Robustness Implies Privacy in Statistical Estimation*

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

Gautam Kamath†
g@csail.mit.edu
University of Waterloo
Waterloo, ON, Canada

Mahbod Majid‡
mahbodm@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Shyam Narayanan§
shyamsn@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

ABSTRACT

We study the relationship between adversarial robustness and differential privacy in high-dimensional algorithmic statistics. We give the first *black-box reduction from privacy to robustness* which can produce private estimators with optimal tradeoffs among sample complexity, accuracy, and privacy for a wide range of fundamental high-dimensional parameter estimation problems, including mean and covariance estimation. We show that this reduction can be implemented in polynomial time in some important special cases. In particular, using nearly-optimal polynomial-time robust estimators for the mean and covariance of high-dimensional Gaussians which are based on the Sum-of-Squares method, we design the first polynomial-time private estimators for these problems with nearly-optimal samples-accuracy-privacy tradeoffs. Our algorithms are also robust to a nearly optimal fraction of adversarially-corrupted samples.

CCS CONCEPTS

• **Theory of computation** → **Sample complexity and generalization bounds**; **Semidefinite programming**; • **Mathematics of computing** → **Multivariate statistics**; *Probabilistic algorithms*; • **Security and privacy** → *Information-theoretic techniques*.

KEYWORDS

Robustness, Differential Privacy, Gaussians, Parameter Estimation

ACM Reference Format:
Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. 2023. Robustness Implies Privacy in Statistical Estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3564246.3585115

*Full version of the paper available at https://arxiv.org/abs/2212.05015
†Supported by an NSERC Discovery Grant, an unrestricted gift from Google, and a University of Waterloo startup grant.
‡Supported by an NSERC Discovery Grant. Work done as a student at the University of Waterloo.
§Supported by an NSF Graduate Fellowship, the NSF TRIPODS Program (award DMS-2022448), and a Google Fellowship.

STOC '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 978-1-4503-9913-5/23/06.
https://doi.org/10.1145/3564246.3585115

1 INTRODUCTION

Parameter estimation is a fundamental statistical task: given samples $X_1, \dots, X_n$ from a distribution $p_\theta(X)$ belonging to a known family of distributions $\mathcal{P}$ and indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^D$, and for a given a norm $\|\cdot\|$, the goal is find $\hat{\theta}$ such that $\|\theta - \hat{\theta}\|$ is as small as possible. Two important desiderata for parameter estimation algorithms are:

Robustness: If an η-fraction of $X_1, \dots, X_n$ are adversarially corrupted, we would nonetheless like to estimate θ. This *strong contamination model* for robust parameter estimation dates from the 1960's, but has recently been under intense study from an algorithmic perspective, especially in the high-dimensional setting where $X_1, \dots, X_n \in \mathbb{R}^d$ for large d. Thanks to these efforts, we now know efficient algorithms for a wide range of high-dimensional parameter estimation problems which enjoy optimal or nearly-optimal accuracy/sample complexity guarantees.

Privacy: A *differentially private (DP)* [20] algorithm protects the privacy of individuals represented in a dataset $X_1, \dots, X_n$ by guaranteeing that the distribution of outputs of the algorithm given $X_1, \dots, X_n$ is statistically close to the distribution it would generate given $X_1', \dots, X_n'$, where $X_1', \dots, X_n'$ differs from $X_1, \dots, X_n$ on any one sample X_i.

Privacy and robustness are intuitively related: both place requirements on the behavior of an algorithm when one or several inputs are adversarially perturbed. Already by 2009, Dwork and Lei recognized that "robust statistical estimators present an excellent starting point for differentially private estimators" [19]. More recent works continue to leverage ideas from robust estimation to design private estimation procedures [9, 11, 23, 26, 32, 37, 40, 44, 45] – these works address both sample complexity and computationally efficient algorithms.

Despite robustness being useful as a tool in privacy, the relationship between robustness and privacy remains murky. Consequently, for many high-dimensional estimation tasks, we know polynomial-time algorithms which obtain (nearly) optimal tradeoffs among accuracy, sample complexity, and robustness, but known private algorithms either require exponential time or give suboptimal tradeoffs among accuracy, sample complexity, and privacy. Indeed, this is the case even for *learning the mean of a high-dimensional (sub-)*

Gaussian distribution, and for learning a high-dimensional Gaussian in total variation distance.

We contribute a new technique to design private estimators using robust ones, leading to:

The first black-box reduction from private to robust estimation: Prior works using robust estimators to design private ones are *white box,* relying on properties of those estimators beyond robustness. Black-box privacy techniques such as the Gaussian and Laplace mechanisms are widely used, but so far do not yield private algorithms for high-dimensional estimation tasks with optimal accuracy-samples-privacy tradeoffs, even when applied to optimal robust estimators. *For tasks including mean and covariance estimation and regression, using any robust estimator with an optimal accuracy-samples-robustness tradeoff, our reduction gives a private estimator with optimal accuracy-samples-privacy tradeoff.*

Our basic black-box reduction yields estimators satisfying *pure* DP, which work assuming Θ is bounded, and which don't necessarily admit efficient algorithms. Two additional properties of an underlying robust estimator can lead to potential improvements in the resulting private estimator:

(1) If Θ is convex and the robust estimator is based on the *Sum of Squares* (SoS) method, the resulting private estimator can often be implemented in polynomial time.
(2) If the robust estimator satisfies a stronger *worst-case* robustness property, satisfied by many high-dimensional robust estimators, we can remove the assumption that Θ is bounded, at the additional (necessary) expense of weakening from pure to *approximate* DP guarantees.

The first polynomial-time algorithms to learn high-dimensional Gaussian distributions with nearly-optimal sample complexity subject to differential privacy: Using SoS-based robust algorithms and our privacy-to-robustness reduction, we obtain polynomial-time estimators with nearly-optimal accuracy-samples-privacy tradeoffs, for both pure and approximate DP, for learning the mean and/or covariance of a high-dimensional Gaussian, and for learning a high-dimensional Gaussian in total variation. In addition, our private algorithms enjoy near-optimal levels of robustness. Prior private polynomial-time estimators have sub-optimal samples-accuracy-privacy tradeoffs, losing polynomial factors in the dimension d and/or privacy parameter $\log 1/\delta$.

Our methods also yield a polynomial-time algorithm for private mean estimation under a bounded-covariance assumption, recovering the main result of [26] with slightly improved sample complexity. We expect them to generalize to other estimation problems where Θ is convex and nearly-optimal robust SoS algorithms are known – e.g., linear regression [34] and mean estimation under other bounded-moment assumptions [27, 35].

Conclusions on Robust versus Private Estimation: Recent work [23] shows that private algorithms *with very high success probabilities* are robust simply by virtue of their privacy guarantees. This complements our results, which show a converse – from robust estimators with optimal samples-accuracy-robustness tradeoffs we get analogous private estimators (with very high success probabilities). Together, these hint at a potential *equivalence* between robust and private parameter estimation, which can be made algorithmic in the context of SoS-based algorithms. Our results show such an equivalence for "nice enough" parameter estimation problems, but the broader relationship between privacy and robustness is more subtle; in Section 2 we discuss situations where optimal robust estimators don't necessarily yield optimal private ones, at least in a black-box way.

1.1 Results

We first recall the definitions of differential privacy and the strong contamination model.

Definition 1.1 (Differential Privacy (DP) [18, 20]). Let $\mathcal{X}$ be a set of *inputs* and $\mathcal{X}^*$ be all finite-length strings of inputs. Let $\mathcal{O}$ be a set of *outputs.* A randomized map ("mechanism") $M : \mathcal{X}^* \to \mathcal{O}$ satisfies (ε, δ)-DP if for every *neighboring* $X, X' \in \mathcal{X}^*$ with Hamming distance 1 and every subset $S \subseteq \mathcal{O}$, $\mathbb{P}(M(X) \in S) \le e^\varepsilon\, \mathbb{P}(M(X') \in S) + \delta$. If $\delta = 0$, we say that M satisfies *pure* DP, otherwise M satisfies *approximate* DP.

Definition 1.2 (Strong Contamination Model). For a probability distribution D and $\eta > 0$, $Y_1, \dots, Y_n$ are η-corrupted samples from D if $X_1, \dots, X_n \overset{i.i.d.}{\sim} D$ and $Y_i = X_i$ for at least $(1-\eta)n$ indices i.

1.1.1 *Learning High-Dimensional Gaussian Distributions in TV Distance.*

We begin with our results on learning Gaussians in total variation distance.

