Private Robust Estimation by Stabilizing Convex Relaxations

Pravesh K. Kothari *
praveshk@cs.cmu.edu

Pasin Manurangsi†
pasin@google.com

Ameya Velingker †
ameyav@google.com

December 6, 2021

Abstract

We give the first polynomial time and sample (ε, δ)-differentially private algorithm to estimate the mean, covariance and higher moments in the presence of a constant fraction of adversarial outliers. Our algorithm handles an absolute constant fraction of adversarial outliers in the data and succeeds for families of distributions that satisfy two well-studied properties in prior works on robust estimation: *certifiably subgaussianity* of directional moments and *certifiably hypercontractivity* of degree 2 polynomials. Our recovery guarantees hold in the "right affine-invariant norms": Mahalanobis distance for mean, multiplicative spectral and relative Frobenius distance guarantees for covariance and injective norms for higher moments. Prior works obtained private robust algorithms for mean estimation of subgaussian distributions with bounded covariance. For covariance estimation, ours is the first efficient algorithm (even in the absence of outliers) that succeeds without any condition-number assumptions.

Our algorithms are obtained via a new framework that provides a general blueprint modifying convex relaxations for robust estimation to satisfy strong *worst-case stability* guarantees in the *appropriate parameter norms* whenever the algorithms produce *witnesses of correctness* in their run. We verify such guarantees for a slight modification of standard sum-of-squares (SoS) semidefinite programming relaxations for robust estimation. Our privacy guarantees are obtained by combining our stability guarantees with a new "estimate dependent" noise injection mechanism that adds noise with magnitude that scales with the eigenvalues of the estimated covariance. We believe this framework will be useful more generally in obtaining differentially private counterparts of results in robust statistics.

Independently of our work, Ashtiani and Liaw [AL21] also obtained a polynomial time and sample private robust estimation algorithm for Gaussian distributions.

*Carnegie Mellon University
†Google Research

1 Introduction

In this work, we consider the problem of efficiently estimating the mean, covariance and, more generally, the higher moments of an unknown high-dimensional probability distribution on $\mathbb{R}^d$, given a sample $y_1, y_2, \ldots, y_n \in \mathbb{R}^d$, under two design constraints: outlier robustness and privacy. The first demands that we build estimators for such basic parameters of probability distributions that tolerate a fixed (dimension-independent) constant fraction of adversarial outliers in the input data. The second demands that our estimators preserve the privacy of individual points y_is (that we model as being contributed by different individuals) participating in our input data.

Sans privacy constraints, the problem of robustly estimating the basic parameters of an unknown distribution has been the focus of intense research in algorithmic robust statistics starting with the pioneering works of [DKK+16, LRV16] from 2016. In addition to new (and often, information-theoretically optimal) algorithms for several basic robust estimation tasks [KS17b, KS17a, HL18, BK20a, DHKK20], this line of work has led to a deeper understanding of the properties of the underlying distribution (algorithmic certificates of analytic properties such as subgaussianity, hypercontractivity and anti-concentration, resilience [SCV18]) that make robust estimation possible along with general frameworks such as outlier filtering and the sum-of-squares (SoS) method for attacking algorithmic problems in robust statistics.

Sans outlier robustness constraints, the task of *private estimation* of the mean and covariance of probability distributions has also seen considerable progress in the recent years. *Differential privacy* [DMNS06] has emerged as a widely-used standard for providing strong individual privacy guarantees. Under differential privacy, a single sample is not allowed to have too significant of an impact on the output distribution of an algorithm that operates on a dataset. Differential privacy has now been deployed in a number of production systems, including those at Google [EPK14, BEM+17], Microsoft [DKY17], Apple [Gre16, App17], and the US Census Bureau [Abo18]. While initial approaches to estimating the mean and covariance under differential privacy required *a priori* bounds on the support of the samples, a more recent work [KV18] managed to obtain the first private mean estimation algorithm for samples with unbounded support. Subsequent works have built on this progress to obtain differentially private algorithms for mean estimation and covariance estimation (under assumptions on the condition number of the unknown covariance) of Gaussian and heavy-tailed distributions [KLSU19, BS19, BKSW19, CWZ19, BDKU20, KSU20, DFM+20, WXDX20, AAK21, BGS+21].

In this paper, we focus on the task of finding efficient estimation algorithms for mean, covariance and, more generally, higher moments with recovery guarantees in multiplicative spectral distance (i.e., an affine invariant guarantee necessary, for example, to whiten the data or put a set of points in approximate isotropic position) and relative Frobenius distance (necessary for obtaining total variation close estimates of an unknown high-dimensional Gaussian). A very recent work of Liu, Kong, Kakade and Oh [LKKO21] found the first private and robust algorithm for mean estimation under natural distributional assumptions with bounded covariance. However, their techniques do not appear to extend to covariance estimation. Informally, this is because in order to obtain privacy guarantees, we need robust estimation algorithms that are *stable*, i.e., whose output suffers from a bounded perturbation when a single data point is changed arbitrarily. When the unknown

covariance is bounded, one can effectively assume that the change in a single data point is bounded. However, in general, the covariance of the unknown distribution can be exponentially (in the underlying dimension) varying eigenvalues which precludes such a method (even in the outlier-free regime).

This work In this paper, we give the first algorithms for differentially private robust moment estimation with polynomial time and sample complexity. Our algorithms, in fact, provide a general blueprint for transforming any robust estimation algorithm into a differentially private robust moment estimation algorithm with similar accuracy guarantees as long as the robust estimation algorithm satisfies two key properties: 1) the algorithm is *"witness-producing,"* i.e., the algorithm finds a sequence of "weights" on the input corrupted sample that induce a distribution with a relevant property of the unknown distribution family (such as certifiable subgaussianity or hypercontractivity) and 2) the algorithm allows for finding weights that minimize a natural strongly convex objective function in polynomial time. Such properties are naturally satisfied by robust estimation algorithms based on sum-of-squares semidefinite programs. Our main technical result is a simple framework that transforms such an algorithm into one that satisfies *worst-case stability* under input perturbation *in the relevant norms on the parameters*. The final ingredient in our framework is a new noise injection mechanism that uses the stability guarantees so obtained to derive privacy guarantees. This mechanism allows obtaining privacy guarantees even though the distribution of the noise being added depends on the unknown quantity being estimated. In particular, such a subroutine allows us to obtain private robust covariance estimation without any assumptions on the condition number. We note that even without the robustness constraints, a private covariance estimation algorithm without any assumptions on the condition number was not known prior to our work.

Robustness implies privacy? Our blueprint presents an intuitively appealing picture—that robustness, when obtained by estimators that satisfy some additional but generic conditions, implies privacy via a generic transformation. This connection might even appear natural: privacy follows by "adding noise" to the estimates obtained via algorithms that are insensitive or stable with respect to changing any single point in the input, while robustness involves finding estimators that are *insensitive* to the effects of even up to a constant fraction of outliers. Despite this apparent similarity, there are two key differences that prevent such an immediate connection from being true: 1) privacy is a worst-case guarantee while robustness guarantees are only sensible under distributional assumptions, and, 2) privacy guarantees need insensitivity even against "inliers." Nevertheless, our main result shows that robustness, when obtained via algorithms that satisfy some natural additional conditions, does yield stable (or insensitive) algorithms as required for obtaining differentially private algorithms.

In what follows, we describe our results and techniques in more detail.

1.1 Our Results

Formally, our results provide differentially private robust estimation algorithms in the strong contamination model, which we define below.

Definition 1.1 (Strong Contamination Model)**.** Let $\eta > 0$ be the *outlier rate*. Given a distribution D on $\mathbb{R}^d$ and a parameter $n \in \mathbb{N}$, the strong contamination model with outlier rate η gives access to a set $Y \subseteq \mathbb{R}^d$ of n points generated as follows: 1) Generate $X \subseteq \mathbb{R}^d$, an i.i.d. sample from D of size n, 2) Return any (potentially adversarially chosen) Y such that $|Y \cap X| \geqslant (1-\eta)n$. In this case, we say that Y is an *η-corruption* of X.

In the context of analyzing privacy, we will say that two subsets of n points $Y, Y' \subseteq \mathbb{R}^d$ (a.k.a. *databases*) are *adjacent* if they differ in exactly one point (i.e $|Y \cap Y'| \geqslant n-1$.) We now present our main theorem, which provides a differentially private robust algorithm for moment estimation of an unknown certifiably subgaussian distribution in the strong contamination model.

Our formal guarantees hold for moment estimation of certifiably subgaussian distributions. A distribution D is *C-subgaussian* if for any direction v and any $t \in \mathbb{N}$, $\mathbb{E}_D\langle x - \mu(D), v\rangle^{2t} \leqslant (Ct)^t(\mathbb{E}_D\langle x-\mu(D), v\rangle^2)^t$ where $\mu(D)$ is the mean of the distribution D. Certifiable subgaussianity is a stricter version of such a property that additionally demands that the difference between the two sides of the inequality be a sum-of-squares (SoS) polynomial in the variable v. Gaussian distributions, uniform distributions on product domains, all strongly log-concave distributions and, more generally, any distribution that satisfies a Poincaré inequality with a dimension-independent constant [KS17a] are known to satisfy certifiable subgaussianity. See Definition 3.22 and the preliminaries for a detailed discussion.

Our first result is an algorithm for moment estimation of certifiably subgaussian distributions that runs in polynomial time and has polynomial sample complexity.

Theorem 1.2. *Fix $C_0 > 0$ and $k \in \mathbb{N}$. Then, there exists an $\eta_0 > 0$ such that for any given outlier rate $0 < \eta \leqslant \eta_0$ and $\varepsilon, \delta > 0$, there exists a randomized algorithm* Alg *that takes an input of $n \geqslant n_0 = \widetilde{\Omega}\left(\frac{d^{4k}}{\eta^2}\left(1 + \left(\frac{\ln(1/\delta)}{\varepsilon}\right)^4 + \left(\frac{\ln(1/\delta)}{\varepsilon}\right)^{\frac{2k}{k-1}}\right) \cdot C^{4k}k^{4k+6}\right)$ points $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$ (where $C = 2C_0 + \frac{3\ln(3/\delta)}{\varepsilon} + \frac{9}{\varepsilon} + 1$), runs in time $(Bn)^{O(k)}$ (where B is the bit complexity of the entries of Y) and outputs either "reject" or estimates $\hat{\mu} \in \mathbb{Q}^d$, $\hat{\Sigma} \in \mathbb{Q}^d$, and $\hat{M}^{(t)} \in \mathbb{Q}^{d\times d\times\cdots\times d}$ (for all even $t < 2k$ such that t divides $2k$) satisfying the following guarantees:*

1. ***Privacy:*** *Alg is (ε, δ)-differentially private with respect to the input Y, viewed as a d-dimensional database of n individuals.*

2. ***Utility:*** *Let $X = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample of size $n \geqslant n_0$ from a certifiably C_0-subgaussian distribution $\mathcal{D}$ with mean μ_*, covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)}I$, and moment tensors $M_*^{(t)}$ for $t \geqslant 2$. If $Y = \{y_1, y_2, \ldots, y_n\}$ is an η-corruption of X, then with probability at least $9/10$ over the draw of X and random choices of the algorithm, Alg does not reject and outputs estimates $\hat{\mu} \in \mathbb{Q}^d$, $\hat{\Sigma} \in \mathbb{Q}^{d\times d}$, and $\hat{M}^{(t)} \in \mathbb{Q}^{d\times d\times\cdots\times d}$ (for all $t < 2k$ such that t divides $2k$) satisfying the following guarantees:*

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u\rangle \leqslant O(\sqrt{Ck})\eta^{1-1/2k}\sqrt{u^\top \Sigma_* u}\,,$$

and,

$$\left(1 - O((Ck)^{t/2k}\right) \eta^{1-1/k})\Sigma_* \preceq \hat{\Sigma} \preceq \left(1 + O((Ck)^{t/2k})\eta^{1-1/k}\right) \Sigma_* \,,$$

and, for all even $t < 2k$ such that t divides $2k$,

$$\left(1 - O(Ck)\eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle \leqslant \langle u^{\otimes t}, \hat{M}^{(t)} \rangle \leqslant \left(1 + O(Ck)\eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle \,.$$

In the above and subsequent theorems, we use the $\widetilde{\Omega}$ notation to hide multiplicative logarithmic factors in d, C, k, $1/\eta$, $1/\varepsilon$, and $\ln(1/\delta)$.

Discussion Our algorithm above achieves an error guarantee in the "right" affine-invariant norms similar to the robust moment estimation algorithm of [KS17b]. In particular, the error in the mean in any direction scales proportional to the variance of the unknown distribution providing recovery error bounds in the strong "Mahalanobis error." Similarly, the error in the covariance is multiplicative in the Löwner ordering. Our algorithm succeeds in the standard word RAM model of computation. In particular, the lower bound assumption on the eigenvalue of the unknown covariance in the statement above is entirely an artifact of numerical issues. Such an assumption can be removed (and in particular, we can deal with rank deficient covariances) if we assume that the unknown covariance Σ_* has rational entries with polynomial bit complexity. We choose to make an assumption on the smallest eigenvalue of Σ_* for the sake of simpler exposition.

Our algorithm above is obtained by applying a general blueprint that applies to any robust estimation algorithms that use "one-shot rounding" to produce a differentially private version. We explain our general blueprint in more detail in Section 2.

Applications Our differentially private moment estimation algorithm immediately allows us to obtain a differentially private mechanism to implement an outlier-robust *method of moments*. This allows us to learn parameters of statistical models that rely on the method of moments, such as mixtures of spherical Gaussians with linearly independent means [HK13] (that rely on decomposing 3rd moments) as well as independent component analysis [DLCC07] (that relies on decomposing fourth moments). We direct the reader to the work on robust moment estimation that details such applications [KS17b].

Covariance estimation in relative Frobenius error The above theorem provides a multiplicative spectral guarantee. Such a guarantee, however, only yields a dimension-dependent bound on the Frobenius norm of the error. While this is provably unavoidable for the class of certifiably subgaussian distributions, recent work [BK20b] showed that for distributions that satisfy the stronger property of having certifiably hypercontractive degree 2 polynomials (informally speaking, this is the analog of certifiable subgaussianity for moments of degree 2 polynomials instead of linear polynomials $\langle x, v \rangle$ of the random variable x), one can obtain a *dimension-independent* bound on the Frobenius estimation error that vanishes as the fraction of outliers tends to zero. Their algorithm relies on rounding an SoS relaxation with a slightly different constraint system. By working with

their constraint system and applying our blueprint for obtaining a "stable" version, we obtain a version of the above theorem with the stronger Frobenius estimation guarantee (see Theorem 5.6).

By combining our privacy analysis above with the recent work that shows that the algorithm in [BK20b] gives optimal estimation error when analyzed for corrupted samples from a Gaussian distribution, we obtain the following stronger guarantees for private mean and covariance estimation for Gaussian distributions.

Theorem 1.3 (Mean and Covariance Estimation for Gaussian Distributions)**.** *Fix $\varepsilon, \delta > 0$. Then, there exists an absolute constant $\eta_0 > 0$ such that for any given outlier rate $0 < \eta \leqslant \eta_0$, there exists a randomized algorithm Alg that takes an input of $n \geqslant n_0 = \widetilde{\Omega}\left(\frac{d^8}{\eta^4}\left(1 + \frac{\ln(1/\delta)}{\varepsilon}\right)^4\right)$ points $Y \subseteq \mathbb{Q}^d$, runs in time $(Bn)^{O(1)}$ (where B is the bit complexity of the entries of Y) and outputs either "reject" or estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d\times d}$ with the following guarantees:*

1. ***Privacy:*** *Alg is (ε, δ)-differentially private with respect to the input Y, viewed as a d-dimensional database of n individuals.*

2. ***Utility:*** *Let $X = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample of size $n \geqslant n_0$ from a Gaussian distribution with mean μ_* and covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)} I$ such that Y is an η-corruption of X. Then, with probability at least $9/10$ over the random choices of the algorithm, Alg outputs estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d\times d}$ satisfying the following guarantees [Ameya: I added in the $\log(1/\delta)$ dependence to be explicit. Can you check whether this lets us get rid of the tilde on the O?]:*

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u\rangle \leqslant \widetilde{O}\left(\eta \cdot \frac{\log(1/\delta)}{\varepsilon}\right)\sqrt{u^\top \Sigma_* u}\,,$$

and,

$$\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \preceq \widetilde{O}\left(\eta \cdot \sqrt{\frac{\log(1/\delta)}{\varepsilon}}\right).$$

In particular, $d_{\mathsf{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu_*, \Sigma_*)) < \widetilde{O}(\eta \log(1/\delta)/\varepsilon)$.

1.2 Related Work

Since the works of [DKK+16, LRV16], there has been a spate of works designing additional robust estimation algorithms for a wide variety of problems, including mean and covariance estimation [DKK+17a, DKK+17b, CDGW19, DHL19, HLZ20, Hop20, LY20], mixture models [HL18, KSS18, BK20c, DHKK20, BDH+20], principal component analysis (PCA) [KSKO20, JLT20], etc. (see survey [DK19] for details on recent advances in robust statistics). Furthermore, the criterion of *reslience* formulated in [SCV18] as a sufficient condition for robustly learning a property of a dataset was subsequently generalized in [ZJS19] in order to deal with a more general class of perturbations.

In the setting of high-dimensional parameter estimation, release of statistics can often reveal signficant information about individual data points, which can be problematic in a number of applications in which it is desirable to protect the privacy of individuals while still providing useful aggregate information (e.g., medical data or census data). Attacks exploiting such properties have

been investigated in a long line of works [DN03, BUV14, DSS+15, SU15, DSSU17, SSSS17]. In light of such exploits, there has been much interest in designing statistical algorithms that protect the privacy of individual samples in a dataset.

In the area of differentially privacy, various works have explored private estimation pertaining to Gaussian mixtures [NRS07, KSSU19], identity testing [CKM+20], Markov random fields [ZKKW20], etc.

Concurrent related works The problem of private robust mean and covariance estimation has been the subject of great interest resulting in a few concurrent and independent related works. Kamath, Mouzakis, Singhal, Steinke, and Ullman [KMS+21] give a differentially private (in the outlier-free regime) algorithm for mean and covariance estimation of Gaussian without making condition number assumptions on the covariance. The work of Liu, Kong, and Oh [LKO21] gives a statistical feasibility of private robust estimation with optimal sample complexity via a computationally *inefficient* algorithm. Finally, Hopkins, Kamath, and Majid [HKM21] also use the sum-of-squares semidefinite programs to obtain private mean estimation (in the outlier-free setting) algorithm for bounded covariance distribution with *pure differential privacy*. Our result are most directly related to the work of Asthiani and Liaw [AL21] that also obtains efficient private and robust mean and covariance estimation for Gaussian distributions.

2 Technical Overview

In this section, we give a high-level overview of our general blueprint for obtaining differentially private versions of robust estimation algorithms. As a running example, we will focus on the problem of obtaining private and robust mean and covariance estimators. Specifically, our goal is to design an algorithm that takes input consisting of n points, say $Y \subseteq \mathbb{R}^d$, along with an *outlier rate* η and returns estimates of the mean and covariance. We would like the algorithm to be (ε, δ)-differentially private for every Y (i.e., a "worst-case" guarantee), viewed as a database in which each d-dimensional point in Y is contributed by an individual. We would like the outputs of the algorithm to provide faithful estimates whenever Y is an η-corruption of a i.i.d. sample from a distribution that has C-subgaussian fourth moments.

For the purpose of the first part of this overview, we recommend the reader to ignore the distinction between certifiable subgaussianity and "vanilla" subgaussianity. Recall that a distribution D on $\mathbb{R}^d$ has C-subgaussian fourth moments if for every $v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim D}\langle x - \mu(D), v\rangle^4 \leqslant 4C(\mathbb{E}_{x\sim D}\langle x - \mu(D), v\rangle^2)^2$. It turns out that the uniform distribution on a $O(d^2)$ size i.i.d. sample X from a C-subgussian distribution has $2C$-subgaussian fourth moments.

Stable robust estimation algorithms In order to design differentially private algorithms, we need to find robust moment estimation algorithms that are *stable*. Specifically, a robust moment estimation algorithm Alg is stable if the outputs of Alg on any pair of adjacent inputs Y, Y' (i.e., inputs that differ in at most one point but arbitrarily so) are close. Such a guarantee must hold over *worst-case* pairs Y, Y'—in particular, Y may not be obtained by taking an η-corruption of an

i.i.d. sample from a distribution following our assumptions. This presents a problem at the outset as robust moment estimation algorithms are typically analyzed under *distributional assumptions*. The work of [LKKO21] addresses this issue by "opening up" an iterative filter based algorithm for robust moment estimation and effectively making every step of the algorithm stable.

2.1 A Prototypical Robust Estimator to Privatize

To understand our ideas, it is helpful to work with a "prototypical" but inefficient robust estimation algorithm that we can eventually swap with an efficent one. Let us thus start with a simple (but inefficient) robust estimation algorithm that we call Alg in the discussion below.

Algorithm 2.1. Input: $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ and outlier rate $\eta > 0$.

Output: Estimates $\hat{\mu}, \hat{\Sigma}$ of mean and covariance or "reject."

Operation:

1. Find a *witness* set of n points $X' \subseteq \mathbb{R}^d$ such that the uniform distribution on X' has subgaussian fourth moments and $|Y \cap X'| \geqslant (1-\eta)n$. Reject if no such X' exists.
2. Return the mean and covariance of X'.

Observe that the property of having subgaussian fourth moments requires verifying an inequality for every $v \in \mathbb{R}^d$, and, in general, there is no efficient (or even sub-exponential time) algorithm known (or expected, modulo the small-set expansion hypothesis) for this problem. Nevertheless, in [KS17b] (see Section 2), the authors prove that a variant of the above program (which we discuss this at the end of this overview) produces estimates that are guaranteed to be close to the mean and covariance of D if Y is an η-corruption of an i.i.d. sample X from D. Note that, though inefficient, such a result is sufficient to establish statistical identifiability of mean and covariance of D from $O_\eta(d^2)$ samples. The closeness guarantees in [KS17b] hold from a more general and basic result that is useful to us in this exposition, which we note below:

Fact 2.2 (See Section 2 of [KS17b], Parameter Closeness from Total Variation Closeness)**.** *Suppose D, D' are two distributions such that 1) both have subgaussian fourth moments and 2) the total variation distance between D, D' is at most β. Then, for every $v \in \mathbb{R}^d$, $\langle \mu(D') - \mu(D), v\rangle \leqslant O(\eta^{3/4})\sqrt{v^\top(\Sigma(D) + \Sigma(D'))v}$ and $v^\top(\Sigma(D') - \Sigma(D))v \leqslant O(\sqrt{\eta})v^\top(\Sigma(D) + \Sigma(D'))v$. We will say that the means (covariances, respectively) of D, D' are close to within $O(\eta^{3/4})$ ($O(\eta^{1/2})$, respectively) in Mahalanobis distance, to summarize such a guarantee.*

This fact effectively says that if two distributions both have bounded fourth moments and happen to be close in total variation distance, then their parameters (mean and covariance) must be close. In fact, the closeness is in strong *affine-invariant* norms—often called the Mahalanobis distance for mean and covariance.

