Robustness Implies Privacy in Statistical Estimation*

Samuel B. Hopkins†
MIT EECS

Gautam Kamath‡
University of Waterloo

Mahbod Majid§
University of Waterloo

Shyam Narayanan¶
MIT EECS

December 12, 2022

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 constant fraction of adversarially-corrupted samples.

*Authors are listed in alphabetical order.

†samhop@mit.edu.

‡g@csail.mit.edu. Supported by an NSERC Discovery Grant, an unrestricted gift from Google, and a University of Waterloo startup grant.

§m2majid@uwaterloo.ca. Supported by an NSERC Discovery Grant.

¶shyamsn@mit.edu. Supported by an NSF Graduate Fellowship, the NSF TRIPODS Program (award DMS-2022448), and a Google Fellowship.

Contents

1 Introduction **1**
- 1.1 Results . . . 2
- 1.2 Related Work . . . 4

2 Techniques **6**
- 2.1 Black-Box Reduction from Privacy to Robustness . . . 6
- 2.2 Algorithms . . . 11

3 Preliminaries **14**

4 A General Private Sampling Algorithm **15**
- 4.1 Sampling and volume computation with an imperfect oracle . . . 17
- 4.2 Proof of Theorem 4.1 . . . 18
- 4.3 Proof of Theorem 4.2 . . . 20

5 Estimating the Mean of a Gaussian **21**
- 5.1 Main Theorem . . . 21
- 5.2 Resilience of First and Second Moments . . . 22
- 5.3 Robust Algorithm . . . 23
- 5.4 Score Function and its Properties . . . 23
- 5.5 Proof of Theorem 5.1 . . . 28
- 5.6 The approx-DP setting . . . 29

6 Preconditioning the Gaussian **30**
- 6.1 Main Theorems . . . 30
- 6.2 Resilience of Moments . . . 31
- 6.3 Robust Algorithm . . . 31
- 6.4 Score Function and its Properties . . . 32
- 6.5 Proof of Theorem 6.1 . . . 38
- 6.6 The approx-DP setting . . . 38

7 Learning a Gaussian in Total Variation Distance **39**
- 7.1 Robust Algorithm . . . 40
- 7.2 Score Function and its Properties . . . 40
- 7.3 Proof of Theorem 7.1 . . . 44
- 7.4 Proof of Theorems 1.3 and 1.4 . . . 45

A Omitted proofs for Private Sampling **51**
- A.1 Preliminaries . . . 51
- A.2 Sampling from a well-rounded convex body with an imperfect oracle . . . 52

B Sum-of-squares proofs **58**
B.1 Proofs of Accuracy Lemmas 59
B.2 SoS bounds for arbitrary samples: Covariance estimation 62
B.3 SoS bounds for arbitrary samples: Mean estimation 65

C Computing Score Functions **68**

D High-Probability Bound for Stability of Covariance **76**
D.1 Preliminaries 76
D.2 Main Probability Bound 77
D.3 Proof of Lemma 6.3 80

E Mean Estimation in ℓ_∞ **81**

1 Introduction

Parameter estimation is a fundamental statistical task: given samples $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n \in \mathbb{R}^d$ 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)* [DMNS06] algorithm protects the privacy of individuals represented in a dataset $X_1, \ldots, X_n$ by guaranteeing that the distribution of outputs of the algorithm given $X_1, \ldots, X_n$ is statistically close to the distribution it would generate given $X'_1, \ldots, X'_n$, where $X'_1, \ldots, X'_n$ differs from $X_1, \ldots, X_n$ 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" [DL09]. More recent works continue to leverage ideas from robust estimation to design private estimation procedures [BKSW19, KSU20, BGS+21, RC21, KMV22, LKO22, HKM22, GH22, RJC22] – 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 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 [HKM22] 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 [KKM18] and mean estimation under other bounded-moment assumptions [HL18a, KSS18].

Conclusions on Robust versus Private Estimation: Recent work [GH22] 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) [DMNS06, DKM+06])**.** 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 $S \subseteq \mathcal{O}$, $\mathbb{P}(M(X) \in S) \leq e^{\varepsilon}\,\mathbb{P}(M(X') \in S) + \delta$. 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 $\eta > 0$, $Y_1, \ldots, Y_n$ are η-corrupted samples from D if $X_1, \ldots, X_n \sim D$ i.i.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 learning Gaussians in total variation distance.

Theorem 1.3 (Learning Arbitrary Gaussians, Pure DP, Subsection 7.4)**.** *Assume that $0 < \alpha, \beta, \varepsilon < 1$, $0 < \eta < \eta^*$ for some absolute constant η^*, and $K, R > 0$. 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 $\|\mu\| \leq R$ 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 $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \leq \alpha + \widetilde{O}(\eta)$ with probability at least $1 - \beta$, if*[1]

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

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 [KLSU19],[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 [AL22, KMS+22b, KMV22, TCK+22] (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 polynomial-time robust and private estimator; prior works require $\Omega(d^{3.5})$ samples [AL22, TCK+22].

Theorem 1.4 (Learning Arbitrary Gaussians, (ε, δ)-DP, Subsection 7.4)**.** *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 $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \leq \alpha + \widetilde{O}(\eta)$ with probability at least $1 - \beta$, if*

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

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] With more careful analysis, we expect the error bound can be tightened to $\alpha + O(\eta \log 1/\eta)$, which is expected to be tight for statistical query algorithms [DKS17]; the same goes for our other results on learning Gaussians.

[2] replacing the Gaussian mechanism with the Laplace mechanism

[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 $\log K$.

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 [KV18, KLSU19, BS19, BKSW19, KSU20, LKKO21, BGS+21, LKO22, HKM22]. We focus on the following simple problem: given (corrupted) samples from $\mathcal{N}(\mu, I)$, find $\hat{\mu}$ such that $\|\mu - \hat{\mu}\| \leq \alpha$. 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 [BKSW19, KSU20]. 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 [KLSU19, HKM22] (see Table 2). We give the first nearly-sample-optimal pure-DP algorithm:

Theorem 1.5 (Estimating the Mean of a Spherical Subgaussian Distribution, Theorem 5.1)**.** *Assume that $0 < \alpha, \beta, \varepsilon < 1$, $0 < \eta < \eta^*$ for some absolute constant η^*, and $R > 0$. 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 $\|\mu\| \leq R$, 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 $\|\mu - \hat{\mu}\| \leq \alpha + \widetilde{O}(\eta)$ with probability at least $1 - \beta$, as long as*

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

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 [HL19].

There is also an analogue (Theorem 5.2) for polynomial-time mean estimation subject to (ε, δ)-DP without the $\|\mu\| \leq R$ 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 [EMN22].

Finally, we note that Theorems 1.3 and 1.5 are known to be near-optimal from standard packing lower bounds [BKSW19], and Theorem 1.4 and Theorem 5.2 are also known to be near-optimal, via the technique of fingerprinting [KLSU19, KMS22a], 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, 4.2, 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 [RSS18]), and on high-dimensional algorithmic robust statistics (see [DK19]). We discuss other works at the intersections of these areas.

Private and Robust Estimators: [DL09] first used robust statistics primitives to design private algorithms, a tradition continued by [BKSW19, KSU20, LKO22, BGS+21, RC21, KMV22, HKM22]. Other works from the Statistics community also investigate connections between robustness and privacy [AM20, AM21, RJC22, SM22], including local differential privacy [LBY22]. Our black-box reduction from privacy to robustness can be seen as a generalization of methods of [BKSW19, KSU20],

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[KV18]	$\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\min(\log K, \log \delta^{-1})}{\varepsilon}$, $d = 1$	No	Yes	Pure/Approximate
[KLSU19]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d^{3/2}\sqrt{\log K \log \delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[BKSW19]	$\frac{d^2}{\alpha^2} + \frac{d^2 \log K}{\alpha\varepsilon}$	Optimal	No	Pure
[AAK21]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[LKO22]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[KMS+22b]	$\frac{d^2}{\alpha^2} + \left(\frac{d^2}{\alpha\varepsilon} + \frac{d^{5/2}}{\varepsilon}\right) \cdot (\log \delta^{-1})^{O(1)}$	No	Yes	Approximate
[KMV22]	$\frac{d^8}{\alpha^4} \cdot \left(\frac{\log \delta^{-1}}{\varepsilon}\right)^6$	Suboptimal	Yes	Approximate
[AL22, TCK+22]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d \log \delta^{-1}}{\varepsilon}$	No	Yes	Approximate
[AL22, TCK+22]	$\frac{d^{3.5} \log \delta^{-1}}{\alpha^3\varepsilon}$	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 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.

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 [BKSW19, GKMN21, LKKO21, EMN22, AL22, KMV22, TCK+22, LKO22]; see Tables 1, 2.

Recall that [GH22] 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: [HKM22] and [KMV22] pioneer the use of SoS for private algorithm design. [HKM22] gives a polynomial-time algorithm for pure-DP mean estimation under a bounded covariance assumption, using $\frac{d}{\alpha^2\varepsilon}$ samples, and [KMV22] gives a $\approx d^8$-sample (ε, δ)-DP algorithm for learning d-dimensional Gaussians. [GH22] uses SoS for private *sparse* mean estimation.

On a technical level, our work most resembles [HKM22]; we also employ SoS SDPs as score functions and leverage tools from log-concave sampling. However, there are fundamental roadblocks to using [HKM22]'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 converting a much wider range of SoS-based robust algorithms to private ones.

Inverse Sensitivity Mechanism: In [AD20b, AD20a], 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

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[KV18]	$\frac{1}{\alpha^2}+\frac{1}{\alpha\varepsilon}+\frac{\min(\log R,\log\delta^{-1})}{\varepsilon}$, $d=1$	No	Yes	Pure/Approximate
[KLSU19]	$\frac{d}{\alpha^2}+\frac{d\sqrt{\log\delta^{-1}}}{\alpha\varepsilon}+\frac{\sqrt{d\log R\log\delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[BKSW19]	$\frac{d}{\alpha^2}+\frac{d\log R}{\alpha\varepsilon}$	Optimal	No	Pure
[KSU20]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{d\log R}{\varepsilon}$	Optimal	No	Pure
[AAK21]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[LKKO21]	$\frac{d}{\alpha^2}+\frac{d^{3/2}\log\delta^{-1}}{\alpha\varepsilon}$	Optimal	Yes	Approximate
[BKSW19, LKO22]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[HKM22]	$\frac{d}{\alpha^2\varepsilon}+\frac{d\log R}{\varepsilon}$	Suboptimal	Yes	Pure
Theorem 1.5	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{d\log R}{\varepsilon}$	Optimal	Yes	Pure
Theorem 5.2	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\varepsilon}$	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.

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 $\mathrm{poly}(1/\varepsilon, \log 1/\delta, 1/\alpha)$-factors more samples than those we present [AKT+22].

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 $\theta \in \mathbb{R}^d$, 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 $X_1, \ldots, X_n \sim p_\theta \in \mathcal{P}$, for every $\eta \in [0,1]$, given any η-corruption of $X_1, \ldots, X_n$, 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 $X_1, \ldots, X_n$, 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 [MT07]: Given $\varepsilon > 0$, $X_1, \ldots, X_n$ and a threshold $\eta_0 \in [0,1]$, the mechanism picks a random $\theta \in \Theta + \alpha(\eta_0) \cdot B_{\|\cdot\|}$ with:

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

[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.

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 [HT10]. 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 [AD20b]. 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 $\eta_0 = \Theta(\alpha^2)$ [DKK+17, SCV18].

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.

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*

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

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 $d(X, X') \le \eta$. 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 $\geq \eta_0 n$, the overall probability of seeing some θ with score greater than η_0 is at most

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

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 $\|\theta - \theta^*\| \leq 2\alpha(\eta_0)$. □

Consequences of Lemma 2.1: Applied to robust mean estimators with optimal error rates under bounded k-th moment assumptions, for any $k \geq 2$, Lemma 2.1 gives optimal pure-DP estimators under those same assumptions, recovering the main results of [KSU20], applied to robust linear regression (with known covariance) [DKS19], it yields a pure-DP analogue of the nearly-optimal regression result of [LKKO21], 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 [LKKO21, BGS+21], 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 $\|\hat{\mu} - \mu\| \leq O(\alpha + \eta)$, if $\eta < 1/2$. For ε-DP guarantees, we need to restrict to the case of $\|\mu\| \leq R$ for some (large) $R > 0$; we will assume that even for $\eta \geq 1/2$, $\|\hat{\mu}\| \leq R$.