THEOREM 1.3 (LEARNING ARBITRARY GAUSSIANS, PURE DP). *Assume that $0 < \alpha, \beta, \varepsilon < 1$, $0 < \eta < \eta^*$ for some absolute constant η^*, and $K, R > 1$. There is a polynomial-time $(\varepsilon, 0)$-DP algorithm with the following guarantees for every $d \in \mathbb{N}$ and every $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ such that $\|\mu\| \le R$ and $\frac{1}{K} \cdot I \preceq \Sigma \preceq K \cdot I$. Given n η-corrupted samples from $\mathcal{N}(\mu, \Sigma)$, the algorithm returns $\hat{\mu}, \hat{\Sigma}$ such that $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \le \alpha + \widetilde{O}(\eta)$ with probability at least $1 - \beta$, if*[1]

$$n \ge \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2} + \frac{d^2 + \log(1/\beta)}{\alpha\varepsilon} + \frac{d^2 \log K}{\varepsilon} + \frac{d \log R}{\varepsilon}\right).$$

We are unaware of prior computationally efficient pure-DP algorithms for learning high-dimensional Gaussians in TV distance; we believe that state of the art is based on the techniques of [29],[2] which would give an algorithm requiring $n \gg d^3$ samples (and lack robustness).

Pure-DP necessitates the *a priori* upper bounds R and K on μ and Σ in Theorem 1.3. Under (ε, δ)-DP these bounds are avoidable. But, obtaining a polynomial-time (ε, δ)-DP algorithm to learn Gaussians with optimal samples-accuracy-privacy tradeoffs and without assumptions on μ, Σ has been a significant challenge, with progress in several recent works [3, 31, 37, 48] (see Table 1). These algorithms require a number of samples exceeding the information-theoretic optimum by polynomial factors in either d, $\log(1/\delta)$, or both.

We give the first polynomial-time (ε, δ)-DP algorithm for learning an arbitrary high-dimensional Gaussian distribution with nearly-optimal sample complexity with respect to *all* of: dimension, accuracy, privacy, and corruption rate. Ours is the first $\tilde{O}(d^2)$-sample

[1]With more careful analysis, we expect that the error bound can be tightened to $\alpha + O(\eta \log 1/\eta)$, which is expected to be tight for statistical query algorithms [15]; the same goes for our other results on learning Gaussians.
[2]replacing the Gaussian mechanism with the Laplace mechanism

polynomial-time robust and private estimator; prior works require $\Omega(d^{3.5})$ samples [3, 48].

Theorem 1.4 (Learning Arbitrary Gaussians, (ε,δ)-DP). *Assume that $0 < \alpha, \beta, \delta, \varepsilon < 1$, and $0 < \eta < \eta^*$ for some absolute constant η^*. There is a polynomial-time (ε,δ)-DP algorithm with the following guarantees for every $d \in \mathbb{N}$, $\mu \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succ 0$.*[3] *Given n η-corrupted samples from $\mathcal{N}(\mu,\Sigma)$, the algorithm returns $\hat{\mu}, \hat{\Sigma}$ such that $d_{TV}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\hat{\mu},\hat{\Sigma})) \le \alpha + \widetilde{O}(\eta)$ with probability at least $1-\beta$, if*

$$n \ge \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2} + \frac{d^2 + \log(1/\beta)}{\alpha\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right).$$

The sample-complexity guarantees of Theorems 1.3 and 1.4 are information-theoretically tight up to logarithmic factors in d, α, ε, and $\log 1/\delta$. The $\log(1/\beta)/\alpha\varepsilon$ term in each is potentially improvable to $\min(\log(1/\beta), \log(1/\delta))/\alpha\varepsilon$, and the $\log^2(1/\beta)$ term is potentially improvable to $\log(1/\beta)$. However, this still means our algorithms succeed with exponentially small (e^{-d}) failure probability, with no blowup in the sample complexity.

1.1.2 Estimating the Mean of a Subgaussian Distribution. Mean estimation in high dimensions subject to differential privacy has also received substantial recent attention [9, 11, 12, 26, 29, 32, 33, 39, 40]. We focus on the following simple problem: given (corrupted) samples from $\mathcal{N}(\mu, I)$, find $\hat{\mu}$ such that $\|\mu - \hat{\mu}\| \le \alpha$. In the pure-DP setting, exponential-time estimators are known which achieve this guarantee using $n \approx \frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon}$ samples [11, 32]. Existing polynomial-time estimators require $n \gg \min(\frac{d}{\alpha^2\varepsilon}, \frac{d^{1.5}}{\varepsilon})$ samples or satisfy a weaker privacy guarantee [26, 29] (see Table 2). We give the first nearly-sample-optimal pure-DP algorithm:

Theorem 1.5 (Estimating the Mean of a Spherical Subgaussian Distribution). *Assume that $0 < \alpha, \beta, \varepsilon < 1$, $0 < \eta < \eta^*$ for some absolute constant η^*, and $R > 1$. There is a polynomial-time $(\varepsilon, 0)$-DP algorithm with the following guarantees for every $d \in \mathbb{N}$, every $\mu \in \mathbb{R}^d$ with $\|\mu\| \le R$, and every subgaussian distribution D on $\mathbb{R}^d$ with mean μ and covariance I. Given n η-corrupted samples from D, the algorithm returns $\hat{\mu}$ such that $\|\mu - \hat{\mu}\| \le \alpha + \widetilde{O}(\eta)$ with probability at least $1-\beta$, as long as*

$$n \ge \widetilde{O}\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{d + \log(1/\beta)}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon}\right).$$

It is natural to ask whether the identity-covariance assumption can be removed from Theorem 1.5, since information-theoretically the assumption of covariance $\Sigma \preceq I$ is enough to obtain the same guarantees. Removing this assumption while retaining polynomial running time and high-probability privacy guarantees would improve over state-of-the-art algorithms for robust mean estimation which have withstood significant efforts at improvement [28].

There is also an analogue for polynomial-time mean estimation subject to (ε,δ)-DP without the $\|\mu\| \le R$ assumption, using $\tilde{O}(\frac{d}{\alpha\varepsilon} + \frac{d}{\alpha^2} + \frac{\log 1/\delta}{\varepsilon})$ samples. We obtain this result from our approx-DP framework similar to proving Theorem 1.4: one could alternatively combine Theorem 1.5 with an (ε,δ)-DP procedure that obtains an $O(d)$-accurate estimate, such as [22]. The analogue is formally stated and proven as Theorem 5.2 in the full version of this paper.

Finally, we note that Theorems 1.3 and 1.5 are known to be near-optimal from standard packing lower bounds [11], and Theorem 1.4 and its approx-DP analogue are also known to be near-optimal, via the technique of fingerprinting [29, 30], except, as in Theorems 1.3 and 1.4, that $\log(1/\beta)/\alpha\varepsilon$ is potentially improvable to $\min(\log(1/\beta), \log(1/\delta))/\alpha\varepsilon$. All our algorithmic results are applications of Theorems 4.1 and 4.2 in the full version of the paper, which give general tools for turning SoS-based robust estimators into private ones.

1.2 Related Work

Our work joins three bodies of literature too large to survey here: on private and high-dimensional parameter estimation, on high-dimensional statistics via SoS (see [42]), and on high-dimensional algorithmic robust statistics (see [14]). We discuss other works at the intersections of these areas.

Private and Robust Estimators: [19] first used robust statistics primitives to design private algorithms, a tradition continued by [9, 11, 26, 32, 37, 40, 44]. Some of these works attempt to give generic recipes for converting robust algorithms to private ones [37, 40], though do not give a black-box reduction as we do in Lemmas 2.1 and 2.2. Other works from the Statistics community also investigate connections between robustness and privacy [7, 8, 45, 46], including local differential privacy [38]. Our black-box reduction from privacy to robustness can be seen as a generalization of methods of [11, 32], which also instantiate the exponential mechanism with a score function counting the minimum point changes to achieve some accuracy guarantee, but for specific robust estimators. A recent line of work focuses on simultaneously private and robust estimators for high-dimensional statistics [3, 11, 22, 24, 37, 39, 40, 48]; see Tables 1, 2.

Recall that [23] observes that pure-DP algorithms which succeed *with sufficiently high probability over the internal coins of the algorithm* are automatically robust to a constant fraction of corrupted inputs. While optimal inefficient private estimators often satisfy this high-probability requirement, most existing polynomial-time private estimators do not. Our private estimators have not only (nearly) optimal sample complexity but also (nearly) optimal success probability.

Private Estimators via SoS: [26] and [37] pioneer the use of SoS for private algorithm design. [26] gives a polynomial-time algorithm for pure-DP mean estimation under a bounded covariance assumption, using $\frac{d}{\alpha^2\varepsilon}$ samples, and [37] gives a $\approx d^8$-sample (ε,δ)-DP algorithm for learning d-dimensional Gaussians. [23] uses SoS for private *sparse* mean estimation.

