probability 0.9. However by Le Cam's lemma (Proposition 25), every testing algorithm has probability of failure at least $\frac{1}{2}\left(1 - D_{\mathrm{TV}}(Q_1^{\otimes n}, Q_2^{\otimes n})\right)$. In order for that probability of failure to be less than 0.1 we need $n > \frac{1}{1-\varepsilon} e^{\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$. ■

Proof [Proof of Lemma 39] Fix the threshold $t := \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{\delta}$ through this proof. Let $p_1(x), p_2(x)$ denote the pdfs of P_1, P_2. We define the functions q_1, q_2 as shown below:

$$q_1(x) = \begin{cases} p_1(x) & x \in (0, \infty) \\ p_2(x) & x \in [-t, 0] \\ (1-\varepsilon)p_1(x) & x \in (-\infty, -t) \end{cases} \qquad q_2(x) = \begin{cases} (1-\varepsilon)p_2(x) & x \in (t, \infty) \\ p_1(x) & x \in [0, t] \\ p_2(x) & x \in (-\infty, 0) \end{cases}$$

Claim 40 *It holds* $(1-\varepsilon)p_1(x) \le q_1(x) \le p_1(x)$ *and* $(1-\varepsilon)p_2(x) \le q_2(x) \le p_2(x)$ *for all* $x \in \mathbb{R}$.

Proof We do the check for the first part of the claim involving p_1 and q_1. The check for the second part is identical. The only non-trivial part of check is showing that $p_2(x) \ge (1-\varepsilon)p_1(x)$ for all $x \in [-t, 0]$.

Recall that p_2 is the pdf of $\mathcal{N}(0, \delta/2)$ and p_1 is the pdf of $\mathcal{N}(-\delta/2, 1)$. We thus want to solve for $(1-\varepsilon)p_1(x) \le p_2(x)$. Plugging in the pdf of the two Gaussians

$$\exp\left(-\frac{(x-\delta/2)^2}{2} + \frac{(x+\delta/2)^2}{2}\right) \ge 1-\varepsilon$$

Solving the above yields $x \ge -\log(1+\frac{\varepsilon}{1-\varepsilon})/\delta$. Therefore, $p_2(x) \ge (1-\varepsilon)p_1(x)$ for all $x \in [-t, 0]$. ■

Since $(1-\varepsilon)p_1(x) \le q_1(x) \le p_1(x)$ the function q_1 induces a definition of an ε-corruption of P_1 according to Definition 1. That is, a sample from Q_1 is generated according to the following procedure: With probability $\int_{\mathbb{R}} q_1(x)\mathrm{d}x$ the sample is drawn $q_1(x)/\int_{\mathbb{R}} q_1(x)\mathrm{d}x$, and with probability $1 - \int_{\mathbb{R}} q_1(x)\mathrm{d}x$ the sample is set to the special symbol $\perp$.

Similarly, $(1-\varepsilon)p_2(x) \le q_2(x) \le p_2(x)$ and q_2 induces an ε-corrupted version of P_1, that we denote by Q_2. Samples from Q_2 are generated as follows: with probability $\int_{\mathbb{R}} q_2(x)\mathrm{d}x$ the sample is drawn from $q_2(x)/\int_{\mathbb{R}} q_2(x)\mathrm{d}x$, and with probability $1 - \int_{\mathbb{R}} q_2(x)\mathrm{d}x$ the sample is set to the special symbol $\perp$.

By symmetry of our setup, the probability of the sample not being deleted (set to $\perp$) is the same $\int_{\mathbb{R}} q_1(x)\mathrm{d}x = \int_{\mathbb{R}} q_2(x)\mathrm{d}x$. Denote by α this probability. Also denote by $\widetilde{Q}_1$ the conditional distribution of Q_1 conditioned on the sample not being $\perp$ and let $\widetilde{Q}_2$ denote the corresponding conditional distribution for Q_2.