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 $n \geq O(\frac{d+\log(1/\beta)}{\alpha\varepsilon})$, and in the latter $\alpha(1) = R$, so we get the additional requirement $n \geq \frac{d \log R}{\varepsilon}$, 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 [HT10, BKSW19]. 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 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.

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

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 starting with 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 privacy, as well as accuracy as in Lemma 2.1.

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 [LKO22], and the approach of [BGS+21], 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 $\to \Theta \cup \{\textsc{reject}\}$ 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 nondecreasing function, $n \in \mathbb{N}$, and $\eta_0, \eta^* \in [0,1]$. We continue to employ $\textsc{score}_X(\theta)$ as defined in (1). Suppose as before that with probability $1-\beta$ over samples $X_1, \dots, X_n \sim p_\theta \in \mathcal{P}$, for every $\eta < \eta^*$, given any η-corruption of $X_1, \dots, X_n$, $\|\hat{\theta} - \theta\| \le \alpha(\eta)$. And, suppose that $\hat{\theta}$ has the following worst-case robustness property: for *any* input $X = X_1, \dots, X_n$, if $\hat{\theta}(X) \ne \textsc{reject}$, then for every $\eta < \eta^*$, given any η-corruption X' of X, either $\hat{\theta}(X') = \textsc{reject}$, 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 $X_1, \dots, X_n \sim p_{\theta^*}$ 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.

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

Proof. Since there exists some θ such that $\textsc{score}_X(\theta) < 0.2\eta^* n$, there's some $Y \sim_{0.2\eta^*} X$ such that $\hat{\theta}(Y) \ne \textsc{reject}$: this is because we can consider any such Y which has $\textsc{score}_Y(\theta) = 0$, and thus $\hat{\theta}(Y)$ outputs an element of Θ and not reject. Similarly, for any other θ' with $\textsc{score}_X(\theta') \le 0.4\eta^* n$, 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 $t \ge \log(1/\delta)/\varepsilon + 0.1\eta^* n$. But for t^* to lie in the stated interval, we need $\log(1/\delta)/\varepsilon + 0.1\eta^* n \le t \le 0.2\eta^* n$, 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 $T, T' \le 0.2\eta^* n$, 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

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

To see why, first note that for any θ we have $|s_X(\theta) - s_{X'}(\theta)| \le 1$. This implies that $e^{-\varepsilon s_X(\theta)} \le e^{\varepsilon} e^{-\varepsilon s_{X'}(\theta)}$. Similarly, if $s_X(\theta) \le 0.3\eta^* n$, it also implies that at least one of the following must be true (potentially both): $s_{X'}(\theta) \le 0.3\eta^* n$ 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

$$\begin{aligned} \mathop{\mathbb{P}}_{X}(\theta \in S) &\le e^{\varepsilon} \cdot \frac{\int_{\theta \in S} e^{-\varepsilon s_{X'}(\theta)} \cdot [\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) + \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])]}{\int_{\theta \in \Theta + \alpha(\eta_0) B_{\|\cdot\|}} e^{-\varepsilon s_X(\theta)} \cdot \mathbf{1}(s_X(\theta) < 0.3\eta^* n)} \\ &\le e^{\varepsilon} \cdot \frac{\int_{\theta \in S} e^{-\varepsilon s_{X'}(\theta)} \cdot [\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) + \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])]}{e^{-\varepsilon} \int_{\theta \in \Theta + \alpha(\eta_0) B_{\|\cdot\|}} e^{-\varepsilon s_{X'}(\theta)} \cdot [\mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) - \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])]} \end{aligned}$$

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

We next simplify the denominator. 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 Claim 2.3 to upper-bound the magnitude of the second term in the denominator, $\int e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]) \leq \exp(-\varepsilon \cdot 0.25\eta^* n) \cdot V_{2\alpha(\eta^*)}$, 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

$$\begin{aligned}\mathop{\mathbb{P}}_{X}(\theta \in S) &\leq \frac{e^{2\varepsilon}}{1-\delta} \cdot \left(\frac{\int_{\theta \in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n) + \int_{\theta \in S} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n])}{\int_{\theta \in \Theta + \alpha(\eta_0) B_{\|\cdot\|}} e^{-\varepsilon s_{X'}(\theta)} \cdot \mathbf{1}(s_{X'}(\theta) < 0.3\eta^* n)} \right) \\ &= \frac{e^{2\varepsilon}}{1-\delta} \cdot \left(\mathop{\mathbb{P}}_{X'}(\theta \in S) + \mathop{\mathbb{P}}_{X'}(s_{X'}(\theta) \in [0.25\eta^* n, 0.35\eta^* n]) \right).\end{aligned}$$

Using Claim 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.

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 inequlities 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 $\tilde{\mathbf{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ is a degree *d pseudoexpectation* if $\tilde{\mathbf{E}}1 = 1$ and $\tilde{\mathbf{E}}p^2 \geq 0$ for any p of degree at most $d/2$. A pseudoexpectation $\tilde{\mathbf{E}}$ *satisfies* a system of polynomial inequalities $p_1 \geq 0, \ldots, p_m \geq 0$, written $\tilde{\mathbf{E}} \models p_1 \geq 0, \ldots, p_m \geq 0$, if for every $S \subseteq [m]$ and every p, we have $\tilde{\mathbf{E}} \prod_{i \in S} p_i \cdot p^2 \geq 0$ 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 $X = X_1, \dots, X_n$ from p_{θ^*}, define a degree-$O(1)$ system of polynomial inequalities $\mathcal{A}(X, \theta, z)$ where $\theta = \theta_1, \dots, \theta_D, z = z_1, \dots, z_{(nD)^{O(1)}}$ 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 $\mathcal{A}(X, \theta, z) \vdash_{O(1)} \langle \theta - \theta^*, v \rangle \le \alpha$ for every v in the dual ball of $\|\cdot\|$.

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

Lemma 2.6 (Informal, implicit in [KMZ22])**.** *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 $\alpha = \tilde{O}(\eta)$.*

2.2.1 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

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

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 $X'_1, \dots, X'_n, w_1, \dots, w_n, \theta'$, and consider

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

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

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

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 $X_1, \dots, X_n \sim \mathcal{N}(\theta^*, I)$ and $\tilde{\mathbf{E}} \models \mathcal{B}_t$ for $t = \eta n$, then $\|\tilde{\mathbf{E}}\theta' - \theta^*\| \le \tilde{O}(\eta)$. This is slightly stronger than what we already know

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

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 [KMZ22] readily generalize to show this.

In fact, [KMZ22]'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 in Sections 5, 6, and 7, we remove this indirection for simplicity.

On "Satisfies": An important technical difference between our score function and that of [HKM22] is that the $\tilde{\mathbf{E}}$s it involves must have $\tilde{\mathbf{E}} \models \sum_{i=1}^n w_i = n - t$, rather than something weaker, like $\tilde{\mathbf{E}} \sum_{i=1}^n w_i = n - t$. 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 $\tilde{\mathbf{E}} \sum_{i=1}^n w_i = n - t$, we could have $\tilde{\mathbf{E}}$ which satisfies the rest of $\mathcal{B}_t$ but has $\|\tilde{\mathbf{E}}\theta' - \theta^*\| \geq \Omega(R)$, just by taking $\tilde{\mathbf{E}}$ 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 $\tilde{\mathbf{E}}$ satisfying a given system of polynomials is known – existing techniques to find $\tilde{\mathbf{E}}$s work best in the context of *satisfiable* polynomial systems [RW17]. We sidestep this challenge by generalizing a technique from the robust statistics literature, which searches for $\tilde{\mathbf{E}}$ which *approximately* satisfies a system of polynomials, to the setting where those polynomials may be unsatisfiable – see Appendix C. 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, we could sample from $\mathbb{P}(\theta) \propto \exp(-\varepsilon \cdot \text{SoS-score}_X(\theta))$ with log-concave sampling techniques, as in [HKM22]. 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* $V_i = \text{Vol}(S_{t_i})$, 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 sublevel sets are convex, we are able

[7] It seems to be folklore that volume computation algorithms, e.g. the seminal [DFK91], work given only weak membership oracles, as opposed to e.g. weak separation oracles. For completeness, in Appendix A, we analyze a hit-and-run sampling algorithm which uses a weak membership oracle, tracking the numerical errors this creates.

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 $t_1, \dots, t_{n^{O(1)}}$ *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 (Lemma 4.7).

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 capture this strategy formally.

3 Preliminaries

First, we note a few notational conventions. We will use $\mathbf{0}$ to denote the origin in $\mathbb{R}^d$ (or in Euclidean space generally). For $x \in \mathbb{R}^d$ and $r \geq 0$, we define $B(x, r)$ to be the ℓ_2-ball of radius r around x.

We note a series of important definitions that we will use in our analysis.

Definition 3.1 (sensitivity)**.** We say that a function $f(\theta, \mathcal{X})$ has sensitivity Δ with respect to $\mathcal{X}$ if for all θ and all neighboring datasets $\mathcal{X}, \mathcal{X}'$ (i.e., datasets that differ in exactly one data point), $|f(\theta, \mathcal{X}) - f(\theta, \mathcal{X}')| \leq \Delta$. We will implicitly assume that sensitivity is with respect to the dataset.

Definition 3.2 (quasi-convexity)**.** A function $f : S \to \mathbb{R}$, defined on a convex subset S of a real vector space is quasi-convex if for all $x, y \in S$ and $\lambda \in [0, 1]$ we have

$$f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}.$$

Next, we note some important distance metrics for mean vectors and covariance matrices. We will use $\|\cdot\|_F$ to denote Frobenius norm and $\|\cdot\|_{op}$ to denote the operator norm (a.k.a. spectral norm) of a matrix.

Definition 3.3 (Mahalanobis distance)**.** Given two vectors $\mu, \mu' \in \mathbb{R}^d$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, we define the *Mahalanobis* distance between μ and μ' with respect to Σ, written as $\|\mu - \mu'\|_\Sigma$, to equal $\|\Sigma^{-1/2}(\mu - \mu')\|_2$.

In addition, given two covariance matrices $\Sigma, \Sigma' \in \mathbb{R}^{d \times d}$, we define the *Mahalanobis* distance between Σ and Σ' to equal $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F$.

Note that there are two different definitions of Mahalanobis distance, though which definition we are using will be clear from context.

It is well known that Mahalanobis distance captures total variation distance. Namely, if $\|\mu - \mu'\|_\Sigma = \alpha \leq 1$, then $d_{\mathrm{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma)) = \Theta(\alpha)$, and if Σ, Σ' have Mahalanobis distance $\alpha \leq 1$, then $d_{\mathrm{TV}}(\mathcal{N}(\mathbf{0}, \Sigma), \mathcal{N}(\mathbf{0}, \Sigma')) = \Theta(\alpha)$.

It is well-known that Mahalanobis distance between covariance matrices is roughly symmetric: namely, $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F = \Theta(\|\Sigma'^{-1/2}\Sigma\Sigma'^{-1/2} - I\|_F)$ if either is at most 0.5. In addition, $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F = \|\Sigma'^{1/2}\Sigma^{-1}\Sigma'^{1/2} - I\|_F$, and $\|\Sigma'^{-1/2}\Sigma\Sigma'^{-1/2} - I\|_F = \|\Sigma^{1/2}\Sigma'^{-1}\Sigma^{1/2} - I\|_F$.

Definition 3.4 (Spectral distance)**.** Given two covariance matrices $\Sigma, \Sigma' \in \mathbb{R}^{d \times d}$, we define the *spectral* distance between Σ and Σ' to equal $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_{op}$.