2.2 Robustness Implies Weak Stability of Alg with a Randomized Outlier Rate

Let us now consider the stability of the above inefficient algorithm. We are seemingly in trouble at the outset: as written, there must be two adjacent Y, Y' such that Alg rejects on Y but not on Y'. Let us introduce our first simple idea and show how to patch the algorithm to prevent it from displaying such "drastic" change in its behavior.

Randomizing the outlier rate The following is a simple but useful observation: If Alg does not reject on input Y with outlier rate η, then, Alg must also not reject on Y' outlier rate $\eta + 1/n$. To see why, let X be the set of points with subgaussian fourth moments that intersects Y in $(1-\eta)n$ points. Then, since Y and Y' differ in at most one point, Y' must intersect X in at least $(1-\eta)n - 1 = (1-(\eta+1/n))n$ points. Thus, if, instead of a fixed outlier rate η, we ran Alg above with an appropriately "randomized" outlier rate, we might expect the rejection probabilities of Alg on Y, Y' to be similar. Such an argument can be made formal with a simple truncated Laplace noise injection procedure.

Robustness implies weak stability in Mahalanobis norms We now address the issue of whether the estimates computed on Y and Y' (assuming Alg does not reject on either of Y, Y') are close. We first observe that the fact that Alg is outlier-robust already guarantees a *weak stability* property. Specifically, suppose X, X' are the sets of size n generated by Alg when run on inputs Y, Y'. Then, since $Y \cap Y'$ is of size $n-1$, $|X \cap X'| \geqslant (1-2\eta)n - 1$. Next, observe that intersection bound above is equivalent to the uniform distributions on X, X' having a total variation distance of at most $2\eta + 1/n$. Thus, from Fact 2.2, we know that the parameters of X, X' are $O(\eta^{O(1)})$ close in the relative Mahalanobis distance defined above. Observe that this argument gives stability properties in the *right norms* directly! However, this is a weak stability guarantee since it only provides a fixed constant distance guarantee instead of $o_n(1)$ that one might expect given that Y and Y' differ in at most 1 out of n points. Nevertheless, our discussion shows that *robustness, via the inefficient algorithm above, immediately implies weak stability.*

2.3 A Simple Private Robust Mean Estimator from Weak Stability

Can we derive private algorithms from the weak stability guarantees? If the unknown covariance happens to be *spherical* (i.e., has all of its eigenvalues equal to each other), then the Mahalanobis distance guarantees are in fact equivalent (up to constant factor scaling) to Euclidean distance guarantees. As a result, simply adding Gaussian noise calibrated to the sensitivity bounds yields a private robust mean estimation algorithm! Indeed, 1) randomizing the outlier rate, 2) working with the SoS relaxation of the above program and 3) adding Gaussian noise to the resulting estimate, immediately yields a simple, straightforward private robust mean estimator that gives essentially optimal sample complexity guarantees (i.e., matching those of the known non-private robust estimators).

Weak stability is not enough for covariance estimation The challenge in using weak stability to obtain private robust covariance estimators arise when the covariance is non-spherical (e.g., is rank deficient or has eigenvalues of vastly different scales), in which case our Mahalanobis or multiplicative spectral stability guarantee does not translate into Euclidean/spectral norm distance guarantees. In particular, if we were to add Gaussian noise, we would end up *scrambling all small eigenvalues up* and end up with no non-trivial recovery guarantee.

Indeed, the aforementioned challenge necessitates a rethink of noise injection mechanisms for covariance estimation in general—standard noise addition mechanisms do not appear meaningful in faithfully preserving eigenvalues of different scales. Prior works (e.g., [KLSU19]) deal with this by iteratively computing some approximate preconditioning matrices. We have not investigated robust variants of their method. We instead explore *one-shot, blackbox* noise injection mechanisms that still provide us the right guarantees for covariance estimation.

2.4 Noise Injection in Estimate-Dependent Norms

If we wanted to faithfully preserve all eigenvalues (of varying scales) of the unknown covariance, a natural mechanism would be to add noise *linearly transformed with respect to the computed estimate*. For example, if $\hat{\Sigma}$ is the computed estimate, we would like to consider the mechanism that returns $\hat{\Sigma} + \hat{\Sigma}^{1/2} Z \hat{\Sigma}^{1/2}$ where Z is a matrix of random Gaussians. The upshot of such a mechanism is that it adds noise that is scaled relative to the eigenvalues of the estimate $\hat{\Sigma}$—directions where $v^\top \hat{\Sigma} v$ is small get a smaller additive noise as against directions where the same quadratic form is large.

However, the distribution of the added noise in this mechanism *depends on the non-privately estimated quantity itself*. Thus, *a priori*, it provides no useful privacy guarantee!

Key Observation: Nevertheless, our main idea to rescue the above plan is to note that the mechanism above does indeed provide meaningful privacy guarantees (by standard computations from the celebrated Gaussian mechanism) if we are able to guarantee that on any adjacent inputs Y, Y', the non-privately computed estimates are $o_n(1)$ close in relative Frobenius distance! This follows from elementary arguments and is presented in Lemmas 4.20 and 4.21.

The observation above crucially needs the distance between covariances (in relative Frobenius norm) to tend to 0 as $n \to \infty$; in fact, we need the rate to be inverse polynomial to achieve polynomial sample complexity. Our weak stability guarantee above, however, guarantees only a weak $O(\eta^{1/2})$ bound on multiplicative spectral distance which translates into a relative Frobenius bound of $O(\eta^{1/2}\sqrt{d})$—not only does this not tend to 0 as $n \to \infty$ but it, in fact, explodes as $d \to \infty$.

Thus, in order to use the above mechanism for covariance estimation, we must come up with significantly stronger (and asymptotically vanishing) stability guarantees. Let us investigate how to obtain such guarantees next.

2.5 Strong Stability for Robust Estimation Algorithms

Lack of stability because of multiple differing solutions There is an important barrier that prevents Alg from offering the strong stability guarantees we need in the covariance estimation mechanism above. Consider the case when Y is an i.i.d. sample from a one-dimensional standard

Gaussian distribution with mean 0 and variance 1 without any outliers added to it. Then, $\mathcal{N}(0, 1 \pm c\eta)$ for a small enough constant c is η-close in total variation distance to $\mathcal{N}(0, 1)$. By a straighforward argument, this implies that we can choose X' to be an i.i.d. sample of size n from $\mathcal{N}(0, 1 \pm c\eta)$— if n is large enough, then X' will have subgaussian fourth moments and will intersect Y in $(1-\eta)n$ points. The two difference distributions (and the corresponding samples X') however, have variances differing by an additive $O(\eta)$—a fixed constant independent of the sample size n. This shows that even in one dimension, Alg has feasible solutions with variance both $(1-O(\eta))$ and $1+O(\eta)$. Observe that this issue concerns the output of Alg itself, which can belong to a range that is significantly larger than what we can tolerate—we have not yet touched upon the issue of what happens when we change Y to an adjacent Y'.

Convexification and entropy surrogates In order to modify Alg to output a canonical solution (and with an eye for satisfying the stronger stability property), we wish to make the feasible solution space of Alg belong to a convex set (instead of the discrete set of solutions X' that intersect with Y in $(1-\eta)n$ points). With no fear of computational complexity, this is easy to do in a canonical way: we search instead for a *probability distribution* over X' that satisfy the constraints that Alg imposes. Unlike X', distributions on X' that satisfy the constraints are easily seen to form a convex set.

Given such a convex set, we can resolve our difficulty of not having canonical solutions for any given Y by simply finding a solution (i.e., a probability distribution ζ over X') that minimizes an appropriate strongly convex objective function. Specifically, for any X', let w_i be the 0-1 indicator of those indices i where $x_i = y_i$. Then, the constraints in Alg force $\sum_i w_i \geqslant (1-\eta)n$, and the distribution ζ can be thought to be over (X', w) in a natural way.

In order to ensure that Alg finds a canonical solution, a natural idea is to search over distributions ζ over (X', w) while minimizing some strongly convex function. We choose the simplest: $\|\mathbb{E}_\zeta[w]\|_2^2$ [1]. We think of this objective as a surrogate for finding "maximum entropy solutions" as, when viewing $\mathbb{E}_\zeta[w_i]$ as defining a probability distribution over y_i, minimizing the ℓ_2 norm favors "spread-out" or high entropy solutions. Since $\|\mathbb{E}_\zeta[w]\|_2^2$ is a convex function being minimized over convex set of expectations with respect to ζ, we expect that the minimizing solution $\mathbb{E}_{\zeta_*}[w]$ should be unique.

This is not immediately true, however, as our Alg as stated outputs the mean of X' (there could be "multiple" X' with the same intersection with Y, in principle).

Modifying the output of Alg In order to fit our framework better, we modify the above blueprint in Alg to instead output the weighted average of points in Y instead of X'. While such a procedure is not directly analyzed in [KS17b], the methods there can be naturally adapted without much hiccup. As a result we obtain the following modified version of Alg that we can now work with:

Algorithm 2.3. Input: $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ and an outlier rate $\eta > 0$.

[1] The exponent of the polynomials appearing in our sample complexity bounds improve if we use a strongly convex function with respect to 1-norm such as $\|x\|_q^2$ for $q = 1 + 1/\log d$. Our interest is in presenting a general "privatizing" blueprint so we continue with the simpler choice above in this work.

Output: Estimates $\hat{\mu}, \hat{\Sigma}$ of mean and covariance or "reject."

Operation:

1. Find a probability distribution ζ over a *witness* set of n points $X' \subseteq \mathbb{R}^d$ and intersection indicator $w \in \{0,1\}^n$ that minimizes $\|\mathbb{E}_\zeta[w]\|_2^2$ and is supported on (X', w) such that 1) the uniform distribution on X' has subgaussian fourth moments and 2) $\sum_i w_i \geqslant (1-\eta)n$. Reject if no such ζ exists.
2. Return $\hat{\mu} = \frac{1}{Z}\sum_i \mathbb{E}_\zeta[w_i] y_i$, $\hat{\Sigma} = \frac{1}{Z}\sum_i \mathbb{E}_\zeta[w_i](y_i - \hat{\mu})(y_i - \hat{\mu})^\top$ where $Z = \sum_i \mathbb{E}_\zeta[w_i]$.

With this modification, Alg outputs a canonical single solution on any given Y (or rejects).

Stability of Alg from the stability of the entropy potential We now return to the issue of stability. What happens if we switch the input Y of Alg above to Y'? The strongly convex objective we imposed in the above discussion comes in handy here! Namely, by basic convex analysis (see Proposition 3.20), it follows that if optimum entropy potential values of Alg on Y and Y' are say, $O(1)$-close, then, the vectors $\mathbb{E}_\zeta[w](Y)$ and $\mathbb{E}_\zeta[w](Y')$ are themselves $O(1)$ close. Recall that each $\mathbb{E}_\zeta[w_i]$ is a number in $[0,1]$ and that these numbers add up to 1. Hence, intuitively speaking, $O(1)$-closeness of $\|\mathbb{E}_\zeta[w]\|_2^2$ corresponds to constant perturbation in a constant number of coordinates.

Thus, working with the strongly convex objective above reduces our stability analysis of Alg to simply understanding how much can our entropy potential change when changing a single point in Y.

Unfortunately, this change can be large in general. $\|\mathbb{E}_\zeta[w]\|_2^2$ varies between $(1-\eta)n$ and $(1-\eta)^2 n$. The additive difference between these two extremes is $O(\eta n) \gg O(1)$.

Stabilizing the entropy potential: private stable selection Before describing our key idea, we first make a simple observation: Fix an input Y and consider the optimum value of the entropy potential of Alg when run with outlier rate η. What happens if we change η to $\eta + 1/n$? Clearly, the potential cannot *increase*: any solution ζ with outlier rate η is also a solution for outlier rate $\eta + 1/n$. The potential can decrease arbitrarily though.

More specifically, we show the following: in order to make the entropy potential stable under a change of Y to an adjacent Y', it is enough to run Y with an outlier rate $\eta' = O(\eta)$ such that the entropy potential of Alg on Y for any outlier rate in the interval $[\eta' - L/n, \eta' + L/n]$ is within an additive $\widetilde{O}(L/n)$ of any other.

To see why this claim could be true, informally speaking, observe that if Y' is obtained from Y by changing at most a single point, then a solution ζ with outlier rate η' can be modified into a solution ζ' for Y' with outlier rate $\eta' + 1/n$ by simplying zeroing out the w_i for the index i where Y' and Y differ. This allows us to relate the potentials for neighboring outlier rates on Y and Y'. Under the above assumption, the potential remains stable in an interval around η' on Y. This allows us to conclude that the same must be true for Y' for the interval $[\eta' - L/n + 1, \eta' + L/n - 1]$.

The above reasoning allows us to obtain strong stability guarantees if we can 1) show that a stable interval as above exists and 2) find such an interval via a stable process.

A stable selection procedure via the exponential mechanism We show that a stable interval as above (for $L = \widetilde{O}_n(1)$) exists via a simple Markov-like argument. Using an appropriate scoring rule, we show that the standard exponential mechanism can then be used to produce a stable interval like above via a stable algorithm (see Section 3.6.4).

Putting things together Altogether, we obtain a version of Alg that outputs a sequence of weights (i.e., $\mathbb{E}_\zeta[w_i]$) that are stable under the modification of a single point in Y. When viewed as a distribution on Y, the stability guarantee we obtain corresponds to an ℓ_1-stability of $\widetilde{O}(1/\sqrt{n})$ compared to the $O(\eta)$ (a fixed constant) stability that follows from any naive robust estimation algorithm.

We note that $\widetilde{O}(1/\sqrt{n})$ can be upgraded to $\widetilde{O}(1/n)$ if we work with a more sophisticated potential function $\|x\|_q^2$ for $q = 1 + 1/\log n$.

By applying Fact 2.2, we immediately get that if Alg does not reject on Y, Y', then the parameters of the respective inputs must be close in the Mahalanobis distance up to a *polynomially vanishing function* of n, as desired. This allows us to implement the estimate-dependent noise injection mechanism for covariance estimation!

We note that the discussion above can be formalized into an *information-theoretic private identifiability algorithm* (i.e., an inefficient private robust algorithm). We next discuss how to transform the above blueprint result into an efficient algorithm.

2.6 From Ideal Algorithms to Efficient Algorithms

Let us now go back and summarize 1) facts about the idealized inefficient algorithm and 2) our general blueprint for making such an algorithm Alg private.

1. **Witness Production:** We have used that the fact that Alg searches over witnesses X' that share the relevant property of the distributional model we have chosen (e.g., subgaussianity of fourth moments in the above discussion).

2. **Strongly Convex Entropy Potential:** We have minimized a strongly convex potential function in order to ensure that Alg outputs a canonical solution.

3. **Stable Outlier Rate Selection:** We have implemented a randomized stable selection scheme (via the exponential mechanism) for the outlier rate in order to argue that the optimum entropy potential of Alg is stable under the modification of a single point in the input Y.

We can apply this scheme to any algorithm that outputs a sequence of weights on the input sample Y, subject to the constraint that 1) the weights induce the relevant property of the distributional model, and 2) they minimize a strongly convex potential function.

Witness-producing SoS-based robust estimation algorithms It turns out that we can ensure all the above properties for *efficient* robust estimation algorithms based on "one-shot rounding" of

convex relaxations. We specifically rely on the algorithms for robust estimation based on SoS semidefinite programs in this work.

The SoS-based algorithms in the prior works that we use [BK20a, KS17b] almost fit our requirements except with two technical constraints:

1. The algorithms in the aforementioned prior works do not output weights on Y explicitly. However, we are able to show that a natural modification that outputs such weights on Y can be analyzed by the same methods.

2. The algorithms in the aforementioned prior works were analyzed under distributional assumptions on Y without the need to explicitly argue that the weights induce good witnesses (which we desire in our above analysis). Indeed, arguing that these algorithms produce such witnesses on worst-case datasets Y (whenever they don't reject) appears challenging. However, we are able to get by without such a statement by observing that we can adapt the analyses of the algorithms in the prior works to infer the following statement: if the algorithm returns a good witness on Y, then under a small perturbation of the parameters, it must also return a good witness on an adjacent Y'.

While verifying the properties makes our transformation not entirely blackbox at the moment, we strongly believe that our blueprint demonstrates a conceptually appealing connection between robust algorithm design and private algorithm design. Concretly, we expect our blueprint to be useful in designing more private (and robust) estimation algorithms. Indeed, we believe our techniques immediately extend to other problems where SoS-based robust estimation algorithms are known, such as linear regression [KKM18, BP20] and clustering spherical and non-spherical mixtures [DHKK20, BK20a, HL18, KS17c, FKP19].

3 Preliminaries

In this work, we will deal with algorithms that operate on numerical inputs. In all such cases, we will rely on the standard word RAM model of computation and assume that all the numbers are rational represented as a pair of integers describing the numerator and the denominator. In order to measure the running time of our algorithms, we will need to account for the length of the numbers that arise during the run of the algorithm. The following definition captures the size of the representations of rational numbers:

Definition 3.1 (Bit Complexity)**.** The bit complexity of an integer $p \in \mathbb{Z}$ is $1 + \lceil \log_2 p \rceil$. The bit complexity of a rational number p/q where $p, q \in \mathbb{Z}$ is the sum of the bit complexities of p and q.

For any finite set X of points in $\mathbb{R}^d$, we will use $\mu(X), \Sigma(X), M^{(t)}(X)$ to denote the mean, covariance and the t-th moment tensor of the uniform distribution on X.

3.1 Pseudo-Distributions

Pseudo-distributions are generalizations of probability distributions and form dual objects to sum-of-squares proofs in a precise sense that we will describe below.

Definition 3.2 (Pseudo-distribution, Pseudo-expectations, Pseudo-moments)**.** A *degree-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x) f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$. (Here, the summations are over the support of μ.)

The *pseudo-expectation* of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted $\widetilde{\mathbb{E}}_{D(x)} f(x)$, as

$$\widetilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x) \ . \tag{3.1}$$

In particular, the *mean* μ of a pseduo-distribution is defined naturally as the pseudo-expectation of $f(x) = x$, i.e., $\mu \widetilde{\mathbb{E}}_{D(x)} x$.

The degree-ℓ moment tensor of a pseudo-distribution μ is the tensor $\mathbb{E}_{\mu(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of every monomial of degree at most ℓ in x.

Observe that if a pseudo-distribution μ satisfies, in addition, that $\mu(x) \geqslant 0$ for every x, then it is a mass function of some probability distribution. Further, a straightforward polynomial-interpolation argument shows that every degree-∞ pseudo-distribution satisfies $\mu \geqslant 0$ and is thus an actual probability distribution. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree-d pseudo-distributions is also convex.

We now define what it means for $\widetilde{\mathbb{E}}$ to (approximately) satisfy constraints.

Definition 3.3 (Satisfying constraints)**.** For a polynomial g, we say that a degree-k $\widetilde{\mathbb{E}}$ satisfies the constraint $\{g = 0\}$ exactly if for every polynomial p of degree $\leqslant k - \deg(g)$, $\widetilde{\mathbb{E}}[pg] = 0$ and τ-approximately if $|\widetilde{\mathbb{E}}[pg_j]| \leqslant \tau \|p\|_2$. We say that $\widetilde{\mathbb{E}}$ satisfies the constraint $\{g \geqslant 0\}$ exactly if for every polynomial p of degree $\leqslant k/2 - \deg(g)/2$, it holds that $\widetilde{\mathbb{E}}[p^2 g] \geqslant 0$ and τ-approximately if $\widetilde{\mathbb{E}}[p^2 g] \geqslant -\tau \|p\|_2^2$.

The following fact describes the precise sense in which pseudo-distributions are duals to sum-of-squares proofs.

Fact 3.4 (Strong Duality, [JH16], see Theorem 3.70 in [FKP19] for an exposition)**.** *Let $p_1, p_2, \ldots, p_k$ be real-coefficient polynomials in $x_1, x_2, \ldots, x_n$. Suppose there is a degree-d sum-of-squares refutation of the system $\{p_i(x) \geqslant 0\}_{i \leqslant k}$. Then, there is no pseudo-distribution μ of degree $\geqslant d$ satisfying $\{p_i(x) \geqslant 0\}_{i \leqslant k}$. On the other hand, suppose that there is a pseudo-distribution μ of degree d consistent with $\{p_i(x) \geqslant 0\}_{i \leqslant k}$. Suppose further that the set $\{p_1, p_2, \ldots, p_k\}$ contains the quadratic polynomial $R - \sum_i x_i^2$ for some $R > 0$. Then, there is no degree-d sum-of-squares refutation of the system $\{p_i(x) \geqslant 0\}_{i \leqslant k}$.*

Basic sum-of-squares (SoS) proofs

Fact 3.5 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix with rational entries with numerators and denominators upper-bounded by 2^B and v be a vector in $\mathbb{R}^d$. Then, for every $\varepsilon \geqslant 0$,*

$$\vdash_2^v \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 + \varepsilon \right\}$$

The total bit complexity of the proof is $\mathrm{poly}(B, d, \log 1/\varepsilon)$.

Fact 3.6 (SoS Hölder's Inequality)**.** *Let f_i, g_i for $1 \leqslant i \leqslant s$ be indeterminates. Let p be an even positive integer. Then,*

$$\left|\frac{f,g}{p^2}\right. \left\{\left(\frac{1}{s}\sum_{i=1}^{s} f_i g_i^{p-1}\right)^p \leqslant \left(\frac{1}{s}\sum_{i=1}^{s} f_i^p\right)\left(\frac{1}{s}\sum_{i=1}^{s} g_i^p\right)^{p-1}\right\}.$$

The total bit complexity of the SoS proof is $s^{O(p)}$.

Observe that using $p = 2$ yields the SoS Cauchy-Schwarz inequality.

Fact 3.7 (SoS Almost Triangle Inequality)**.** *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then,*

$$\left|\frac{f_1,f_2,\ldots,f_r}{2t}\right. \left\{\left(\sum_{i\leqslant r} f_i\right)^{2t} \leqslant r^{2t-1}\left(\sum_{i=1}^{r} f_i^{2t}\right)\right\}.$$

The total bit complexity of the SoS proof is $r^{O(t)}$.