In the following we will show that $D_{\mathrm{TV}}(Q_1^{\otimes n}, Q_2^{\otimes n}) \le \frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$. By Fact 21 it suffices to find a coupling Π between the two joint distributions $Q_1^{\otimes n}, Q_2^{\otimes n}$ with probability of disagreement at most $\frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$. That is, we need to define a joint distribution Π on two sets of n samples $((X_1, \ldots, X_n), (Y_1, \ldots, Y_n))$ such that (i) the marginals are $(X_1, \ldots, X_n) \sim \widetilde{Q}_1^{\otimes n}$ and $(Y_1, \ldots, Y_n) \sim \widetilde{Q}_2^{\otimes n}$ respectively (i.e., it is a valid coupling) and (ii) the probability of disagreement is $\mathbb{P}_{\Pi}[(X_1, \ldots, X_n) \ne (Y_1, \ldots, Y_n)] \le \frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$.

We define the coupling by defining the data generation process for $((X_1, \ldots, X_n), (Y_1, \ldots, Y_n))$ bellow. In the construction bellow, we will assume that we already have a coupling Π_0 for the conditional distributions of single samples, i.e., a distribution Π_0 such that if $(X, Y) \sim \Pi_0$ it holds $X \sim \widetilde{Q}_1, Y \sim \widetilde{Q}_2$ (i.e., Π_0 is a coupling between $\widetilde{Q}_1$ and $\widetilde{Q}_2$) and $\mathbb{P}_{(X,Y)\sim\Pi_0}[X \ne Y] \le \frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$. We will show why P_0 exists at the end; for now we will conclude the construction of Π using Π_0. We define the sample generation process for Π as follows:

1. Draw $c_i \sim \mathrm{Ber}(\alpha)$ for $i \in [n]$ (recall that α is the probability of "missing" samples).
2. For each $i \in [n]$:
   (a) If $c_i = 1$, then draw $(X_i, Y_i) \sim \Pi_0$.
   (b) Else, set $(X_i, Y_i) = (\perp, \perp)$.

Note that in the above construction, each X_i is distributed according to Q_1 and each Y_i follows Q_2 thus the above defines a valid coupling between $Q_1^{\otimes n}, Q_2^{\otimes n}$. For the probability of disagreement we have the following:

$$\begin{aligned}
\mathbb{P}[(X_1,\dots,X_n) \neq (Y_1,\dots,Y_n)] &\leq \sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i] \\
&= \sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i | c_i = 1]\,\mathbb{P}[c_i = 1] + \mathbb{P}[X_i \neq Y_i | c_i = 0]\,\mathbb{P}[c_i = 0] \\
&\leq \sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i | c_i = 1] \\
&= \sum_{i=1}^{n} \mathop{\mathbb{P}}_{(X_i,Y_i)\sim\Pi_0}[X_i \neq Y_i] \\
&\leq \frac{n}{1-\varepsilon} \exp\left(-\Omega\left(\frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{\delta}\right)^2\right).
\end{aligned}$$

It suffices to show that the coupling Π_0 with $\mathbb{P}_{(X,Y)\sim\Pi_0}[X \neq Y] \leq \frac{n}{1-\varepsilon} e^{-\Omega(\log(1+\varepsilon/(1-\varepsilon))/\delta)^2}$ exists. We show this by bounding the TV distance between the conditional distributions $\widetilde{Q}_1, \widetilde{Q}_2$ and defining Π_0 to be the maximal coupling (cf. Fact 21).

$$\begin{aligned}
D_{\mathrm{TV}}(\widetilde{Q}_1, \widetilde{Q}_2) &= \frac{1}{2}\int_{-\infty}^{+\infty} \left| \frac{q_1(x)}{\int_{\mathbb{R}} q_1(x)\mathrm{d}x} - \frac{q_2(x)}{\int_{\mathbb{R}} q_2(x)\mathrm{d}x} \right| && (17) \\
&= \frac{1}{2\alpha}\int_{-\infty}^{\infty} |q_1(x) - q_2(x)|\,\mathrm{d}x && (18) \\
&= \frac{1}{2\alpha}\left(\int_{-\infty}^{-t} \varepsilon p_1(x)\mathrm{d}x + \int_{t}^{+\infty} \varepsilon p_2(x)\mathrm{d}x\right) && \text{(by definition of } q_1, q_2) \\
&= \frac{1}{2\alpha}\left(\int_{-\infty}^{-t} \varepsilon\phi(x+\delta/2)\mathrm{d}x + \int_{t}^{+\infty} \varepsilon\phi(x+\delta/2)\mathrm{d}x\right) && (\phi \text{ is the pdf of } \mathcal{N}(0,1)) \\
&\leq \frac{1}{2\alpha} 2e^{-\Omega\left(\frac{1}{\delta}\log(1+\frac{\varepsilon}{1-\varepsilon})-\delta/2\right)^2} && \text{(see below)} \\
&\leq \frac{1}{1-\varepsilon} e^{-\Omega\left(\frac{1}{\delta}\log(1+\frac{\varepsilon}{1-\varepsilon})\right)^2}. && \text{(see below)}
\end{aligned}$$