Similarly, we have $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_{op} = \|\Sigma'^{1/2}\Sigma^{-1}\Sigma'^{1/2} - I\|_{op}$, which are asymptotically equal to $\|\Sigma'^{-1/2}\Sigma\Sigma'^{-1/2} - I\|_{op} = \|\Sigma^{1/2}\Sigma'^{-1}\Sigma^{1/2} - I\|_{op}$ if either is at most 0.5.

Finally, we define the notions of flattening and tensor powers.

Definition 3.5 (Tensor power)**.** Given two vectors $x \in \mathbb{R}^d$, $y \in \mathbb{R}^{d'}$, the tensor product $x \otimes y$ is the vector in $\mathbb{R}^{d \cdot d'}$, with entries indexed by $(i, j) \in [d] \times [d']$, such that $(x \otimes y)_{ij} = x_i \cdot y_j$.

We also will define $x^{\otimes 2} := x \otimes x$.

Definition 3.6 (Flattening)**.** Given a matrix $M \in \mathbb{R}^{d \times d'}$, we define the *flattening* $M^\flat$ to be the vector in $\mathbb{R}^{d \cdot d'}$ with $(M^\flat)_{ij} = M_{i,j}$.

Note that for any vectors x, y, $x \otimes y$ equals $(xy^T)^\flat$.

To represent linear functionals and polynomials, we look at the value of the linear functional over monomials.

Definition 3.7 (monomial vector)**.** A monomial vector of degree d is a $n^{O(d)}$-dimensional vector $v_d(x)$ indexed by multisets $S \subseteq [n]$, $|S| \leq d$, where the entry $v_d(x)_S$ is the monomial

$$v_d(x)_S := \prod_{i \in S} x_i.$$

Remark. The definition of n for number of variables and d for degree is a slight abuse of notation, as in the rest of the paper n represents the number of data points and d is the dimension of the data points. We will only use the former definition here and in Appendix C.

Linear functionals over the set of polynomials of up to degree d over $\mathbb{R}^n$ form an $n^{O(d)}$-dimensional vector space and we can represent them as follows numerically.

Definition 3.8 (numerical representation of linear functionals and polynomials)**.** Suppose $\mathcal{L}$ is a linear functional over polynomials of up to degree d over $\mathbb{R}^n$. We define the representation of $\mathcal{L}$, $\mathcal{R}(\mathcal{L}) \in \mathbb{R}^{n^{O(d)}}$ indexed by multisets $S \subseteq [n]$, $|S| \leq d$, as

$$\mathcal{R}(\mathcal{L})_S = \mathcal{L}(v_d(x)_S).$$

Similarly, for a polynomial q, we define its representation $\mathcal{R}(q) \in \mathbb{R}^{n^{O(d)}}$ to be

$$\mathcal{R}(q)_S = \text{coefficient of } x^S \text{ in } q.$$

4 A General Private Sampling Algorithm

In this section, we prove two general theorems showing that if one has a score function corresponding to a robust algorithm for parameter estimation from samples, with a few important properties, then one can construct a differentially private algorithm. The results can either generate a pure-DP algorithm (Theorem 4.1), or an approx-DP algorithm (Theorem 4.2), depending on the properties we assume about the robust algorithm.

Assuming the robust algorithm and score function can be computed efficiently, and we have another property that we call *quasi-convexity*, the private algorithms also run in polynomial time. One can also generate analogous statements by removing these assumptions, but the algorithm no longer runs in polynomial time. To avoid rewriting, we color certain parts of Theorems 4.1 and 4.2 in blue: one can read the same theorems and ignore what is written in blue to obtain an inefficient private algorithm arising from an inefficient robust algorithm.

We first state our theorem for creating a pure-DP algorithm.

Theorem 4.1. *Let $0 < \eta, r < 1 < R$ be fixed parameters. Suppose we have a score function $\mathcal{S}(\theta, \mathcal{Y}) \in [0, n]$ that takes as input a dataset $\mathcal{Y} = \{y_1, \ldots, y_n\}$ and a parameter $\theta \in \Theta \subset \mathbb{R}^d$ (where Θ is convex and contained in a ball of radius R), with the following properties:*

- *(Bounded Sensitivity) For any two adjacent datasets $\mathcal{Y}, \mathcal{Y}'$ and any $\theta \in \Theta$, $|\mathcal{S}(\theta, \mathcal{Y}) - \mathcal{S}(\theta, \mathcal{Y}')| \le 1$.*
- *(Quasi-Convexity) For any fixed dataset $\mathcal{Y}$, any $\theta, \theta' \in \Theta$, and any $0 \le \lambda \le 1$, we have that $\mathcal{S}(\lambda\theta + (1-\lambda)\theta', \mathcal{Y}) \le \max(\mathcal{S}(\theta, \mathcal{Y}), \mathcal{S}(\theta', \mathcal{Y}))$.*
- *(Efficiently Computable) For any given $\theta \in \Theta$ and dataset $\mathcal{Y}$, we can compute $\mathcal{S}(\theta, \mathcal{Y})$ up to error γ in $\mathrm{poly}(n, d, \log R, \log \gamma^{-1})$ time for any $\gamma > 0$.*
- *(Robust algorithm finds low-scoring point) For a given dataset $\mathcal{Y}$, let $T = \min_{\theta_0} \mathcal{S}(\theta_0, \mathcal{Y})$. Then, we can find some point θ such that for all θ' within distance r of θ, $\mathcal{S}(\theta', \mathcal{Y}) \le T + 1$, in time $\mathrm{poly}(n, d, \log \frac{R}{r})$.*
- *(Volume) For any given dataset $\mathcal{Y}$ and $\eta' \ge \eta$, let $V_{\eta'}(\mathcal{Y})$ represent the d-dimensional volume of points $\theta \in \Theta \subset \mathbb{R}^d$ with score at most $\eta' n$. (Note that $V_1(\mathcal{Y})$ is the full volume of Θ).*

Then, we have a pure ε-DP algorithm $\mathcal{A}$ on datasets of size n, that runs in $\mathrm{poly}(n, d, \log \frac{R}{r})$ time, with the following property. For any dataset $\mathcal{Y}$, if there exists θ with $\mathcal{S}(\theta, \mathcal{Y}) \le \eta n$ and if $n \ge \Omega\left(\max_{\eta' : \eta \le \eta' \le 1} \frac{\log(V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y})) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right)$, then $\mathcal{A}(\mathcal{Y})$ outputs some $\theta \in \Theta$ of score at most $2\eta n$ with probability $1 - \beta$.

We remark that this theorem has several important conditions. The bounded sensitivity of the score is important as it ensures that if we sample according to the exponential mechanism, the sampling probability of any θ does not change significantly between adjacent datasets. The conditions of quasi-convexity, computability, and finding a low-scoring point are only required for the algorithm to run in polynomial time. Indeed, the latter two of these conditions are important for the robust algorithm to succeed, and the quasi-convexity assumption generalizes a convexity assumption on the score, which roughly corresponds to sampling from log-concave distributions. Finally, the sample complexity is dictated both by the number of samples needed for the robust algorithm to succeed and by bounds on the volume of low versus high scoring points.

Along with a general result for pure-DP algorithms, we also prove a similar result for approx-DP algorithms, which we now state.

Theorem 4.2. *Let $0 < \eta < 0.1$ and $r < 1 < R$ be fixed parameters. Suppose we have a score function $\mathcal{S}(\theta, \mathcal{Y}) \in \mathbb{R}$ that takes as input a dataset $\mathcal{Y} = \{y_1, \ldots, y_n\}$ and a parameter $\theta \in \Theta \subset \mathbb{R}^d$ (where Θ is convex and contained in a ball of radius R), with the same properties as in Theorem 4.1.*

In addition, fix some parameter $\eta^* \in [10\eta, 1]$. Suppose that $n \ge \Omega\left(\frac{\log(1/\delta) + \log(V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}))}{\varepsilon \cdot \eta^*}\right)$ for all $\mathcal{Y}$ such that there exists θ with $\mathcal{S}(\theta, \mathcal{Y}) \le 0.7\eta^* n$. Then, we have an (ε, δ)-DP algorithm $\mathcal{A}$ that runs in $\mathrm{poly}(n, d, \log \frac{R}{r})$ time, such that for any dataset $\mathcal{Y}$, if there exists θ with $\mathcal{S}(\theta, \mathcal{Y}) \le \eta n$ and if $n \ge \Omega\left(\max_{\eta' : \eta \le \eta' \le \eta^*} \frac{\log(V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y})) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right)$, then $\mathcal{A}(\mathcal{Y})$ outputs some $\theta \in \Theta$ of score at most $2\eta n$ with probability $1 - \beta$.

The main difference in the approx-DP setting is that we set some threshold η^*, and only consider volumes of points of score up to $\eta^* \cdot n$. This is because, roughly speaking, we will sample using a truncated exponential mechanism until score roughly $\eta^* n$. (In reality, we need to be more careful about how we truncate.) But because of this truncation, the volume bound will be crucial for not only bounding sample complexity but also ensuring privacy, to make sure the probability of sampling a point near the threshold score is low.

We will only prove Theorems 4.1 and 4.2 for the efficient case. In the proofs, one can verify that the requirements of quasi-convexity, efficient computability, and efficiently finding a low-scoring point, as well as the promise that Θ is convex and bounded, are only needed for our sampling algorithms to run in polynomial time. Hence, the inefficient algorithm results also follow.

4.1 Sampling and volume computation with an imperfect oracle

To prove the main results of this section, we heavily rely on the theory of sampling and volume computation for convex bodies, given only membership oracle access (as opposed to membership and separation oracle access). While one may wish to directly apply these techniques, we cannot afford to do so, because, to the best of our knowledge, all such results have been written assuming infinite-precision arithmetic and perfect membership oracles. In our setting, we must show such results are possible even if we only have bounded precision arithmetic and imperfect membership oracles. This will be crucial because we assume we cannot perfectly compute the score function, but can only approximately compute it. We now formally define approximate membership oracles.

Definition 4.3. Given two nested convex bodies $K_1 \subset K_2$, a (K_1, K_2)**-membership oracle** O is an oracle that, if given an input $x \in K_1$, outputs YES, if given an input $x \notin K_2$, outputs NO, and if given an input $x \in K_2 \backslash K_1$, may output either YES or NO.

In addition, we will wish for multiplicative approximations for the sake of pure-DP, meaning each point (in a sufficiently fine net) in the convex body should be sampled in a way that is point-wise close to uniform, as opposed to close to uniform in total variation distance. While one could use the techniques of [MV22] to achieve the point-wise guarantee, they still make an assumption of using perfect membership oracles and infinite-precision arithmetic.

To deal with the issues of precision and imperfect oracles, we apply the known analyses of hit-and-run sampling, made discrete in an appropriate fashion, and make sure that the probability of ever being near the boundary of the convex body, where the membership oracle may be incorrect, is low. To ensure the multiplicative approximation, we make a final step where we slightly perturb and then discretize the sample further, and show that this is sufficient. Since most of the analysis derives from known results, we defer the proofs to Appendix A, and here we simply state the results we need.

Lemma 4.4. ***(Main convex body sampling lemma)*** *Fix any parameter $\gamma_6 \leq d^{-100}$ and $r < 1 < R$. Let K_1, K_2 be convex bodies such that $B(\mathbf{0}, r) \subset K_1 \subset K_2 \subset B(\mathbf{0}, R)$, and $\mathrm{vol}(K_2) - \mathrm{vol}(K_1) \leq \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$, for some γ_1 such that $\log \gamma_1^{-1} = \mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_6^{-1})$. Suppose we have a (K_1, K_2)-membership oracle O. Then, in $\mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_6^{-1})$ time and queries to O, we can output a point z that is $(1 \pm \gamma_6)$-pointwise close to uniform on the set of points in $\mathbb{R}^d$ with all coordinates integer multiples of γ_5 that are accepted by O, for $\gamma_5 = \frac{r \cdot \gamma_6}{d^3}$.*

Lemma 4.5. ***(Volume sampling)*** *Set $\gamma_6 = \frac{\varepsilon}{d^{100}\log(R/r)}$, and set γ_1, γ_5, along with $r, R, K_1, K_2, \mathcal{O}$, as in Lemma 4.4. Fix any $\varepsilon < 0.5$. Then, for any $\gamma < 1$, in $\mathrm{poly}(d, \log\frac{R}{r}, \frac{1}{\varepsilon}, \log\gamma^{-1})$ time and oracle accesses, we can approximate the number of points in $\mathbb{R}^d$ with all coordinates integer multiples of γ_5 that are accepted by $\mathcal{O}$, up to a $1 \pm \varepsilon$ multiplicative factor, with failure probability γ.*

Remark. Our parameters skip to γ_5 and γ_6 since we define auxiliary parameters $\gamma_2, \gamma_3, \gamma_4$ in the proofs of Lemmas 4.4 and 4.5.