On a technical level, our work most resembles [26]; we also employ SoS SDPs as score functions and leverage tools from log-concave sampling. However, there are fundamental roadblocks to using [26]'s strategy for converting SoS proofs into private algorithms in settings beyond mean estimation under bounded covariance, as we discuss in Section 2. We provide a blueprint for

[3] We suppress running-time dependence on $\log K$, where K is the condition number of Σ; logarithmic dependence on the condition number orthogonal to $\ker(\Sigma)$ is necessary for learning Gaussians in TV, regardless of privacy or robustness. Note that the sample complexity has no such dependence on $\log K$.

Table 1: Private covariance estimation of Gaussians in Mahalanobis distance, omitting logarithmic factors. Optimal robustness means the algorithm succeeds even with $\tilde{\Omega}(\alpha)$-fraction of corruptions.

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[33]	$\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\min(\log K, \log \delta^{-1})}{\varepsilon}, d = 1$	No	Yes	Pure/Approximate
[29]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d^{3/2}\sqrt{\log K \log \delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[11]	$\frac{d^2}{\alpha^2} + \frac{d^2 \log K}{\alpha\varepsilon}$	Optimal	No	Pure
[1]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[40]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[31]	$\frac{d^2}{\alpha^2} + \left(\frac{d^2}{\alpha\varepsilon} + \frac{d^{5/2}}{\varepsilon}\right) \cdot (\log \delta^{-1})^{O(1)}$	No	Yes	Approximate
[37]	$\frac{d^8}{\alpha^4} \cdot \left(\frac{\log \delta^{-1}}{\varepsilon}\right)^6$	Suboptimal	Yes	Approximate
[3, 48]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d \log \delta^{-1}}{\varepsilon}$	No	Yes	Approximate
[3, 48]	$\frac{d^{3.5} \log \delta^{-1}}{\alpha^3 \varepsilon}$	Optimal	Yes	Approximate
Thm 1.3	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{d^2 \log K}{\varepsilon}$	Optimal	Yes	Pure
Thm 1.4	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	Yes	Approximate

Table 2: Private mean estimation of identity-covariance Gaussians in ℓ_2-norm, omitting logarithmic factors. Optimal robustness means the algorithm succeeds even with $\tilde{\Omega}(\alpha)$ fraction of corruptions.

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[33]	$\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\min(\log R, \log \delta^{-1})}{\varepsilon}, d = 1$	No	Yes	Pure/Approximate
[29]	$\frac{d}{\alpha^2} + \frac{d\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{\sqrt{d \log R \log \delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[11]	$\frac{d}{\alpha^2} + \frac{d \log R}{\alpha\varepsilon}$	Optimal	No	Pure
[32]	$\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon}$	Optimal	No	Pure
[1]	$\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[39]	$\frac{d}{\alpha^2} + \frac{d^{3/2} \log \delta^{-1}}{\alpha\varepsilon}$	Optimal	Yes	Approximate
[11, 40]	$\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[26]	$\frac{d}{\alpha^2\varepsilon} + \frac{d \log R}{\varepsilon}$	Suboptimal	Yes	Pure
Theorem 1.5	$\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon}$	Optimal	Yes	Pure
Theorem 1.5+[22]	$\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	Yes	Approximate

converting a much wider range of SoS-based robust algorithms to private ones.

Inverse Sensitivity Mechanism: In [4, 5], Asi and Duchi design private polynomial-time algorithms for statistical problems with an *inverse sensitivity mechanism* which is closely related to our black-box reduction, as described in (1). However, the focus of their work is rather different, as they investigate applications to instance-optimal private estimation, whereas our goal is to understand private estimation through the lens of robustness. Furthermore, their study is centered on one-dimensional statistics, and their analysis is not black-box.

Contemporaneous work: In independent and simultaneous work, Alabi, Kothari, Tankala, Venkat, and Zhang also design efficient robust and private algorithms for learning high-dimensional Gaussians with nearly-optimal sample complexity with respect to dimension; however, their algorithms require poly$(1/\varepsilon, \log 1/\delta, 1/\alpha)$-factors more samples than those we present [2]. In another independent and simultaneous work, Asi, Ullman, and Zakynthinou introduce the same black-box transformation from robustness to privacy [6]. To contrast the two works: we go beyond this inefficient reduction, and also design efficient algorithms for Gaussian estimation. On the other hand, they show the transformation gives the optimal error for low-dimensional problems, showing tightness of the robustness-privacy connection in certain settings. Finally, two works subsequent to ours give computationally-efficient algorithms for mean estimation in Mahalanobis distance while requiring only a near-linear number of samples [10, 17], improving on the exponential time algorithm of [9]. Both new works are based on "stable" estimators for mean and covariance, where stability is a notion of robustness different from the one we consider in this work.

2 TECHNIQUES

2.1 Black-Box Reduction from Privacy to Robustness

Consider a deterministic[4] robust estimator $\hat{\theta}$: datasets $\to \Theta$ for a parameter space $\Theta \subset \mathbb{R}^D$, a distribution family $\mathcal{P}$, and a norm $\|\cdot\|$, with the following guarantee: for a non-decreasing function $\alpha : [0,1] \to \mathbb{R}$ and some $n \in \mathbb{N}$, with probability $1-\beta$ over samples $X_1, \dots, X_n \sim p_\theta \in \mathcal{P}$, for every $\eta \in [0,1]$, given any η-corruption of $X_1, \dots, X_n$, the estimator obtains $\|\hat{\theta} - \theta\| \le \alpha(\eta)$. That is, α is a function that quantifies the error achieved by the estimator for every corruption level η. Let X denote an n-vector dataset $X_1, \dots, X_n$, and $d(X, X')$ be the Hamming distance between the datasets X, X'.

Our key conceptual contribution is the following instantiation of the exponential mechanism [41]: Given $\varepsilon > 0, X_1, \dots, X_n$ and a threshold $\eta_0 \in [0,1]$, the mechanism picks a random $\theta \in \Theta + \alpha(\eta_0) \cdot B_{\|\cdot\|}$ with:

$$\mathbb{P}(\theta) \propto \exp(-\varepsilon \cdot \text{score}_X(\theta)) \text{ where}$$
$$\text{score}_X(\theta) = \min\{d(X, X') : \|\hat{\theta}(X') - \theta\| \le \alpha(\eta_0)\}, \quad (1)$$

where $B_{\|\cdot\|}$ is the unit ball of $\|\cdot\|$. In words: the mechanism assigns each θ within distance $\alpha(\eta_0)$ of Θ a score given by the number of input samples which would have to be changed to obtain a dataset X' for which the robust estimator $\hat{\theta}(X')$ is close to θ, and samples θ with probability $\propto \exp(-\varepsilon \cdot \text{score}_X(\theta))$. If Θ is unbounded these probabilities are not well defined; in that case pure-DP guarantees are not obtainable anyway, due to packing lower bounds [25]. Later, we use a *truncated* version of (1) to allow unbounded Θ with (ε, δ)-DP.

The general idea to instantiate the exponential mechanism where the score of some θ is the number of inputs which must be changed to make some function $\hat{\theta}$ take the value (approximately) θ appears to be folklore; see for instance the *inverse sensitivity mechanism* of [5]. Our contribution is (a) to show that for (1) to have nontrivial utility guarantees, it suffices for $\hat{\theta}$ to be robust to adversarial corruptions, and (b) to show how to implement variants of (1) in polynomial time.

To elucidate the role of and how to set the threshold parameter η_0: if the target bound on the error of our private estimator is some value α, we can think of η_0 as the maximum amount of contamination a robust estimator could tolerate if the goal was to achieve the same error α. This will depend on the distribution class $\mathcal{P}$; for example, if we consider the class of distributions with bounded covariance $\Sigma \preceq I$, then the appropriate setting is $\eta_0 = \Theta(\alpha^2)$ [13, 47].

The exponential mechanism enjoys $(2\varepsilon, 0)$-DP, but the question of utility remains. Suppose that $X_1, \dots, X_n \sim p_{\theta^*}$. How small is $\|\theta - \theta^*\|$? The following lemma bounds this quantity in terms of the robustness of $\hat{\theta}$. Despite its simplicity, we are not aware of a similar result in the literature.

[4] If we are not concerned with running time, the deterministic assumption is without loss of generality, as any randomized estimator can be converted to a deterministic one with at most a constant-factor loss in accuracy, by enumerating over all choices of the estimator's internal random coins and selecting an output which is contained in a ball which contains at least 50% of the mass of the estimator's output distribution.

Lemma 2.1. *Suppose a dataset $X_1, \dots, X_n \sim p_{\theta^*}$, where the parameter vector $\theta^* \in \Theta \subseteq \mathbb{R}^D$. For any threshold $\eta_0 \in [0,1]$, a random θ drawn according to (1) has $\|\theta - \theta^*\| \le 2\alpha(\eta_0)$ with probability at least $1 - 2\beta$, if*

$$n \ge \max_{\eta_0 \le \eta \le 1} \frac{D \cdot \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log(1/\beta) + O(\log \eta n)}{\eta \varepsilon}. \quad (2)$$

Observe that the $O(\log \eta n)$ term in (2) is negligible compared to $D \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} \ge D \log 2$ if $n \ll 2^D$.