Where the second to last line uses the standard Gaussian tail bound $\mathbb{P}_{z\sim\mathcal{N}(0,1)}[z > r] \leq e^{-r^2/2}$ for every $r \geq 0$. We are using this with $r := \frac{1}{\delta}\log(1+\frac{\varepsilon}{1-\varepsilon}) - \delta/2$. Note that this is non-negative because $\varepsilon \in (0,1)$ and $\delta^2 \leq \log(1+\varepsilon/(1-\varepsilon))$. The final line uses $\alpha := 1-\varepsilon$ and $\varepsilon \in (0,1)$ and $\delta^2 \leq \log(1+\varepsilon/(1-\varepsilon))$.

■

Appendix F. Sample Complexity Upper Bound

We restate the main result for the sample complexity of one dimensional estimation below. We will then use this together with a cover argument to show a multi-variate estimator in Theorem 45.

Theorem 41 (Sample complexity upper bound) *There exists a computationally efficient algorithm such that the following holds for any $\varepsilon \in (0,1)$, $\delta \in (0, \log^{1/2}(1+\frac{\varepsilon}{1-\varepsilon}))$ and $\tau \in (0,1)$. The algorithm takes as input $\varepsilon, \delta, \tau$, draws $n = \exp\left(O\left(\frac{\log(1+\frac{\varepsilon}{1-\varepsilon})}{\delta}\right)^2\right)\frac{\log(1/\tau)}{\varepsilon^2(1-\varepsilon)}$ samples from an ε-corrupted version of $\mathcal{N}(\mu,1)$ under the contamination model of Definition 1, and it returns $\widehat{\mu}$ such that it holds $|\widehat{\mu}-\mu| \le \delta$ with probability at least $1-\tau$.*

Remark 42 *Some remarks follow:*

- *If ε is not known to the algorithm it can be easily estimated by taking the fraction of samples that are equal to $\perp$.*
- *The algorithm's sample complexity matches the lower bound of Theorem 37 and the sample complexity upper bound of Ma et al. (2024) up to the ε^{-2} factor.*

We start with the claim that if the cdfs of two Gaussians are multiplicatively close to each other, then the means of the Gaussians must also be appropriately close. The following structural lemma quantifies this.

Lemma 43 *Let ξ, t, ε be real numbers with $\xi/2 > 0$, $t < -\xi/2$ and $\varepsilon \in (0,1)$. Let two Gaussians $P_+ = \mathcal{N}(\xi/2, 1)$ and $P_- = \mathcal{N}(-\xi/2, 1)$ and denote by $F_+(x)$ and $F_-(x)$ their cumulative distribution functions (cdfs). If t is a point for which $F_+(t) \ge (1-\varepsilon)F_-(t)$, then $\xi \le \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|t|}$.*

Proof

It suffices to prove the claim for the extreme case, i.e., that $F_+(t) = (1-\varepsilon)F_-(t)$ implies $\xi \le \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|t|}$. Let $\Phi(x)$ denote the cdf of $\mathcal{N}(0,1)$. Then the equation $F_+(t) = (1-\varepsilon)F_-(x)$ is equivalent to $\Phi(t-\xi/2)/\Phi(t+\xi/2) = 1-\varepsilon$. Taking logarithms on both sides and doing some further rewriting, we have

$$\begin{aligned}\log(1-\varepsilon) &= \log\left(\frac{\Phi(t-\xi/2)}{\Phi(t+\xi/2)}\right) = \log(\Phi(t-\xi/2)) - \log(\Phi(t+\xi/2)) \\ &= -\int_{t-\xi/2}^{t+\xi/2} \frac{\mathrm{d}}{\mathrm{d}y}\log\Phi(y)\mathrm{d}y = -\int_{t-\xi/2}^{t+\xi/2}\frac{\phi(y)}{\Phi(y)}\mathrm{d}y, \end{aligned} \tag{19}$$