Fact 3.8 (SoS AM-GM Inequality, see Appendix A of [BKS15])**.** *Let $f_1, f_2, \ldots, f_m$ be indeterminates. Then,*

$$\left\{f_i \geqslant 0 \mid i \leqslant m\right\} \left|\frac{f_1,f_2,\ldots,f_m}{m}\right. \left\{\left(\frac{1}{m}\sum_{i=1}^{m} f_i\right)^m \geqslant \Pi_{i\leqslant m} f_i\right\}.$$

The total bit complexity of the SoS proof is $\exp(O(m))$.

We will also use the following two consequence of the SoS AM-GM inequality:

Proposition 3.9. *Let a, b be indeterminates. Then,*

$$\left|\frac{a,b}{2t}\right. \left\{a^{2j}b^{2t-2j} \leqslant ja^{2t} + (t-j)b^{2t}\right\}.$$

The total bit complexity of the SoS proof is $\exp(O(t))$.

Proof. We apply the SoS AM-GM inequality with $f_i = a^2$ for $i = 1, \ldots, j$ and $f_i = b^2$ for $i = j+1, \ldots, t$. We thus obtain:

$$\left|\frac{a,b}{2t}\right. \left\{(j/t\,a^2 + (1-j/t)b^2)^t \geqslant a^{2j}b^{2t-2j}\right\}$$

By the SoS Almost Triangle inequality, we have:

$$\left|\frac{a,b}{2t}\right. \left\{(j/t\,a^2 + (1-j/t)b^2)^t \leqslant (ja^{2t} + (t-j)b^{2t})\right\}$$

Combining the above two claims completes the proof. The total bit complexity of the SoS proof follows immediately by using the bounds for the two constituent inequalities used in the proof above. □

Proposition 3.10. *Let a, b be indeterminates. Then, for any positive integers i, t such that i is odd and $2t \geqslant i$, we have:*

$$\left|\frac{a,b}{2t}\right. \left\{a^i b^{2t-i} \leqslant \frac{1}{2}(a^{i-1}b^{2t-i+1} + a^{i+1}b^{2t-i-1})\right\}.$$

The total bit complexity of the SoS proof is $\exp(O(t))$.

Proof. Write $i = 2r - 1$ for some $r \geqslant 1$. Then, we have: $a^i b^{2t-1} = a^r b^{t-r} a^{r-1} b^{t-r+1}$. By the SoS AM-GM inequality with $f_1 = a^r b^{t-r}$ and $f_2 = a^{r-1} b^{t-r+1}$, we thus have:

$$\left| \frac{a,b}{2t} \left\{ a^i b^{2t-i} = a^r b^{t-r} a^{r-1} b^{t-r+1} \leqslant \frac{1}{2}(a^{i-1} b^{2t-i+1} + a^{i+1} b^{2t-i-1}) \right\} .$$

□

Fact 3.11 (Cancellation within SoS, Constant RHS [BK20b])**.** *Suppose A is indeterminate and $t \geqslant 1$. Then,*

$$\left\{ A^{2t} \leqslant 1 \right\} \left| \frac{A}{2t} \right. \left\{ A^2 \leqslant 1 \right\}$$

Further, the total bit complexity of the SoS proof is at most $2^{O(t)}$.

Lemma 3.12 (Cancellation within SoS [BK20b])**.** *Suppose A and C are indeterminates and $t \geqslant 1$. Then,*

$$\left\{ A \geqslant 0 \cup A^t \leqslant C A^{t-1} \right\} \left| \frac{A,C}{2t} \right. \left\{ A^{2t} \leqslant C^{2t} \right\} .$$

Further, the total bit complexity of the SoS proof is at most $2^{O(t)}$.

3.2 Algorithms and Numerical Accuracy

The following fact follows by using the ellipsoid algorithm for semidefinite programming. The resulting algorithm to compute pseudo-distributions approximately satisfying a given set of polynomial constraints is called the *sum-of-squares algorithm*.

Fact 3.13 (Computing pseudo-distributions consistent with a set of constraints [Sho87, Par00, Nes00, Las01])**.** *There is an algorithm with the following properties: The algorithm takes input $B \in \mathbb{N}$, $\tau > 0$, and polynomials $p_1, p_2, \ldots, p_k$ of degree ℓ with rational coefficients of bit complexity B. If there is a pseudo-distribution of degree d consistent with the constraints $\{p_i(x) \geqslant 0\}_{i \leqslant k}$, the algorithm in time $(Bn)^{O(d)} \operatorname{poly} \log(1/\tau)$ outputs a pseudo-distribution μ of degree d that τ-approximately satisfies $\{p_i(x) \geqslant 0\}_{i \leqslant k}$.*

3.3 Tensors

Since we will deal with higher moments of distributions, which are naturally represented as tensors, we will need to define some related notation and conventions for the sake of clarity in our exposition.

Let $[n] = \{1, 2, \ldots, n\}$ for any natural number n. We define the following.

Definition 3.14. Suppose we have an $m \times n$ matrix M and an $m' \times n'$ matrix N. We define $M \otimes N$ to be the standard $mm' \times nn'$ matrix given by the *Kronecker product* of M and N.

Moreover, for an $m \times n$ matrix M, we denote by $M^{\otimes t}$ the *t-fold Kronecker product* $\underbrace{M \otimes M \otimes \cdots \otimes M}_{t \text{ times}}$ (of dimension $tm \times tn$).

Given an $m \times n$ matrix M, we will also find it convenient to index $M^{\otimes t}$ as follows: for any $1 \leqslant i_1, i_2, \ldots, i_t \leqslant m$ and $1 \leqslant j_1, j_2, \ldots, j_t \leqslant n$, we can refer to the term $M^{\otimes t}_{(i_1, i_2, \ldots, i_t), (j_1, j_2, \ldots, j_t)} = \prod_{k=1}^{t} M_{i_t, j_t}$.

We also define a useful *flattening* operation on tensors:

Definition 3.15. Given an $m_1 \times m_2 \times \cdots \times m_t$ tensor M, we define the *flattening*, or *vectorization*, of M to be the $(m_1 m_2 \cdots m_t)$-dimensional vector, denoted $\mathrm{vec}(M)$, whose entries are precisely the entries of M appearing in the natural lexicographic order on $[m_1] \times [m_2] \times \cdots \times [m_t]$. In other words, the entry $M_{i_1,i_2,\ldots,i_t}$ appears before $M_{j_1,j_2,\ldots,j_t}$ (where $i_k, j_k \in [m_k]$ for $k = 1, 2, \ldots, t$) in $\mathrm{vec}(M)$ if and only if there exists some $1 \leqslant k \leqslant t$ such that $i_k < j_k$ and $i_l = j_l$ for all $l < k$.

Definition 3.16. Given an n-dimensional vector u and an $\underbrace{n \times n \times \cdots \times n}_{t \text{ times}}$-dimensional tensor M, we define $\langle u^{\otimes t}, M \rangle$ to be $\langle \mathrm{vec}(u^{\otimes t}), \mathrm{vec}(M) \rangle_{\mathbb{R}^d}$, i.e., the value of the standard inner product (on n^t-dimensional vectors) between the flattenings of $u^{\otimes t}$ and M.

A convenient fact we will use is a so-called "mixed product" property for matrices.

Fact 3.17. *Given an $m \times n$ matrix A, $m' \times n'$ matrix B, and $n \times n'$ matrix V, we have that*

$$AVB^T = (A \otimes B)\,\mathrm{vec}(V),$$

where the above is expressed as matrix-vector product.

Finally, we define the moment tensor for a probability distribution.

Definition 3.18. Given a probability distribution $\mathcal{D}$ on $\mathbb{R}^d$ and an integer $t > 1$, we define the t^{th} *moment tensor* M to be a $\underbrace{d \times d \times \cdots \times d}_{t \text{ times}}$ tensor whose entries are given by $M_{i_1,i_2,\ldots,i_t} = \mathbb{E}_{X \sim \mathcal{D}}[X_{i_1} X_{i_2} \cdots X_{i_t}]$ for $i_1, i_2, \ldots, i_t \in [d]$.

3.4 Basic Convexity

We will use the following basic propositions about convexity in our analysis.

Proposition 3.19 (Neighborhoods of minimizers of convex functions)**.** *Let K be a closed convex subset of $\mathbb{R}^N$. Let f be a smooth convex function on $\mathbb{R}^N$. Let x be a minimizer of f on K. Then, for every $y \in K$, $\langle y - x, \nabla f(x) \rangle \geqslant 0$.*

Proof. If not, then for a small enough positive λ, $f(x + \lambda(y - x)) < f(x)$. But, $x + \lambda(y - x) = (1 - \lambda)x + \lambda y \in K$. □

Proposition 3.20 (Pythagorean theorem from strong convexity w.r.t 2 norm)**.** *Let K be a convex subset of $\mathbb{R}^d$ for $d \in \mathbb{N}$. Let x be a minimizer of the convex function $f(x) = \|x\|_2^2$ on K. Let $y \in K$. Then, $f(y) - f(x) \geqslant \|y - x\|_2^2$.*

Proof. We have: $\|y\|_2^2 = \|y - x\|_2^2 + \|x\|_2^2 + 2\langle y - x, x \rangle$. The proposition follows by applying Proposition 3.19 to observe that $\langle y - x, x \rangle \geqslant 0$. □

We will also need the following basic bound:

Lemma 3.21. *Suppose $x, y \in [0,1]^n$ such that $\sum_i x_i, \sum_i y_i \geqslant n/2$ and $\|x - y\|_1 \leqslant \beta n$ for $\beta \leqslant 1/10$. Let $\bar{x} = \frac{x}{\|x\|_1}$ and $\bar{y} = \frac{y}{\bar{y}_1}$ be the normalized versions of x, y. Then,*

$$\|\bar{x} - \bar{y}\|_1 \leqslant 6\beta\,.$$

Proof. Suppose, without loss of generality, that $\|x\|_1 = c_1 n \geqslant c_2 n = \|y\|_1$ for $c_1, c_2 \geqslant 1/2$. Then, we know that $\|y\|_1 = c_2 n \geqslant (c_1 - \beta)n$. Thus, $\|\bar{x} - \bar{y}\|_1 \leqslant \frac{1}{c_1 c_2 n^2}(\|x\|y\|_1 - y\|x\|_1\|_1 \leqslant \frac{1}{c_1 c_2 n^2}(c_1 n\|x - y\|_1 + \beta n^2) \leqslant 6\beta$. □

3.5 Certifiable Subgaussianity

Definition 3.22 (Certifiable Subgaussianity)**.** A distribution D on $\mathbb{R}^d$ with mean μ_* is said to be $2k$-certifiably C-subgaussian if there is a degree $2k$ sum-of-squares proof of the following polynomial inequality in d-dimensional vector-valued indeterminate v:

$$\mathop{\mathbb{E}}_{x\sim D}\langle x - \mu_*, v\rangle^{2k} \leqslant (Ck)^k \left(\mathop{\mathbb{E}}_{x\sim D}\langle x - \mu_*, v\rangle^2\right)^k\,.$$

Furthermore, we say that D is certifiable C-subgaussian if it is $2k$-certifiably C-subgaussian for every $k \in \mathbb{N}$.

A finite set $X \subseteq \mathbb{R}^d$ is said to be $2k$-certifiable C-subgaussian if the uniform distribution on X is $2k$-certifiably C-subgaussian.

Fact 3.23 (Consequence of Theorem 1.2 in [KS17b])**.** *Let Y be a collection of n points in $\mathbb{R}^d$. Let $p, p' \in [0,1]^n$ be weight vectors satisfying $\|p\|_1, \|p'\|_1 = 1$, and $\|p - p'\|_1 = \tau$. Suppose that the distributions on Y where the probability of i is p_i (p'_i, respectively) is $2k$-certifiably C_1 (C_2, respectively) subgaussian. Let $\mu_p = \sum_i p_i y_i$, $\Sigma_p = \sum_i p_i (y_i - \mu_p)(y_i - \mu_p)^\top$, and $M_p^{(t)} = \sum_i p_i y_i^{\otimes t}$ for every $t \in \mathbb{N}$ be the mean, covariance and t-th moment tensor of distribution defined p. Define $\mu_{p'}, \Sigma_{p'}, M_{p'}^{(t)}$ similarly for the distribution corresponding to p'.*

Then, for every $\tau \leqslant \eta_0$ for some absolute constant η_0, for every $u \in \mathbb{R}^d$, $C' = C_1 + C_2$ and $t \leqslant k$:

$$\langle \mu_p - \mu_{p'}, u\rangle \leqslant \tau^{1-1/2k} \cdot O(\sqrt{Ck})\sqrt{u^\top \Sigma_p u})\,,$$

$$(1 - O(C'k)\tau^{1-1/k})\Sigma_p \preceq \Sigma_{p'} \preceq (1 + O(C'k)\tau^{1-1/k})\Sigma_{p'}\,,$$

$$(1 - O(C'^{t/2}k^{t/2})\tau^{1-t/2k})\langle u^{\otimes t}, M_p^{(t)}\rangle \leqslant \langle u^{\otimes t}, \hat{M}_{p'}^{(t)}\rangle \leqslant \langle u^{\otimes t}, M_p^{(t)}\rangle\,,$$

3.6 Differential Privacy

In this section, we state a few tools from differential privacy (DP) literature that will be used in our algorithms. We start by recalling the definition of DP:

Definition 3.24 (Differential Privacy [DMNS06])**.** An algorithm $\mathcal{M} : \mathcal{Y} \to \mathcal{O}$ is said to be (ε, δ)-*differentially private* (or (ε, δ)-*DP*) for $\varepsilon, \delta > 0$ iff, for every $S \subseteq \mathcal{O}$ and every neighboring datasets Y, Y', we have

$$\mathbb{P}[\mathcal{M}(Y) \in S] \leqslant e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(Y') \in S] + \delta.$$

Throughout this work, our set Y will consist of $y_1, \ldots, y_n \in \mathbb{R}^d$. $Y = (y_1, \ldots, y_n)$ and $Y' = (y'_1, \ldots, y'_n)$ are neighbors iff they differ on a single data point, i.e., $y'_j = y_j$ for all $j \neq i$. Note that this is the so-called *substitution* variant of DP; another popular variant is the *add/remove* DP where a neighboring Y' results from adding or removing an example from Y. We remark that it is not hard to extend our algorithm to the add/remove DP setting, by first computing a DP estimate $\hat{n}$ of n and either throwing away random elements or adding zero vectors to arrive at an n-size dataset on which our algorithm can be applied.

3.6.1 Laplace Mechanism and Its Variants

The Laplace mechanism [DMNS06] is among the most widely used mechanisms in differential privacy. It works by adding a noise drawn from the Laplace distribution (defined below) to the output of the function one wants to privatize.

Definition 3.25 (Laplace Distribution)**.** The Laplace distribution with mean μ and parameter b on $\mathbb{R}$, denoted by $\mathrm{Lap}(\mu, b)$, has the PDF $\frac{1}{2b} e^{-|x-\mu|/b}$.

We will also use the "truncated" version of the Laplace mechanism where the noise distribution is shifted and truncated to be non-negative. The precise definition of the noise distribution and its guarantee is given below. For completeness, we provide the DP analysis (Lemma 3.27) in Appendix A.1.

Definition 3.26 (Truncated Laplace Distribution)**.** The (negatively) truncated Laplace distribution with mean μ and parameter b, denoted by $\mathrm{tLap}(\mu, b)$ is defined as $\mathrm{Lap}(\mu, b)$ conditioned on the value being negative.

Lemma 3.27 (Truncated Laplace Mechanism)**.** *Let $f : \mathcal{Y} \to \mathbb{R}$ be any function with sensitivity at most Δ. Then the algorithm that adds* $\mathrm{tLap}\left(-\Delta\left(1 + \frac{\ln(1/\delta)}{\varepsilon}\right), \Delta/\varepsilon\right)$ *to f satisfies (ε, δ)-DP.*

Finally, we also state a bound on the tail probability of the truncated Laplace distribution which will be useful in our subsequent analysis.

Lemma 3.28. *Suppose $\mu < 0$ and $b > 0$. Let $X \sim \mathrm{tLap}(\mu, b)$. Then, for $y < \mu$, we have that*

$$\mathbb{P}[X < y] = \frac{e^{(y-\mu)/b}}{2 - e^{\mu/b}}.$$

3.6.2 Composition Theorem

It will be convenient to also consider DP algorithms whose privacy guarantee holds only against subsets of inputs. Specifically, we define:

Definition 3.29 (Differential Privacy Under Condition)**.** An algorithm $\mathcal{M} : \mathcal{Y} \to \mathcal{O}$ is said to be (ε, δ)*-differentially private under condition* Ψ (or (ε, δ)*-DP under condition* Ψ) for $\varepsilon, \delta > 0$ iff, for every $S \subseteq \mathcal{O}$ and every neighboring datasets Y, Y' both satisfying Ψ, we have

$$\mathbb{P}[\mathcal{M}(Y) \in S] \leqslant e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(Y') \in S] + \delta.$$

It is not hard to see that an analogue of the basic composition theorem still holds in this setting, which we formalize below. We remark that this is similar to the composition theorem derived in [DL09, Section 5]. However, since our composition theorem is slightly different, we provide its proof in Appendix A.2.

Lemma 3.30 (Composition for Algorithm with Halting)**.** *Let* $\mathcal{M}_1 : \mathcal{Y} \to \mathcal{O}_1 \cup \{\perp\}, \mathcal{M}_2 : \mathcal{O}_1 \times \mathcal{Y} \to \mathcal{O}_2 \cup \{\perp\}, \ldots, \mathcal{M}_k : \mathcal{O}_{k-1} \times \mathcal{Y} \to \mathcal{O}_k \cup \{\perp\}$ *be algorithms. Furthermore, let* $\mathcal{M}$ *denote the algorithm that proceeds as follows (with* o_0 *being empty): For* $i = 1, \ldots, k$*, compute* $o_i = \mathcal{M}_i(o_{i-1}, Y)$ *and, if* $o_i = \perp$*, halt and output* $\perp$*. Finally, if the algorithm has not halted, then output* o_k*.*

Suppose that:

- *For any* $1 \leqslant i < k$*, we say that* Y *satisfies the condition* Ψ_i *if running the algorithm on* Y *does not result in halting after applying* $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_i$*.*
- $\mathcal{M}_1$ *is* $(\varepsilon_1, \delta_1)$*-DP.*
- $\mathcal{M}_i$ *is* $(\varepsilon_i, \delta_i)$*-DP (with respect to neighboring datasets in the second argument) under condition* Ψ_{i-1} *for all* $i = \{2, \ldots, k\}$*.*

Then, $\mathcal{M}$ *is* $\left(\sum_{i\in[k]} \varepsilon_i, \sum_{i\in[k]} \delta_i\right)$*-DP.*

3.6.3 Hockey-Stick Divergence

It will be convenient in our analysis to use an equivalent definition of DP based on the *hockey-stick divergence*. For ease of notation, let $[a]_+ = \max\{a, 0\}$ for all $a \in \mathbb{R}$.

Definition 3.31 (Hockey-Stick Divergence)**.** Let $p(x), q(x)$ be probability density functions on $\mathbb{R}^d$, and α a non-negative real number. The Hockey-stick divergence $D_\alpha(p, q)$ between p, q is defined as:

$$D_{e^\varepsilon}(p, q) = \int_{x\in\mathbb{R}^d} [p(x) - \alpha \cdot q(x)]_+ dx\,.$$

The following fact is simple to derive from the definition of DP and is often used in literature.

[Pasin: We probably need the fact that $\mathcal{A}(Y)$ defines a PDF below? I.e. the hockey-stick divergence as defined above is not well-defined for discrete mechanisms. But maybe this is already clear from context?]

Fact 3.32 ((ε, δ)-DP from Hockey-Stick Divergence Bounds)**.** *Let* $\mathcal{M} : \mathcal{Y} \to \mathbb{R}^d$ *be a randomized algorithm.* $\mathcal{M}$ *is* (ε, δ)*-DP under condition* Ψ *iff for any neighboring pair of databases* Y, Y' *both satisfying* Ψ*, we have* $D_{e^\varepsilon}(\mathcal{M}(Y), \mathcal{M}(Y')) \leqslant \delta$*.*

We will need to bound the hockey-stick divergence between two distributions in terms of the hockey-stick divergences to a third distribution. Unfortunately, the hockey-stick divergence does not define a metric and, therefore, does not admit the usual triangle inequality. However, it is possible to prove a looser inequality, which we will find useful:

Lemma 3.33. *Suppose $p(x), q(x), r(x)$ are probability density functions on $\mathbb{R}^d$. Then,*

$$D_{e^\varepsilon}(p, r) \leqslant D_{e^{\varepsilon/2}}(p, q) + e^{\varepsilon/2} \cdot D_{e^{\varepsilon/2}}(q, r).$$

We remark that such a bound is already implicit in the so-called *group differential privacy* (see e.g. [Vad17, Lemma 2.2]). Nonetheless, we provide a (short) proof in Appendix A.3.

3.6.4 Approximate-DP Selection

Finally, we will also use a DP algorithm for the *selection* problem, where the goal is to pick from a (public) set of candidates one which has a high "score". This problem can be solved using the *exponential mechanism* [MT07]. The version of the algorithm we use deviates slightly from this traditional version in that we also include a check (via truncated Laplace mechanism) to make sure that the score is at least a certain threshold κ; otherwise, the algorithm's properties are summarized below. Its proof is deferred to Appendix A.4.

Theorem 3.34. *Suppose $\varepsilon, \delta \in (0, 1]$. Let $\mathcal{C}$ be a set of candidates and let* score $: \mathcal{C} \times \mathcal{Y}$ *be a scoring function for candidates as a function of the databases $Y \in \mathcal{Y}$, such that its sensitivity (w.r.t. Y) is at most Δ. There exists an algorithm Selection that satisfies the following properties:*

1. *Selection is (ε, δ)-DP.*
2. *If the output of Selection is $c^* \neq \perp$, then* $\text{score}(c^*, Y) \geqslant \kappa$.
3. *If there exists $c \in \mathcal{C}$ such that* $\text{score}(c, Y) \geqslant \kappa + O\left(\frac{\Delta}{\varepsilon} \cdot \log\left(\frac{|\mathcal{C}|}{\beta\delta}\right)\right)$, *then Selection output $\perp$ with probability at most β.*

4 Differentially Private Robust Moment Estimation

In this section, we describe a differentially private robust moment estimation algorithm. The following is our main technical result:

Theorem 4.1 (Differentially Private Robust Moment Estimation)**.** *Fix $C_0 > 0$ and $k \in \mathbb{N}$. Then, there exists an $\eta_0 > 0$ such that for any given outlier rate $0 < \eta \leqslant \eta_0$ and $\varepsilon, \delta > 0$, there exists a randomized algorithm Alg that takes an input of $n \geqslant n_0 = \widetilde{\Omega}\left(\frac{d^{4k}}{\eta^2}\left(1 + \left(\frac{\ln(1/\delta)}{\varepsilon}\right)^4 + \left(\frac{\ln(1/\delta)}{\varepsilon}\right)^{\frac{2k}{k-1}}\right) \cdot C^{4k} k^{4k+6}\right)$ points $Y \subseteq \mathbb{Q}^d$ (where $C = C_0 + \frac{3\ln(3/\delta)}{\varepsilon} + \frac{9}{\varepsilon} + 1$), runs in time $(Bn)^{O(k)}$ (where B is the bit complexity of the entries of Y) and outputs either "reject" or estimates $\hat{\mu} \in \mathbb{Q}^d$, $\hat{\Sigma} \in \mathbb{Q}^{d \times d}$, and $\hat{M}^{(t)} \in \mathbb{Q}^{d \times d \times \cdots \times d}$ (for all even $t < 2k$ such that t divides $2k$) with the following guarantees[2]:*

[2]The $\widetilde{\Omega}$ notation hides multiplicative logarithmic factors in $d, C, k, 1/\eta, 1/\varepsilon$, and $\ln(1/\delta)$.