4.2 Proof of Theorem 4.1

Our algorithm will roughly sample each θ based on the exponential mechanism, where each θ is sampled proportional to $e^{-\varepsilon\cdot\mathcal{S}(\theta,\mathcal{Y})}$. In the following lemma, we apply Lemmas 4.4 and 4.5 to obtain a desired sampling procedure.

We note the following fact, which we will prove in Appendix A, while proving Lemma 4.5.

Fact 4.6. *Suppose $K \subset \mathbb{R}^d$ is a convex body that contains a ball of radius r. Suppose γ is a parameter which is at most $\frac{r}{d^3}$. Then, the number of points in K that have all coordinates integral multiples of γ is $(1 \pm O(1/d)) \cdot \mathrm{vol}(K)/\gamma^d$.*

Lemma 4.7. *Set $\gamma_1, \gamma_5, \gamma_6$ as in Lemma 4.5, and let $\tilde{\Theta}$ be the set of points in Θ with all coordinates integral multiples of γ_5. Then, in $\mathrm{poly}(n, d, \frac{1}{\varepsilon}, \log\frac{R}{r})$ time, we can sample from each $\theta \in \tilde{\Theta}$ with probability proportional to $e^{-\varepsilon\cdot\mathcal{S}(\theta,\mathcal{Y})} \cdot e^{\pm O(\varepsilon)}$.*

Proof. First, we set some additional parameters. Finally, we define $\gamma_8 := \frac{\varepsilon}{2} \cdot e^{-n} \cdot (\gamma_5/2R)^d$ and $\gamma_7 := \gamma_8 \cdot \left(\frac{\gamma_1 \cdot r}{6d}\right)^d / (2(2R)^d)$.

We now describe our algorithm. Define $T := \max_{\theta_0} \mathcal{S}(\theta_0, \mathcal{Y})$. Even if T is unknown, in time $\mathrm{poly}(n, d, \log\frac{1}{r})$, we can find some point θ such that $\mathcal{S}(\theta', \mathcal{Y}) \le T + 1$ for all θ' within r of θ. We can also get some estimate T' between T and $T+1$. We pick a uniformly random number between $T'+1$ and $T'+2$ that is an integral multiple of γ_7. Call this number $\hat{T}$: note that $\hat{T} \le \min_\theta \mathcal{S}(\theta_0, \mathcal{Y}) + 3$. Now, for any point $\theta \in \tilde{\Theta}$, let $t(\theta)$ be the smallest nonnegative integer t such that the **estimate** (where the estimate has accuracy γ_7) of the score $\mathcal{S}(\theta, \mathcal{Y})$ is at most $\hat{T} + t$.

Our goal will be to produce an $e^{\pm O(\varepsilon)}$-pointwise sample from the distribution proportional to $e^{-\varepsilon\cdot t(\theta)}$. For each integer $t \ge 0$, define $K_1^{(t)}$ to be the convex body of points in Θ with (true) score at most $\hat{T} + t - \gamma_7$, and $K_2^{(t)}$ to be the convex body of points in Θ with (true) score at most $\hat{T} + t + \gamma_7$. We will apply Lemmas 4.4 and 4.5, with $\mathcal{O}$ as the $(K_1^{(t)}, K_2^{(t)})$-oracle that accepts if the estimate of the score is at most $\hat{T} + t$, i.e., if $t(\theta) \le t$. (Note that while $K_1^{(t)}$ may not contain $\mathbf{0}$, it contains a ball of radius r around an efficiently computable point θ, which is sufficient.) Also, let $S^{(t)}, N^{(t)}$ to be the set of and number of points in $\tilde{\Theta}$, respectively, such that $t(\theta) \le t$. Since $t(\theta) \in \{0, 1, \ldots, n\}$, we can write $\sum_{\theta\in\tilde{\Theta}} e^{-\varepsilon t(\theta)} = N^{(0)} + \sum_{t=1}^{n} e^{-\varepsilon t}(N^{(t)} - N^{(t-1)}) = \sum_{t=0}^{n-1} \left(e^{-\varepsilon t}(1 - e^{-\varepsilon})N^{(t)}\right) + e^{-\varepsilon n}N^{(n)}$. Assuming that $\mathrm{vol}(K_2^{(t)}) - \mathrm{vol}(K_1^{(t)}) \le \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$ for all t, then we can provide a $e^{\pm\varepsilon}$-factor approximation $\tilde{N}^{(t)}$ for each $N^{(t)}$, with failure probability at most γ_8, in time $\mathrm{poly}(d, \log\frac{R}{r}, \frac{1}{\varepsilon}, \log\gamma_8^{-1})$, by Lemma 4.5.

Our final algorithm will sample each number $t \in \{0, 1, \ldots, n-1\}$ with probability proportional to $e^{-\varepsilon t}(1 - e^{-\varepsilon})\tilde{N}^{(t)}$ and $t = n$ with probability proportional to $e^{-\varepsilon n}\tilde{N}^{(n)}$. Then, we use Lemma 4.4 to sample $(1 \pm \gamma_6)$-pointwise close to uniform from the set $S^{(t)}$ in time $\mathrm{poly}(d, \log\frac{R}{r}, \log\gamma_6^{-1})$.

Overall, assuming that $\mathrm{vol}(K_2^{(t)}) - \mathrm{vol}(K_1^{(t)}) \le \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$ for all t, since $\gamma_6 < \varepsilon$, we obtain an $e^{\pm 2\varepsilon}$-pointwise approximation to sampling from the distribution proportional to $e^{-\varepsilon \cdot t(\theta)}$ for $\theta \in \tilde{\Theta}$, with failure probability at most γ_8. In addition, note that $t(\theta) = \mathcal{S}(\theta, \mathcal{Y}) - T \pm O(1)$, which means in fact we are sampling proportional to $e^{-\varepsilon \cdot \mathcal{S}(\theta, \mathcal{Y})}$ up to a $e^{\pm O(\varepsilon)}$ pointwise approximation. There are two ways for this to fail: if either there is some t with $\mathrm{vol}(K_2^{(t)}) - \mathrm{vol}(K_1^{(t)}) > \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$ or in the γ_8 probability event that some estimate $\tilde{N}^{(t)}$ is incorrect. Note however, that this volume represents the set of points with score between $T' + t + 1 + u - \gamma_7$ and $T' + t + 1 + u + \gamma_7$, where $u \in [0, 1)$ is chosen at random to be an integer multiple of γ_7. Therefore, the expectation $\mathbb{E}_u[\mathrm{vol}(K_2^{(t)}) - \mathrm{vol}(K_1^{(t)})]$ is at most $2 \cdot \gamma_7$ times the volume difference of points with score at least $T' + t + 2$ and $T' + t + 1$, which is at most $\mathrm{vol}(\Theta) \le (2R)^d$. So, by Markov's inequality, $\mathrm{vol}(K_2^{(t)}) - \mathrm{vol}(K_1^{(t)}) > \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$ with probability at most $\left(2\gamma_7 \cdot (2R)^d\right) / \left(\frac{\gamma_1 \cdot r}{6d}\right)^d = \gamma_8$.

Therefore, with probability at least $1 - 2\gamma_8$, we are sampling $\theta \in \tilde{\Theta}$ from a distribution proportional to $e^{-\varepsilon \cdot \mathcal{S}(\theta, \mathcal{Y})} \cdot e^{\pm O(\varepsilon)}$. However, note that the number of points in $\tilde{\Theta}$ is at most $\mathrm{vol}(\Theta)/(\gamma_5)^d \cdot (1 + o(1)) \le (2R/\gamma_5)^d$ by Fact 4.6, so each point in $\tilde{\Theta}$ is selected with probability at least $\Omega(e^{-n} \cdot (\gamma_5/2R)^d)$. So, since we set $\gamma_8 = \frac{\varepsilon}{2} \cdot e^{-n} \cdot (\gamma_5/2R)^d$, we are still sampling each element with probability proportional to $e^{-\varepsilon \cdot \mathcal{S}(\theta, \mathcal{Y})} \cdot e^{\pm O(\varepsilon)}$. □

Proof of Theorem 4.1. The algorithm is the same as in Lemma 4.7. To see why this implies a private algorithm, for any two adjacent datasets $\mathcal{Y}, \mathcal{Y}'$, the score of any point changes by at most 1, which means the distribution does not change by more than a $e^{\pm O(\varepsilon)}$ factor multiplicatively for any fixed θ between $\mathcal{Y}$ and $\mathcal{Y}'$. So, if we could approximately sample from this distribution, the distribution still does not change by more than a $e^{\pm O(\varepsilon)}$ factor multiplicatively. This ensures the algorithm will be $O(\varepsilon)$-DP.

The runtime has already been verified, with the fact that $n \ge \Omega(1/\varepsilon)$ is already known, so we can ignore polynomial runtime dependencies on $\frac{1}{\varepsilon}$.

Finally, we check accuracy. Assume there exists a $\theta \in \Theta$ with score at most ηn. By Fact 4.6, if $\gamma_5 \le \frac{r}{d^3}$, then for any convex body K containing a ball of radius r, $\mathrm{vol}(K) = (1 \pm o(1)) \cdot (\gamma_5)^d \cdot N_K$ if N_K is the number of points in $K \cap \tilde{\Theta}$. Now, for any $j \ge 1$, we bound the probability that we select a $\theta \in \tilde{\Theta}$ with score between $2^j \cdot \eta n$ and $2^{j+1} \cdot \eta n$. If we consider K_j to be the convex body of points in Θ with score at most $2^{j+1} \cdot \eta n$, then the probability of sampling a point with a score between $2^j \cdot \eta n$ and $2^{j+1} \cdot \eta n$ is proportional to at most $e^{-\varepsilon \cdot 2^j \cdot \eta n} \cdot \mathrm{vol}(K_j)/(\gamma_5)^d \cdot (1 + o(1))$. However, the set of points with score at most $\eta n + 1$ contains a ball of radius r, so the probability of sampling such a point is proportional to at least $e^{-\varepsilon \cdot (\eta n + 1)} \cdot V_\eta/(\gamma_5)^d \cdot (1 - o(1))$.

So, to select a point with score at most $2\eta n$ with probability $1 - O(\beta)$, it suffices to check that $\sum_{j=1}^{\lceil \log_2(1/\eta) \rceil} e^{-\varepsilon \cdot (2^j - 1) \cdot \eta n} \cdot \mathrm{vol}(K_j)/V_\eta \le \beta$. Now, by setting $\eta' = 2^{j+1} \cdot \eta$, we have that $\mathrm{vol}(K_j)/V_\eta = V_{\eta'}/V_\eta$, and $e^{-\varepsilon \cdot (2^j - 1) \cdot \eta n} \le e^{-\varepsilon \cdot \eta' \cdot n/4}$. Thus, if $n \ge 8 \log(V_{\eta'}/V_\eta) \cdot \frac{1}{\eta' \cdot \varepsilon}$, then $e^{-\varepsilon \cdot \eta' \cdot n/8} \le \frac{V_\eta}{V_{\eta'}}$. Also, if $n \ge \frac{8 \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}$, then $e^{-\varepsilon \cdot \eta' \cdot n/8} \le \beta \cdot \eta$. Therefore, $e^{-\varepsilon \cdot \eta' \cdot n/4} \le \frac{V_\eta}{V_{\eta'}} \cdot \beta \cdot \eta$, which means $\sum_{j=1}^{\lceil \log_2(1/\eta) \rceil} e^{-\varepsilon \cdot (2^j - 1) \cdot \eta n} \cdot \frac{\mathrm{vol}(K_j)}{V_\eta} \le \sum_{j=1}^{\lceil \log_2(1/\eta) \rceil} \frac{V_\eta}{\mathrm{vol}(K_j)} \cdot \beta \cdot \eta \cdot \frac{\mathrm{vol}(K_j)}{V_\eta} \le \beta$. Thus, the algorithm is accurate with $1 - \beta$ probability. □

4.3 Proof of Theorem 4.2

In this subsection, we prove Theorem 4.2. We start by describing the algorithm.