where $\phi(y)$ denotes the pdf of $\mathcal{N}(0,1)$. Recall the Mills ratio inequality (cf. Fact 19):

$$x \le \frac{\phi(x)}{1-\Phi(x)} \le x + \frac{1}{x} \qquad \forall x > 0.$$

However, due to our assumption $t < -\xi/2$, the variable y inside the integral in Equation (19) is always negative. We can obtain a version of Mill's ratio inequality for negative reals by using the symmetry properties $\phi(x) = \phi(-x)$, $1 - \Phi(x) = \Phi(-x)$:

$$-y \le \frac{\phi(y)}{\Phi(y)} \le -y - \frac{1}{y} \qquad \forall y < 0.$$

Combining the left part of the above inequality with Equation (19), we obtain

$$\int_{t-\xi/2}^{t+\xi/2} y \mathrm{d}y \ge -\int_{t-\xi/2}^{t+\xi/2} \frac{\phi(y)}{\Phi(y)} \mathrm{d}y = \log(1-\varepsilon).$$

Using $\int_{t-\xi/2}^{t+\xi/2} y \mathrm{d}y = \frac{1}{2}((t+\xi/2)^2 - (t-\xi/2)^2) = 2t\xi/2 = -2|t|\xi/2$ and rearranging the above inequality, we finally obtain

$$\xi \le \frac{-\log(1-\varepsilon)}{|t|} = \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|t|}.$$

■

We will use the contrapositive and shifted version of Lemma 43 that is stated below. This is contrapositive because it is saying that large difference in the mean of two Gaussians translates to large multiplicative gap of their cdfs, and it is shifted because it includes an arbitrary shift μ in the means of both Gaussians.

Corollary 44 *Let μ, ξ, t, ε be reals with, $\xi > 0$, $t + \xi/2 < 0$ and $\varepsilon \in (0,1)$. If $P_1 = \mathcal{N}(\mu, 1)$ and $P_2 = \mathcal{N}(\mu + \xi, 1)$ are two Gaussians with $\xi > \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|t|}$, then their cdfs F_1, F_2 satisfy $F_2(t + \mu + \xi/2) < (1-\varepsilon)F_1(t + \mu + \xi/2)$.*

Proof

First, with $\frac{n}{1-\varepsilon}$ samples one can learn an approximation $\widehat{F}$ to the cumulative distribution function F of the corrupted distribution of the non-missing samples (Fact 20). That is, with probability at least $1 - 2e^{-2n\eta^2}$ we have

$$\left|\widehat{F}(x) - F(x)\right| \le \eta \tag{20}$$

for all $x \in \mathbb{R}$. For the remainder of the proof fix $t := \frac{\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{\delta}$. We will use $\eta := \varepsilon\Phi(-t)$ (where Φ denotes the cdf of $\mathcal{N}(0,1)$) and we will set n a sufficiently large multiple of $\eta^{-2}\log(1/\tau)$ so that the probability of failure is at most τ.

Let $u := \widehat{F}^{-1}(\Phi(-t))$, i.e, the point for which it holds $\widehat{F}(u) = \Phi(-t)$. Consider the Gaussian distribution $\mathcal{N}(\mu_0, 1)$ where $\mu_0 := u + t$. This is exactly the Gaussian whose cdf F_0 satisfies $F_0(u) = \Phi(-t)$. The algorithm then is this: We simply return $\widehat{\mu} = \mu_0 = u + t$.

Given Theorem 44 it is easy to see why this has accuracy $O(\delta)$: Let F^* be the cdf of the ground truth Gaussian (inlier distribution). By the fact that $\widehat{F}(u) = \Phi(-t)$, Equation (20) and the definition of our contamination model, we have that $(1 - O(\varepsilon))\Phi(-t) \le F^*(u) \le (1 + O(\varepsilon))\Phi(-t)$. By