1. ***Privacy:*** *Alg is (ε, δ)-differentially private with respect to the input Y, viewed as a d-dimensional database of n individuals.*

2. ***Utility:*** *Suppose there exists a $2k$-certifiably C_0-subgaussian set $X \subseteq \mathbb{Q}^d$ of $n \geqslant n_0$ points such that $|Y \cap X| \geqslant (1-\eta)n$ with mean μ_*, covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)} I$, and t-th moments $M_*^{(t)}$ for $2 \leqslant t \leqslant k$. Then, with probability at least $9/10$ over the random choices of the algorithm, Alg outputs estimates $\hat{\mu} \in \mathbb{Q}^d$, $\hat{\Sigma} \in \mathbb{Q}^{d \times d}$, and $M^{(t)} \in \mathbb{Q}^{d \times d \times \cdots \times d}$ (for all even $t < 2k$ such that t divides $2k$) satisfying the following guarantees:*

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u \rangle \leqslant O\left(\sqrt{Ck}\right) \eta^{1-1/2k} \sqrt{u^\top \Sigma_* u}\,,$$

and,

$$\left(1 - O((Ck)^{t/2k})\eta^{1-1/k}\right)\Sigma_* \preceq \hat{\Sigma} \preceq \left(1 + O((Ck)^{t/2k})\eta^{1-1/k}\right)\Sigma_*\,,$$

and, for every even $t < 2k$ such that t divides $2k$,

$$\left(1 - O(Ck)\eta^{1-t/2k}\right)\langle u^{\otimes t}, M_*^{(t)} \rangle \leqslant \langle u^{\otimes t}, \hat{M}^{(t)} \rangle \leqslant \left(1 + O(Ck)\eta^{1-t/2k}\right)\langle u^{\otimes t}, M_*^{(t)} \rangle\,.$$

Moreover, the algorithm succeeds (i.e., does not reject) with probability at least $9/10$ over the random choices of the algorithm.

Observe that the privacy guarantees of the algorithm are (necessarily) *worst-case*. The utility guarantees, however, hold only under the assumption that Y is an η-corruption of a good set X.

The above theorem can also be translated into utility guarantees for points sampled from a given distribution by recalling the well-known fact that points sampled from a certifiably subgaussian distribution are good with high probability:

Fact 4.2 (See Section 5 in [KS17b])**.** *Suppose $\mathcal{D}$ is a certifiably C-subgaussian distribution with mean μ_* and covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)} I$ and t-moment tensors $M^{(t)}$ for $t \in \mathbb{N}$. For any $k \in \mathbb{N}$, let $X = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample from $\mathcal{D}$ of size $n \geqslant n_0 = O(d^{2k}/\eta^2)$. Then, for any $t \in \mathbb{N}$ such that t divides k, with probability at least 0.99 over the draw of X, the following all hold:*

1. *X is $2k$-certifiably $2C$-subgaussian.*
2. *$\left\|\Sigma_*^{-1/2}(\mu(X) - \mu_*)\right\|_2 \leqslant \eta$.*
3. *$\Sigma(X) \in (1 \pm \eta)\Sigma_*$.*
4. *$\left|\frac{v}{2k}\right\rangle \left\{\langle v^{\otimes t}, M^{(t)}(X)\rangle \in (1 \pm \eta)\langle v^{\otimes t}, M_*^{(t)}\rangle\right\}$.*

We note that our main theorem for private robust moment estimation, Theorem 1.2, is an immediate consequence of Theorem 4.1 and Fact 4.2.

For the rest of the section, we will work to prove Theorem 4.1. In Section 4.1, we will introduce a witness-producing robust moment estimation algorithm that will be used as a subroutine for our main algorithm and present relevant utility guarantees. In Section 4.2, we will then introduce our main algorithm. After that, we will prove the necessary privacy guarantees in Section 4.3. Finally, we will put together the pieces to prove our main theorem, Theorem 4.1, in Section 4.4.

4.1 Witness-Producing Version of Robust Moment Estimation Algorithm

As a key building block, we will use the following (non-private) version of the robust moment estimation algorithm of [KS17b] that uses the same constraint system $\mathcal{A}$ as in [KS17b]. Our algorithm itself, however, makes one key change (we call our version "witness-producing" for reasons that will soon become clear) to that of [KS17b] in order to obtain a private robust moment estimation algorithm. Instead of outputting estimates of the moments of the unknown distribution, our algorithm outputs a sequence of non-negative weights $p_1, p_2, \ldots, p_n$ forming a probability distribution on the input set of points Y. The estimates can then be obtained by taking moments of the finite set Y with respect to the probability distribution on Y defined by the weights p_is. This simple change is crucial to our *worst-case* analysis of the resulting algorithm (i.e. even when the distributional assumption that Y is an η-corruption of some good set X is not met) and obtaining our privacy guarantees. As we discuss, our blueprint for modifying convex optimization based robust estimation algorithms appears to broadly applicable beyond the specific setting of robust moment estimation.

The underlying constraint system $\mathcal{A}$ is shown below, and the witness-producing robust moment estimation algorithm is shown as Algorithm 4.3.

$\mathcal{A}_{C,k,\eta,n}(\{y_1, y_2, \ldots, y_n\})$**: Constraint System for η-Robust Moment Estimation**

1. $w_i^2 = w_i$ for each $1 \leqslant i \leqslant n$,
2. $\sum_{i=1}^n w_i \geqslant (1-\eta)n$,
3. $\mu' = \frac{1}{n}\sum_i x'_i$,
4. $w_i(x'_i - y_i) = 0$ for $1 \leqslant i \leqslant n$,
5. $\frac{1}{n}\sum_{i=1}^n \langle x'_i - \mu', v\rangle^k \leqslant (Ck)^{k/2}\left(\frac{1}{n}\sum_{i=1}^n \langle x'_i - \mu', v\rangle^2\right)^{k/2}$.

Algorithm 4.3 (Witness-Producing Robust Moment Estimation)**.**

Given: A set of points $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$, $\eta > 0$, a parameter $k \in \mathbb{N}$.

Output: Either "reject" or non-negative weights $p_1, p_2, \ldots, p_n$ s.t. $p_i \leqslant \frac{1}{(1-\eta)n}\ \forall i$ and $\sum_i p_i = 1$.

Operation:

1. Find a pseudo-distribution $\tilde{\zeta}$ of degree $O(k)$ ([Ameya: Is this $2k$?]) satisfying the constraint system $\mathcal{A}_{C,k,\eta,n}(Y)$. If such a pseudo-distribution does not exist, then return "reject."
2. Output weights $p \in [0,1]^n$ defined by $p_i = \frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]}{\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]}$ for each i.

Analysis of the witness-producing robust estimation algorithm Robust estimation algorithms that rely on the use of semidefinite programming are all analyzed under distributional assumptions on the input set of points. Roughly speaking, such algorithms search over set of points that have a large enough intersection with the input corrupted sample and satisfy certain relevant property of the underlying family of distributions. In order to obtain privacy guarantee that holds for worst-case inputs, we need to upgrade the analyses of such algorithms so that they not only provide estimates of the target parameters, but also explicitly produce "witnesses"—these are subsets of the input corrupted sample that define distributions with the estimated parameters and further, satisfy the relevant property of the underlying family of distributions.

In this section, we verify that such a stronger guarantee can be obtained for robust moment estimation algorithm of [KS17b]. Formally, their algorithm succeeds as long as the input is an η-corruption of a certifiably subgaussian set.

The following guarantees for the algorithm above were shown in [KS17b].

Fact 4.4 (Lemmas 4.4, 4.5, and 4.8 in [KS17b])**.** *Let $X \subseteq \mathbb{R}^d$ be a set of size n that is $2k$-certifiably C-subgaussian with mean μ_*, covariance Σ_* and t-th moment $M_*^{(t)}$ for t evenly dividing $2k$. Let Y be an η-corruption of X. Then, for $\mu' = \frac{1}{n}\sum_i x'_i$, $\Sigma' = \frac{1}{n}\sum_i (x_i - \mu')(x_i - \mu')^\top$, and $M^{(t)'} = \frac{1}{n}\sum_i x_i'^{\otimes t}$, we have:*

$$\mathcal{A} \left|\frac{u}{2k}\right. \left\{\langle \mu' - \mu_*, u\rangle^{2k} \leqslant O(C^k k^k) u^\top \Sigma_* u^k\right\},$$

$$\mathcal{A} \left|\frac{u}{2k}\right. \left\{\langle \Sigma' - \Sigma_*, u^{\otimes 2}\rangle^{k} \leqslant O(C^k k^k) u^\top \Sigma_* u^k\right\},$$

$$\mathcal{A} \left|\frac{u}{2k}\right. \left\{\langle M^{(t)'} - M_*^{(t)}, u^{\otimes t}\rangle^{2k/t} \leqslant O(C^k k^k) u^\top \Sigma_* u^k\right\}.$$

Lemma 4.5 (Guarantees for Witness-Producing Robust Moment Estimation Algorithm)**.** *Given a subset of of n points $Y \subseteq \mathbb{Q}^d$ whose entries have bit complexity B, Algorithm 4.3 runs in time $(Bn)^{O(k)}$ and either (a.) outputs "reject," or (b.) returns a sequence of weights $0 \leqslant p_1, p_2, \ldots, p_n$ satisfying $p_1 + p_2 + \cdots + p_n = 1$.*

Moreover, if $X \subseteq \mathbb{R}^d$ is $2k$-certifiably C-subgaussian with mean μ_*, covariance Σ_* and in general, t-th moment tensor $M^{(t)_*}$ such that $|Y \cap X| \geqslant (1-\eta)n$, then Algorithm 4.3 never rejects, and the corresponding estimates $\hat{\mu} = \frac{1}{n}\sum_i p_i y_i$ and $\hat{\Sigma} = \sum_{i=1}^n p_i (y_i - \hat{\mu})(y_i - \hat{\mu})^\top$ satisfy the following guarantees for $\beta_t = O(C^{t/2} k^{t/2}) \eta^{1-t/2k}$ for $t \leqslant k$:

1. ***Mean Estimation:***

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u\rangle \leqslant O(\sqrt{Ck})\eta^{1-1/2k}\sqrt{u^\top \Sigma_* u},$$

2. ***Covariance Estimation:***

$$(1-\beta_2)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\beta_2)\Sigma_*,$$

3. ***Moment Estimation:*** *For all even $t < 2k$ such that t divides $2k$,*

$$\forall u \in \mathbb{R}^d,\ (1-\beta_t)\langle u^{\otimes t}, M_*^{(t)}\rangle \leqslant \langle u^{\otimes t}, \hat{M}^{(t)}\rangle \leqslant (1+\beta_t)\langle u^{\otimes t}, M_*^{(t)}\rangle$$

4. ***Witness:*** *For $C' \leqslant C(1 + O(\eta^{1-1/k}))$,*

$$\vdash \left\{ \frac{1}{n}\sum_{i=1}^{n} p_i \langle y_i - \hat{\mu}\rangle^{2k} \leqslant (C'k)^k \left(\frac{1}{n}\sum_{i=1}^{n} p_i \langle y_i - \hat{\mu}\rangle^2\right)^k \right\}$$

The first three properties follow easily from an analysis similar to the one in [KS17b]. We verify the last property below.

Lemma 4.6. *Let $\tilde{\zeta}$ be a pseudo-distribution of degree $O(k)$ consistent with $\mathcal{A}$ on input Y with outlier rate $\eta \ll 1/k$. Suppose there exists a $2k$-certifiably C_1-subgaussian distribution $X \subseteq \mathbb{R}^d$ with mean μ_* of size n such that $|Y \cap X| \geqslant (1-\eta)n$. Then, for $\eta \leqslant \eta_0$ for some absolute constant η_0 and for $\hat{\mu} = \frac{1}{W}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] y_i$ where $W = \sum_{i=1}^{n} \widetilde{\mathbb{E}}[w_i]$, we have:*

$$\left|\frac{u}{2k}\right. \left\{ \frac{1}{W}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]\langle y_i - \hat{\mu}, u\rangle^{2k} \leqslant (C'k)^k \left(\frac{1}{W}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]\langle y_i - \hat{\mu}, u\rangle^2\right)^k \right\},$$

for $C' \leqslant C(1 + O(\eta^{1-1/2k})k) \leqslant C + 1$ for small enough η.

Proof. We have:

$$\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]\langle y_i - \hat{\mu}, u\rangle\right)^{2k} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i \langle x_i' - \hat{\mu}, u\rangle^{2k}] \leqslant \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu' + \mu' - \hat{\mu}, u\rangle^{2k}]$$

The first term on the right-hand side above is at most $(Ck)^k \widetilde{\mathbb{E}}_{\tilde{\zeta}}[(\frac{1}{n}\sum_{i=1}^{n}\langle x_i' - \mu', u\rangle^2)^k] \leqslant (C(1 + O(\eta^{1-1/2k})k)^k u^\top \Sigma_* u^k$ using certifiable subgaussianity constraints and Fact 5.2.

Let us analyze the 2nd term above.

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu' + \mu' - \hat{\mu}, u\rangle^{2k}] \tag{4.1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k}] + 2k\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k-2}\langle \mu' - \hat{\mu}, u\rangle^2 \tag{4.2}$$

$$+ \sum_{j=2}^{2k} \binom{2k}{j} \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k-j}\langle \mu' - \hat{\mu}, u\rangle^j] \tag{4.3}$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k}] + 2k\frac{1}{n}\sum_{i=1}^{n} (\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k})^{(2k-2)/2k} (\widetilde{\mathbb{E}}_{\tilde{\zeta}}\langle \mu' - \hat{\mu}, u\rangle^{2k})^{1/2k} \tag{4.4}$$

$$+ \sum_{j=2}^{2k} \binom{2k}{j} \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k-j}\langle \mu' - \hat{\mu}, u\rangle^j] \tag{4.5}$$

Here, in the 2nd inequality, we used the Hölder's inequality for pseudo-distributions. Let us analyze the 2nd term in the right-hand side above by observing the following that uses the bounds from Fact 5.2:

$$\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle \mu' - \hat{\mu}, u\rangle^{2k}] \leqslant 2^{2k}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle \mu' - \mu_*, u\rangle^{2k}] + 2^{2k}\langle \mu_* - \hat{\mu}, u\rangle^{2k} \tag{4.6}$$

$$\leqslant 2^{2k}(Ck)^k \eta^{2k-1} u^\top \Sigma_* u^k + 2^{2k}(Ck)^k \eta^{2k-1}(1+\beta_2)^k u^\top \Sigma_* u^k \tag{4.7}$$

This allows us to infer that the 2nd term in (4.5) is at most $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\frac{1}{n}\sum_{i=1}^n \langle x_i' - \mu', u\rangle^{2k}]^{(2k-2)/2k} \cdot (5Ck)^{1/2}\eta^{1-1/2k}\sqrt{u^\top \Sigma_* u} \leqslant O(k)(Ck)^k \eta^{1-1/2k} u^\top \Sigma_* u^k$ using certifiable subgaussianity constraints and Fact 5.2.

Let's now analyze the terms corresponding to $j \geqslant 2$ in the right-hand side of (4.5). Each of these terms corresponds to a "mixed monomial" in $\langle x_i' - \mu', u\rangle$ and $\langle \mu' - \mu, u\rangle$. Let us first analyze the even individual degree terms.

First observe that by Hölder's inequality for pseudo-distributions again, we have:

$$\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k-2}\langle \mu' - \hat{\mu}, u\rangle^2] \leqslant \langle x_i' - \mu', u\rangle^{2k})^{(k-1)/k}(\widetilde{\mathbb{E}}_{\tilde{\zeta}}\langle \mu' - \hat{\mu}, u\rangle^{2k})^{1/k}. \tag{4.8}$$

By an analysis similar to the case of the first term on the right-hand side of (4.5) above, we obtain that the right-hand side is at most: $O(1)(Ck)^k(\eta^{1-1/2k})^2 u^\top \Sigma_* u^k$.

Next, let's analyze all terms corresponding to even j. By Proposition 3.9, we have:

$$\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x_i' - \mu', u\rangle^{2k-2j}\langle \mu' - \hat{\mu}, u\rangle^{2j}] \leqslant \frac{1}{n}\sum_{i=1}^n (\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle \mu' - \hat{\mu}, u\rangle^2\langle x_i' - \mu', u\rangle^{2k-2j}\langle \mu' - \hat{\mu}, u\rangle^{2j-2}]$$

$$\leqslant \frac{2k}{n}\sum_{i=1}^n (\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle \mu' - \hat{\mu}, u\rangle^2(\langle x_i' - \mu', u\rangle^{2k-2} + \langle \mu' - \hat{\mu}, u\rangle^{2k-2})]$$

The first term can now be upper bounded by the bound for (4.8) and the 2nd term by an application of Fact 5.2.

The case of odd terms is similar with the first step using Proposition 3.10.

Altogether, we obtain an upper bound of $(C(1+O(\eta^{1-1/2k})k)^k u^\top \Sigma_* u^k$.

On the other hand, using the sum-of-squares version of the Cauchy-Schwarz inequality along with the almost triangle inequality and invoking Fact 5.2 we have:

$$\mathcal{A}\vdash_{\frac{u}{2k}} \left\{\left(\frac{1}{n}\sum_{i=1}^n (1-w_i)\langle x_i' - \hat{\mu}, u\rangle^2\right)^2 \leqslant \left(\frac{1}{n}\sum_{i=1}^n (1-w_i)^2\right)\frac{1}{n}\sum_{i=1}^n \langle x_i' - \hat{\mu}, u\rangle^4\right.$$

$$\left.\leqslant 16\eta C^2\left(\frac{1}{n}\sum_{i=1}^n \langle x_i' - \mu', u\rangle^2 + \frac{1}{n}\sum_{i=1}^n \langle \mu' - \hat{\mu}, u\rangle^2\right)^2 \leqslant 20\eta C^2(1+\beta_2)^2 u^\top \Sigma_* u^2\right\}$$

Thus,

$$\mathcal{A}\vdash_{\frac{u}{2k}} \left\{\left(\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i\langle y_i - \hat{\mu}, u\rangle^2\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle y_i - \hat{\mu}, u\rangle^2\right)^2] - \left(\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[(1-w_i)\langle y_i - \hat{\mu}, u\rangle^2\right)^2 n$$

$$\geqslant (1 - O(Ck)\eta^{1-1/k} - 80\eta C^2)u^\top \Sigma_* u \,.\Big\}$$

The lemma now follows immediately for small enough fixed constant η.

□

4.2 Private Robust Moment Estimation

We are now ready to present our main algorithm for private robust moment estimation. Our algorithm uses the witness-producing algorithm (Algorithm 4.3) as a major building block while augmenting it to search for pseudo-distributions that, in addition to satisfying the relevant set of constraints, also minimize an appropriate strongly convex potential function. We define the relevant potential function Pot below in Definition 4.7.

Definition 4.7 (Potential Function)**.** Let $C > 0$ and $n, k \in \mathbb{N}$. For any pseudo-distribution $\tilde{\zeta}$ of degree 2 consistent with $\mathcal{A}_{C,k,\eta,n}(Y)$ for outlier rate η and input $Y \subseteq \mathbb{R}^d$, let $\mathrm{Pot}^{C,k,n}_{\eta,\tilde{\zeta}}(Y)$ be defined as $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2$. Furthermore, let $\mathrm{Pot}^{C,k,n}_{\eta}(Y) = \min_{\tilde{\zeta} \text{ sat } \mathcal{A}_{C,k,\eta,n}(Y)} \mathrm{Pot}^{C,k,n}_{\eta,\tilde{\zeta}}(Y)$ be the minimum value of the potential as $\tilde{\zeta}$ ranges over all pseudo-distributions of degree $2t$ consistent with $\mathcal{A}_{C,k,\eta,n}(Y)$. If no such pseudo-distribution exists, set $\mathrm{Pot}_\eta(Y) = \infty$.

When C, n, k are understood from context, we may suppress these parameters and simply write Pot_η and $\mathrm{Pot}_{\eta,\tilde{\zeta}}$.

Now, we are ready to describe our main private robust moment algorithm, which is listed as Algorithm 4.8. The algorithm consists of three main steps. In the first step, the randomized DP selection algorithm (Theorem 3.34) is used to pick an outlier rate (according to a suitable scoring function, as defined below in Definition 4.12). The second step invokes the witness-producing algorithm (Algorithm 4.3) with the outlier rate chosen in step 1, after which one checks that the outputted weights induce a certifiably subgaussian distribution on the input dataset Y. Finally, in the last step, one takes the estimates of the mean, covariance, and higher moments provided by the resulting weight vector and adds suitable noise to guarantee differential privacy.

Algorithm 4.8 (Private Robust Moment Estimation)**.**

Given: A set of points $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$, parameters $C, \eta, \varepsilon, \delta > 0$, $L, k \in \mathbb{N}$.

Output: Estimates $\hat{\mu}$, $\hat{\Sigma}$, and $\hat{M}^{(t)}$ $(3 \leqslant t \leqslant k)$ for mean, covariance, and t-moments.

Operation:

1. **Stable Outlier Rate Selection:** Use the $(\varepsilon/3, \delta/3)$-DP Selection with $\kappa = L/2$ to sample an integer $\tau \in [\eta n]$ with the scoring function as defined in Definition 4.12. If $\tau = \perp$, then reject and halt. Otherwise, let $\eta' = \tau/n$.