The sample complexity in (2) is a maximum over the parameter η; we pay a cost in samples depending on the underlying robust estimator's robustness profile, taking the worst case over all corruption levels η. The price at each η scales roughly as the log-volume of the set of solutions which satisfy the robust estimator's accuracy level under η-corruptions. The more robust the estimator is, the smaller this volume will be, matching the intuition that settings which permit more robust estimation also are easier to privatize.

A *robust* analogue of Lemma 2.1, in which the dataset $X_1, \dots, X_n$ is a *contamination* of i.i.d. samples from p_{θ^*}, follows by a similar proof.

Proof. Condition on the $(1-\beta)$-probable event that the robustness guarantees of $\hat{\theta}$ hold with respect to X. Consider θ with score ηn. By definition, $\|\theta - \hat{\theta}(X')\| \le \alpha(\eta_0)$ for some X' with $d(X, X') \le \eta \cdot n$. By robustness, $\|\hat{\theta}(X') - \theta^*\| \le \alpha(\eta)$. Using triangle inequality, $\|\theta - \theta^*\| \le \alpha(\eta_0) + \alpha(\eta) \le 2\alpha(\eta)$, assuming $\eta \ge \eta_0$. In summary, any θ with score ηn is within distance $2\alpha(\eta)$ of θ^*.

Let V_r be the volume of a radius r $\|\cdot\|$-ball. Any θ such that $\|\theta - \hat{\theta}(X)\| \le \alpha(\eta_0)$ has score 0. The normalizing factor implicit in (1) can be lower bounded by the contribution due to these points, or $V_{\alpha(\eta_0)} \cdot \exp(-\varepsilon \cdot 0) = V_{\alpha(\eta_0)}$. Combining this with the argument above, the probability of seeing θ with score ηn with $\eta > \eta_0$ in a draw from (1) is at most $\frac{V_{2\alpha(\eta)}}{V_{\alpha(\eta_0)}} \exp(-\varepsilon \eta n)$. Summing over all scores $\ge \eta_0 n$, the overall probability of seeing some θ with score greater than η_0 is at most

$$\begin{aligned}
&\sum_{t=\eta_0 n}^{n} \frac{V_{2\alpha(t/n)}}{V_{\alpha(\eta_0)}} \cdot \exp(-\varepsilon t) \\
&= \sum_{t=\eta_0 n}^{n} \frac{V_{2\alpha(t/n)}}{V_{\alpha(\eta_0)}} \cdot \exp(-\varepsilon t) \cdot t^2 \cdot 1/t^2 \\
&\le O(1) \cdot \max_{\eta_0 \le \eta \le 1} \left\{ (\eta n)^2 \cdot \frac{V_{2\alpha(\eta)}}{V_{\alpha(\eta_0)}} \cdot \exp(-\varepsilon \eta n) \right\},
\end{aligned}$$

where the inequality is Hölder's. This quantity is at most β for n as in (2). So, with probability at least $1-\beta$ the random θ will have score at most $\eta_0 n$, meaning $\|\theta - \theta^*\| \le 2\alpha(\eta_0)$. At the beginning, we conditioned on a $(1-\beta)$-probable event, so the overall failure probability is at most 2β. □

Consequences of Lemma 2.1: Applied to robust mean estimators with optimal error rates under bounded k-th moment assumptions, for any $k \ge 2$, Lemma 2.1 gives optimal pure-DP estimators under those same assumptions, recovering the main results of [32]; applied to robust linear regression (with known covariance) [16], it yields a pure-DP analogue of the nearly-optimal regression result of [39]; and so on. The same argument can be adapted to perform covariance-aware mean estimation[5] and covariance-aware linear regression, recovering pure-DP versions of the results of [9, 39], using a robust estimator of mean and covariance.

To illustrate, we apply Lemma 2.1 to Gaussian mean estimation. With $n \gg d/\alpha^2$ samples from a d-dimensional Gaussian $\mathcal{N}(\mu, I)$, it is possible to estimate the mean under η-contamination with error $\|\hat{\mu} - \mu\| \le O(\alpha + \eta)$, if $\eta < 1/2$. For ε-DP guarantees, we need to restrict to the case of $\|\mu\| \le R$ for some (large) $R > 0$; we will assume that even for $\eta \ge 1/2$, $\|\hat{\mu}\| \le R$.

Plugging such a robust $\hat{\mu}$ into Lemma 2.1, and choosing $\eta_0 = \alpha$, there are two interesting cases: $\eta = O(\eta_0)$ and $\eta = 1$. In the former, $\alpha(2\eta_0)/\alpha(\eta_0) = O(1)$, so we get the requirement $n \ge O(\frac{d+\log(1/\beta)}{\alpha\varepsilon})$, and in the latter $\alpha(1) = R$, so we get the additional requirement $n \ge \frac{d \log R}{\varepsilon}$, meaning that we obtained an ε-DP estimator with accuracy $O(\alpha)$ using n samples,

$$n \gg \frac{d + \log(1/\beta)}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon} + \frac{d}{\alpha^2}.$$

This is tight up to constants [11, 25]. Similarly tight results can be derived for mean estimation under bounded covariance, covariance estimation, linear regression, and more. We remind that the resulting private algorithms are *not* computationally efficient, though we will see how this approach can be made efficient for several interesting cases.

When Is Lemma 2.1 Loose? More refined analyses of the construction (1) are possible. In particular, if the robust estimator $\hat{\theta}$ enjoys the property that the *volume* of the sets of possible values it assumes under η-corrupted inputs are substantially smaller than $V_{2\alpha(\eta)}$, the bound in Lemma 2.1 can be improved accordingly (at the cost of breaking black-box-ness in the analysis.)

As an example, consider estimating the mean of a Gaussian $\mathcal{N}(\mu, I)$ to ℓ_∞ error α. Using a similar argument as in the ℓ_2 example above, Lemma 2.1 gives a sample-complexity upper bound of $\frac{\log d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon}$. But, because $d_{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \approx \|\mu - \mu'\|_2$, it's possible to construct a robust estimator $\hat{\mu}$ such that under η-corruptions, $\|\hat{\mu} - \mu\|_\infty$ can only be as large as η if $\|\hat{\mu} - \mu\|_2 \approx \|\hat{\mu} - \mu\|_\infty$; otherwise $\|\hat{\mu} - \mu\|_\infty$ is much smaller. This affords better control over the volumes of candidate outputs with a given score ηn than the η-radius ℓ_∞ ball would offer. Using this, we show in Appendix E in the full version of the paper that $\tilde{O}(\frac{\log d}{\alpha^2} + \frac{d^{2/3}}{\alpha\varepsilon^{2/3}} + \frac{\sqrt{d}}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon})$ samples are enough, in the pure-DP setting.

From Robustness to (ε, δ)-DP: If $\hat{\theta}$ has a nontrivial *breakdown point* – i.e., a fraction of corruptions η beyond which it admits no error guarantees, then Lemma 2.1 doesn't give a nontrivial private estimator. For example, in the Gaussian mean estimation setting, if we remove the assumption $\|\mu\| \le R$, then when $\eta \ge 1/2$ no estimator has a finite accuracy guarantee (i.e., $\alpha(\eta)$ is unbounded for such η).

By relaxing from pure to (ε, δ)-DP, however, we can design private estimators even from robust estimators $\hat{\theta}$ which have a breakdown point. Our reduction in this case, however, requires $\hat{\theta}$ to satisfy a *worst-case* robustness property, because we will need to appeal to robustness to ensure not only accuracy, as in Lemma 2.1, but also privacy, which is inherently a worst-case guarantee.

[5] a.k.a., mean estimation in Mahalanobis distance

Simple adaptations of standard robust estimators of mean and covariance, and robust regression algorithms, have such worst-case robustness guarantees. This approach gives an alternative to the high-dimensional propose-test-release framework of [40], and the approach of [9], for building approx-DP estimators from robust estimation primitives; we can recover their results on covariance-aware mean estimation and linear regression with (ε, δ)-DP guarantees. This approach carries the advantages of black-box-ness and potential polynomial-time implementability, since SoS-based robust estimators for mean and covariance have the required worst-case behavior.