Theorem 44, if $\widetilde{F}$ is the cdf of a unit-variance Gaussian with mean larger than $\mu_0 + \frac{C\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{u-\mu_0-\delta/2}$, then it would hold $\widetilde{F}(t) < (1-C\varepsilon)\Phi(-t)$. Thus, this Gaussian could not be the ground truth one as its cdf evaluated on point t falls ouside of the interval $[(1-O(\varepsilon))\Phi(-t), \leq (1+O(\varepsilon))\Phi(-t)]$. Similarly we can rule out any Gaussian with mean smaller than $\mu_0 - \frac{C\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|u-\mu_0-\delta/2|}$ from being the ground truth. This means that the point μ_0 is within $\frac{2C\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{|u-\mu_0-\delta/2|} = \frac{2C\log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)}{t+\delta/2} = O(\delta)$ from the ground truth Gaussian's mean, where in the last step we used that $t = \log\left(1+\frac{\varepsilon}{1-\varepsilon}\right)/\delta$ and $\delta^2 \leq \log(1+\frac{\varepsilon}{1-\varepsilon})$. By adjusting the constants we can turn this $O(\delta)$ into just δ. ■

We now show the extension of the algorithm to multiple dimensions.

Theorem 45 (Multivariate estimator) *Let $d \in \mathbb{Z}_+$ denote the dimension, and C be a sufficiently large absolute constant. Let $\varepsilon \in (0,1)$, be a contamination parameter and $\delta \in (0, \log^{1/2}\left(1+\frac{\varepsilon}{1-\varepsilon}\right))$ be an accuracy parameter. Let $\mu \in \mathbb{R}^d$ be an (unknown) vector. There exists an algorithm that takes as input ε, δ, draws $n = \exp\left(O\left(\frac{\log(1+\frac{\varepsilon}{1-\varepsilon})}{\delta}\right)^2\right)\frac{d+\log(1/\tau)}{\varepsilon^2(1-\varepsilon)}$ points from an ε-corrupted version of $\mathcal{N}(\mu, I_d)$ under the contamination model of Definition 1, outputs a $\widehat{\mu}$ such that $\|\widehat{\mu}-\mu\|_2 \leq \delta$ with probability at least $1-\tau$. Moreover, it runs in time $2^{O(d)}\,\mathrm{poly}(n,d)$.*

Proof [Proof of Theorem 45] Denote by $T = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$ the points from the ε-corrupted version of $\mathcal{N}(\mu, I)$ and denote by $\mathcal{C}$ the the cover set of Theorem 23. The algorithm is the following: First, using the algorithm from Theorem 41, calculate a m_v for each $v \in \mathcal{C}$ such that $|m_v - v^\top\mu| \leq \delta/8$ (see next paragraph for more details on this step). Then, output the solution of the following linear program (note that the program always has a solution, as it is satisfied by $\widehat{\mu} = \mu$):

$$\begin{aligned}&\text{Find } \widehat{\mu} \in \mathbb{R}^d \text{ s.t.}\\ &|v^\top\widehat{\mu} - m_v| \leq \delta/4, \forall v \in \mathcal{C}\ .\end{aligned}$$

The claim is that this solution $\widehat{\mu}$ is indeed close to the target μ, since

$$\begin{aligned}\|\mu-\widehat{\mu}\|_2 &\leq 2\max_{v\in\mathcal{C}}|v^\top(\mu-\widehat{\mu})| && \text{(using Theorem 23)}\\ &\leq 2\max_{v\in\mathcal{C}}(|v^\top\mu - m_v| + |m_v - v^\top\widehat{\mu}|)\\ &\leq 2(\varepsilon/8+\varepsilon/4) < \varepsilon\ . && (21)\end{aligned}$$

We now explain how to obtain the approximations m_v with the guarantee $|m_v - v^\top\mu| \leq \delta/8$. Fixing a direction $v \in \mathcal{C}$, we note that $v^\top x \sim \mathcal{N}(v^\top\mu, 1)$ thus $\{v^\top x_i\}_{i=1}^m$ is a set of samples from an ε-corrupted version of $\mathcal{N}(v^\top\mu, 1)$. Thus, if we apply algorithm from Theorem 41 with probability of failure $\tau' = \tau/|\mathcal{C}|$, the event $|m_v - v^\top\mu| \leq \delta/8$ will hold with probability at least $1-\tau/|\mathcal{C}|$. By union bound, the probability all the events for $v \in \mathcal{C}$ hold simultaneously is at least $1-\tau$. The number of samples for this application of Theorem 41 is $2^{O(\log(1+\varepsilon/(1-\varepsilon)/\delta)^2}\frac{\log(1/\tau')}{\varepsilon^2(1-\varepsilon)} = 2^{O(\log(1+\varepsilon/(1-\varepsilon)/\delta)^2}\frac{\log(|\mathcal{C}|/\tau)}{\varepsilon^2(1-\varepsilon)} = 2^{O(\log(1+\varepsilon/(1-\varepsilon)/\delta)^2}\frac{d+\log(1/\tau)}{\varepsilon^2(1-\varepsilon)}$.