2. **Witness Checking:** Compute a pseudo-distribution $\tilde{\zeta}$ of degree $2k$ satisfying $\mathcal{A}_{C,k,\eta',n}(Y)$ and minimizing $\mathrm{Pot}_{\eta',\tilde{\zeta}}(Y)$. Let $\gamma \sim \mathrm{tLap}\left(-\left(1+\frac{3\ln(3/\delta)}{\varepsilon}\right), 3/\varepsilon\right)$ and $C' = C + \gamma$. Check that the weight vector $p = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]$ induces a C'-certifiably subgaussian distribution on Y. If not, reject immediately. Otherwise, let $\widetilde{\mu} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\mu]$, $\widetilde{\Sigma} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\Sigma]$, and $\widetilde{M}^{(t)} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[M^{(t)}]$ (for all even $t < 2k$ such that t divides $2k$) be the mean, covariance, and t^{th} moment estimates, respectively, that are induced by the pseudo-distribution $\tilde{\zeta}$.
3. **Noise Addition:** Let $\gamma_1 = O(C'k)(L/n)^{\frac{1}{2}\left(1-\frac{1}{2k}\right)}$ and $\gamma_t = O((C'k)^{t/2})(L/n)^{\frac{1}{2}\left(1-\frac{t}{2k}\right)}$ for $t \geqslant 2$. Let $z \sim \mathcal{N}(0,\sigma_1)^d$ and $Z \sim \mathcal{N}(0,\sigma_2)^{\binom{d+1}{2}}$, where we interpret Z as a symmetric $d \times d$ matrix with i.i.d. entries in the upper triangular portion. Similarly, for $t \geqslant 2$, let $Z^{(t)} \sim \mathcal{N}(0,\sigma_t)^{\binom{d+(t-1)}{t}}$, where we interpret Z as a symmetric $\underbrace{d \times d \times \cdots d}_{t \text{ times}}$ tensor with $\binom{d+(t-1)}{t}$ independent "upper-triangular" entries. Moreover, let

$$\begin{cases} \sigma_j = 6k\varepsilon^{-1}\gamma_j d^{\frac{t-1}{2}}\sqrt{2\ln(7.5k/\delta)}, & \text{for } j = 1,2 \\ \sigma_j = 6k\varepsilon^{-1}\gamma_j (C'k)^t d^{\frac{t-1}{2}}\sqrt{2\ln(7.5k/\delta)}, & \text{for } j > 2 \end{cases}.$$

Then, output:

- $\hat{\mu} = \widetilde{\mu} + \widetilde{\Sigma}^{1/2} z$.
- $\hat{\Sigma} = \widetilde{\Sigma} + \widetilde{\Sigma}^{1/2} Z \widetilde{\Sigma}^{1/2}$.
- $\hat{M}^{(t)} = \widetilde{M}^{(t)} + ((\widetilde{\Sigma} + \widetilde{\mu}\widetilde{\mu}^T)^{1/2})^{\otimes t} Z^{(t)}$, for all even $t < 2k$ such that t divides $2k$.

4.3 Privacy Analysis

Our analysis of the privacy of Algis based on a sequence of claims about each of the steps of Algthat cumulatively establish the stability of the behavior of Algon adjacent inputs Y, Y'. We will rely on the following simple but key observation in our analysis. It is easy to verify using the definition of pseudo-distributions.

Lemma 4.9 (Adjacent Pseudo-distributions)**.** *Let $\tilde{\zeta}$ be a pseudo-distribution of degree $2k$ that satisfies all the constraints in 5 on input $Y = \{y_1, y_2, \ldots, y_n\}$ with outlier rate η. Let $Y' \subseteq \mathbb{R}^d$ be adjacent to Y. Define an* adjacent pseudo-distribution *$\tilde{\zeta}'$ (that "zeroes out w_i") by $\widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w_S p(X', \cdots)] = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S p(X', \cdots)]$ if $i \notin S$ and $\widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w_S p(X', \cdots)] = 0$ if $i \in S$ for every polynomial p in X' and other auxiliary indeterminates in $\mathcal{A}$. Then, $\tilde{\zeta}'$ is a pseudo-distribution of degree $2k$ that satisfies all the constraints in 5 on both inputs Y' and Y with outlier parameter $\eta + 1/n$.*

This allows us to conclude the following basic calculus of our potential function:

Lemma 4.10 (Basic Facts about Pot)**.** *Suppose that for some $Y \subseteq \mathbb{R}^d$ of size n, some $t \in \mathbb{N}$ and $\eta' \in [0, \eta_0/4]$, there is a pseudo-distribution of degree $2t$ consistent with $\mathcal{A}$ on input Y. Then, for every $\eta \geqslant \eta'$, the following holds:*

1. ***Monotonicity:*** $\mathrm{Pot}_{\eta+1/n}(Y) \leqslant \mathrm{Pot}_{\eta}(Y)$. *In particular,* Pot *is monotonically decreasing as its subscript increases.*

2. ***Lower Bound:*** $\mathrm{Pot}_{\eta}(Y) \geqslant (1-\eta)^2 n$.

3. ***Upper Bound:*** $\mathrm{Pot}_{\eta}(Y) \leqslant (1-\eta)n$.

Proof. The first fact follows immediately from Lemma 4.9. For the second, observe that any pseudo-distribution $\tilde{\zeta}$ of degree $2t$ consistent with $\mathcal{A}$ on input Y with outlier rate η must satisfy $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \geqslant (1-\eta)n$. Thus, by Cauchy-Schwarz inequality, $\sum_{i=1}^n \widetilde{\mathbb{E}}[w_i]^2 \geqslant (\sum_{i=1}^n \widetilde{\mathbb{E}}[w_i])^2/n = (1-\eta)^2 n$. This completes the proof. For the last part, observe that $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \leqslant 1$ for every i. Thus, $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]^2 \leqslant \sum_{i=1}^n \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] = (1-\eta)n$. □

Analysis of stable outlier rate selection The goal of the first step of Algis to find an outlier rate η' such that the strongly convex potential function $\mathrm{Pot}(\tilde{\zeta})$ on the pseudo-distribution we will eventually compute (in Step 3) is close on adjacent input points Y, Y'. We will later use the strong convexity of the Pot and the closeness guarantee on Pot on Y, Y' to infer that the weight vector $p(Y)$ and $p(Y')$ output by the algorithm themselves are close.

Our key algorithmic trick to ensure the closeness of the strongly convex potential Pot is to find a "stable interval" $[\eta' - 0.5L/n, \eta' + 0.5L/n]$ of outlier rates η'' such that strongly convex potential function at near-optimal solutions must vary slowly as η'' varies in the the interval. We find such an interval via a variant of the exponential mechanism.

Definition 4.11 (Stability)**.** Fix $L \in \mathbb{N}$. Let $\tau, \gamma \in \{0, \ldots, n\}$ such that $\gamma \leqslant \tau, n-\tau$. Suppose for some $Y \subseteq \mathbb{R}^d$ of size n, the constraint system $\mathcal{A}((\tau-\gamma)/n)$ is feasible. We define the stability of the 2γ length interval centered at τ to be

$$\mathrm{stab}_Y(\tau, \gamma) = \mathrm{Pot}_{(\tau-\gamma)/n}(Y) - \mathrm{Pot}_{(\tau+\gamma)/n}(Y)$$

Observe that if there is a pseudo-distribution consistent with $\mathcal{A}$ on Y with outlier rate $(\tau-\gamma)/n$ then there is a pseudo-distribution consistent with $\mathcal{A}$ on Y with any outlier rate $\geqslant (\tau-\gamma)/n$. Thus, stability above is well-defined.

Definition 4.12 (Score Function)**.** Fix $n, k \in \mathbb{N}$ and $C > 0$. Let $Y \subseteq \mathbb{R}^d$ be a set of size n. For a parameter L, we define the following score function for every integer $\tau \in [n]$:

$$\mathrm{score}_{n,C,k}(\tau, Y) = \begin{cases} 0 & \text{if } \mathsf{Alg}(Y, \tau/n) \text{ is infeasible,} \\ \max_{\substack{\gamma \\ \mathcal{A}_{C,k,(\tau-\gamma)/n,n}(Y) \text{ is feasible}}} \min\{\gamma, 20L - \mathrm{stab}_Y(\tau,\gamma)\} & \text{otherwise.} \end{cases}$$

In the second case, we define $\gamma^*_Y(\tau) := \arg\max_{\substack{\gamma \\ \mathcal{A}_{C,k,(\tau-\gamma)/n,n}(Y) \text{ is feasible}}} \min\{\gamma, L - \mathrm{stab}_Y(\tau,\gamma)\}$.

Lemma 4.13. *Let $\tau, \gamma \in [n]$ such that $\gamma \leqslant \tau, n-\tau$. Suppose for some $Y \subseteq \mathbb{R}^d$ of size n, the constraint system $\mathcal{A}((\tau-\gamma)/n)$ is feasible for both Y. Let Y' be any collection of n points in $\mathbb{R}^d$ differing from Y in at most one point. Then, for any τ, γ,*

$$\mathrm{stab}_{Y'}(\tau, \gamma-1) \leqslant \mathrm{stab}_Y(\tau, \gamma)$$

Proof. Using Lemma 4.9 and noting that if $\tilde{\zeta}'$ is adjacent to $\tilde{\zeta}$ then $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w]\right\|_2^2 \leqslant \left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2$, we have:

$$\mathrm{Pot}_{(\tau-\gamma+1)/n}(Y') \leqslant \mathrm{Pot}_{(\tau-\gamma)/n}(Y),$$

and

$$\mathrm{Pot}_{(\tau+\gamma)/n}(Y) \leqslant \mathrm{Pot}_{(\tau+\gamma-1)/n}(Y').$$

Combining the two equations yields

$$\begin{aligned}\mathrm{stab}_{Y'}(\tau,\gamma-1) &= \mathrm{Pot}_{(\tau-\gamma+1)/n}(Y') - \mathrm{Pot}_{(\tau+\gamma-1)/n}(Y') \\ &\leqslant \mathrm{Pot}_{(\tau-\gamma)/n}(Y) - \mathrm{Pot}_{(\tau+\gamma)/n}(Y) = \mathrm{stab}_{Y}(\tau,\gamma).\end{aligned}$$

□

Lemma 4.14 (Sensitivity of Score Function)**.** *Let Y, Y' be set of n points in $\mathbb{R}^d$ differing at most in one point, and $\tau \in [n]$. Then, for every $\tau > 0$,*

$$|\mathrm{score}(\tau, Y) - \mathrm{score}(\tau, Y')| \leqslant 2. \tag{4.9}$$

Proof. It suffices to prove that $\mathrm{score}(\tau, Y') \geqslant \mathrm{score}(\tau, Y) - 2$. A symmetric argument then proves that $\mathrm{score}(\tau, Y) \geqslant \mathrm{score}(\tau, Y') - 2$, which establishes (4.9).

Consider the following two cases:

- $\mathsf{Alg}(Y, (\tau-1)/n)$ is infeasible for Y or Y'. In this case, we have $\mathrm{score}(\tau, Y) \leqslant 2$, which implies the desired bound.
- $\mathsf{Alg}(Y, (\tau-1)/n)$ is feasible for both Y and Y'.

Let $\gamma^* := \gamma^*_Y(\tau)$. From Lemma 4.13, we know that $\mathrm{stab}(\tau, \gamma^*-1, Y') \leqslant \mathrm{stab}_Y(\tau, \gamma^*) + 2$. Thus, it follows that

$$\begin{aligned}\mathrm{score}(\tau, Y') &\geqslant \min\{\gamma^* - 1, 20L - \mathrm{stab}_{Y'}(\tau, \gamma^* - 1)\} \\ &\geqslant \min\{\gamma^* - 1, 20L - \mathrm{stab}_{Y}(\tau, \gamma^*)\} \qquad\qquad \geqslant \mathrm{score}_Y(\tau) - 1,\end{aligned}$$

as desired. □

Lemma 4.15 (Existence of a Good Stable Interval)**.** *Suppose $\mathcal{A}(\eta/2)$ is feasible on Y. For every $L \in [0, 0.25\eta n]$, there is a $\tau \in [0, \eta n]$ such that* $\mathrm{score}(\tau, Y) \geqslant L$.

Proof. Consider $\mathrm{Pot}_{\eta/2}, \mathrm{Pot}_{\eta/2+2L/n}, \ldots, \mathrm{Pot}_{\eta/2+2Lr/n}$ where $r := \lfloor 0.25\eta n/L \rfloor$. Observe that $\mathrm{Pot}_{\eta/2}(Y) - \mathrm{Pot}_{\eta}(Y) \leqslant (1-\eta/2)n - (1-\eta)^2 n \leqslant 1.5\eta n$. Therefore, there must exists $r^* \in [r]$ such that

$$\mathrm{Pot}_{\eta/2+2L(r^*-1)/n} - \mathrm{Pot}_{\eta/2+2Lr^*/n} \leqslant \frac{1.5\eta n}{r} \leqslant 12L.$$

Let $\tau = \eta/2 + (2Lr^* - 1)/n$ and $\gamma = L$. Then, we have $\mathrm{stab}(\tau, \gamma) \leqslant 12L$ and, thus,

$$\mathrm{score}(\tau, Y) \geqslant \max\{\gamma, 20L - 12L\} \geqslant L.$$

□

Lemma 4.16 (Utility of Score Function)**.** *Suppose $\mathcal{A}(\eta/2)$ is feasible on Y. Let $\varepsilon, \delta, \beta \in (0,1]$. For every $L \in [0, 0.25\eta n]$, if $L \geqslant O\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\beta\delta}\right)\right)$, then with probability $1-\beta$, Theorem 3.34, invoked with the score function in Definition 4.12 and $\kappa = L/2$, does not reject, and the output τ satisfies $\mathrm{stab}_Y(\tau, L/2) < 20L$.*

Proof. This follows from the guarantee of Selection (Theorem 3.34), Lemma 4.15 and the definition of score. □

Lemma 4.17 (Potential Stability Under Good Coupling)**.** *Let $\eta, \varepsilon, \delta > 0$ and $k, L \in \mathbb{N}$ be given input parameters such that $0.25\eta n \geqslant L = \Omega\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\beta\delta}\right)\right)$. Let Y, Y' be adjacent subsets of $\mathbb{Q}^d$. Suppose Alg does not halt and chooses $\eta' = \tau/n$ in Step 1 on input Y and Y'. Then,*

$$\left|\mathrm{Pot}_{\eta'}(Y) - \mathrm{Pot}_{\eta'}(Y')\right| \leqslant 20L\,.$$

Consequently, if p, p' are scalings of $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]$ and $\widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w]$ so that $\|p\|_1 = \|p'\|_1 = 1$, then,

$$\|p - p'\|_1 \leqslant 120\sqrt{L/n}\,.$$

Proof. It is enough to prove that $\mathrm{Pot}_{\eta'}(Y) - \mathrm{Pot}_{\eta'}(Y') \leqslant 20L$ as a symmetric argument proves the other direction and completes the proof.

Let $\tilde{\zeta}$ be the pseudo-distribution that minimizes $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2$ while satisfying $\mathcal{A}$ on Y' with outlier rate η' (computed in Step 3 of the algorithm on input Y'). Suppose Y and Y' differ on i-th sample point. Let $\tilde{\zeta}_{adj}$ be the adjacent pseudo-distribution obtained by zeroing out w_i. Then, from Lemma 4.9, we know that $\tilde{\zeta}_{adj}$ is consistent with $\mathcal{A}$ on input Y with outlier rate $\eta' + 1/n$. Further, $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}_{adj}}[w]\right\|_2^2 \leqslant \left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2$. Thus, $\mathrm{Pot}_{\eta'+1/n}(Y) \leqslant \mathrm{Pot}_{\eta'}(Y')$. Further, Lemma 4.16 implies that $\left|\mathrm{Pot}_{\eta'+1/n}(Y) - \mathrm{Pot}_{\eta'}\right| \leqslant 20L$. Therefore, we have $\mathrm{Pot}_{\eta'}(Y) - \mathrm{Pot}_{\eta'}(Y') \leqslant 20L$ as desired.

Now, by Cauchy-Schwarz inequality, we immediately obtain that:

$$\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w] - \widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w]\right\|_1^2 \leqslant 20nL$$

Thus, from Lemma 3.21, we have that:

$$\|p - p'\|_1 \leqslant 120\sqrt{L/n}\,.$$

□

Parameter closeness from potential stability The following lemma observes that if a sequence of weights $p_i(Y)$ induces a $2k$-certifiably C'-subgaussian distribution on Y and $p'_i(Y)$ is a sequence of weights on an adjacent Y such that $p_i(Y)$ is not too far from $p_i(Y')$, then, $p_i(Y')$ must also induce a $2k$-certifiably $C'+1$-subgaussian distribution on Y'.

Lemma 4.18. *Let $0 \leqslant p_i(Y) \leqslant \frac{1}{(1-2\eta')}$ be a sequence of non-negative weights adding up to n that induce a $2k$-certifiable C'-subgaussian distribution on Y. Let $p_i(Y')$ be a sequence of non-negative weights adding up to n on Y' adjacent to Y such that $\|p(Y) - p(Y')\|_1 \leqslant \beta$ for $\beta \leqslant \eta_0$. Then, for small enough absolute constant $\eta' > 0$, $p_i(Y')$ induces a $2k$-certifiable $(C'+1)$-subgaussian distribution on Y.*

Proof Sketch. Let's first describe the idea of the proof: the proof of Lemma 4.6 requires the existence of a certifiably subgaussian distribution that was close (in total variation distance) to the input Y. Since Y is adjacent to Y', the $2k$-certifiably C'-subgaussian distribution is $1-\beta-2/n$-close (the additive $2/n$ comes from "removing" the index of the point where Y and Y' differ) in total variation distance to Y. Thus, the idea is to use the certifiably subgaussian distribution supported on Y in lieu of X to repeat the argument. In order to apply Lemma 4.6, we need a "flat" distribution—but this is easily achieved. Given a distribution with weights (without loss of generality, say, rational numbers r_i/s), we can consider a sample expansion to ns samples that has r_i copies of sample y_i for each i and an analogous transformation to Y'. And finally, given a pseudo-distribution on $w_1, w_2, \ldots, w_n$ on $Y \cap Y'$, we can transform to a pseudo-distribution on ns variables by each "copying" w_i for i such that $y_i = y'_i$ r_i times. □

As an immediate corollary of Lemma 4.17 and Lemma 4.18, we obtain:

Corollary 4.19 (Parameter Closeness from Stability of Potential)**.** *Let $\eta, \varepsilon, \delta > 0$ and $k, L \in \mathbb{N}$ be given input parameters to Algorithm 4.8 such that $0.25\eta n \geqslant L = \Omega\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\beta\delta}\right)\right)$. Also, let Y, Y' be adjacent subsets of $\mathbb{Q}^d$. Suppose Alg does not reject in any of the 3 steps, uses the constant C' in Step 2 and chooses η' in Step 1 on input Y and Y'.*

Then, for every $u \in \mathbb{R}^d$ and $\theta = \sqrt{L/n}$, we have:

$$\langle \mu_p - \mu_{p'}, u \rangle \leqslant O(C'k)\theta^{1-1/2k}\sqrt{u^\top \Sigma_p u}\,,$$

$$(1 - O(C'k)\theta^{1-1/k})\Sigma_p \preceq \Sigma_{p'} \preceq (1 + O(C'k)\theta^{1-1/k})\Sigma_p\,,$$

and, for every $t \leqslant k$ such that t divides $2k$,

$$(1 - O(C'^{t/2}k^{t/2}))\theta^{1-t/2k})\langle u^{\otimes t}, M_p^{(t)}\rangle \leqslant \langle u^{\otimes t}, M_{p'}^{(t)}\rangle \leqslant (1 + O(C'^{t/2}k^{t/2})\theta^{1-t/2k})\langle u^{\otimes t}, M_p^{(t)}\rangle\,,$$

Proof. Let $\tilde{\zeta}_{adj}$ be the adjacent pseudo-distribution of degree $2k$ to $\tilde{\zeta}$ obtained by zeroing out w_i where i is the index of the point that Y and Y' differ on. Then, from Lemma 4.9, we know that $\tilde{\zeta}_{adj}$ satisfies $\mathcal{A}$ on both inputs Y, Y' with outlier rate $\eta' + 1/n$ and $\left|\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}_{adj}}[w] - \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2\right| \leqslant 1$, $\left|\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}_{adj}}[w] - \widetilde{\mathbb{E}}_{\tilde{\zeta}'}[w]\right\|_2^2\right| \leqslant 1$. Let p_{adj} be the scaling of $\widetilde{\mathbb{E}}_{\tilde{\zeta}_{adj}}[w]$ so that $\|p_{adj}\|_1 = 1$. Then, clearly, $\|p - p_{adj}\|_1 \leqslant 2/n$ (since $\eta' \ll 1/2$). Further, applying Lemma 4.17 and triangle inequality, we have that $\|p_{adj} - p'\|_1 \leqslant O(\sqrt{L/n})$. Applying Fact 3.23 to p_{adj} and p on Y and p_{adj} and p' on Y' and using triangle inequality completes the proof. □

Noise injection in estimate-dependent norms Our final ingredient for obtaining privacy guarantees for our robust estimation algorithms is a new noise injection mechanism where the distribution of noise depends on the covariance estimated by our algorithm.

Lemma 4.20. *Suppose $\varepsilon, \delta > 0$. Let A be an invertible $d \times d$ matrix that satisfies $(1-\beta)I \preceq AA^T \preceq (1+\beta)I$, where $\beta \leqslant \frac{\varepsilon}{3d\ln(d/\delta)}$. Let $z \in \mathbb{R}^d$ be a vector whose entries are i.i.d. from $\mathcal{N}(0,1)$. Then,*

$$D_{e^\varepsilon}(z, Az) \leqslant \delta.$$

Proof. Note that the probability distribution function of Az at $u \in \mathbb{R}^d$ is

$$\frac{1}{\det(A)} \frac{1}{(\sqrt{2\pi})^d} e^{-\|A^{-1}u\|_2^2/2}.$$

Moreover, $\det(A) \leqslant (1+\beta)^{d/2}$, since $\det(A)^2 = \det(A)\det(A^T) = \det(AA^T) \leqslant (1+\beta)^d$. Thus, the ratio of the probability densities of z and Az at u is

$$\begin{aligned}
\det(A) e^{\|A^{-1}u\|_2^2/2 - \|u\|_2^2/2} &\leqslant (1+\beta)^{d/2} e^{\|A^{-1}u\|_2^2/2 - \|u\|_2^2/2} \\
&\leqslant (1+\beta)^{d/2} e^{\frac{1}{2}u^T((AA^T)^{-1} - I)u} \\
&\leqslant (1+\beta)^{d/2} e^{\frac{1}{2}(1-\beta)^{-1}\|u\|_2^2 - \frac{1}{2}\|u\|_2^2} \\
&\leqslant (1+\beta)^{d/2} e^{\frac{\beta}{2(1-\beta)}\|u\|_2^2}.
\end{aligned}$$

Thus, note that if $\|u\|_\infty \leqslant \sqrt{2\ln(d/\delta)}$, then $\|u\|_2 \leqslant \sqrt{d} \cdot \|u\|_\infty \leqslant \sqrt{2d\ln(d/\delta)}$, and so,

$$\det(A) e^{\|A^{-1}u\|_2^2/2 - \|u\|_2^2/2} \leqslant (1+\beta)^{d/2} e^{\frac{\beta}{1-\beta} d\ln(d/\delta)} < e^{\varepsilon},$$

since $\beta \leqslant \frac{\varepsilon}{3d\ln(d/\delta)}$.