Consider again a deterministic robust estimator $\hat{\theta}$: datasets $\to \Theta \cup \{\text{REJECT}\}$ for a parameter $\theta \in \mathbb{R}^d$, which takes n inputs and returns either some element of Θ or REJECT. Let $\mathcal{P}$ be a distribution family, $\|\cdot\|$ be a norm, $\alpha : [0, 1] \to \mathbb{R}$ be a non-decreasing function, $n \in \mathbb{N}$, and $\eta_0, \eta^* \in [0, 1]$. We continue to employ $\text{SCORE}_X(\theta)$ as defined in (1). Suppose as before that with probability $1 - \beta$ over samples $X_1, \ldots, X_n \sim p_\theta \in \mathcal{P}$, for every $\eta < \eta^*$, given any η-corruption of $X_1, \ldots, X_n$, $\|\hat{\theta} - \theta\| \le \alpha(\eta)$. And, suppose that $\hat{\theta}$ has the following worst-case robustness property: for *any* input $X = X_1, \ldots, X_n$, if $\hat{\theta}(X) \ne \text{REJECT}$, then for every $\eta < \eta^*$, given any η-corruption X' of X, either $\hat{\theta}(X') = \text{REJECT}$, or $\|\hat{\theta}(X') - \hat{\theta}(X)\| \le \alpha(\eta^*)$.

LEMMA 2.2. *Let $\eta_0 < \eta^* \in [0, 1]$ be such that $\eta^* n$ is a sufficiently large constant. For every $\varepsilon, \delta > 0$, there is an $(O(\varepsilon), O(e^{2\varepsilon}\delta))$-DP mechanism which, for any θ^*, takes $X_1, \ldots, X_n \sim p_{\theta^*}$ and with probability $1 - \beta$ outputs θ such that $\|\theta - \theta^*\| \le 2\alpha(\eta_0)$, if*

$$n \ge O\left(\max_{\eta_0 \le \eta \le \eta^*} \frac{D \cdot \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log(1/\beta) + \log \eta n}{\eta\varepsilon} + \frac{\log(1/\delta)}{\eta^* \varepsilon} \right).$$

Before proving the lemma, we need a preliminary claim.

PROPOSITION 2.3. *Suppose for a dataset X there exists θ such that $\text{SCORE}_X(\theta) < 0.2\eta^* n$. Then there exists a ball of radius $2\alpha(\eta^*)$ which contains every θ' with $\text{SCORE}_X(\theta') < 0.4\eta^* n$.*

PROOF. Since there exists some θ such that $\text{SCORE}_X(\theta) < 0.2\eta^* n$, there's some $Y \sim_{0.2\eta^*} X$ such that $\hat{\theta}(Y) \ne \text{REJECT}$: this is because we can consider any such Y which has $\text{SCORE}_Y(\theta) = 0$, and thus $\hat{\theta}(Y)$ outputs an element of Θ and not REJECT. Similarly, for any other θ' with $\text{SCORE}_X(\theta') \le 0.4\eta^* n$, there's some $Z \sim_{0.4\eta^*} X$ such that $\|\theta' - \hat{\theta}(Z)\| \le \alpha(\eta_0)$. By triangle inequality, $Z \sim_{0.6\eta^*} Y$, so by worst-case robustness of $\hat{\theta}$, $\|\theta' - \hat{\theta}(Y)\| \le \|\theta' - \hat{\theta}(Z)\| + \|\hat{\theta}(Z) - \hat{\theta}(Y)\| \le \alpha(\eta_0) + \alpha(\eta^*) \le 2\alpha(\eta^*)$. □

PROOF OF LEMMA 2.2. First, let $g : \mathbb{Z} \to \mathbb{R}$ be a function with the following properties: for $t < 0.1\eta^* n$, $g(t) = 1$, for $t > 0.2\eta^* n$, $g(t) = 0$, and for all t, $e^{-\varepsilon} g(t + 1) - \delta \le g(t) \le e^{\varepsilon} g(t + 1) + \delta$. Such a function exists since $n \gg \log \frac{1}{\delta} / \eta^* \varepsilon$.

This is not hard to show: one could, for example, consider the function which, for t over the interval $[0.1\eta^* n, 0.2\eta^* n]$, first decreases by a multiplicative factor of $e^{-\varepsilon}$ (i.e., $g(t + 1) = e^{-\varepsilon} g(t)$) until some point t^* when $g(t^*) \le \delta$. Then, we set $g(t) = 0$ for all $t > t^*$. This satisfies the requirements on the function for all $t \le t^*$ with $\delta = 0$, and for $t > t^*$ with $\varepsilon = 0$. We need that $\delta \ge \exp(-(t - 0.1\eta^* n)\varepsilon)$ is satisfied by some t in the interval

$[0.1\eta^* n, 0.2\eta^* n]$ (roughly speaking, to allow enough multiplicative $e^{-\varepsilon}$ decreases to accumulate in order to cancel out the remainder with a subtractive δ shift), which we can take to be t^*. Rearranging the inequality, we get $t \geq \log(1/\delta)/\varepsilon + 0.1\eta^* n$. But for t^* to lie in the stated interval, we need $\log(1/\delta)/\varepsilon + 0.1\eta^* n \leq t \leq 0.2\eta^* n$, which is satisfied as long as $n \gg \log(1/\delta)/\eta^*\varepsilon$, as claimed.

The mechanism is as follows. Given $X = X_1, \ldots, X_n$, let $T = \min_{\theta\in\Theta} \textsc{score}_X(\theta)$. First, output REJECT with probability $1 - g(T)$. If REJECT is not output, output a sample from the distribution on $\Theta + \alpha(\eta_0)B_{\|\cdot\|}$ where

$$\mathbb{P}(\theta) \propto \begin{cases} \textsc{score}_X(\theta) & \text{if } \textsc{score}_X(\theta) < 0.3\eta^* n \\ 0 & \text{otherwise} \end{cases}$$

and $B_{\|\cdot\|}$ is the unit ball for the norm $\|\cdot\|$.

Proof of privacy: The REJECT phase of the mechanism clearly satisfies (ε, δ)-DP, because $\textsc{score}_X(\theta)$ can change by at most 1 when X is replaced with neighboring X', and based on the definition of g.

Now we turn to the sampling phase. Let X, X' differ on one sample. Let T, T' be the numbers computed in the REJECT phase of the mechanism; we may assume $T, T' \leq 0.2\eta^* n$, since otherwise on both X, X' the mechanism outputs REJECT with probability at least $1 - \delta$. We show that the mechanism above, conditioned on not rejecting, satisfies $(O(\varepsilon), O(e^{2\varepsilon}\delta))$-DP; then the overall result follows by composition.

For brevity, we abbreviate $\textsc{score}_X$ to s_X. For any $S \subseteq \Theta + \alpha(\eta_0) \cdot B_{\|\cdot\|}$, we can bound its associated weight via

$$\begin{aligned}&\int_{\theta\in S} e^{-\varepsilon s_X(\theta)} \cdot \mathbf{1}(s_X(\theta) < 0.3\eta^* n) \\ &\leq e^{\varepsilon} \int_{\theta\in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \big[\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \\ &\qquad + \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]\big].\end{aligned}$$

To see why, first note that for any θ we have $|s_X(\theta) - s_{X'}(\theta)| \leq 1$. This implies that $e^{-\varepsilon s_X(\theta)} \leq e^{\varepsilon} e^{-\varepsilon s_{X'}(\theta)}$. Similarly, if $s_X(\theta) \leq 0.3\eta^* n$, it also implies that at least one of the following must be true (potentially both): $s_{X'}(\theta) \leq 0.3\eta^* n$ or $s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]$ (we use the fact that $\eta^* n$ is at least a sufficiently large constant).

Normalizing to get a probability, we have

$$\mathbb{P}_X(\theta \in S) \leq e^{\varepsilon} \cdot \frac{g}{h} \leq e^{\varepsilon} \cdot \frac{g}{h'},$$

where

$$\begin{aligned} g &= \int_{\theta\in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \big[\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \\ &\qquad + \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])\big], \\ h &= \int_{\theta\in\Theta+\alpha(\eta_0)B_{\|\cdot\|}} e^{-\varepsilon s_X(\theta)} \cdot \mathbf{1}(s_X(\theta) < 0.3\eta^* n), \\ h' &= e^{-\varepsilon} \int_{\theta\in\Theta+\alpha(\eta_0)B_{\|\cdot\|}} e^{-\varepsilon s_{X'}(\theta)} \cdot \big[\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \\ &\qquad - \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])\big]. \end{aligned}$$

The denominator h' is split into two terms with a similar argument as used for the numerator g.

We next simplify the denominator h'. Because, by assumption, there is θ' such that $\textsc{score}_{X'}(\theta') < 0.2\eta^* n$, there is a ball of radius $\alpha(\eta_0)$, contained in $\Theta + \alpha(\eta_0) \cdot B_{\|\cdot\|}$, of points with score at most $0.2\eta^* n$; we can hence lower-bound the first term $\int e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \geq \exp(-\varepsilon \cdot 0.2\eta^* n) \cdot V_{\alpha(\eta_0)}$, where $V_{\alpha(\eta_0)}$ is the volume of a $\|\cdot\|$-ball of radius $\alpha(\eta_0)$.