First, we define the function $g : \mathbb{Z} \to [0, 1]$ as follows. First, for $t < 0.3\eta^* n$, we let $g(t) = 1$, and for $t \geq 0.7\eta^* n$, we let $g(n) = 0$. For $0.3 \cdot \eta^* n \leq t \leq 0.5\eta^* n$, we let $g(t) = \max\left(\frac{1}{2}, 1 - \delta \cdot e^{\varepsilon(t-0.3\eta^* n)}\right)$, and for $0.5\eta^* n \leq t \leq 0.7\eta^* n$, we let $g(t) = \min\left(\frac{1}{2}, \delta \cdot e^{\varepsilon(0.7\eta^* n - t)}\right)$. The first step of the algorithm is to compute some $\hat{T}$, which equals $\min_{\theta\in\Theta} \mathcal{S}(\theta, \mathcal{Y})$ up to additive error 1. The first part of the algorithm, which we call $\mathcal{A}_1$, will *accept* the dataset $\mathcal{Y}$ with probability $g(\hat{T})$.

If $\mathcal{A}_1$ rejects $\mathcal{Y}$, the overall algorithm $\mathcal{A}$ outputs nothing. If $\mathcal{A}_1$ accepts $\mathcal{Y}$, the algorithm proceeds to the second phase. The second phase, at a high level, attempts to sample a θ proportional to $e^{-\varepsilon\cdot\mathcal{S}(\theta,\mathcal{Y})}$ as long as $\mathcal{S}(\theta, \mathcal{Y}) \leq 0.9\eta^* n$. This may be impossible as we cannot perfectly compute θ. Instead, if we define the function $h(t)$ to equal $e^{-\varepsilon t}$ for $t \leq 0.9\eta^* n$ and 0 for $t > 0.9\eta^* n$, we prove the following.

Lemma 4.8. *Let $\tilde{\Theta}$ be as in Lemma 4.7. Then, in time $\mathrm{poly}(n, d, \frac{1}{\varepsilon}, \log\frac{R}{r})$ time and with failure probability at most $\min(\beta, \delta)$, we can sample from each $\theta \in \tilde{\Theta}$ with probability proportional to $h'(\theta)$, where $h'(\theta)$ is a function satisfying $h(\mathcal{S}(\theta, \mathcal{Y}) + O(1)) \cdot e^{-O(\varepsilon)} \leq h'(\theta) \leq h(\mathcal{S}(\theta, \mathcal{Y}) - O(1)) \cdot e^{O(\varepsilon)}$.*

Proof. The proof is nearly identical to that of Lemma 4.7. We again define $t(\theta)$ to be the smallest nonnegative t such that our estimate of $\mathcal{S}(\theta, \mathcal{Y})$ is at most $\hat{T}+t$. This time, rather than approximately sampling with probability proportional to $e^{-\varepsilon t(\theta)}$, we approximately sample proportional to $h(t(\theta)+\hat{T})$. (Note that $t(\theta) + \hat{T} = \mathcal{S}(\theta, \mathcal{Y}) \pm O(1)$.) As in Lemma 4.7, we define $S^{(t)}, N^{(t)}$ to be the set of and number of points in $\tilde{\Theta}$, respectively, such that $t(\theta) \leq t$. We can write $\sum_{\theta\in\tilde{\Theta}} h(t(\theta) + \hat{T}) = h(\hat{T}) \cdot N^{(0)} + \sum_{t=1}^{n} h(\hat{T} + t)(N^{(t)} - N^{(t-1)}) = \sum_{t=0}^{n-1} \left(h(\hat{T} + t) - h(\hat{T} + t + 1)\right) N^{(t)}$, since $h(\hat{T} + n) = 0$.

Again, we can compute each $N^{(t)}$ up to a $e^{\pm\varepsilon}$ multiplicative factor (to get estimates $\tilde{N}^{(t)}$, choose t proportional to $\tilde{N}^{(t)} \cdot (h(\hat{T} + t) - h(\hat{T} + t + 1))$, and then sample $(1 \pm \gamma_6)$-pointwise close to uniform on $S^{(t)}$, where we ensure $\gamma_6 \leq \varepsilon$. The algorithm fails with probability $2\gamma_8$. This time we cannot charge this error to multiplicative error (since some points may have large enough score that they will be sampled with probability 0), so we additionally ensure that $\gamma_8 \leq \frac{\min(\beta,\delta)}{2}$ as well. (I.e., we set $\gamma_8 := \min\left(\frac{\beta}{2}, \frac{\delta}{2}, \frac{\varepsilon}{2} \cdot e^{-n} \cdot (\gamma_5/2R)^d\right)$.) Note that as long as $n \geq \log\delta^{-1} + \log\beta^{-1}$, the runtime is still $\mathrm{poly}(n, d, \frac{1}{\varepsilon}, \log\frac{R}{r})$. □

Proof of Theorem 4.2. The algorithm is as described, where the second phase samples (with failure probability at most $\min(\beta, \delta)$) proportional to $h'(\theta)$. We recall that by Fact 4.6, for the convex body $N^{(t)}$ of points with score at most $t + \hat{T}$, $\mathrm{vol}(K^{(t)}) = (1 \pm o(1)) \cdot \gamma_5^d \cdot N^{(t)}$.

First, we check privacy. It is clear that the first phase of the algorithm is $(O(\varepsilon), O(\delta))$-DP as long as $n \gg \frac{\log\delta^{-1}}{\varepsilon\cdot\eta^*}$, since for any two adjacent datasets, $\max_{\theta\in\Theta} \mathcal{S}(\theta, \mathcal{Y})$ changes by at most 1, and our estimate for this maximum is accurate up to error 1. So, $\hat{T}$ changes by at most $O(1)$ between adjacent datasets $\mathcal{Y}$ and $\mathcal{Y}'$, which is sufficient. For the second phase, we sample each θ proportional to $e^{-\varepsilon\cdot(\mathcal{S}(\theta,\mathcal{Y})\pm O(1))}$ if $\mathcal{S}(\theta, \mathcal{Y}) \leq 0.9\eta^* n - O(1)$, and proportional to 0 if $\mathcal{S}(\theta, \mathcal{Y}) \geq 0.9\eta^* n + O(1)$. So, the sampling probability stays proportional between adjacent datasets, unless $\mathcal{S}(\theta, \mathcal{Y}) = 0.9\eta^* n \pm O(1)$. So, we need to make sure the probability of sampling such a dataset is at most $O(\delta)$, so that the overall algorithm is $(O(\varepsilon), O(\delta))$-DP.

To see why this is true, the probability of sampling a point $\theta \in \tilde{\Theta}$ with score in the range $0.9\eta^* n \pm O(1)$ is proportional to at most $e^{-\varepsilon\cdot(0.9\eta^* n - O(1))}(1+o(1)) \cdot \gamma_5^{-d} \cdot V_{0.9\eta^*+O(1/n)}(\mathcal{Y}) \le O(1) \cdot e^{-\varepsilon\cdot 0.9\eta^* n} \cdot \gamma_5^{-d} \cdot V_{\eta^*}(\mathcal{Y})$. Conversely, since we didn't reject, we know that $\hat{T} \le 0.7\eta^* n$, which means the probability that we sample a point $\theta \in \tilde{\Theta}$ with score at most $0.85\eta^* n$ is proportional to at least $e^{-\varepsilon\cdot(0.85\eta^* n + O(1))}(1-o(1)) \cdot \gamma_5^{-d} \cdot V_{0.85\eta^* - O(1/n)}(\mathcal{Y}) \ge \Omega(1) \cdot e^{-\varepsilon\cdot 0.85\eta^* n} \cdot \gamma_5^{-d} \cdot V_{0.8\eta^*}(\mathcal{Y})$. So, it suffices to show that $\frac{e^{-\varepsilon\cdot 0.9\eta^* n}\cdot\gamma_5^{-d}\cdot V_{\eta^*}(\mathcal{Y})}{e^{-\varepsilon\cdot 0.85\eta^* n}\cdot\gamma_5^{-d}\cdot V_{0.8\eta^*}(\mathcal{Y})} \le \delta$, which is true as long as $n \ge \frac{\log(1/\delta)\cdot\log(V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}))}{\varepsilon\cdot\eta^*}$. Note that this only has to be true for datasets $\mathcal{Y}$ such that $\min_\theta \mathcal{S}(\theta,\mathcal{Y}) \le 0.7\eta^* n$, since otherwise the algorithm would have already rejected in the first phase.

To check efficiency, note that we can compute $\hat{T}$ in $\mathrm{poly}(n, d, \log\frac{R}{r})$ time, using the condition that the robust algorithm can find a low-scoring point up to error 1. Then, in time $\mathrm{poly}(n, d, \log\frac{R}{r})$, since $n \ge \Omega(\frac{1}{\varepsilon})$, we can sample proportional to $h'(\theta)$, using Lemma 4.8.

Checking accuracy will be very similar to as in the proof of Theorem 4.1. Suppose there exists θ such that $\mathcal{S}(\theta,\mathcal{Y}) \le \eta n$. Then, $\hat{T} \le \eta n + O(1) \le 0.1\eta^* n + O(1)$, which means the first part of the algorithm will succeed. For the second phase, we sample each $\theta \in \tilde{\Theta}$ with probability proportional to $h'(\theta)$, with failure probability at most β. The probability of sampling a point with score at most $\eta n + 1$ is proportional to at least $e^{-\varepsilon\cdot(\eta n + O(1))} \cdot V_\eta(\mathcal{Y})/(\gamma_5)^d \cdot (1-o(1))$. Also, if we define K_j to be the convex body of points in Θ with score at most $2^{j+1}\cdot\eta n$, then the probability of selecting a $\theta \in \tilde{\Theta}$ with score between $2^j\cdot\eta n$ and $2^{j+1}\cdot\eta n$ is proportional to at most $e^{-\varepsilon\cdot(2^j\cdot\eta n - O(1))}\cdot\min(\mathrm{vol}(K_j), V_{\eta^*}(\mathcal{Y}))/(\gamma_5)^d\cdot(1+o(1))$, by Lemma 4.8 and the definition of h.