Moreover, by standard tail bounds of the normal distribution, we have that $\|z\|_\infty > \sqrt{2\ln(d/\delta)}$ with probability at most δ. This proves the claim. [Pasin: This seems a tad strange; can't we use tail bound directly on Euclidean norm, and not have a d inside ln factor?][Ameya: Indeed, I think we can use concentration of norm properties to get that $\|z\|_2 < c(\sqrt{d} + \sqrt{\ln(1/\delta)})$ with probability $\geqslant 1-\delta$, and this will allow us to get away with $\beta < \frac{\varepsilon}{d+\ln(1/\delta)}$ instead of $\beta < \frac{\varepsilon}{d\ln(d/\delta)}$. In the end, this will give a looser upper bound condition on γ_2 that needs to be satisfied (see the sentence before (4.23)); however, even with this improvement, the bottleneck will still be the condition on γ_2 coming from the covariance noise (see (4.28)).] □

Lemma 4.21. *Suppose $\varepsilon, \delta > 0$. Let A be a $d \times d$ matrix that satisfies $\left\|AA^T - I\right\|_2 \leqslant \beta$.*

Let $t \in \mathbb{N}$. Moreover, let $Z \in \mathbb{R}^{d^t}$ be a random vector indexed by $[d]^t$, whose entries $Z_{i_1,i_2,\ldots,i_t}$, for $1 \leqslant i_1 \leqslant i_2 \leqslant \cdots \leqslant i_t \leqslant d$, are i.i.d. from $\mathcal{N}(0,1)$, and moreover, $Z_{i_1,i_2,\ldots,i_t} = Z_{i_{\pi(1)},i_{\pi(2)},\ldots,i_{\pi(t)}}$ for any $i = (i_1, i_2, \ldots, i_d)$ and permutation π.

If $\beta \leqslant \frac{\varepsilon}{8t^2 d^t \ln(d^t/\delta)}$, then

$$D_{e^\varepsilon}(Z, A^{\otimes t}Z) \leqslant \delta.$$

Proof. Let $K = \sqrt{2\ln(d^t/\delta)}$. By standard tail bounds, note that

$$\mathbb{P}[\|Z\|_\infty > K] \leqslant \delta. \tag{4.10}$$

Let S be the subspace of $\mathbb{R}^{d^t}$ consisting of all symmetric tensors, i.e.,

$$S = \left\{ u \in \mathbb{R}^{d^t} : u_{i_1,i_2,\ldots,i_t} = u_{(i_{\pi(1)},i_{\pi(2)},\ldots,i_{\pi(t)})},\ \forall i = (i_1, i_2, \ldots, i_t) \in [d]^t, \pi \text{ a permutation on } [t] \right\}.$$

Note that S is an d'-dimension invariant subspace of $A^{\otimes t}$, where $d' = \binom{d+t-1}{t} \leqslant d^t$. Moreover, let $R \subseteq [d]^t$ be a representative set of indices of size $|R| = d'$, i.e., R satisfies the property that for any $(i_1, i_2, \ldots, i_t) \in [d]^t$, there exists a permutation π on $[t]$ such that $(i_{\pi(1)}, i_{\pi(2)}, \ldots, i_{\pi(t)}) \in R$.

Now, let $M = A^{\otimes t}|_S$ be the restriction of $A^{\otimes t}$ to the subspace S. Moreover, let $Z_R \in \mathbb{R}^{d'}$ denote the projection of Z to indices in R.

Note that the probability distribution of Z can be equivalently viewed as the probability distribution of Z_R, since Z is uniquely determined by the projection Z_R. Let p be the probability density function of Z_R over $\mathbb{R}^{d'}$. Then, note that the probability distribution of MZ_R is q, given by

$$q(v) = \frac{1}{\det(M)} \cdot p(M^{-1}v).$$

for $v \in \mathbb{R}^{d'}$. By standard properties, we know that the i^{th} singular value of M is bounded from above by the i^{th} singular value of $A^{\otimes t}$ and bounded from below by the $(i + d^t - d')^{\text{th}}$ singular value of $A^{\otimes t}$. Moreover, by $\left\|AA^T - I\right\|_2 \leqslant \beta$, we know that the singular values of $A^{\otimes t}$ lie in $[(1-\beta)^{t/2}, (1+\beta)^{t/2}]$. Hence, the singular values of M also lie in $[(1-\beta)^{t/2}, (1+\beta)^{t/2}]$, which, together with $\beta t \leqslant \frac{1}{4}$, implies that

$$\left\|MM^T - I\right\|_2 \leqslant 2\beta t \tag{4.11}$$

and so,

$$\det(M)^2 = \det(M)\det(M^T) = \det(MM^T) \leqslant (1+2\beta t)^{t \cdot d'} \leqslant (1+2\beta t)^{td^t}, \tag{4.12}$$

and so, $\det(M) \leqslant (1+2\beta t)^{td^t/2}$.

Let $u \in \mathbb{R}^{d^t}$. Note that $\|u\|_\infty \leqslant K$ if and only if $v \in \mathbb{R}^{d'}$ given by $v = u|_R$ also satisfies $\|v\|_\infty \leqslant K$. Moreover, note that if $\|v\|_\infty \leqslant K$, then

$$\begin{aligned}
\frac{p(v)}{q(v)} &\leqslant \det(M) \cdot \frac{p(v)}{p(M^{-1}v)} \\
&\leqslant (1+2\beta t)^{td^t/2} \cdot \exp\left(\frac{1}{2}\left(\left\|M^{-1}v\right\|_2^2 - \|v\|_2^2\right)\right) \\
&\leqslant e^{\beta t^2 d^t} \cdot \exp\left(\frac{1}{2}\left(v^T((MM^T)^{-1} - I)v\right)\right) \\
&\leqslant e^{\beta t^2 d^t} \cdot \exp\left(\frac{1}{2}\|v\|_2^2 \cdot \left\|(MM^T)^{-1} - I\right\|_2\right)
\end{aligned} \tag{4.13}$$

By (4.11), we have that $\left\|(MM^T)^{-1} - I\right\|_2 \leqslant \frac{2\beta t}{1-2\beta t} \leqslant 4\beta t$, since $\beta t \leqslant \frac{1}{4}$. Therefore, (4.13) is at most

$$e^{\beta t^2 d^t} \cdot \exp\left(2K^2 d^t \beta t\right).$$

Thus, if $\beta \leqslant \frac{\varepsilon}{2K^2 t d^t} = \frac{\varepsilon}{8t^2 d^t \ln(d^t/\delta)}$, the above quantity is at most e^ε. This, combined with (4.10), proves the desired claim. □

Remark 4.22. Note that Lemma 4.21 uses an assumption on the *spectral norm* of $\left\|AA^T - I\right\|_2$. However, it is also possible to obtain a version of the lemma under an assumption on the *Frobenius norm*, $\left\|AA^T - I\right\|_F$. In particular, if we assume that, instead, $\left\|AA^T - I\right\|_F \leqslant \beta$, then Eq. (4.12) instead

becomes $\det(M) \leqslant \left(1 + \frac{\beta}{\sqrt{d}}\right)^{td^t/2} \leqslant e^{\beta t d^{t-\frac{1}{2}}/2}$: This follows from the fact that (a.) the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ of AA^T satisfy $\sum_{i=1}^d (\lambda_i - 1)^2 \leqslant \beta^2$, (b.) under the aforementioned constraint, $\lambda_1 \lambda_2 \cdots \lambda_d$ is maximized when $\lambda_1 = \lambda_2 = \cdots = \lambda_d = 1 + \frac{\beta}{\sqrt{d}}$, (c.) the eigenvalues of $(AA^T)^{\otimes t}$ are precisely the d^t t-fold products of eigenvalues of AA^T.

Putting things together Now, we are ready to prove the main privacy guarantee provided by our robust moment estimation algorithm, Algorithm 4.8.

Lemma 4.23 (Privacy Guarantee)**.** *Suppose $C, \eta, \varepsilon, \delta > 0$ and $k \in \mathbb{N}$. Suppose $n \geqslant n_0 = \widetilde{\Omega}\left(\left(\frac{Ck^4d^k}{\varepsilon}\left(\ln(6kd^k/\delta) + \frac{\varepsilon}{6k}\right)\right)^{\frac{2k}{k-1}}\right)$. Then, Alg (given by Algorithm 4.8), invoked with $L = O(\log(n/\delta)/\varepsilon)$, is (ε, δ)-DP.*

Proof. Let $\varepsilon' = \varepsilon/3$ and $\delta' = \delta/3$. By our adaptive composition theorem under halting (Lemma 3.30), it suffices to show that each step of the algorithm is (ε', δ')-DP (given the outputs of the previous steps as parameter[3]). Let Y and Y' be any neighboring datasets.

- **Stable Outlier Rate Selection.** Since this step invokes the (ε', δ')-DP Selection algorithm (Selection), it immediately follows from Theorem 3.34 that this step is (ε', δ')-DP.
- **Witness Checking.** [Pasin: I'd suggested phrasing the DP guarantee of this step and the next step in terms of Definition 3.29 because our composition theorem is stated in that term.] Let $C^*(Y)$ denote the smallest C^* for which $p_i(Y)$ induces a $2k$-certifiable C^*-subgaussian distribution on Y. Lemma 4.18 ensures that $|C^*(Y) - C^*(Y')| \leqslant \Delta$ for $\Delta = 1$. Therefore, we may apply Lemma 3.27 with DP parameters ε', δ' to conclude that this step is also (ε', δ')-DP.
- **Noise Addition.** Since the algorithm has not halted in the previous step and the truncated Laplace noise is negative, $p_i(Y)$ and $p_i(Y')$ must induce $2k$-certifiable C'-subgaussian distributions on Y and Y' respectively. Let $\widetilde{\mu}$ and $\widetilde{\mu}'$ denote the corresponding mean estimates under $p_i(Y)$ and $p_i(Y')$, respectively, and, similarly, let $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ denote the corresponding covariance estimates. By Corollary 4.19, we have that, for all $u \in \mathbb{R}^d$,

$$\langle \widetilde{\mu} - \widetilde{\mu}', u \rangle \leqslant \gamma_1 \sqrt{u^\top \widetilde{\Sigma} u} \tag{4.14}$$

$$(1 - \gamma_2)\widetilde{\Sigma} \preceq \widetilde{\Sigma}' \preceq (1 + \gamma_2)\widetilde{\Sigma} \tag{4.15}$$

  and, for all $2 \leqslant t \leqslant k$,

$$(1 - \gamma_t)\langle u^{\otimes t}, \widetilde{M^{(t)}} \rangle \leqslant \langle u^{\otimes t}, \widetilde{M'^{(t)}} \rangle \leqslant (1 + \gamma_t)\langle u^{\otimes t}, \widetilde{M^{(t)}} \rangle, \tag{4.16}$$

  where $\theta = \sqrt{L/n}$, $\gamma_1 = O(C'k)\theta^{1-1/2k}$, and $\gamma_t = O((C'k)^{t/2})\theta^{1-t/2k}$ for $2 \leqslant t \leqslant k$. Moreover, let $B = \widetilde{\Sigma}^{-1/2}\widetilde{\Sigma}'^{1/2}$.

[3]Note that we may also assume that the algorithm has not halted from the previous steps.

Note that in order to show that the noise addition step is (ε', δ')-DP, it suffices to show that

$$D_{e^{\varepsilon''}}(\widetilde{\mu} + \widetilde{\Sigma}^{1/2} z, \widetilde{\mu}' + \widetilde{\Sigma}'^{1/2} z) \leqslant \delta'' \tag{4.17}$$

$$D_{e^{\varepsilon''}}(\widetilde{\Sigma} + \widetilde{\Sigma}^{1/2} Z \widetilde{\Sigma}^{1/2}, \widetilde{\Sigma}' + \widetilde{\Sigma}'^{1/2} Z \widetilde{\Sigma}'^{1/2}) \leqslant \delta'' \tag{4.18}$$

$$\forall\, 2 < t \leqslant k, \quad D_{e^{\varepsilon''}}(\widetilde{M^{(t)}} + (\widetilde{\Sigma}^{1/2})^{\otimes t} Z^{(t)}, \widetilde{M'^{(t)}} + (\widetilde{\Sigma}'^{1/2})^{\otimes t} Z^{(t)}) \leqslant \delta'' \tag{4.19}$$

for $\varepsilon'' = \varepsilon'/k$ and $\delta'' = \delta'/k$, since Fact 3.32 and standard DP composition [DKM+06] then imply that the entire noise addition step is (ε', δ')-DP. We now establish each of the above inequalities.

Noise addition for mean: We first show (4.17). Note that

$$\begin{aligned}
D_{e^{\varepsilon''}}(\widetilde{\mu} + \widetilde{\Sigma}^{1/2} z, \widetilde{\mu}' + \widetilde{\Sigma}'^{1/2} z) &= D_{e^{\varepsilon''}}(\widetilde{\Sigma}^{1/2} z, (\widetilde{\mu}' - \widetilde{\mu}) + \widetilde{\Sigma}'^{1/2} z) \\
&= D_{e^{\varepsilon''}}(z, \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu}) + Bz) \\
&= D_{e^{\varepsilon''/2}}(z, z + \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu})) \\
&\quad + e^{\varepsilon''/2} D_{e^{\varepsilon''/2}}(z + \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu}), \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu}) + Bz) \qquad (4.20) \\
&= D_{e^{\varepsilon''/2}}(z, z + \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu})) + D_{e^{\varepsilon''/2}}(z, Bz), \qquad (4.21)
\end{aligned}$$

where (4.20) follows from Lemma 3.33. For the first term on the right-hand side of (4.21), we note that $\left\|\widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu})\right\|_2 \leqslant \gamma_1$ (which follows from plugging in $u = \widetilde{\Sigma}^{-1}(\widetilde{\mu} - \widetilde{\mu}')$ into (4.14)). Thus, by the standard hockey-stick divergence calculation for the Gaussian mechanism [DR14, Appendix A], we have that

$$D_{e^{\varepsilon''/2}}(z, z + \widetilde{\Sigma}^{-1/2}(\widetilde{\mu}' - \widetilde{\mu})) < \delta''/2, \tag{4.22}$$

provided that

$$\sigma_1 \geqslant \frac{2\gamma_1 \sqrt{2\ln(2.5/\delta'')}}{\varepsilon''},$$

For the second term in (4.21), note that (4.15) implies that $(1-\gamma_2)I \preceq BB^T \preceq (1+\gamma_2)I$. Moreover, $\gamma_2 \leqslant \frac{\varepsilon''}{3d\ln(2d/\delta'')}$, by the condition $n \geqslant n_0$. Thus, by Lemma 4.20,

$$D_{e^{\varepsilon'/2}}(z, Bz) \leqslant \delta''/2. \tag{4.23}$$

Therefore, (4.22), (4.23),and (4.21) imply (4.17), as desired.

Noise addition for covariance: Next, we establish (4.18). Observe that

$$\begin{aligned}
D_{e^{\varepsilon''}}(\widetilde{\Sigma} + \widetilde{\Sigma}^{1/2} Z \widetilde{\Sigma}^{1/2}, \widetilde{\Sigma}' + \widetilde{\Sigma}'^{1/2} Z \widetilde{\Sigma}'^{1/2}) &= D_{e^{\varepsilon''}}(I + Z, BB^T + BZB^T) \\
&= D_{e^{\varepsilon''}}(Z, (BB^T - I) + BZB^T) \\
&\leqslant D_{e^{\varepsilon''/2}}(Z, Z + (BB^T - I)) \\
&\quad + e^{\varepsilon''/2} D_{e^{\varepsilon''/2}}(Z + (BB^T - I), (BB^T - I) + BZB^T)) \qquad (4.24) \\
&\leqslant D_{e^{\varepsilon''/2}}(Z, Z + (BB^T - I)) + e^{\varepsilon''/2} D_{e^{\varepsilon''/2}}(Z, BZB^T)), \qquad (4.25)
\end{aligned}$$

where (4.24) follows from Lemma 3.33. To bound the right-hand side of (4.25), note that the first term is precisely the hockey-stick divergence computation corresponding to the Gaussian mechanism (restricted to the upper triangular portion). Moreover, by (4.15),

$$\left\|BB^T - I\right\|_F \leqslant \sqrt{d} \cdot \left\|BB^T - I\right\|_2 \leqslant \gamma_2\sqrt{d}. \tag{4.26}$$

Therefore ([DR14, Appendix A]), as long as

$$\sigma_2 \geqslant \frac{2\gamma_2\sqrt{2d\ln(2.5/\delta'')}}{\varepsilon''},$$

we have that

$$D_{e^{\varepsilon''/2}}(Z, Z + (BB^T - I)) \leqslant \delta''/2. \tag{4.27}$$

For the second term in (4.25), note that $\sigma_2^{-1}Z$ has entries distributed in $\mathcal{N}(0,1)$. Moreover, let $Z' = \mathrm{vec}(Z)$ be the d^2-dimensional vector given by the flattening of Z (see Definition 3.15). By Fact 3.17, we know that $BZB^T = B^{\otimes 2}Z'$. Thus, by Lemma 4.21 applied with $t = 2$,

$$\begin{aligned} D_{e^{\varepsilon''/2}}(Z, BZB^T)) &= D_{e^{\varepsilon''/2}}(Z', B^{\otimes 2}Z') \\ &= D_{e^{\varepsilon''/2}}(\sigma_2^{-1}Z', B^{\otimes 2}(\sigma_2^{-1}Z')) \\ &\leqslant \delta''/2e^{e^{\varepsilon''/2}}, \end{aligned} \tag{4.28}$$

as long as

$$\gamma_2 < \frac{\varepsilon''}{32d^2\ln(2d^2e^{\varepsilon''/2}/\delta'')},$$

which is true, since $n \geqslant n_0$ by the conditions of the theorem. Thus, (4.27) and (4.28) imply that (4.25) is at most $\delta''/2 + e^{\varepsilon''/2}(\delta''/2e^{\varepsilon''/2}) = \delta''/2$, which establishes (4.18).

Noise addition for higher-order moments: Let $2 < t \leqslant k$. We write $R = \widetilde{\Sigma} + \widetilde{\mu}\widetilde{\mu}^T$ and $R' = \widetilde{\Sigma}' + \widetilde{\mu}'\widetilde{\mu}'^T$ for simplicity.

Observe that the injective/spectral norm $\|\cdot\|_\sigma$ of $(R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)$ can be bounded as

$$\begin{aligned} \left\|(R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right\|_\sigma &= \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \left|\left(v^{\otimes t}\right)^T (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right| \\ &\leqslant \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \left|\langle (R^{-1/2}v)^{\otimes t}, \widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\rangle\right| \\ &\leqslant \gamma_t \cdot \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \left|\langle (R^{-1/2}v)^{\otimes t}, \widetilde{M^{(t)}}\rangle\right| \\ &\leqslant \gamma_t \cdot (C'k)^t \cdot \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \left|\left(\left(R^{-1/2}v\right)^T R\left(R^{-1/2}v\right)\right)^{t/2}\right| \\ &= \gamma_t \cdot (C'k)^t \cdot \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \|v\|_2^t \end{aligned} \tag{4.29}$$

$$= \gamma_t \cdot (C'k)^t,$$

where (4.29) follows from the C'-subgaussianity property of the distribution induced by the weight vector at the end of Step 2. Therefore, the Frobenius norm (or Hilbert-Schmidt norm) can be bounded as (see Corollary 4.10 of [WDFS17])

$$\left\|(R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right\|_F \leqslant d^{\frac{t-1}{2}} \cdot \left\|(R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right\|_\sigma \leqslant \gamma_t \cdot (C'k)^t \cdot d^{\frac{t-1}{2}}. \tag{4.30}$$

Moreover, letting $W = R^{-1/2}R'^{1/2}$, we have

$$\begin{aligned}
D_{e^{\varepsilon''}}(\widetilde{M^{(t)}} + (R^{1/2})^{\otimes t}Z^{(t)}, \widetilde{M'^{(t)}} + (R'^{1/2})^{\otimes t}Z^{(t)}) &= D_{e^{\varepsilon''}}\left((R^{1/2})^{\otimes t}Z^{(t)}, \widetilde{M'^{(t)}} - \widetilde{M^{(t)}} + (R'^{1/2})^{\otimes t}Z^{(t)}\right) \\
&= D_{e^{\varepsilon''}}\left(Z^{(t)}, (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right) + W^{\otimes t}Z^{(t)}\right) \\
&\leqslant D_{e^{\varepsilon''/2}}\left(Z^{(t)}, Z^{(t)} + (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right) \\
&\quad + e^{\varepsilon''/2}D_{e^{\varepsilon''/2}}(Z^{(t)} + (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right), \\
&\quad (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right) + W^{\otimes t}Z^{(t)})) \qquad (4.31) \\
&\leqslant D_{e^{\varepsilon''/2}}\left(Z^{(t)}, Z^{(t)} + (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right) \\
&\quad + e^{\varepsilon''/2}D_{e^{\varepsilon''/2}}(Z^{(t)}, W^{\otimes t}Z^{(t)}), \qquad (4.32)
\end{aligned}$$

where again we have used Lemma 3.33 in (4.31). In order to bound the right-hand side of (4.32), note that the first term is again the hockey-stick divergence computation corresponding to the Gaussian mechanism (restricted according to symmetry conditions). Recalling (4.30), we see that ([DR14, Appendix A]) as long as

$$\sigma_t \geqslant \frac{2\gamma_t(C'k)^t d^{\frac{t-1}{2}}\sqrt{2\ln(2.5/\delta'')}}{\varepsilon''},$$

we have that

$$D_{e^{\varepsilon''/2}}\left(Z^{(t)}, Z^{(t)} + (R^{-1/2})^{\otimes t}\left(\widetilde{M'^{(t)}} - \widetilde{M^{(t)}}\right)\right) \leqslant \delta''/2. \tag{4.33}$$

For the second term in (4.32), note that $\sigma_t^{-1}Z^{(t)}$ has entries distributed in $\mathcal{N}(0,1)$. Moreover, note that $\left\|WW^T - I\right\|_F \leqslant \left\|BB^T - I\right\|_F \leqslant \gamma_2\sqrt{d}$ by (4.26) and the fact that (4.15) implies

$$(1-\gamma_2)R \preceq R' \preceq (1+\gamma_2)R.$$

Thus, by Lemma 4.21, we have that

$$\begin{aligned}
D_{e^{\varepsilon''/2}}\left(Z^{(t)}, W^{\otimes t}Z^{(t)}\right) &= D_{e^{\varepsilon''/2}}\left(\sigma_t^{-1}Z^{(t)}, W^{\otimes t}(\sigma_t^{-1}Z^{(t)})\right) \\
&\leqslant \delta''/2e^{\varepsilon''/2}, \qquad (4.34)
\end{aligned}$$

as long as

$$\gamma_2 < \frac{\varepsilon''}{16t^2d^t\ln(2d^te^{\varepsilon''/2}/\delta'')},$$

which is true since $n \geqslant n_0$, by the hypothesis of the lemma. Thus, (4.33) and (4.34) imply that (4.32) is at most $\delta''/2 + e^{\varepsilon''/2}(\delta''/2e^{\varepsilon''/2}) = \delta''$, which establishes (4.19), as desired. □

4.4 Proof of Theorem 4.1

We are now ready to prove our main theorem, Theorem 4.1.