We can use Proposition 2.3 to upper-bound the magnitude of the second term in the denominator,

$$\begin{aligned}&\int e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]) \\ &\leq \exp(-\varepsilon \cdot 0.25\eta^* n) \cdot V_{2\alpha(\eta^*)},\end{aligned}$$

which is at most δ times the lower bound on the first term, under our hypotheses on the lower bound for n. Overall, we obtain

$$\mathbb{P}_X(\theta \in S) \leq \frac{e^{2\varepsilon}}{1-\delta} \cdot \frac{A+B}{C} \leq \frac{e^{2\varepsilon}}{1-\delta} \cdot D,$$

where

$$\begin{aligned} A &= \int_{\theta\in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \\ B &= \int_{\theta\in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]) \\ C &= \int_{\theta\in\Theta+\alpha(\eta_0)B_{\|\cdot\|}} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) \\ D &= \big(\mathbb{P}_{X'}(\theta \in S) + \mathbb{P}_{X'}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])\big). \end{aligned}$$

Using Proposition 2.3 in the same fashion to bound the last term, this is at most $e^{2\varepsilon}\,\mathbb{P}_{X'}(\theta \in S) + O(e^{2\varepsilon}\delta)$, which completes the privacy proof.

Proof of accuracy: Observe that with probability at least $1 - \beta$ over samples $X_1, \ldots, X_n$, the REJECT phase of the mechanism accepts with probability 1. Conditioned on it doing so, the remainder of the accuracy proof parallels the proof of Lemma 2.1, except instead of allowing $\eta \in [\eta_0, 1]$ we can now limit it to $\eta \in [\eta_0, \eta^*]$. ◻

2.2 Algorithms

Even if the robust estimator $\hat{\theta}$ can be computed in polynomial time, the sampling problem in (1) lacks an obvious polynomial-time algorithm, for two reasons. First, computing the score of a single $\theta \in \Theta$ given an input dataset X appears to require solving a minimization problem over all other datasets X'. Second, even if computing the scores were somehow made efficient, the resulting sampling problem might still be computationally hard. Our main technical contribution is to overcome both of these hurdles in the context of learning high-dimensional Gaussian distributions.

2.2.1 *Background: Sum of Squares and Robust Estimation.*

The Sum of Squares method (*SoS*) uses convex programming to solve multivariate systems of polynomial inequalities. It is extremely useful for designing polynomial-time robust estimators.

Definition 2.4 (SoS Proof). Let $p_1(x) \geq 0, \ldots, p_m(x) \geq 0$ be a system of polynomial inequalities in variables $x_1, \ldots, x_n$. An inequality $q(x) \geq 0$ has a *degree d SoS proof* from $p_1 \geq 0, \ldots, p_m \geq 0$, written $\{p_1 \geq 0, \ldots, p_m \geq 0\} \vdash_d^x q \geq 0$, if for each multiset $S \subseteq [m]$ there exists a sum of squares polynomial $q_S(x)$, such that $\deg(q_S(x) \cdot \prod_{i \in S} p_i(x)) \leq d$ and such that

$$q(x) = \sum_{S \subseteq [m]} q_S(x) \cdot \prod_{i \in S} p_i(x) \,.$$

SoS proofs form a convex set described by a semidefinite program (SDP), so they have duals:

Definition 2.5 (Pseudoexpectation). Let $\mathbb{R}[x]_{\leq d}$ be the set of degree at most d polynomials in variables $x_1, \ldots, x_n$. A linear operator $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ is a degree d *pseudoexpectation* if $\tilde{\mathbb{E}}1 = 1$ and $\tilde{\mathbb{E}}p^2 \geq 0$ for any p of degree at most $d/2$. A pseudoexpectation $\tilde{\mathbb{E}}$ *satisfies* a system of polynomial inequalities $p_1 \geq 0, \ldots, p_m \geq 0$, written $\tilde{\mathbb{E}} \models p_1 \geq 0, \ldots, p_m \geq 0$, if for every $S \subseteq [m]$ and every p, we have $\tilde{\mathbb{E}} \prod_{i \in S} p_i \cdot p^2 \geq 0$ when the degree of this polynomial is at most d, where $\|p\|$ is the ℓ_2-norm of the vector of coefficients of p in the monomial basis.

The by-now standard approach to use SoS to robustly estimate a D-dimensional parameter θ in a norm $\| \cdot \|$ works as follows. For η-corrupted $X = X_1, \ldots, X_n$ from p_{θ^*}, define a degree-$O(1)$ system of polynomial inequalities $\mathcal{A}(X, \theta, z)$ where $\theta = \theta_1, \ldots, \theta_D, z = z_1, \ldots, z_{(nD)^{O(1)}}$ are some indeterminates. With high probability, $\mathcal{A}(X, \theta, z)$ should (a) be satisfied by some choice of z when $\theta = \theta^*$, and (b) should have $\mathcal{A}(X, \theta, z) \vdash_{O(1)} \langle \theta - \theta^*, v \rangle \leq \alpha$ for every v in the dual ball of $\| \cdot \|$.

To give a robust estimation algorithm, on input η-corrupted X, we can obtain $\tilde{\mathbb{E}}$ which satisfies $\mathcal{A}(X, \theta, z)$ using semidefinite programming,[6] and then output $\hat{\theta} = \tilde{\mathbb{E}}\theta$. Applying $\tilde{\mathbb{E}}$ to the SoS proofs $\mathcal{A} \vdash_{O(1)}^{\theta, z} \langle \theta - \theta^*, v \rangle \leq \alpha$, we get $\|\tilde{\mathbb{E}}\theta - \theta^*\| \leq \alpha$.

Lemma 2.6 (Informal, implicit in [36]). *There exists $\mathcal{A}$ with the above properties with respect to $n \gg d/\eta^2$ η-corrupted samples from $\mathcal{N}(\theta^*, I)$, for any $\theta^* \in \mathbb{R}^d$, where $\| \cdot \| = \ell_2$, and $\alpha = \tilde{O}(\eta)$.*

2.2.2 Robustness to Privacy, Algorithmically.

For this technical overview, we focus on mean estimation in the pure-DP setting; similar ideas extend to covariance estimation and (ε, δ)-DP. Even for the SoS-based robust mean estimation algorithm described above, which we call KMZ, given X we do not know how to efficiently compute

$$\text{score}_X(\theta) = \min\{d(X, X') : \|\text{KMZ}(Y) - \theta\| \leq \alpha\} \,, \qquad (3)$$

much less sample from the distribution (1). At a very high level, will tackle these challenges by using the polynomial system $\mathcal{A}(X, \theta, z)$ underlying KMZ to design an SoS-based relaxation of the above score function, SoS-score$_X(\theta)$, which has favorable enough convexity properties that we will be able to both efficiently compute it and sample from the distribution it induces (both up to small error). The SoS robustness proofs which $\mathcal{A}$ enjoys will be enough for us to apply an argument like Lemma 2.1 to prove accuracy of the resulting estimator, and it will be private by construction.

First, we describe an attempt at an SoS relaxation of SoS-score, which will have several flaws we'll fix later. We can introduce more indeterminates $X_1', \ldots, X_n', w_1, \ldots, w_n, \theta'$, and consider

$$\mathcal{B}_t = \left\{ w_i^2 = w_i, \ \sum_{i=1}^n w_i = n - t, \ w_i X_i = w_i X_i', \right\} \cup \mathcal{A}(X', \theta', z) \,, \qquad (4)$$

which is satisfied when X' is a dataset with $d(X, X') \leq t$ and $\mathcal{A}(X', \theta', z)$ is satisfied. Let

$$\text{SoS-score}_X(\theta) = \min \ t \ \text{ s.t. } \exists \text{ degree } O(1) \ \tilde{\mathbb{E}} \text{ in variables } X', w, \theta', z, \ \tilde{\mathbb{E}} \models \mathcal{B}_t, \ \|\tilde{\mathbb{E}}\theta' - \theta\| \leq \alpha \,. \qquad (5)$$

Privacy and Accuracy for SoS-score: Suppose for a moment that SoS-score solves our computational problems. Does it lead to a good private estimator, when we sample from the distribution $\mathbb{P}(\theta) \propto \exp(-\varepsilon \cdot \text{SoS-score}_X(\theta))$? Standard arguments show privacy; the main question is accuracy.