We conclude with the runtime analysis. The runtime to find the m_v's is $O(|\mathcal{C}|\,\mathrm{poly}(nd)) = 2^{O(d)}\,\mathrm{poly}(nd)$ since for each fixed $v \in \mathcal{C}$ we need $\mathrm{poly}(nd)$ time to calculate the projection $\{x_i^\top v\}$ of our dataset onto v and $\mathrm{poly}(n)$ time to run the one-dimensional estimator. The linear program can be solved using the ellipsoid algorithm. Consider the separation oracle that exhaustively checks all $2^{O(d)}$ constraints. We need $\mathrm{poly}(d)\log(\frac{R}{r})$ calls to that separation oracle, where R, r are the radii of the bounding spheres of the feasible region. First, $R \le \varepsilon$, because we have already shown in (21) that the feasible set belongs in a ball of radius ε around μ. Regarding the upper bound r, note that all $\widehat{\mu}$ inside a ball of radius $\delta/8$ around μ are feasible since $|v^\top\widehat{\mu} - m_v| \le |v^\top\widehat{\mu} - v^\top\mu| + |v^\top\mu - m_v| \le \|\widehat{\mu} - \mu\| + \delta/8 \le \varepsilon/4$. This means that $r = \varepsilon/4$. Hence the total runtime for solving the LP is $2^{O(d)}\,\mathrm{poly}(d)$ or simply $2^{O(d)}$.

■

Appendix G. Hardness in Other Restricted Models of Computation

We give a brief summary of known information-computation gaps for Non-Gaussian Component Analysis problem Problem 6 for different restricted models of computation. Since we have already shown that our problem is an instance of NGCA, we obtain immediately corollaries for Low-Degree Polynomials (Corollary 48) and PTFs (Corollary 50). The result for SoS has a few mild but cumbersome to verify conditions, so we refer the readers directly to Diakonikolas et al. (2024b) for the formal statements.

G.1. Hardness in the Low-Degree Polynomial Class of Algorithms

We start with the Low-Degree Polynomial (LDP) model, which we describe in more detail. We will consider tests that are thresholded polynomials of low-degree, i.e., output H_1 if the value of the polynomial exceeds a threshold and H_0 otherwise. We need the following notation and definitions. For a distribution D over $\mathcal{X}$, we use $D^{\otimes n}$ to denote the joint distribution of n i.i.d. samples from D. For two functions $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{X} \to R$ and a distribution D, we use $\langle f, g\rangle_D$ to denote the inner product $\mathbb{E}_{X\sim D}[f(X)g(X)]$. We use $\|f\|_D$ to denote $\sqrt{\langle f, f\rangle_D}$. We say that a polynomial $f(x_1, \ldots, x_n) : \mathbb{R}^{n\times d} \to \mathbb{R}$ has sample-wise degree (r, ℓ) if each monomial uses at most ℓ different samples from $x_1, \ldots, x_n$ and uses degree at most r for each of them. Let $\mathcal{C}_{r,\ell}$ be linear space of all polynomials of sample-wise degree (r, ℓ) with respect to the inner product defined above. For a function $f : \mathbb{R}^{n\times d} \to \mathbb{R}$, we use $f^{\le r,\ell}$ to be the orthogonal projection onto $\mathcal{C}_{r,\ell}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{D_0^{\otimes n}}$. Finally, for the null distribution D_0 and a distribution P, define the likelihood ratio $\overline{P}^{\otimes n}(x) := P^{\otimes n}(x)/D_0^{\otimes n}(x)$.