Hence, we wish to check that $\sum_{j=1}^{\lceil\log_2(\eta^*/\eta)\rceil} e^{-\varepsilon\cdot(2^j-1)\cdot\eta n}\cdot\min(\mathrm{vol}(K_j), V_{\eta^*}(\mathcal{Y}))/V_\eta(\mathcal{Y})$ is at most β: it suffices to show that $e^{-\varepsilon\cdot 2^j\cdot\eta n/2}\cdot\min(\mathrm{vol}(K_j), V_{\eta^*}(\mathcal{Y}))/V_\eta(\mathcal{Y}) \le \beta\cdot\eta$ for any $1 \le j \le \lceil\log_2(\eta^*/\eta)\rceil$. If $2^{j+1}\cdot\eta \le \eta^*$, then by setting $\eta' = 2^{j+1}$, we are assuming that $n \ge \Omega\left(\frac{\log(V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}))+\log(1/(\beta\cdot\eta))}{\varepsilon\cdot\eta'}\right)$. This means $e^{-\varepsilon\cdot 2^j\cdot\eta n/2}\cdot\min(\mathrm{vol}(K_j), V_{\eta^*}(\mathcal{Y}))/V_\eta(\mathcal{Y}) \le e^{-\varepsilon\cdot\eta' n/4}\cdot\mathrm{vol}(K_j)/V_\eta(\mathcal{Y}) \le \frac{V_\eta(\mathcal{Y})}{V_{\eta'}(\mathcal{Y})}\cdot\beta\cdot\eta\cdot\frac{\mathrm{vol}(K_j)}{V_\eta(\mathcal{Y})} = \beta\cdot\eta$. If $2^{j+1}\cdot\eta > \eta^*$, then by setting $\eta' = \eta^*$, we are assuming that $n \ge \Omega\left(\frac{\log(V_{\eta^*}(\mathcal{Y})/V_\eta(\mathcal{Y}))+\log(1/(\beta\cdot\eta))}{\varepsilon\cdot\eta^*}\right)$. This means $e^{-\varepsilon\cdot 2^j\cdot\eta n/2}\cdot\min(\mathrm{vol}(K_j), V_{\eta^*}(\mathcal{Y}))/V_\eta(\mathcal{Y}) \le e^{-\varepsilon\cdot\eta^* n/4}\cdot V_{\eta^*}(\mathcal{Y})/V_\eta(\mathcal{Y}) \le \frac{V_\eta(\mathcal{Y})}{V_{\eta^*}(\mathcal{Y})}\cdot\beta\cdot\eta\cdot\frac{V_{\eta^*}(\mathcal{Y})}{V_\eta(\mathcal{Y})} = \beta\cdot\eta$. Hence, the algorithm is accurate. □

5 Estimating the Mean of a Gaussian

5.1 Main Theorem

Our main theorem in this section is a polynomial time and pure-DP algorithm for private mean estimation of an identity-covariance Gaussian, with optimal sample complexity.

Theorem 5.1 (Private Mean Estimation of a (Sub-)Gaussian)**.** *Assume that $0 < \alpha, \beta, \varepsilon < 1$ and $R > 0$. Let $\mu \in \mathbb{R}^d$, where $\|\mu\|_2 \le R$, be unknown. There is an ε-DP algorithm that takes n i.i.d. samples from $\mathcal{N}(\mu, I)$ (or in general, a subgaussian distribution with mean μ and covariance I) and with probability $1-\beta$ outputs μ such that $\|\mu - \hat{\mu}\|_2 \le \alpha$, where*

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

Here, $\widetilde{O}$ only hides logarithmic factors in $1/\alpha$. Moreover, this algorithm runs in time $\mathrm{poly}(n, d)$, and succeeds with the same accuracy even if $\eta = \tilde{\Omega}(\alpha)$ fraction of the samples are adversarially corrupted, assuming $\eta \le \eta^*$ for some universal constant η^*.

For pure-DP algorithms, the $\frac{d \log R}{\varepsilon}$ term is required by a standard packing lower bound. However, in the approximate-DP setting, we can replace this term with $\frac{\log(1/\delta)}{\varepsilon}$, as we now state.

Theorem 5.2 (Private Mean Estimation of a (Sub-)Gaussian with Approx-DP)**.** *Let $\mu \in \mathbb{R}^d$, where $\|\mu\|_2 \le R$, be unknown. There is an (ε, δ)-DP algorithm that takes n i.i.d. samples from $\mathcal{N}(\mu, I)$ (or a subgaussian distribution with mean μ and covariance I) and with probability $1 - \beta$ outputs μ such that $\|\mu - \hat{\mu}\|_2 \le \alpha$, where*

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

Moreover, this algorithm runs in time $\mathrm{poly}(n, d, \log R)$, and still succeeds with the same accuracy even if $\eta = \tilde{\Omega}(\alpha)$ fraction of the samples are adversarially corrupted.

Note that the runtime dependence on $\log R$ is required as even reading the input up to $O(1)$-precision requires $\log R$ time.

We note that one could alternatively prove Theorem 5.2 by combining Theorem 5.1 with [EMN22, Corollary 5] (or alternatively [GKM21, TCK+22]), which allows us to learn μ up to radius $O(d)$ first. However, this method is slightly suboptimal in that the final term would be $\frac{\log(1/\delta)\cdot\log(1/\beta)}{\varepsilon}$.

The rest of this section is devoted to proving Theorem 5.1 and Theorem 5.2.

5.2 Resilience of First and Second Moments

In this subsection, we note some known concentration inequalities for subgaussian random variables (commonly known as resilience or stability conditions) that will be crucial for our analysis.

Lemma 5.3 (Resilience of First and Second Moments, Proposition 3.3 in [DK22])**.** *Let $n \ge O((d + \log(1/\beta))/\alpha^2)$, for some $\alpha = \widetilde{O}(\eta)$. Let $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{D}$, where $\mathcal{D}$ is a subgaussian random variable with mean $\mu \in \mathbb{R}^d$ and covariance I. Then, with probability $1-\beta$, for all vectors $b \in [0,1]^n$ such that $\mathbf{E}_i\, b_i \ge 1-\eta$ and all unit vectors $v \in \mathbb{R}^d$, we have*

$$\left|\mathbf{E}_i\, b_i \langle v, x_i - \mu\rangle\right| \le \alpha.$$

In addition,

$$\left|\mathbf{E}_i\, b_i \langle v, x_i - \mu\rangle^2 - 1\right| \le \alpha.$$

Corollary 5.4. *Let $\mu, \mathcal{D}, \{x_i\}, \alpha, \eta$ be as in Lemma 5.3. Then, with probability $1 - \beta$, the following all hold for all unit vectors v simultaneously.*

1. $\left|\mathbf{E}_i \langle x_i - \mu, v\rangle\right| \le \alpha.$
2. $\left|\mathbf{E}_i \langle x_i - \mu, v\rangle^2 - 1\right| \le \alpha.$
3. *For any real values $a_1, \ldots, a_n \in [0,1]$ such that $\sum_{i=1}^n a_i \le \eta \cdot n$, $\left|\mathbf{E}_i\, a_i \langle x_i - \mu, v\rangle\right| \le \alpha$ and $\left|\mathbf{E}_i\, a_i \langle x_i - \mu, v\rangle^2\right| \le \alpha$.*

4. $\mathbf{E}_i \left|\langle x_i - \mu, v\rangle\right| \le O(1)$.

Proof. Fix a vector v and let $z_i := \langle x_i - \mu, v\rangle$. Suppose the events of Lemma 5.3 hold.

Parts 1 and 2 are immediate from Lemma 5.3, by setting $b_i = 1$ for all i. Part 3 follows by setting $a_i = 1 - b_i$, and then noticing that $\left|\frac{1}{n}\sum_{i=1}^n a_i z_i\right| \le \left|\frac{1}{n}\sum_{i=1}^n z_i\right| + \left|\frac{1}{n}\sum_{i=1}^n b_i z_i\right| \le \widetilde{O}(\eta)$.

Finally, to check part 4, we may consider $\eta = 0.1$ and then apply part 3, to obtain that $\left|\frac{1}{n}\sum_{i=1}^n a_i z_i\right| \le O(1)$ for any $a \in [0,1]^n$ with $\sum a_i \le 0.1n$. Since every vector in $[-1,1]^n$ can be written as a sum of at most 20 vectors $a \in [0,1]^n$ with $\sum a_i \le 0.1n$, we thus have that $\left|\frac{1}{n}\sum_{i=1}^n c_i z_i\right| \le O(1)$ for all choices of $c_i \in \{-1,1\}^n$ simultaneously. Thus, $\frac{1}{n}\sum_{i=1}^n |z_i| \le O(1)$. □

Remark. The conditions in Corollary 5.4 will be the only conditions we will require about the samples we draw. So in fact, our algorithm will output a point close to μ if given an η-corrupted version of $\mathcal{X}$ for any $\mathcal{X}$ satisfying Corollary 5.4.

We also note that if $\mu, \{x_i\}$ satisfy Corollary 5.4, then for all symmetric H with $\|H - I\|_{op} \le \alpha$, $H\mu, \{Hx_i\}$ also satisfies Corollary 5.4 (up to replacing α with $O(\alpha)$). To see why, assume without loss of generality that $\mu = \mathbf{0}$. Then, using Condition 1, for all unit vectors v, $|\mathbf{E}_i\langle Hx_i, v\rangle| = |\mathbf{E}_i\langle x_i, Hv\rangle| \le \alpha \cdot \|Hv\|_2 \le \alpha \cdot (1+\alpha) \le 2\alpha$. We can repeat the same argument for the 2nd, 3rd, and 4th conditions.

5.3 Robust Algorithm

Here, we describe the robust algorithm that will inspire our score function to generate a differentially private algorithm. The robust algorithm, as well as the algorithms used in the covariance settings, are essentially the same as in [KMZ22].

Suppose $\{x_i\}_{i=1}^n$ are samples from $\mathcal{N}(\mu, I)$ (or a subgaussian distribution with mean μ and covariance I). Let $\{y_i\}$ be an arbitrary η-corruption of the $\{x_i\}$. Consider the following pseudo-expectation program with input points $\{y_i\}$ and domain the degree-4 pseudo-expectations with $\{w_i\}, \{x_i\}, \{M_{i,j}\}$ as indeterminates. ($M = \{M_{i,j}\}$ will represent a $d \times d$-matrix of indeterminates.)

$$
\begin{aligned}
&\text{find } \tilde{\mathbf{E}}\\
&\text{such that } \tilde{\mathbf{E}} \text{ satisfies } w_i^2 = w_i,\\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } \sum w_i \ge (1-\eta)n,\\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } w_i x_i' = w_i y_i,\\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } \frac{1}{n}\sum (x_i' - \mu')(x_i' - \mu')^\mathsf{T} + MM^\mathsf{T} = (1 + \widetilde{O}(\eta))I, \text{where } \mu' = E_i x_i'
\end{aligned}
$$

It can be proven that if n is as in Lemma 5.3, with probability $1 - \beta$ over the choice of $\{x_i\}$ and for any η-corruption $\{y_i\}$ of $\{x_i\}$, then $\|\tilde{\mathbf{E}}\mu' - \mu\|_2 = \widetilde{O}(\eta)$ for *any* feasible pseudo-expectation $\tilde{\mathbf{E}}$.

5.4 Score Function and its Properties

Our goal is to use Theorem 4.1, but to do so, we need to design a suitable score function. Our score function will be very similar to the robust algorithm, but modified to deal with precision issues.

Before we define our score function, we make a definition of *certifiable means*, which modifies the pseudoexpectation program in Section 5.3 to deal with approximate pseudoexpectations.

Definition 5.5 (Certifiable Mean)**.** Let $\alpha, \tau, \phi, T \in \mathbb{R}^{\geq 0}$, $y_1, \dots y_n \in \mathbb{R}^d$ (with $\mathcal{Y} := \{y_1, \dots, y_n\}$), and $\widetilde{\mu} \in \mathbb{R}^d$. We call the point $\widetilde{\mu}$ an (α, τ, ϕ, T)-*certifiable mean* for $\mathcal{Y}$ if and only if there exists a linear functional $\mathcal{L}$ over the set of polynomials in indeterminates $\{w_i\}, \{x'_{i,j}\}, \{M_{j,k}\}$ of degree at most 6 such that

1. $\mathcal{L}1 = 1$,
2. for every polynomial p, where $\|\mathcal{R}(p)\|_2 \leq 1$ (where we recall that $\mathcal{R}(p)$ is the vector of monomial coefficients of p):
   (a) $\mathcal{L}p^2 \geq -\tau \cdot T$,
   (b) $\forall i, \mathcal{L}(w_i^2 - w_i)p^2 \in [-\tau \cdot T, \tau \cdot T]$,
   (c) $\mathcal{L}(\sum w_i - n + T)p^2 \geq -5\tau \cdot T \cdot n$,
   (d) $\forall i, j, \mathcal{L}w_i(x'_{i,j} - y_{i,j})p^2 \in [-\tau \cdot T, \tau \cdot T]$,
   (e) $\forall j, k : \mathcal{L}\left(\left[\frac{1}{n}\sum_i (x'_i - \mu')(x'_i - \mu')^{\mathsf{T}} + MM^{\mathsf{T}} - (1+\alpha)I\right]_{j,k} p^2\right) \in [-\tau \cdot T, \tau \cdot T]$, where $x'_i = \{x'_{i,j}\}_{1 \leq j \leq d}$, and $\mu' = \mathbf{E}_i\, x'_i$. Note that here $[\dots]_{j,k}$ denotes the (j,k) entry of a matrix, which is a polynomial in indeterminates $\{w_i\}, \{x'_i\}, \{M_{j,k}\}$. We write in this format for the sake of conciseness.
3. $\forall i, \mathcal{L}\mu'_i - \widetilde{\mu}_i \in [-\phi - \tau \cdot T, \phi + \tau \cdot T]$.