Proof of Theorem 4.1. Choose $\beta = 1/30$. Choose $L = \Omega\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\beta\delta}\right)\right)$ (according to the condition in Lemma 4.16). Moreover, let $C = C_0 + \frac{3\ln(3/\delta)}{\varepsilon} + \frac{9}{\varepsilon} + 1$. Then, we claim that setting Alg to be Algorithm 4.8 with parameters $C, \eta, \varepsilon, \delta, L, k$ satisfies the desired conditions, as long as $\eta \leqslant \eta_0$, where we set η_0 later.

Note that the desired privacy guarantees follow immediately from Lemma 4.23.

It remains to prove the utility guarantees. Suppose that there indeed exists a good set $X \subseteq \mathbb{Q}^d$ with mean μ_*, covariance Σ_*, and t-th moments $M_*^{(t)}$ for $3 \leqslant t \leqslant k$, such that $|Y \cap X| \geqslant (1-\eta)n$.

By Theorem 3.34, we have that Step 1 (stable outlier rate selection) rejects and halts with probability at most $\beta = 1/30$, and the resulting output τ satisfies $\mathrm{score}(\tau, Y) \geqslant L/2$. In particular, the latter condition implies that there exists some $\gamma \geqslant L/2$ for which $\mathcal{A}\left(\frac{\tau-\gamma}{n}\right)$ is feasible. By monotonicity, $\mathcal{A}\left(\eta'\right)$ is also feasible, where we let $\eta' = \tau/n$.

Hence, the invocation of Algorithm 4.3 in Step 2 does not yield "reject." Moreover, note that by Lemma 3.28, we have that $C' = C + \gamma \geqslant C_0$ with probability at least $29/30$. In this case, the computed weight vector p induces a C'-certifiably subgaussian distribution on Y. Hence, the probability of rejection in Step 2 is at most $1/30$.

Let $\widetilde{\mu} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\mu]$, $\widetilde{\Sigma} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\Sigma]$, and $\widetilde{M}^{(t)} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\Sigma]$ (for $2 \leqslant t \leqslant k$) be the estimates of the mean, covariance, and t-th moments, respectively, that are outputted by the Algorithm 4.3 subroutine in Step 2 of Algorithm 4.8. Then, by Lemma 4.5, we have

$$\forall u \in \mathbb{R}^d,\ \langle \widetilde{\mu} - \mu_*, u\rangle \leqslant O\left(\sqrt{Ck}\right)\eta^{1-1/2k}\sqrt{u^\top \Sigma_* u}$$
$$(1-\beta_2)\Sigma_* \preceq \widetilde{\Sigma} \preceq (1+\beta_2)\Sigma_*$$

and, for all even $2 \leqslant t \leqslant k$ such that t divides $2k$,

$$\forall u \in \mathbb{R}^d,\ (1-\beta_t)\langle u^{\otimes t}, M_*^{(t)}\rangle \leqslant \langle u^{\otimes t}, \widetilde{M}^{(t)}\rangle \leqslant (1+\beta_t)\langle u^{\otimes t}, M_*^{(t)}\rangle,$$

where $\beta_t = \beta_t(\eta) = O((Ck)^{t/2})\eta^{1-t/2k}$. We now set η_0 such that $\beta_t(\eta_0) \leqslant \frac{1}{2}$ for all aforementioned t. Note that this guarantees that $\beta_t = \beta_t(\eta) \leqslant \frac{1}{2}$, since we are assuming $\eta \leqslant \eta_0$.

Now, consider the noise addition step, i.e., Step 3 of Algorithm 4.8. Note that by the Cauchy-Schwarz Inequality, for any $u \in \mathbb{R}^d$, we have

$$\begin{aligned}
\langle \widetilde{\Sigma}^{1/2} z, u\rangle &= \langle \Sigma_*^{-1/2}\widetilde{\Sigma}^{1/2} z, \Sigma_*^{1/2} u\rangle \\
&\leqslant \left\|\Sigma_*^{-1/2}\widetilde{\Sigma}^{1/2} z\right\|_2 \cdot \left\|\Sigma_*^{1/2} u\right\|_2 \\
&= (z^T \widetilde{\Sigma}^{1/2}\Sigma_*^{-1}\widetilde{\Sigma}^{1/2} z) \cdot \sqrt{u^\top \Sigma_* u} \\
&\leqslant \|z\|_2^2 \cdot \left(1 + \left\|\widetilde{\Sigma}^{1/2}\Sigma_*^{-1}\widetilde{\Sigma}^{1/2} - I\right\|_2\right) \cdot \sqrt{u^\top \Sigma_* u} \\
&\leqslant \|z\|_2^2 \cdot \left(1 + 2\beta_2\right) \cdot \sqrt{u^\top \Sigma_* u} \\
&\leqslant 2\|z\|_2^2 \cdot \sqrt{u^T \Sigma_* u},
\end{aligned}$$

since $\beta_2 \leqslant \frac{1}{2}$ by our choice of η_0. Now, note that with probability at least $1 - \frac{1}{30k}$, we have that $\|z\|_2 \leqslant O\left(\sigma_1\sqrt{d\ln(kd)}\right)$, in which case it follows that

$$\|z\|_2^2 = O(d)\cdot\sigma_1^2\ln(kd) = O(\sqrt{Ck})\eta^{1-1/2k},$$

by our choice of $n \geqslant n_0$. Thus, the mean estimate $\hat{\mu}$ outputted by the Step 3 satisfies

$$\begin{aligned}\langle\hat{\mu}-\mu_*, u\rangle &= \langle\hat{\mu}-\widetilde{\mu}, u\rangle + \langle\widetilde{\mu}-\mu_*, u\rangle \\ &= \langle\widetilde{\Sigma}^{1/2}z, u\rangle + \langle\widetilde{\mu}-\mu_*, u\rangle \\ &= O\left(\sqrt{Ck}\right)\eta^{1-1/2k}\sqrt{u^T\Sigma_* u}. \end{aligned} \tag{4.35}$$

Next, we consider the utility guarantee for the covariance. Note that $\|Z\|_2 \leqslant \nu_2 = O\left(\sigma_2\sqrt{d\ln(kd^2)}\right)$ with probability at least $1-\frac{1}{30k}$ (this follows from standard spectral properties of Wigner matrices; see, for instance, [Tao12]), in which case, it follows that $-\nu_2\widetilde{\Sigma} \preceq \widetilde{\Sigma}^{1/2}Z\widetilde{\Sigma}^{1/2} \preceq \nu_2\widetilde{\Sigma}$. Moreover, by our choice of $n \geqslant n_0$ as well as η_0, we have that $\nu_2 \leqslant \beta_2 \leqslant \frac{1}{2}$. Thus, it follows that

$$\begin{aligned}\hat{\Sigma} &\preceq (1+\beta_2)\widetilde{\Sigma} \\ &\preceq (1+\beta_2)^2\Sigma_* \\ &\preceq \left(1+\frac{5}{2}\beta_2\right)\Sigma_* \\ &\preceq \left(1+O(Ck)\cdot\eta^{1-1/k}\right)\Sigma_*,\end{aligned}$$

and by a similar argument, we also have $\hat{\Sigma} \succeq \left(1-O(Ck)\cdot\eta^{1-1/k}\right)\Sigma_*$, thus implying that

$$\left(1-O(Ck)\cdot\eta^{1-1/k}\right)\Sigma_* \preceq \hat{\Sigma} \preceq \left(1+O(Ck)\cdot\eta^{1-1/k}\right)\Sigma_*. \tag{4.36}$$

Finally, we consider the utility guarantee for moment estimation. Suppose $2 \leqslant t \leqslant k$ and t is an even number dividing $2k$. Let $A = \widetilde{\Sigma} + \widetilde{\mu}\widetilde{\mu}^T$ and $A_* = \Sigma_* + \mu_*\mu_*^T$. Note that for any $2 < t \leqslant k$, we have $\left\|Z^{(t)}\right\|_F = O\left(\sigma_t d^{t/2}\sqrt{\ln(kd^t)}\right)$ with probability at least $1-\frac{1}{30k}$. In this case, note that for any $u \in \mathbb{R}^d$, we have the following (recall that $\|\cdot\|_\sigma$ indicates the injective norm of a tensor):

$$\begin{aligned}\langle u^{\otimes t}, (A^{1/2})^{\otimes t}Z^{(t)}\rangle &= \langle (A^{1/2}u)^{\otimes t}, Z^{(t)}\rangle \\ &\leqslant \left\|Z^{(t)}\right\|_\sigma \cdot \left\|A^{1/2}u\right\|_2^t \\ &\leqslant \left\|Z^{(t)}\right\|_F \cdot \left\|A^{1/2}u\right\|_2^t \\ &= O(\sigma_t d^{t/2}\sqrt{\ln(kd^t)})\cdot\left\|A^{1/2}u\right\|_2^t \\ &= O(\sigma_t d^{t/2}\sqrt{\ln(kd^t)})\cdot(u^TAu)^{t/2} \\ &= O(\sigma_t(d^{t/2}\sqrt{\ln(kd^t)})\cdot((1+\beta_2)u^TA_*u)^{t/2} \\ &= O(\sigma_t(d(1+\beta_2))^{t/2}\sqrt{\ln(kd^t)})\cdot\left\|A_*^{1/2}u\right\|_2^t\end{aligned}$$

$$= O(\sigma_t (de^{\beta_2})^{t/2} \sqrt{\ln(kd^t)}) \cdot \langle u^{\otimes t}, M_*^{(t)} \rangle \tag{4.37}$$

$$= O((Ck)^{t/2}) \eta^{1-t/2k} \cdot \langle u^{\otimes t}, M_*^{(t)} \rangle, \tag{4.38}$$

where (4.37) follows from Jensen's Inequality, and (4.38) follows from our choice of $n \geqslant n_0$. Thus, the moment estimate $\hat{M}^{(t)} = \widetilde{M}^{(t)} + (A^{1/2})^{\otimes t} Z^{(t)}$ outputted by our algorithm satisfies

$$\begin{aligned}
\langle u^{\otimes t}, \hat{M}^{(t)} \rangle &\leqslant \langle u^{\otimes t}, \widetilde{M}^{(t)} \rangle + \langle u^{\otimes t}, (A^{1/2})^{\otimes t} Z^{(t)} \rangle \\
&\leqslant (1 + \beta_t) \langle u^{\otimes t}, M_*^{(t)} \rangle + O((Ck)^{t/2}) \eta^{1-t/2k} \cdot \langle u^{\otimes t}, M_*^{(t)} \rangle \\
&= \left(1 + O((Ck)^{t/2}) \eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle.
\end{aligned}$$

In a similar fashion, we also get that $\langle u^{\otimes t}, \hat{M}^{(t)} \rangle \geqslant \left(1 - O((Ck)^{t/2}) \eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle$, thus implying that

$$\left(1 - O((Ck)^{t/2}) \eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle \leqslant \langle u^{\otimes t}, \hat{M}^{(t)} \rangle \leqslant \left(1 + O((Ck)^{t/2}) \eta^{1-t/2k}\right) \langle u^{\otimes t}, M_*^{(t)} \rangle. \tag{4.39}$$

Hence, (4.35), (4.36), and (4.39) imply the desired utility guarantees.

Moreover, recall that the rejection probabilities at Steps 1 and 2 are each at most $\frac{1}{30}$, and it is not possible to reject in Step 3. Moreover, the $\leqslant k$ utility guarantees each fail with probability at most $\frac{1}{30k}$. Thus, by a union bound, it follows that the algorithm does not reject and, moreover, outputs estimates satisfying the desired utility guarantees with probability at least $1 - \frac{1}{30} - \frac{1}{30} - k \cdot \frac{1}{30k} = \frac{9}{10}$.

Finally, note that the running time of $(Bn)^{O(k)}$ follows from the time complexity guarantee in Lemma 4.5, as the invocation of Algorithm 4.3 in Step 2 is the bottleneck. Steps 1 and 3 are easily seen to run in $(Bn)^{O(k)}$ time. This completes the proof. □

5 Robust Mean and Covariance Estimation for Certifiably Hypercontractive Distributions

In this section, we observe that we can upgrade our guarantees from the previous section for robust estimation of moments of distributions that have certifiably hypercontractive degree 2 polynomials.

Definition 5.1. A distribution cD on $\mathbb{R}^d$ with mean μ_* and covariance Σ_* is said to have $2h$-certifiably C-hypercontractive degree 2 polynomials if for a $d \times d$ matrix-valued indeterminate Q and $\bar{x} = x - \mu_*$,

$$\vdash_{2h}^{Q} \left\{ \mathop{\mathbb{E}}_{x \sim D} (\bar{x}^\top Q \bar{x} - \mathop{\mathbb{E}}_{x \sim D} \bar{x}^\top Q \bar{x})^{2h} \leqslant (Ch)^{2h} \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2h} \right\}.$$

The Gaussian distribution [KOTZ14], uniform distribution on the hypercube and more generally other product domains and their affine transforms are known to satisfy $2t$-certfiably C-hypercontractivity with an absolute constant C for every t.

In order to derive this conclusion, we note the following analog of the witness-producing algorithm and its guarantees:

Witness-producing version of the robust moment estimation algorithm We will use the following (non-private) guarantees for the robust moment estimation algorithm in the previous section that hold for a strengthening of the constraint system $\mathcal{A}$ with certifiable hypercontractivity constraints. Using the analysis of [DKK+16], the following guarantees were recently shown in [KMZ21] for the case when the unknown distribution is Gaussian.

For any $d \times d$ matrix-valued indeterminate Q, let $\bar{x'_i}^\top Q \bar{x'_i} = x'^\top Q x' - \frac{1}{n}\sum_{i=1}^n x_i'^\top Q x'$.

$\mathcal{A}$: Constraint System for η-Robust Moment Estimation

1. $w_i^2 = w_i$ for each $1 \leqslant i \leqslant n$,
2. $\Pi^2 = \frac{1}{n}\sum_{i=1}^n (x'_i - \mu')(x'_i - \mu')^\top$,
3. $\sum_{i=1}^n w_i \geqslant (1-\eta)n$,
4. $\mu' = \frac{1}{n}\sum_i x'_i$,
5. $w_i(x'_i - y_i) = 0$ for $1 \leqslant i \leqslant n$,
6. $\frac{1}{n}\sum_{i=1}^n \bar{x'_i}^\top Q \bar{x'_i}^2 \leqslant C\,\|\Pi Q \Pi\|_F^2$.

The following guarantees for the algorithm above were shown in [BK20a].

Fact 5.2 ([BK20a]). *Let $X \subseteq \mathbb{R}^d$ be an i.i.d. sample of size $n \geqslant n_0 = \widetilde{O}(d^2/\eta)$ from $\mathcal{N}(\mu_*, \Sigma_*)$. Let Y be an η-corruption of X. Then, for $\mu' = \frac{1}{n}\sum_i x'_i$, $\Sigma' = \frac{1}{n}\sum_i (x_i - \mu')(x_i - \mu')^\top$, we have:*

$$\mathcal{A} \left|\frac{u}{O(k)}\right. \left\{\langle \mu' - \mu_*, u\rangle \leqslant O(\eta^{1-1/2k}) u^\top \Sigma_* u^2\right\},$$

$$\mathcal{A} \left|\frac{u}{O(k)}\right. \left\{\langle u, \Sigma' - \Sigma_*, u\rangle \leqslant O(\eta^{1-1/k}) u^\top \Sigma_* u\right\},$$

$$\mathcal{A} \left|\frac{}{O(k)}\right. \left\{\left\|\Sigma_*^{-1/2}\Sigma'\Sigma_*^{-1/2} - I\right\|_F^2 \leqslant O(\eta^{1-1/k})\right\}.$$

The first two guarantees of the lemma below were shown in [BK20a]. The third guarantee follows from an argument similar to that of Lemma 4.6. Notice that the key difference in the guarantees below (compared to the ones in Lemma 4.5) is the bound on the Frobenius (instead of the weaker spectral) distance between the estimated covariance and true unknown covariance.

Lemma 5.3 (Guarantees for Witness-Producing Robust Moment Estimation Algorithm). *Given a subset of of n points $Y \subseteq \mathbb{Q}^d$ whose entries have bit complexity B, Algorithm 4.3 runs in time $(Bn)^{O(1)}$ and either (a.) outputs "reject," or (b.) returns a sequence of weights $0 \leqslant p_1, p_2, \ldots, p_n$ satisfying $p_1 + p_2 + \cdots + p_n = 1$.*

Moreover, if there exists a set $X \subseteq \mathbb{R}^d$ of points with 4-certifiably C-hypercontractive degree 2 polynomials with mean μ_*, covariance Σ_*, then Algorithm 4.3 does not reject, and the corresponding estimates $\hat{\mu} = \frac{1}{n}\sum_i p_i y_i$ and $\hat{\Sigma} = \sum_{i=1}^n p_i (y_i - \hat{\mu})(y_i - \hat{\mu})^\top$ satisfy the following guarantees:

1. ***Mean Estimation:***

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u\rangle \leqslant O(\sqrt{C})\eta^{3/4}\sqrt{u^\top \Sigma_* u},$$

2. ***Covariance Estimation:***
$$\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \leqslant O(C\eta^{1/2}),$$

3. ***Witness:*** *For* $C' \leqslant C(1 + O(\eta^{1/2}))$,
$$\vdash^{Q} \left\{\frac{1}{n}\sum_{i=1}^{n} p_i \left(\langle y_i - \hat{\mu}, Q(y_i - \hat{\mu})\rangle - \frac{1}{n}\sum_{i=1}^{n} p_i \langle y_i - \hat{\mu}, Q(y_i - \hat{\mu})\rangle\right)^2 \leqslant C' \left\|\hat{\Sigma}^{1/2} Q \hat{\Sigma}^{1/2}\right\|_F^2\right\}$$

We can now use the above witness-producing algorithm to obtain a stronger Frobenius norm estimation guarantee with (ε, δ)-privacy for Gaussian distributions. Notice that the only change from the previous section is in the choice of the constraint system $\mathcal{A}$ and the corresponding change in the witness checking step.

Algorithm 5.4 (Private Robust Moment Estimation)**.**

Given: A set of points $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$, parameters $\eta, \varepsilon, \delta > 0$, $L \in \mathbb{N}$.

Output: Estimates $\hat{\mu}$ and $\hat{\Sigma}$ for mean and covariance.

Operation:

1. **Stable Outlier Rate Selection:** Use the $(\varepsilon/3, \delta/3)$-DP Selection with $\kappa = L/2$ to sample an integer $\tau \in [\eta n]$ with the scoring function as defined in Definition 4.12. If $\tau = \perp$, then reject and halt. Otherwise, let $\eta' = \tau/n$.
2. **Witness Checking:** Compute a pseudo-distribution $\tilde{\zeta}$ of degree $O(1)$ satisfying $\mathcal{A}$ on input Y with outlier rate η' and minimizing $\mathrm{Pot}_{\eta',\tilde{\zeta}}(Y)$. Let $\gamma \sim \mathrm{tLap}\left(-\left(1 + \frac{3\ln(3/\delta)}{\varepsilon}\right), 3/\varepsilon\right)$. Check that the weight vector $p = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]$ induces a distribution on Y that has $(C + \gamma)$-certifiably hypercontractive polynomials. If not, reject immediately. Otherwise, let $\widetilde{\mu} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\mu]$ and $\widetilde{\Sigma} = \widetilde{\mathbb{E}}_{\tilde{\zeta}}[\Sigma]$.
3. **Noise Addition:** Let $\gamma_1 = O(C')(L/n)^{\frac{1}{4}}$ and $\gamma_2 = O(C')(L/n)^{\frac{1}{4}}$. Let $z \sim \mathcal{N}(0, \sigma_1)^d$ and $Z \sim \mathcal{N}(0, \sigma_2)^{\binom{d+1}{2}}$, where we interpret Z has a symmetric $d \times d$ matrix with independent lower-triangular entries, and $\sigma_j = 12\varepsilon^{-1}\gamma_j\sqrt{2\ln(15/\delta)}$ for $1 \leqslant j \leqslant 2$. Then, output:
   - $\hat{\mu} = \widetilde{\mu} + \widetilde{\Sigma}^{1/2} z$.
   - $\hat{\Sigma} = \widetilde{\Sigma} + \widetilde{\Sigma}^{1/2} Z \widetilde{\Sigma}^{1/2}$.

The parameter closeness from potential stability is also upgraded from Corollary 4.19:

Lemma 5.5 (Parameter Closeness from Stability of Potential)**.** *Let* $\eta, \varepsilon, \delta > 0$ *and* $L \in \mathbb{N}$ *be given input parameters to Algorithm 5.4 such that* $0.25\eta n \geqslant L = \Omega\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\beta\delta}\right)\right)$*. Also, let* Y, Y' *be adjacent subsets of* $\mathbb{Q}^d$*. Suppose* Alg *does not reject in any of the 3 steps, uses the constant* C' *in Step 2 and chooses* η' *in Step 1 on input* Y *and* Y'*.*

Then, for every $u \in \mathbb{R}^d$ and $\theta = \sqrt{L/n}$, we have:

$$\langle \mu_p - \mu_{p'}, u \rangle \leqslant O(C')\theta^{3/4}\sqrt{u^\top \Sigma_p u}$$

and

$$\left\| \Sigma_p^{-1/2} \Sigma_{p'} \Sigma_p^{-1/2} - I \right\|_F \leqslant O(C')\theta^{1/2} .$$

The following theorem summarizes our privacy and utility guarantees for the algorithm above. We specialize to the "base case assumption" of 4-certifiable C-hypercontractivity of degree 2 polynomials in order to derive explicit bounds here. Our analysis of the algorithm above follows *mutatis mutandis* with the key upgrade being the stronger Frobenius norm guarantees in Lemma 4.18 that hold under certifiably hypercontractivity constraints in our constraint system $\mathcal{A}$ (this requires us to use a version of Lemma 4.21 that makes use of a bound on $\left\|AA^T - I\right\|_F$ instead of $\left\|AA^T - I\right\|_2$; see the remark at the end of Lemma 4.21). As before, the $\widetilde{\Omega}$ notation hides logarthmic multiplicative factors in d, C, $1/\eta$, $1/\varepsilon$, and $\ln(1/\delta)$.