Definition 46 (n-sample τ-distinguisher) *For the hypothesis testing problem between D_0 (null distribution) and D_1 (alternate distribution) over $\mathcal{X}$, we say that a function $p : \mathcal{X}^n \to \mathbb{R}$ is an n-sample τ-distinguisher if $|\mathbb{E}_{X\sim D_0^{\otimes n}}[p(X)] - \mathbb{E}_{X\sim D_1^{\otimes n}}[p(X)]| \ge \tau\sqrt{\mathrm{Var}_{X\sim D_0^{\otimes n}}[p(X)]}$. We call τ the* advantage *of the polynomial p.*

Note that if a function p has advantage τ, then the Chebyshev's inequality implies that one can furnish a test $p' : \mathcal{X}^n \to \{D_0, D_1\}$ by thresholding p such that the probability of error under the null distribution is at most $O(1/\tau^2)$. We will think of the advantage τ as the proxy for the inverse of the probability of error and we will show that the advantage of all polynomials up to a certain

degree is $O(1)$. It can be shown that for hypothesis testing problems of the form of Problem 6, the best possible advantage among all polynomials in $\mathcal{C}_{r,\ell}$ is captured by the low-degree likelihood ratio (see, e.g., Brennan et al. (2021); Kunisky et al. (2022)):

$$\left\| \mathop{\mathbb{E}}_{v\sim\mathcal{U}(S)} \left[\left(\overline{P}_{A,v}^{\otimes n} \right)^{\leq r,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} ,$$

where in our case $D_0 = \mathcal{N}(0, I)$.

It has been known by Brennan et al. (2021) that a lower bound for the SQ dimension translates to an upper bound for the low-degree likelihood ratio. Given this, one can obtain the corollary regarding the hardness of NGCA:

Theorem 47 (Information-computation gap for NGCA in LDP) *Let c be a sufficiently small positive constant and consider the hypothesis testing problem of Problem 6 the distribution A matches the first t moments with $\mathcal{N}(0, I)$. For any $d \in \mathbb{Z}_+$ with $d = t^{\Omega(1/c)}$, any $n \leq \Omega(d)^{(t+1)/10}/\chi^2(A, \mathcal{N}(0, I))$ and any even integer $\ell < d^c$, we have that*

$$\left\| \mathop{\mathbb{E}}_{v\sim\mathcal{U}(S)} \left[\left(\overline{P}_{A,v}^{\otimes n} \right)^{\leq \infty,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} \leq 1 .$$

The interpretation of this result is that unless the number of samples used n is greater than $\Omega(d)^{(t+1)/10}/\chi^2(A, \mathcal{N}(0, I))$, any polynomial of degree roughly up to d^c fails to be a good test (note that any polynomial of degree ℓ has sample-wise degree at most (ℓ, ℓ)).

We now show the corollary for the robust mean estimation problem of this paper. Recall that the hypothesis testing problem of Theorem 3 includes as a special case the NGCA problem with A being the ε-corrupted version of $N(\delta v, I)$ where v is a unit vector and the corruption adversary from Definition 1 uses as $f(x)$ the function from Lemma 11. For this distribution A we have that (i) it matches the first $\Omega(\gamma^2/\log\gamma)$ moments with $\mathcal{N}(0, 1)$, where $\gamma := \frac{1}{\delta}\log(1 + \frac{\varepsilon/2}{1-\varepsilon/2})$ and (ii) $\chi^2(A, \mathcal{N}(0, 1) = O(\frac{1}{1-\varepsilon})$; this part is not included in Lemma 11 because it was not needed for the SQ lower bound, but it immediately follows by using that $f(x)/\int f(x)\mathrm{d}x \leq \frac{g(x)+p(x)\mathbb{1}(|x|\leq 1)}{1-\varepsilon} \leq \frac{\phi(x-\delta)+\varepsilon}{1-\varepsilon}$ (in the proofs of Lemmata 10 and 11 it can be seen that $g(x)$ bounded by a translated Gaussian and that the polynomial $p(x)$ has small absolute value in $[-1, 1]$).

Corollary 48 (Hardness of mean estimation against Low-Degree Polynomials) *Consider the same hypothesis testing problem as in Theorem 3 and let m be defined as in Theorem 3. There is a way for the adversary of Definition 1 to corrupt the alternative hypothesis $P = \mathcal{N}(\delta u, I)$ into a distribution $\widetilde{P}$ such that the following holds. Let c be a sufficiently small positive constant. For any $d \in \mathbb{Z}_+$ with $d = m^{\Omega(1/c)}$, any $n \leq \Omega(d)^{(m+1)/10}(1-\varepsilon)$ and any even integer $\ell < d^c$, we have that*

$$\left\| \mathop{\mathbb{E}}_{v\sim\mathcal{U}(S)} \left[\left(\widetilde{P}^{\otimes n} \right)^{\leq \infty,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} \leq 1 .$$

This is interpreted as a tradeoff between $d^{\Omega(m)}(1-\varepsilon)$ and super-polynomial runtime.