It turns out the relaxation is tight enough that the proof of Lemma 2.1 still applies! The key step in that proof is to argue via robustness that if θ has low score, then $\|\theta^* - \theta\|$ is small. To establish the corresponding statement for SoS-score, we need to show that if $X_1, \ldots, X_n \sim \mathcal{N}(\theta^*, I)$ and $\tilde{\mathbb{E}} \models \mathcal{B}_t$ for $t = \eta n$, then $\|\tilde{\mathbb{E}}\theta' - \theta^*\| \leq \tilde{O}(\eta)$. This is slightly stronger than what we already know from the SoS proofs associated to $\mathcal{A}$, because now we have *indeterminates* X' which represent η-corrupted samples, rather than a fixed collection of η-corrupted samples, and we need $\mathcal{B}_t \vdash_{O(1)}^{X', \theta', w, z} \langle \theta' - \theta^*, v \rangle \leq \tilde{O}(\eta)$. Luckily, the SoS proofs of [36] readily generalize to show this.

In fact, [36]'s SoS proofs already show this in part because within the "auxiliary" indeterminates z they already use variables like our X' and w. This means that (4), (5), while closely following our black-box reduction strategy, contain an unnecessary layer of indirection. When we implement this strategy in detail (see Sections 5, 6, and 7 in the full version of this paper), we remove this indirection for simplicity.

On "Satisfies": An important technical difference between our score function and that of [26] is that the $\tilde{\mathbb{E}}$s it involves must have $\tilde{\mathbb{E}} \models \sum_{i=1}^n w_i = n - t$, rather than something weaker, like $\tilde{\mathbb{E}} \sum_{i=1}^n w_i = n - t$. While in some applications of SoS this "satisfies" versus "in expectation" distinction is minor, it is actually crucial for our accuracy guarantees – if we only required $\tilde{\mathbb{E}} \sum_{i=1}^n w_i = n - t$, we could have $\tilde{\mathbb{E}}$ which satisfies the rest of $\mathcal{B}_t$ but has $\|\tilde{\mathbb{E}}\theta' - \theta^*\| \geq \Omega(R)$, just by taking $\tilde{\mathbb{E}}$ to be the moments of a distribution which has all $w_i = 0$ with probability $1/t$.

However, this creates two significant technical challenges. First, for bit-complexity reasons, no polynomial-time algorithm to check if there exists $\tilde{\mathbb{E}}$ satisfying a given system of polynomials is known – existing techniques to find $\tilde{\mathbb{E}}$s work best in the context of *satisfiable* polynomial systems [43]. We sidestep this challenge by generalizing a technique from the robust statistics literature, which searches for $\tilde{\mathbb{E}}$ which *approximately* satisfies a system of polynomials, to the setting where those polynomials may be unsatisfiable. Ultimately, we find a further-relaxed score function SoS-score$'_X$, which we evaluate to error τ in $(nd \log 1/\tau)^{O(1)}$ time.

Quasi-Convexity, Sampling, and Weak Membership: The second challenge is that SoS-score$_X(\theta)$ need not be convex in θ – if it were,

[6]This ignores some issues of numerical accuracy which turn out to be important; see below.

we could sample from $\mathbb{P}(\theta) \propto \exp(-\varepsilon \cdot \text{SoS-score}_X(\theta))$ with log-concave sampling techniques, as in [26]. Indeed, consider θ_0 and θ_1 with corresponding scores t_0, t_1 witnessed by $\tilde{\mathbf{E}}_0, \tilde{\mathbf{E}}_1$. The problem is that $\frac{1}{2}(\tilde{\mathbf{E}}_0 + \tilde{\mathbf{E}}_1)$ need not satisfy $\sum_{i=1}^n w_i \geq n - \frac{1}{2}(t_0 + t_1)$, even though it does have $\frac{1}{2}(\tilde{\mathbf{E}}_0 + \tilde{\mathbf{E}}_1)[\sum_{i=1}^n w_i] \geq n - \frac{1}{2}(t_0 + t_1)$.

SoS-score$_X(\theta)$ is *quasi-convex* in θ, meaning that its sub-level sets $S_t = \{\theta : \text{SoS-score}_X(\theta) \leq t\}$ are convex for all t. This is good news: if we discretize the range of possible scores $[0, n]$ into $t_1, \ldots, t_{n^{O(1)}}$ (replacing SoS-score with a version rounded to the nearest t_i), we can hope to compute the *volumes* $V_i = \mathrm{Vol}(S_{t_i})$, as well as sample uniformly from the S_{t_i}s, using standard techniques for sampling from a convex body. Then, we could sample θ by first sampling a score t_i with probability proportional to $e^{-\varepsilon t_i}(1 - e^{-\varepsilon(t_{i+1} - t_i)})V_i$, then drawing uniformly from S_{t_i}.

Approximate sampling and volume algorithms for convex bodies typically access the body via a *weak membership oracle*, meaning that the oracle is allowed to give incorrect answers to query points very near the body's boundary.[7] We have access to an oracle which computes SoS-score$_X(\theta)$ up to exponentially-small errors. Ideally, we'd create a weak membership oracle by answering a query about S_{t_i} by checking if SoS-score$_X(\theta) \leq t_i$, but if SoS-score$_X$ is not Lipschitz, a small error in computing this value may translate to answering a query incorrectly about some θ far from the boundary of S_{t_i}. That is, we may not notice if $S_{t_i + 2^{-n}}$ is much larger than S_{t_i}.

However, because SoS-score$_X$ is bounded in $[0, n]$ and the sub-level sets are convex, we are able to show that $S_{t_i + 2^{-n}}$ could only be much larger than S_{t_i} at a small-measure set of t_is. Thus, if we choose our discretization $t_1, \ldots, t_{n^{O(1)}}$ *randomly*, with very high probability our approximate score oracle for SoS-score$_X$ translates to a weak membership oracle for the S_{t_i}s (see Lemma 4.7 in the full version of the paper).

Putting it Together: Thus, by modifying SoS-score$_X$ by (a) rounding to the nearest threshold t_i, thresholds chosen randomly, and (b) accounting for some numerical errors, we obtain a polynomial-time-samplable proxy for (1). Theorems 4.1 and 4.2 in the full version of the paper capture this strategy formally.

ACKNOWLEDGMENTS

We thank Xiyang Liu, Weihao Kong, and Sewoong Oh for helpful conversations at the beginning of this project. We also thank Lydia Zakynthinou and Pasin Manurangsi for making us aware of prior work on the inverse sensitivity mechanism.

REFERENCES

[1] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. 2021. On the Sample Complexity of Privately Learning Unbounded High-Dimensional Gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT '21)*. JMLR, Inc., 185–216. http://proceedings.mlr.press/v132/aden-ali21a.html

[2] Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. 2023. Privately estimating a Gaussian: Efficient, robust and optimal. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing (STOC '23)*. ACM, New York, NY, USA. https://doi.org/10.48550/arXiv.2211.07503

[3] Hassan Ashtiani and Christopher Liaw. 2022. Private and polynomial time algorithms for learning Gaussians and beyond. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT '22)*. 1075–1076. https://proceedings.mlr.press/v178/ashtiani22a.html

[4] Hilal Asi and John C Duchi. 2020. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems* 33 (2020), 14106–14117. https://proceedings.neurips.cc/paper/2020/hash/a267f936e54d7c10a2bb70dbe6ad7a89-Abstract.html

[5] Hilal Asi and John C Duchi. 2020. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630* (2020). https://doi.org/10.48550/arXiv.2005.10630

[6] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. 2023. From Robustness to Privacy and Back. *arXiv preprint arXiv:2302.01855* (2023). https://doi.org/10.48550/arXiv.2302.01855

[7] Marco Avella-Medina. 2020. The Role of Robust Statistics in Private Data Analysis. *Chance* 33, 4 (2020), 37–42. https://doi.org/10.1080/09332480.2020.1847958

[8] Marco Avella-Medina. 2021. Privacy-preserving parametric inference: a case for robust statistics. *J. Amer. Statist. Assoc.* 116, 534 (2021), 969–983. https://doi.org/10.1080/01621459.2019.1700130

[9] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. 2021. Covariance-Aware Private Mean Estimation Without Private Covariance Estimation. In *Advances in Neural Information Processing Systems 34 (NeurIPS '21)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/42778ef0b5805a96f9511e20b5611fce-Abstract.html

[10] Gavin Brown, Samuel B Hopkins, and Adam Smith. 2023. Fast, Sample-Efficient, Affine-Invariant Private Mean and Covariance Estimation for Subgaussian Distributions. *arXiv preprint arXiv:2301.12250* (2023). https://doi.org/10.48550/arXiv.2301.12250

[11] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. 2021. Private Hypothesis Selection. *IEEE Trans. Inf. Theory* 67, 3, 1981–2000. https://doi.org/10.1109/TIT.2021.3049802

[12] Mark Bun and Thomas Steinke. 2019. Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation. In *Advances in Neural Information Processing Systems 32 (NeurIPS '19)*. Curran Associates, Inc., 181–191. https://proceedings.neurips.cc/paper/2019/hash/3ef815416f775098fe977004015c6193-Abstract.html