In addition we will require $\|\mathcal{R}(\mathcal{L})\|_2 \leq R' + T \cdot \tau$ for some sufficiently large $R' = \mathrm{poly}(n, d, R)$,[8] where we recall that $\mathcal{R}(\mathcal{L})$ is the vector which represents the value of $\mathcal{L}$ applied to each monomial of degree at most 6. (Note that $\mathcal{R}(\mathcal{L})$ has dimension polynomial in the number of variables, which is polynomial in n, d.) This requirement is only needed for computability purposes. For such $\mathcal{L}$, we also say that $\mathcal{L}$ is an (α, τ, ϕ, T)-certificate for $\mathcal{Y}$.

Note that one may think of $\mathcal{L}$ as an approximate pseudo-expectation. In addition, for each constraint 2a) to 2e) we implicitly assume a bound on the degree of p so that $\mathcal{L}$ is applied to a polynomial of degree at most 6.

For our purposes, we will end up setting $\tau = 1/(n \cdot d)^{O(1)}$, for a large enough $O(1)$.

Now we use this definition to define a score function.

Definition 5.6 (Score Function)**.** Let $\mathbb{B}_R^d$ denote the ball of radius R in $\mathbb{R}^d$ centered at the origin. Let $\alpha, \tau, \phi, T \in \mathbb{R}^{\geq 0}$, $y_1, \dots y_n \in \mathbb{R}^d$ (with $\mathcal{Y} = \{y_1, \dots, y_n\}$) and $\widetilde{\mu} \in \mathbb{R}^d$. We define the score function $\mathcal{S} : \mathbb{B}_R^d \to \mathbb{R}$ as

$$\mathcal{S}(\widetilde{\mu}, \mathcal{Y}; \alpha, \tau, \phi) = \min_T \text{ such that } \widetilde{\mu} \text{ is a } (\alpha, \tau, \phi, T) \text{ certifiable mean for } \mathcal{Y} = \{y_1, \dots, y_n\}.$$

In the rest of this section we will prove the following properties for this score function. This will allow us to use Theorem 4.1.

1. Bounded Sensitivity: Score has sensitivity 1 with respect to $\mathcal{Y}$.

[8] See Lemma 5.11 for more details on how large we require R' to be.

2. Quasi-Convexity: Score is quasi-convex as a function of $\widetilde{\mu}$.

3. Accuracy: All points $\widetilde{\mu}$ that have score at most $\eta \cdot n$ have distance at most $\alpha = \widetilde{O}(\eta)$ away from μ. (Robustness for volume/accuracy purposes).

4. Volume: The volume of points that have score at most $\eta \cdot n$ is sufficiently large, and the volume of points with score at most $\eta' \cdot n$ for $\eta' > \eta$ is not too large.

5. Efficient Computability: Score is efficiently computable for any fixed $\widetilde{\mu}, \mathcal{Y}$.

6. Robust algorithm finds low-scoring point: Finding $\widetilde{\mu}$ that minimizes score (up to error 1) for any fixed $\mathcal{Y}$ can be done efficiently.

5.4.1 Sensitivity

Before proving sensitivity we need to prove the following upper bound on the value of the score function.

Lemma 5.7 (score function upper bound)**.** *The value of the score function $\mathcal{S}$ defined in Definition 5.6 is less than or equal to n.*

Proof. It suffices to show that in Definition 5.5 for $T = n$, there exists a linear functional $\mathcal{L}$ such that the constraints of Definition 5.5 are satisfied.

Let's define $\mathcal{L}$. For any monomial p we should assign a value to $\mathcal{L}p$. To begin, let $\mathcal{L}1 = 1$. If p contains w_i or $M_{j,k}$ where $j \neq k$ let $\mathcal{L}p = 0$. Now we need to define $\mathcal{L}p$ for monomials that only contain $x'_{i,j}$ and $M_{j,j}$. For such monomials p, let $\mathcal{L}p$ be equal to $(1+\eta)^{(\beta/2)} \cdot \prod_{j=1}^{d} \widetilde{\mu}_j^{\alpha_j}$, where α_j is equal to the sum of the number of the factors of the form $x'_{i,j}$ over all i in p, and β is equal to the number of $M_{j,j}$ factors in p over all j. Basically, when applying $\mathcal{L}$ to a polynomial we are treating the indeterminates in the problem as if they were scalars and had the assignment $w_i = 0$, $x'_i = \widetilde{\mu}$, and $M = \sqrt{(1+\eta)I}$. In the non-relaxed version of the problem, this assignment would correspond to changing every point to $\widetilde{\mu}$. It is easy to check that all of the constraints would be satisfied under this choice of $\mathcal{L}$, even if $\tau = 0$. Therefore, the value of the score function $\mathcal{S}$ defined in Definition 5.6 is at most n. □

Lemma 5.8 (sensitivity)**.** *The score function $\mathcal{S}$ as defined in Definition 5.6 has sensitivity 1 with respect to its first input.*

Proof. Suppose that $\mathcal{Y}, \mathcal{Y}'$ are two neighboring datasets, and $\widetilde{\mu} \in \mathbb{R}^d$. Moreover, assume $\mathcal{S}(\widetilde{\mu}, \mathcal{Y}) = T$. If we show that $\mathcal{S}(\widetilde{\mu}, \mathcal{Y}') \leq \mathcal{S}(\widetilde{\mu}, \mathcal{Y}) + 1 = T + 1$, by symmetry we are done. Since $\mathcal{S}(\widetilde{\mu}, \mathcal{Y}) = T$, we know that there exists some functional $\mathcal{L}$ such that the constraints of Definition 5.5 are satisfied for $\mathcal{L}, \mathcal{Y}$, and T. If we construct a new functional $\mathcal{L}'$ such that the constraints of Definition 5.5 are satisfied for $\mathcal{L}', \mathcal{Y}'$ and $T+1$, we have shown that $\mathcal{S}(\widetilde{\mu}, \mathcal{Y}') \leq T + 1$ and we are done.

Without loss of generality assume $\mathcal{Y}$ and $\mathcal{Y}'$ differ on index j. In order to construct $\mathcal{L}'$, for any monomial p, let

$$\mathcal{L}'p = \begin{cases} 0 & \text{if } p \text{ has a } w_j \text{ factor,} \\ \mathcal{L}p & \text{otherwise} \end{cases}.$$

Now let's go through all of the constraints and verify them. The first condition holds since by definition, $\mathcal{L}'1 = \mathcal{L}1 = 1$. Now let's prove the conditions in the second set of conditions. Suppose $\|p\|_2 \le 1$ and $p = q + w_j r$, where q does not contain a monomial containing w_j.

- $\mathcal{L}'p^2 \ge -\tau \cdot (T+1)$.

$$\mathcal{L}'p^2 = \mathcal{L}'(q + w_j r)^2 = \mathcal{L}'q^2 = \mathcal{L}q^2 = -\tau \cdot T \ge -\tau \cdot (T+1)$$

  as desired, where we used the fact that $\|q\|_2 \le \|p\|_2 \le 1$.

- $\forall i : \mathcal{L}'(w_i^2 - w_i)p^2 \in [-\tau \cdot (T+1), \tau \cdot (T+1)]$. If $i = j$, this would be zero, if not then we can write p as $q + w_j r$ similar to the previous part and get the desired bounds.

- $\mathcal{L}'(\sum w_i - n + (T+1))p^2 \ge -5\tau \cdot (T+1) \cdot n$.

$$\begin{aligned}
\mathcal{L}'(\sum w_i - n + (T+1))p^2 &= \mathcal{L}'(\sum w_i - n + (T+1))q^2 \\
&= \mathcal{L}'(\sum_{i \ne j} w_i - n + (T+1))q^2 \\
&= \mathcal{L}(\sum_{i \ne j} w_i - n + (T+1))q^2 \\
&= \mathcal{L}(\sum w_i - n + T)q^2 - w_j q^2 + q^2
\end{aligned}$$

  To bound the first term, we have that $\mathcal{L}(\Sigma w_i - n + T)q^2 \ge -5\tau \cdot T \cdot n$. To bound the second and third terms, we have $\mathcal{L}[-w_j q^2 + q^2] = \mathcal{L}[(1 - w_j)q^2] = \mathcal{L}[(1 - w_j)^2 q^2] + \mathcal{L}[(w_j - w_j^2)q^2]$. We know that $\|(1 - w_j)q\|_2 \le 2\|q\|_2 \le 2$, so $\mathcal{L}[(1 - w_j)^2 q^2] \ge -4\tau \cdot T$, and $\mathcal{L}[(w_j - w_j^2)q^2] \ge -\tau \cdot T$. So together, we have a bound of at least $-5\tau \cdot T \cdot n - 5\tau \cdot T$. Therefore it remains to prove that $-5\tau \cdot T \cdot n - 5\tau \cdot T \ge -5\tau \cdot (T+1) \cdot n$, which is trivial by Lemma 5.7.

- $\forall j, k : \mathcal{L}'\left(\left[\frac{1}{n}\sum_i (x_i' - \mu')(x_i' - \mu')^\mathsf{T} + MM^\mathsf{T} - (1+\alpha)I\right]_{j,k} p^2\right) \in [-\tau \cdot T, \tau \cdot T]$, where $\mu' = \mathbf{E}_i\, x_i'$. Similar to previous parts, we just need to plug in $p = q + w_j r$, and we get the desired inequality.

The last condition holds because $\mathcal{L}'\mu_i' - \widetilde{\mu}_i = \mathcal{L}\mu_i' - \widetilde{\mu}_i$. Therefore we showed that there exists a linear functional $\mathcal{L}'$ which satisfies the constraints of Definition 5.5 for y', and $T+1$. Finally, note that $\|\mathcal{R}(\mathcal{L}')\|_2 \le \|\mathcal{R}(\mathcal{L})\|_2$ clearly holds. Therefore the score function $\mathcal{S}$ has sensitivity 1 with respect to its first input. □

5.4.2 Quasi-convexity

Lemma 5.9 (quasi-convexity)**.** *The score function $\mathcal{S}$ as defined in Definition 5.6 is quasi-convex in its second input, $\widetilde{\mu}$.*

Proof. Suppose $\mathcal{S}(\widetilde{\mu}_1, \mathcal{Y}) = T_1, \mathcal{S}(\widetilde{\mu}_2, \mathcal{Y}) = T_2$, and suppose there exists $\mathcal{L}_1$ and $\mathcal{L}_2$ that satisfy the constraints in Definition 5.5 with $\widetilde{\mu}_1, T_1$, and $\widetilde{\mu}_2, T_2$ respectively. If we can construct a functional $\mathcal{L}_3$ such that the constraints in Definition 5.5, are satisfied with $\widetilde{\mu}_3 = \lambda\widetilde{\mu}_1 + (1-\lambda)\widetilde{\mu}_2$, and $T_3 =$

$\max\{T_1, T_2\}$, we are done. Let $\mathcal{L}_3 = \lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2$. Then all of the constraints in Definition 5.5 will be satisfied trivially except for $\mathcal{L}_3(\sum w_i - n + T_3)p^2 \geq -5\tau \cdot T_3 \cdot n$. Let's verify this constraint. Without loss of generality suppose $T_3 = T_2 \geq T_1$, then

$$\begin{aligned}
\mathcal{L}_3\left(\sum w_i - n + T_3\right)p^2 &\geq (\lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2)\left(\sum w_i - n + T_2\right)p^2 \\
&= \lambda\mathcal{L}_1\left(\sum w_i - n + T_1\right)p^2 + (1-\lambda)\mathcal{L}_2\left(\sum w_i - n + T_2\right)p^2 + \lambda(T_2 - T_1)\mathcal{L}_1 p^2 \\
&\geq -5\tau \cdot n(\lambda T_1 + (1-\lambda)T_2) - \lambda(T_2 - T_1) \cdot \tau \cdot T_1 \\
&\geq -5\tau \cdot n(\lambda T_1 + (1-\lambda)T_2 + \lambda(T_2 - T_1)) \qquad (n \geq T_1, \text{Lemma 5.7}) \\
&= -5\tau \cdot T_3 \cdot n,
\end{aligned}$$

as desired. □

5.4.3 Accuracy

We show that any point $\widetilde{\mu}$ of low score with respect to i.i.d. samples from $\mathcal{N}(\mu, I)$ must be close to μ. We remark that because of our sensitivity bound, this will also imply a similar result for corrupted samples.