G.2. Hardness in the PTFs Class of Algorithms

In the LDP class of the previous section, the goodness of a test is quantified by the advantage, defined in Theorem 46. Hardness of a problem in this class are shown by ruling out existence of polynomials with small advantage. That definition is based on the idea that one can obtain a test by thresholding the polynomial in the midpoint of the expectations for the two distributions. Thus rulling out existence of polynomials with small advantage rulls out the construction of such tests. However it still leaves open the possibility of some other kind of thresholded polynomial might still succeed. As such, a more natural class of tests is the one consisting of all possible thresholded polynomials (i.e., with arbitrary thresholdings).

The information-computation gap for NGCA in this class is the one below.

Theorem 49 (Information-computation gap for NGCA in LDP (Diakonikolas et al. (2025c))) *There exists a sufficiently large absolute constant C^* such that the following holds. For any $c^* \in (0, 1/4)$, $d, k, n, m \in \mathbb{Z}_+$ such that (i) m is even, (ii) $\max(k, m) < d^{c^*/C^*}$, and (iii) $n < d^{(1/4-c^*)m}$, we have that if $p : \mathbb{R}^{n\times d} \mapsto \mathbb{R}$ is a degree-k polynomial, and A is a distribution on $\mathbb{R}$ that matches the first m moments with $\mathcal{N}(0, 1)$, then:*

$$\left| \mathop{\mathbb{E}}_{\substack{\mathbf{v}\sim\mathcal{U}(S)\\ x^{(1)},\ldots,x^{(n)}\sim P_{A,v}}} \left[\operatorname{sgn}(p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}))\right] - \mathop{\mathbb{E}}_{x^{(1)},\ldots,x^{(n)}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\operatorname{sgn}(p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}))\right] \right| \leq 0.11. \tag{22}$$

where $P_{A,v}$ denotes the hidden direction distribution from Definition 5, and $\operatorname{sgn} : \mathbb{R} \to \{0, 1\}$ is the sign function with $\operatorname{sgn}(x) = 1$ if and only if $x \geq 0$.

The corollary for robust mean estimation is stated below.

Corollary 50 (Hardness of mean estimation against Low-Degree Polynomials) *Consider the same hypothesis testing problem as in Theorem 3 and let m be defined as in Theorem 3. There is a way for the adversary of Definition 1 to corrupt the alternative hypothesis $P = \mathcal{N}(\delta u, I)$ into a distribution $\widetilde{P}$ such that the following holds. There exists a sufficiently large absolute constant C^* such that the following holds. For any $d, k, n, m \in \mathbb{Z}_+$ such that (i) m is even, (ii) $\max(k, m) < d^{c^*/C^*}$, and (iii) $n < d^{\Omega(m)}$, we have that if $p : \mathbb{R}^{n\times d} \mapsto \mathbb{R}$ is a degree-k polynomial, then:*

$$\left| \mathop{\mathbb{E}}_{\substack{\mathbf{v}\sim\mathcal{U}(S)\\ x^{(1)},\ldots,x^{(n)}\sim P_{A,v}}} \left[\operatorname{sgn}(p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}))\right] - \mathop{\mathbb{E}}_{x^{(1)},\ldots,x^{(n)}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\operatorname{sgn}(p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}))\right] \right| \leq 0.11. \tag{23}$$

where $P_{A,v}$ denotes the hidden direction distribution from Definition 5, and $\operatorname{sgn} : \mathbb{R} \to \{0, 1\}$ is the sign function with $\operatorname{sgn}(x) = 1$ if and only if $x \geq 0$.

For an arbitrary polynomial p, the runtime for this computation is on the order of $\operatorname{poly}((nd)^k)$ and thus, Theorem 49 implies an inherent trade-off between the exponential runtime $(nd)^{d^{\Omega(1)}}$, and the sample complexity $d^{\Omega(m)}$ for the family of PTF tests.