[13] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Being Robust (in High Dimensions) Can Be Practical. In *Proceedings of the 34th International Conference on Machine Learning (ICML '17)*. JMLR, Inc., 999–1008. http://proceedings.mlr.press/v70/diakonikolas17a.html

[14] Ilias Diakonikolas and Daniel M Kane. 2019. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911* (2019). https://doi.org/10.48550/arXiv.1911.05911

[15] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2017. Statistical Query Lower Bounds for Robust Estimation of High-dimensional Gaussians and Gaussian Mixtures. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS '17)*. IEEE Computer Society, Washington, DC, USA, 73–84. https://doi.org/10.1109/FOCS.2017.16

[16] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. 2019. Efficient Algorithms and Lower Bounds for Robust Linear Regression. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '19)*. SIAM, Philadelphia, PA, USA, 2745–2754. https://doi.org/10.1137/1.9781611975482.170

[17] John Duchi, Saminul Haque, and Rohith Kuditipudi. 2023. A Fast Algorithm for Adaptive Private Mean Estimation. *arXiv preprint arXiv:2301.07078* (2023). https://doi.org/10.48550/arXiv.2301.07078

[18] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy via Distributed Noise Generation. In *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT '06)*. Springer, Berlin, Heidelberg, 486–503. https://doi.org/10.1007/11761679_29

[19] Cynthia Dwork and Jing Lei. 2009. Differential Privacy and Robust Statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing (STOC '09)*. ACM, New York, NY, USA, 371–380. https://doi.org/10.1145/1536414.1536466

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC '06)*. Springer, Berlin, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[21] Martin Dyer, Alan Frieze, and Ravi Kannan. 1991. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)* 38, 1 (1991), 1–17. https://doi.org/10.1145/102782.102783

[22] Hossein Esfandiari, Vahab S Mirrokni, and Shyam Narayanan. 2022. Tight and Robust Private Mean Estimation with Few Users. In *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*. JMLR, 16383–16412. https://proceedings.mlr.press/v162/narayanan22a.html

[23] Kristian Georgiev and Samuel B Hopkins. 2022. Privacy Induces Robustness: Information-Computation Gaps and Sparse Mean Estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS '22)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/hash/2d76b6a9f96181ab717c1a15ab9302e1-Abstract-Conference.html

[7] It seems to be folklore that volume computation algorithms, e.g. the seminal [21], work given only weak membership oracles, as opposed to e.g. weak separation oracles. For completeness, in Appendix A in the full version of the paper, we analyze a hit-and-run sampling algorithm which uses a weak membership oracle, tracking the numerical errors this creates.

[24] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. 2021. Robust and Private Learning of Halfspaces. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS '21, Vol. 130)*. PMLR, 1603–1611. http://proceedings.mlr.press/v130/ghazi21a.html

[25] Moritz Hardt and Kunal Talwar. 2010. On the Geometry of Differential Privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing (STOC '10)*. ACM, New York, NY, USA, 705–714. https://doi.org/10.1145/1806689.1806786

[26] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. 2022. Efficient Mean Estimation with Pure Differential Privacy via a Sum-of-Squares Exponential Mechanism. In *Proceedings of the 54th Annual ACM Symposium on the Theory of Computing (STOC '22)*. ACM, New York, NY, USA. https://doi.org/10.1145/3519935.3519947

[27] Samuel B. Hopkins and Jerry Li. 2018. Mixture Models, Robustness, and Sum of Squares Proofs. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing (STOC '18)*. ACM, New York, NY, USA, 1021–1034. https://doi.org/10.1145/3188745.3188748

[28] Samuel B Hopkins and Jerry Li. 2019. How hard is robust mean estimation?. In *Conference on Learning Theory*. PMLR, 1649–1682. http://proceedings.mlr.press/v99/hopkins19a.html

[29] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. 2019. Privately Learning High-Dimensional Distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT '19)*. 1853–1902. http://proceedings.mlr.press/v99/kamath19a.html

[30] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. 2022. New Lower Bounds for Private Estimation and a Generalized Fingerprinting Lemma. In *Advances in Neural Information Processing Systems 35 (NeurIPS '22)*. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9a6b278218966499194491f55ccf8b75-Abstract-Conference.html

[31] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. 2022. A Private and Computationally-Efficient Estimator for Unbounded Gaussians. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT '22)*. 544–572. https://proceedings.mlr.press/v178/kamath22a.html

[32] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. 2020. Private Mean Estimation of Heavy-Tailed Distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT '20)*. 2204–2235. http://proceedings.mlr.press/v125/kamath20a.html

[33] Vishesh Karwa and Salil Vadhan. 2018. Finite Sample Differentially Private Confidence Intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science (ITCS '18)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 44:1–44:9. https://doi.org/10.4230/LIPIcs.ITCS.2018.44

[34] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. 2018. Efficient Algorithms for Outlier-Robust Regression. In *Proceedings of the 31st Annual Conference on Learning Theory (COLT '18)*. 1420–1430. http://proceedings.mlr.press/v75/klivans18a.html

[35] Pravesh Kothari, Jacob Steinhardt, and David Steurer. 2018. Robust Moment Estimation and Improved Clustering via Sum of Squares. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing (STOC '18)*. ACM, New York, NY, USA, 1035–1046. https://doi.org/10.1145/3188745.3188970

[36] Pravesh K. Kothari, Peter Manohar, and Brian Hu Zhang. 2022. Polynomial-Time Sum-of-Squares Can Robustly Estimate Mean and Covariance of Gaussians Optimally. In *International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France (Proceedings of Machine Learning Research, Vol. 167)*, Sanjoy Dasgupta and Nika Haghtalab (Eds.). PMLR, 638–667. https://proceedings.mlr.press/v167/kothari22a.html

[37] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. 2022. Private Robust Estimation by Stabilizing Convex Relaxations. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT '22)*. 723–777. https://proceedings.mlr.press/v178/kothari22a.html

[38] Mengchu Li, Thomas B Berrett, and Yi Yu. 2022. On Robustness and Local Differential Privacy. *arXiv preprint arXiv:2201.00751* (2022). https://doi.org/10.48550/arXiv.2201.00751

[39] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. 2021. Robust and Differentially Private Mean Estimation. In *Advances in Neural Information Processing Systems 34 (NeurIPS '21)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/1fc5309ccc651bf6b5d22470f67561ea-Abstract.html

[40] Xiyang Liu, Weihao Kong, and Sewoong Oh. 2022. Differential privacy and robust statistics in high dimensions. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT '22)*. 1167–1246. https://proceedings.mlr.press/v178/liu22b.html

[41] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*. IEEE Computer Society, Washington, DC, USA, 94–103. https://doi.org/10.1109/FOCS.2007.41

[42] Prasad Raghavendra, Tselil Schramm, and David Steurer. 2018. HIGH DIMENSIONAL ESTIMATION VIA SUM-OF-SQUARES PROOFS. In *Proceedings of the International Congress of Mathematicians (ICM 2018)* (2018). WORLD SCIENTIFIC, 3389–3423. https://doi.org/10.1142/9789813272880_0186

[43] Prasad Raghavendra and Benjamin Weitz. 2017. On the Bit Complexity of Sum-of-Squares Proofs. In *44th International Colloquium on Automata, Languages, and Programming ICALP (LIPIcs, Vol. 80)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 80:1–80:13. https://doi.org/10.4230/LIPIcs.ICALP.2017.80

[44] Kelly Ramsay and Shoja'eddin Chenouri. 2021. Differentially Private Depth Functions and their Associated Medians. *arXiv preprint arXiv:2101.02800* (2021). https://doi.org/10.48550/arXiv.2101.02800

[45] Kelly Ramsay, Aukosh Jagannath, and Shoja'eddin Chenouri. 2022. Concentration of the Exponential Mechanism and Differentially Private Multivariate Medians. *arXiv preprint arXiv:2210.06459* (2022). https://doi.org/10.48550/arXiv.2210.06459

[46] Aleksandra Slavkovic and Roberto Molinari. 2022. Perturbed M-Estimation: A Further Investigation of Robust Statistics for Differential Privacy. In *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*, Alicia L. Carriquiry, Judith M. Tanur, and William F. Eddy (Eds.). Springer, 337–361. https://doi.org/10.1007/978-3-030-75460-0_19

[47] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 2018. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science (ITCS '18)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 45:1–45:21. https://doi.org/10.4230/LIPIcs.ITCS.2018.45

[48] Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. 2022. FriendlyCore: Practical Differentially Private Aggregation. In *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*. JMLR, Inc., 21828–21863. https://proceedings.mlr.press/v162/tsfadia22a.html

Received 2022-11-07; accepted 2023-02-06