Theorem 5.6 (Private Robust Mean and Covariance Estimation for Certifiably Hypercontractive Distributions)**.** *Fix $C_0 > 0$. Then, there exists an $\eta_0 > 0$ such that for any given outlier rate $0 < \eta \leqslant \eta_0$ and $\varepsilon, \delta > 0$, there exists a randomized algorithm* Alg *that takes an input of $n \geqslant n_0 = \widetilde{\Omega}\left(\frac{d^8}{\eta^2}\left(1 + \frac{\ln(1/\delta)}{\varepsilon}\right)^4 \cdot C^4\right)$ points $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$ (where $C = C_0 + \frac{3\ln(3/\delta)}{\varepsilon} + \frac{9}{\varepsilon} + 1$), runs in time $(Bn)^{O(1)}$ (where B is the bit complexity of the entries of Y) and outputs either "reject" or estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d \times d}$ with the following guarantees:*

1. ***Privacy:*** *Alg is (ε, δ)-differentially private with respect to the input Y, viewed as a d-dimensional database of n individuals.*

2. ***Utility:*** *Suppose there exists a 4-certifiably C_0-subgaussian set $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{Q}^d$ such that $|Y \cap X| \geqslant (1 - \eta_0)n$ with mean μ_* and covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)} I$. Then, with probability at least $9/10$ over the random choices of the algorithm, Alg outputs estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d \times d}$ satisfying the following guarantees:*

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u \rangle \leqslant O(\sqrt{C}\eta^{3/4})\sqrt{u^\top \Sigma_* u} ,$$

and,

$$\left\| \Sigma_*^{-1/2} \hat{\Sigma} \Sigma_*^{-1/2} - I \right\|_F \preceq O(C\sqrt{\eta}) .$$

Moreover, the algorithm succeeds (i.e., does not reject) with probability at least $9/10$ over the random choices of the algorithm.

When specialized to Gaussian distributions, the Frobenius guarantee above is suboptimal—the robust estimation algorithms of [DKK+16] allow estimating the mean and covariance of the unknown Gaussian distribution to an error $\widetilde{O}(\eta)$. We can in fact recover the stronger guarantees by relyong on the analysis in [KMZ21][Theorem 1 and 2] of the same constraint system above for the case of Gaussian distributions (in the "utility case"). This yields the following corollary:

Theorem 1.3 (Mean and Covariance Estimation for Gaussian Distributions). *Fix $\varepsilon, \delta > 0$. Then, there exists an absolute constant $\eta_0 > 0$ such that for any given outlier rate $0 < \eta \leqslant \eta_0$, there exists a randomized algorithm* Alg *that takes an input of $n \geqslant n_0 = \widetilde{\Omega}\left(\frac{d^8}{\eta^4}\left(1 + \frac{\ln(1/\delta)}{\varepsilon}\right)^4\right)$ points $Y \subseteq \mathbb{Q}^d$, runs in time $(Bn)^{O(1)}$ (where B is the bit complexity of the entries of Y) and outputs either "reject" or estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d\times d}$ with the following guarantees:*

1. ***Privacy:*** *Alg is (ε, δ)-differentially private with respect to the input Y, viewed as a d-dimensional database of n individuals.*

2. ***Utility:*** *Let $X = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample of size $n \geqslant n_0$ from a Gaussian distribution with mean μ_* and covariance $\Sigma_* \succeq 2^{-\mathrm{poly}(d)} I$ such that Y is an η-corruption of X. Then, with probability at least $9/10$ over the random choices of the algorithm, Alg outputs estimates $\hat{\mu} \in \mathbb{Q}^d$ and $\hat{\Sigma} \in \mathbb{Q}^{d\times d}$ satisfying the following guarantees [Ameya: I added in the $\log(1/\delta)$ dependence to be explicit. Can you check whether this lets us get rid of the tilde on the O?]:*

$$\forall u \in \mathbb{R}^d,\ \langle \hat{\mu} - \mu_*, u\rangle \leqslant \widetilde{O}\left(\eta \cdot \frac{\log(1/\delta)}{\varepsilon}\right)\sqrt{u^\top \Sigma_* u}\,,$$

and,

$$\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \leq \widetilde{O}\left(\eta \cdot \sqrt{\frac{\log(1/\delta)}{\varepsilon}}\right).$$

In particular, $d_{\mathsf{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu_*, \Sigma_*)) < \widetilde{O}(\eta \log(1/\delta)/\varepsilon)$.

References

[AAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 185–216. PMLR, 2021. 2

[Abo18] John M. Abowd. The u.s. census bureau adopts differential privacy. KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery. 2

[AL21] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. *CoRR*, abs/2111.11320, 2021. 1, 7

[App17] Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 2017. 2

[BDH+20] A. Bakshi, I. Diakonikolas, S. B. Hopkins, D. Kane, S. Karmalkar, and P. K. Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 149–159. IEEE, 2020. 6

[BDKU20] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan R. Ullman. Coinpress: Practical private mean and covariance estimation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[BEM+17] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnés, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017*, pages 441–459. ACM, 2017. 2

[BGS+21] Gavin Brown, Marco Gaboardi, Adam D. Smith, Jonathan R. Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *CoRR*, abs/2106.13329, 2021. 2

[BK20a] A. Bakshi and P. Kothari. Outlier-robust clustering of non-spherical mixtures. *CoRR*, abs/2005.02970, 2020. 2, 14, 43

[BK20b] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. 2020. 5, 6, 17

[BK20c] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *CoRR*, abs/2005.02970, 2020. 6

[BKS15] B. Barak, J. A. Kelner, and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method [extended abstract]. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, pages 143–151. ACM, New York, 2015. 16

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 156–167, 2019. 2

[BP20] A. Bakshi and A. Prasad. Robust linear regression: Optimal rates in polynomial time. *arXiv preprint arXiv:2007.01394*, 2020. 14

[BS19] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 181–191, 2019. 2

[BUV14] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, pages 1–10. ACM, 2014. 7

[CDGW19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 727–757. PMLR, 2019. 6

[CKM+20] Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan R. Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 7

[CWZ19] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *CoRR*, abs/1902.04495, 2019. 2

[DFM+20] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *CoRR*, abs/2001.02285, 2020. 2

[DHKK20] Ilias Diakonikolas, Samuel B. Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *CoRR*, abs/2005.06417, 2020. 2, 6, 14

[DHL19] Yihe Dong, Samuel B. Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6065–6075, 2019. 6

[DK19] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. 6

[DKK+16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 655–664, 2016. 2, 6, 43, 45

[DKK+17a] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 2017. 6

[DKK+17b] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. *CoRR*, abs/1704.03866, 2017. 6

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006. 37

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3571–3580, 2017. 2

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 371–380. ACM, 2009. 21, 54

[DLCC07] Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Trans. Signal Process.*, 55(6, part 2):2965–2973, 2007. 5

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006. 2, 20

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM, 2003. 7

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 37, 38, 39

[DSS+15] Cynthia Dwork, Adam D. Smith, Thomas Steinke, Jonathan R. Ullman, and Salil P. Vadhan. Robust traceability from trace amounts. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 650–669. IEEE Computer Society, 2015. 7

[DSSU17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017. 7

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, pages 1054–1067, 2014. 2

[FKP19] Noah Fleming, Pravesh Kothari, and Toniann Pitassi. Semialgebraic proofs and efficient algorithm design. *Foundations and Trends® in Theoretical Computer Science*, 14(1-2):1–221, 2019. 14, 15

[Gre16] Andy Greenberg. Apple's "differential privacy" is about collecting your data – but not your data. *Wired, June*, 13, 2016. 2

[HK13] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*, pages 11–19. ACM, New York, 2013. 5

[HKM21] Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. *CoRR*, abs/2111.12981, 2021. 7

[HL18] S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1021–1034, 2018. 2, 6, 14

[HLZ20] Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[Hop20] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213, 2020. 6

[JH16] Cédric Josz and Didier Henrion. Strong duality in Lasserre's hierarchy for polynomial optimization. *Optim. Lett.*, 10(1):3–10, 2016. 15

[JLT20] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[KKM18] A. Klivans, P. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Proc. 31st Annual Conference on Learning Theory (COLT)*, pages 1420–1430, 2018. 14

[KLSU19] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan R. Ullman. Privately learning high-dimensional distributions. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902. PMLR, 2019. 2, 10

[KMS+21] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan R. Ullman. A private and computationally-efficient estimator for unbounded gaussians. *CoRR*, abs/2111.04609, 2021. 7

[KMZ21] Pravesh K. Kothari, Peter Manohar, and Brian Hu Zhang. Polynomial-time sum-of-squares can robustly estimate mean and covariance of gaussians optimally, 2021. 43, 45

[KOTZ14] Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou. Hypercontractive inequalities via sos, and the frankl-rödl graph. In *SODA*, pages 1644–1658. SIAM, 2014. 42

[KS17a] P. K. Kothari and J. Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017. 2, 4

[KS17b] P. K. Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017. 2, 5, 8, 11, 14, 19, 23, 24, 25, 26

[KS17c] Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017. 14

[KSKO20] Weihao Kong, Raghav Somani, Sham M. Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[KSS18] P. K. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1035–1046, 2018. 6

[KSSU19] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan R. Ullman. Differentially private algorithms for learning mixtures of separated gaussians. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 168–180, 2019. 7

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan R. Ullman. Private mean estimation of heavy-tailed distributions. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 2020. 2

[KV18] Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science*

Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA, volume 94 of *LIPIcs*, pages 44:1–44:9. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. 2

[Las01] Jean B. Lasserre. New positive semidefinite relaxations for nonconvex quadratic programs. In *Advances in convex analysis and global optimization (Pythagorion, 2000)*, volume 54 of *Nonconvex Optim. Appl.*, pages 319–331. Kluwer Acad. Publ., Dordrecht, 2001. 17

[LKKO21] Xiyang Liu, Weihao Kong, Sham M. Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *CoRR*, abs/2102.09159, 2021. 2, 8

[LKO21] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. *CoRR*, abs/2111.06578, 2021. 7

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016. 2, 6

[LY20] Jerry Li and Guanghao Ye. Robust gaussian covariance estimation in nearly-matrix multiplication time. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 94–103. IEEE Computer Society, 2007. 22, 56

[Nes00] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, volume 33 of *Appl. Optim.*, pages 405–440. Kluwer Acad. Publ., Dordrecht, 2000. 17

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In David S. Johnson and Uriel Feige, editors, *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84. ACM, 2007. 7

[Par00] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000. 17

[SCV18] J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Proc. 9th Innovations in Theoretical Computer Science Conference (ITCS)*, pages 45:1–45:21, 2018. 2, 6

[Sho87] N. Z. Shor. Quadratic optimization problems. *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, (1):128–139, 222, 1987. 17

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017. 7

[SU15] Thomas Steinke and Jonathan R. Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1588–1628. JMLR.org, 2015. 7

[Tao12] T. Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Society, 2012. 41

[Vad17] Salil P. Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017. 22

[WDFS17] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S. Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear algebra and its applications*, 520:44–66, 2017. 39

[WXDX20] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10081–10091. PMLR, 2020. 2

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *CoRR*, abs/1909.08755, 2019. 6

[ZKKW20] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning markov random fields. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11129–11140. PMLR, 2020. 7

A Missing Proofs from Section 3.6

A.1 Proof of Lemma 3.27

Proof of Lemma 3.27. Consider any neighboring datasets Y, Y' and let $\mathcal{M}$ denote the truncated Laplace mechanism (with parameter as specified). Let p, q denote the probability density functions of $\mathcal{M}(Y), \mathcal{M}(Y')$. Observe that $p(x) \leqslant e^{\varepsilon} \cdot q(x)$ for all $x < \min\{f(Y), f(Y')\}$. Thus, we have

$$D_{e^{\varepsilon}}(p, q) = \int_{x \in \mathbb{R}} [p(x) - e^{\varepsilon} q(x)]_+ dx$$

$$
\begin{aligned}
&= \int_{x \geqslant \min\{f(Y), f(Y')\}} [p(x) - e^{\varepsilon} q(x)]_+ dx \\
&\leqslant \int_{x \geqslant \min\{f(Y), f(Y')\}} p(x) dx \\
\text{(Since sensitivity of } f \text{ is at most } \Delta) &\leqslant \int_{x \geqslant f(Y) - \Delta} p(x) dx \\
\text{(Lemma 3.28)} &\leqslant \delta,
\end{aligned}
$$

which means that the truncated Laplace mechanism is indeed (ε, δ)-DP. □

A.2 Proof of Lemma 3.30

The proof of the composition lemma follows from that of the standard adaptive composition of approximate DP proof [DL09, Theorem 16]. Below we use the notation $[x]_+$ to denote $\max\{x, 0\}$ and $x \wedge y$ to denote $\min\{x, y\}$.

Proof of Lemma 3.30. It suffices to prove the theorem for $k = 2$ as we may then apply induction to arrive at the statement for any positive integer k. To prove the case $k = 2$, consider any $S \subseteq O_2 \cup \{\perp\}$ and any pair of neighboring datasets Y, Y'.

For any $S_1 \subseteq O_1 \cup \{\perp\}$, we define the measure $\mu(S_1) := [\mathbb{P}[\mathcal{M}_1(Y) \in S_1] - e^{\varepsilon_1} \mathbb{P}[\mathcal{M}_1(Y') \in S_1]]_+$. Note that we have $\mu(O_1) \leqslant \delta_1$ due to our assumption that $\mathcal{M}_1$ is $(\varepsilon_1, \delta_1)$-DP.

Now consider four cases:

- **Both Y, Y' satisfy Ψ_1.** In this case, we may appeal to $(\varepsilon_2, \delta_2)$-DP under Ψ_1 of $\mathcal{M}_2$ which implies

$$
\mathbb{P}[\mathcal{M}_2(o_1, Y) \in S] \leqslant (e^{\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] \wedge 1) + \delta_2. \tag{A.1}
$$

  For ease of notation, let $p_Y : O_1 \to \mathbb{R}^+$ denote the measure obtained by restricting the probability density function of $\mathcal{M}_1(Y)$ to O_1 (note that $\int_{O_1} p_Y(o_1)\, do_1 = 1 - \mathbb{P}[\mathcal{M}_1(Y) = \perp]$). Then, observe that

$$
\begin{aligned}
\mathbb{P}[\mathcal{M}(Y) \in S] &= \mathbf{1}[\perp \in S]\, \mathbb{P}[\mathcal{M}_1(Y) = \perp] + \int_{O_1} \mathbb{P}[\mathcal{M}_2(o_1, Y) \in S] p_Y(o_1)\, do_1 \\
&\overset{(A.1)}{\leqslant} \mathbf{1}[\perp \in S]\, \mathbb{P}[\mathcal{M}_1(Y) = \perp] + \int_{O_1} ((e^{\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] \wedge 1) + \delta_2)\, p_Y(o_1)\, do_1 \\
&\leqslant \mathbf{1}[\perp \in S]\, \mathbb{P}[\mathcal{M}_1(Y) = \perp] + \delta_2 \\
&\quad + \int_{O_1} (e^{\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] \wedge 1) p_Y(o_1)\, do_1 \\
&\leqslant \mathbf{1}[\perp \in S](e^{\varepsilon_1} \mathbb{P}[\mathcal{M}_1(Y') = \perp] + \mu(\{\perp\})) + \delta_2 \\
&\quad + \int_{O_1} (e^{\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] \wedge 1)(e^{\varepsilon_1} p_{Y'}(o_1)\, do_1 + d\mu(o_1)) \\
&\leqslant \mathbf{1}[\perp \in S](e^{\varepsilon_1} \mathbb{P}[\mathcal{M}_1(Y') = \perp]) + \mu(O_1 \cup \{\perp\}) + \delta_2
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \int_{O_1} (e^{\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] \wedge 1)(e^{\varepsilon_1} p_{Y'}(o_1))\, do_1 \\
&\leqslant \mathbf{1}[\perp \in S](e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}_1(Y') = \perp]) + \delta_1 + \delta_2 \\
&\quad + \int_{O_1} e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}_2(o_1, Y') \in S] p_{Y'}(o_1)\, do_1 \\
&\leqslant \delta_1 + \delta_2 + e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}(Y') \in S].
\end{aligned}
$$

- Y **satisfies** Ψ_1 **but** Y' **does not.** In this case, we have $\mathbb{P}[\mathcal{M}(Y') = \perp] = 1$, which implies that

  $$\mathbb{P}[\mathcal{M}(Y) \in S] - e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}(Y') \in S] \leqslant \mathbb{P}[\mathcal{M}(Y) \neq \perp] = \mathbb{P}[\mathcal{M}(Y) \neq \perp] - e^{\varepsilon_1} \mathbb{P}[\mathcal{M}(Y') \neq \perp] \leqslant \delta_1,$$

  where the last inequality follows from the fact that $\mathcal{M}_1$ is $(\varepsilon_1, \delta_1)$-DP.

- Y' **satisfies** Ψ_1 **but** Y **does not.** In this case, we have $\mathbb{P}[\mathcal{M}(Y) = \perp] = 1$, which implies that

  $$
  \begin{aligned}
  \mathbb{P}[\mathcal{M}(Y) \in S] - e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}(Y') \in S] &\leqslant [\mathbb{P}[\mathcal{M}(Y) = \perp] - e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}(Y') = \perp]]_+ \\
  &\leqslant [\mathbb{P}[\mathcal{M}(Y) = \perp] - e^{\varepsilon_1} \mathbb{P}[\mathcal{M}(Y') = \perp]]_+ \\
  &\leqslant \delta_1,
  \end{aligned}
  $$

  where the last inequality once again follows from the fact that $\mathcal{M}_1$ is $(\varepsilon_1, \delta_1)$-DP.

- Neither Y nor Y' satisfy Ψ_1. In this case, both $\mathcal{M}(Y)$ and $\mathcal{M}(Y')$ always output $\perp$. Therefore, we have $\mathbb{P}[\mathcal{M}(Y) \in S] = \mathbb{P}[\mathcal{M}(Y') \in S]$.

Thus, in all cases, we have $\mathbb{P}[\mathcal{M}(Y) \in S] = e^{\varepsilon_1+\varepsilon_2} \mathbb{P}[\mathcal{M}(Y') \in S] + \delta_1 + \delta_2$ as desired. □

A.3 Proof of Lemma 3.33

Proof of Lemma 3.33. Then, note that

$$
\begin{aligned}
D_{e^\varepsilon}(p, r) &= \int_{x \in \mathbb{R}^d} [p(x) - e^\varepsilon r(x)]_+ \, dx \\
&= \int_{x \in \mathbb{R}^d} [(p(x) - e^{\varepsilon/2} q(x)) + (e^{\varepsilon/2} q(x) - e^\varepsilon r(x))]_+ \, dx \\
&\leqslant \int_{x \in \mathbb{R}^d} [(p(x) - e^{\varepsilon/2} q(x))]_+ \, dx + \int_{x \in \mathbb{R}^d} [e^{\varepsilon/2} q(x) - e^\varepsilon r(x))]_+ \, dx \\
&= \int_{x \in \mathbb{R}^d} [(p(x) - e^{\varepsilon/2} q(x))]_+ \, dx + e^{\varepsilon/2} \int_{x \in \mathbb{R}^d} [q(x) - e^{\varepsilon/2} r(x))]_+ \, dx \\
&= D_{e^{\varepsilon/2}}(p, q) + e^{\varepsilon/2} \cdot D_{e^{\varepsilon/2}}(q, r),
\end{aligned}
$$

as desired. □

A.4 Proof of Theorem 3.34

As stated earlier, the proof of Theorem 3.34 follows from applying the exponential mechanism [MT07] and then use the truncated Laplace mechanism (Lemma 3.27) to check that the score indeed exceeds κ.

Proof of Theorem 3.34. Selection works as follows:

1. First, run the $(\varepsilon/2)$-DP exponential mechanism [MT07], i.e. selecting each $c \in C$ with probability proportional to $\exp\left(\frac{\varepsilon}{4\Delta} \cdot \text{score}(c, Y)\right)$. Let c_1 be the output of this procedure.
2. Sample the noise $N \sim \text{tLap}\left(-\Delta\left(1 + \frac{2\ln(1/\delta)}{\varepsilon}\right), \frac{2\Delta}{\varepsilon}\right)$ and compute $\widetilde{\text{score}} = \text{score}(c_1, Y) + N$. If $\widetilde{\text{score}} \geqslant \kappa$, then output c_1. Otherwise, output $\perp$.

We will now prove each of the claimed properties:

1. The first step satisfies $(\varepsilon/2)$-DP via the standard privacy guarantee of the exponential mechanism [MT07]. The second step is $(\varepsilon/2, \delta)$-DP due to Lemma 3.27. Thereby, applying the basic composition theorem implies that Selection is (ε, δ)-DP.
2. Since $N \leqslant 0$, we are guarantee that if the algorithm outputs $c^* \in C$, we must have $\text{score}(c, Y) \geqslant \kappa$ as desired.
3. For any $c \in C$, the standard utility analysis of the exponential mechanism [MT07] implies that, with probability $1 - 0.5\beta$, we have $\text{score}(c_1, Y) \geqslant \text{score}(c, Y) - O\left(\frac{\Delta}{\varepsilon} \ln\left(\frac{|C|}{\beta}\right)\right)$. Moreover, the tail bound of Laplace noise (Lemma 3.28) implies that with probability $1 - 0.5\beta$ we have $N \geqslant -\Delta\left(1 + \frac{\ln(1/\delta)}{\varepsilon}\right) - O\left(\frac{\Delta}{\varepsilon}\ln(1/\beta)\right) \geqslant -O\left(\frac{\Delta}{\varepsilon}\ln\left(\frac{1}{\delta\beta}\right)\right)$. Therefore, if $\text{score}(c, Y) \geqslant \kappa + O\left(\frac{\Delta}{\varepsilon}\ln\left(\frac{|C|}{\delta\beta}\right)\right)$, the probability that the algorithm outputs $\perp$ is at most β, as desired.

□