Lemma 5.10. *Let $\alpha = \widetilde{O}(\eta)$ and suppose α, η are bounded by a sufficiently small constant. Let $n \geq \frac{d+\log(1/\beta)}{\alpha^2}$, and $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\mu, I)$, for $\mu \in \mathbb{R}^d$.*

Then, for any $\alpha^* \leq \alpha$, and assuming $\tau \ll 1/(nd)^{O(1)}$, with probability at least $1-\beta$, every point $\widetilde{\mu} \in \mathbb{R}^d$ that is $(\alpha^*, \tau, \phi, T)$-certifiable for $\mathcal{X}$ with $T = \eta n$ and $\phi \leq \alpha/\sqrt{d}$ must satisfy $\|\widetilde{\mu} - \mu\|_2 \leq O(\alpha)$.

The proof of Lemma 5.10 essentially follows from the same argument as in [KMZ22], with slight modifications to deal with our modified score function. Hence, we defer the proof to Appendix B.

5.4.4 Volume of Good Points

Lemma 5.11. *Let $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\mu, I)$, and let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for any $\tau, \phi \geq 0$ and $T = \eta \cdot n$, with probability at least $1-\beta$, there exists a μ' such that every $\widetilde{\mu}$ such that $\|\widetilde{\mu} - \mu'\|_\infty \leq \phi$ is an (α, τ, ϕ, T)-certifiable mean for $\mathcal{Y}$.*

Proof. Our linear operator $\mathcal{L}$ generalizes pseudo-expectations $\tilde{\mathbf{E}}$. So, it suffices to find a pseudo-expectation on variables $\{w_i\}, \{x'_i\}, \{M_{i,j}\}$ that satisfy the constraints of the robust algorithm. If so, then by setting $\mu' = \frac{1}{n}\sum x'_i$, we have that for all $\widetilde{\mu}$ such that $\|\widetilde{\mu} - \tilde{\mathbf{E}}\mu'\|_\infty \leq \phi$, $\widetilde{\mu}$ is an (α, τ, ϕ, T)-certifiable mean.

Indeed, finding such a pseudo-expectation is quite simple to do: it will actually just be an expectation over a single point. We just set every $w_i = 1$ if $y_i = x_i$ and 0 otherwise, and set every $x'_i = x_i$, so $\mu' = \frac{1}{n}\sum_i x_i$. By Lemma 5.3, we have that $\frac{1}{n}\sum_{i=1}^n \langle x_i - \mu, v\rangle^2 \leq 1 + \widetilde{O}(\eta)$ for all unit vectors v. In addition,

$$\frac{1}{n}\sum_{i=1}^n \langle x_i - \mu, v\rangle^2 = \frac{1}{n}\sum_{i=1}^n \langle (x_i - \mu') + (\mu' - \mu), v\rangle^2 = \langle \mu' - \mu, v\rangle^2 + \frac{1}{n}\sum_{i=1}^n \langle x_i - \mu', v\rangle^2 \geq \frac{1}{n}\sum_{i=1}^n \langle x_i - \mu', v\rangle^2.$$

So, $\frac{1}{n}\sum_{i=1}^n \langle x_i - \mu', v\rangle^2 \le 1 + \widetilde{O}(\eta)$ for all unit vectors v, which means $\frac{1}{n}\sum_{i=1}^n (x_i-\mu)(x_i-\mu)^\top \preccurlyeq (1+\widetilde{O}(\eta))I$. Therefore, there exists a $d\times d$ matrix M such that $\frac{1}{n}\sum_{i=1}^n (x_i-\mu)(x_i-\mu)^\top + MM^\top = (1+\widetilde{O}(\eta))I$.

Finally, we remark that every w_i, $x_{i,j}$, and $M_{j,k}$ is bounded by $R\cdot n$. Therefore, the corresponding linear operator $\mathcal{L}$ satisfies $\|\mathcal{R}(\mathcal{L})\|_2 \le (Rnd)^{O(1)}$. □

Lemma 5.12. *Let $\mathcal{X} = \{x_1, \dots, x_n\} \sim \mathcal{N}(\mu, I)$, and let $\mathcal{Y} = \{y_1, \dots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for every integer $T \in [\eta\cdot n, \eta^*\cdot n]$ for some fixed constant $\eta^* < 1$, with probability at least $1-\beta$, every (α,τ,ϕ,T)-certifiable mean with respect to $\mathcal{Y}$ has distance at most $\widetilde{O}(T/n)$ from μ.*

Proof. Since the score function has sensitivity at most 1 (Lemma 5.8), this means that any (α,τ,ϕ,T)-certifiable mean with respect to $\mathcal{Y}$ is an $(\alpha,\tau,\phi,T+\eta n)$-certifiable mean with respect to $\mathcal{X}$.

Now, define $\eta' := \frac{T+\eta n}{n} = O(\frac{T}{n})$. In this case, by setting $\alpha' = \widetilde{O}(\eta')$ and since $\alpha = \widetilde{O}(\eta) \le \alpha'$, we have that by Lemma 5.10 that any $(\alpha,\tau,\phi,T+\eta n)$-certifiable mean $\widetilde{\mu}$ must satisfy $\|\widetilde{\mu}-\mu\|_2 \le O(\alpha') \le \widetilde{O}(T/n)$. □

If we set $\phi = \alpha/\sqrt{d}$ and $\tau \ll 1/(nd)^{O(1)}$, this means the volume of (α,τ,ϕ,T)-certifiable means for $T=\eta n$ is at least $(\alpha/\sqrt{d})^d$. However, for any $T=\eta' n$ for $\eta \le \eta' \le \eta^*$, the volume of (α,τ,ϕ,T)-certifiable means is at most $(\tilde{O}(\eta'))^d$ times the volume of a d-dimensional sphere, which is $(\tilde{O}(\eta'))^d/\sqrt{d}^d$. Finally, for $T=\eta' n$ with $\eta' > \eta^*$, the volume of Θ, the set of all candidate means $\widetilde{\mu}$ with $\|\widetilde{\mu}\|_2 \le R$, is at most $O(R/\sqrt{d})^d$.

5.4.5 Efficient Computability

Verifying that we can efficiently compute the score roughly follows from the ellipsoid method used in semidefinite programming. We had to modify the score accordingly (relaxing constraints using τ) – however, we show in Theorem C.6, deferred to Appendix C, that for the score in Definition 5.6, defined by the constraints in Definition 5.5, we can compute it up to error γ in time $\mathrm{poly}(n, d, \log R, \log\gamma^{-1})$. Hence, this verifies the "efficiently computable" criterion for Theorem 4.1.

5.4.6 Efficient Finding of Low-Scoring Point

Verifying the "robust algorithm finds low-scoring point" criterion is also direct from Theorem C.6. We simply remove the constraint that $\mathcal{L}\mu'_i - \widetilde{\mu}_i \in [-\phi, \tau\cdot T + \phi + \tau\cdot T]$, and allow for the much broader $\mathcal{L}\mu'_i \in [-R, R]$. We can apply Theorem C.6 in the same way to find some linear operator $\mathcal{L}$ with score at most $\min_{\widetilde{\mu}} \mathcal{S}(\widetilde{\mu}, \mathcal{Y}) + 1$. Then, we can compute $\mathcal{L}[\mu']$ set $r=\phi$, and obtain that every point within ℓ_2 distance ϕ of $\mathcal{L}[\mu']$ has score at most $\min_{\widetilde{\mu}} \mathcal{S}(\widetilde{\mu}, \mathcal{Y}) + 1$.

5.5 Proof of Theorem 5.1

We apply Theorem 4.1, using the score function defined in Definition 5.6. Indeed, for $r=\phi=\alpha/\sqrt{d}$, we have verified all conditions, as long as $n \ge O((d+\log(1/\beta))/\alpha^2)$. Therefore, we have an ε-DP algorithm running in time $\mathrm{poly}(n, d, \log\frac{R}{\alpha}) = \mathrm{poly}(n,d)$ that finds a candidate mean $\widetilde{\mu}$ of score at most $2\eta n$, as long as

$$n \ge O\left(\max_{\eta':\eta\le\eta'\le 1} \frac{\log(V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y})) + \log(1/(\beta\cdot\eta'))}{\varepsilon\cdot\eta'}\right).$$

Using Lemmas 5.11 and 5.12, we have that for $\eta' \le \eta^*$ for some $\eta^* = \Omega(1)$, $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) = (\widetilde{O}(\eta')/\eta)^d \le (O(1/\eta))^d$. For $\eta' > \eta^*$, we have that $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) \le (O(R/\eta))^d$. So overall, it suffices for

$$n \ge O\left(\frac{d+\log(1/\beta)}{\alpha^2}\right) + O\left(\max_{\eta \le \eta' \le \eta^*} \frac{d\log(1/\eta)+\log(1/(\beta\cdot\eta))}{\varepsilon\cdot\eta'} + \max_{\eta^* \le \eta' \le 1} \frac{d\log(R/\eta)+\log(1/(\beta\cdot\eta))}{\varepsilon\cdot\eta'}\right)$$
$$= \widetilde{O}\left(\frac{d+\log(1/\beta)}{\alpha^2} + \frac{d+\log(1/\beta)}{\varepsilon\cdot\alpha} + \frac{d\log R}{\varepsilon}\right).$$

Hence, our algorithm, using this many samples, can find a point $\widetilde{\mu}$ of score at most $2\eta n$. Finally, by replacing η with 2η and applying Lemma 5.10, we have that any point $\widetilde{\mu}$ with score at most $2\eta n$ is within $O(\alpha)$ of μ. While we did not verify Lemma 5.10 for corrupted points, by our bound on sensitivity, we know that for any $\mathcal{Y}$ which is an η-corruption of $\mathcal{X}$, any point with score at most $2\eta n$ with respect to $\mathcal{Y}$ has score at most $3\eta n$ with respect to $\mathcal{X}$, and therefore is within $O(\alpha)$ of μ. This completes the proof.

5.6 The approx-DP setting

In this subsection, we prove Theorem 5.2. In this setting, the score function is identical, but we can afford fewer samples as we apply the algorithm of Theorem 4.2 instead of Theorem 4.1. The main additional thing we must check is that for *any* dataset $\mathcal{Y}$, if $\mathcal{S}(\Sigma, \mathcal{Y}) \le 0.7\eta^* n$ for some Σ, then the volume ratio $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y})$ is not too high.

Before proving our main result of this subsection, we must first establish the following "worst-case robustness" guarantee, which is important for ensuring privacy. We defer the proof to Appendix B.

Lemma 5.13. *Fix η^* to be a sufficiently small constant, and $T = \eta^* n$. Also, suppose $\phi \le \alpha/\sqrt{d}$. Then, for a dataset $\mathcal{Y}$ with every y_i bounded in ℓ_2 norm by $R \cdot d^{100}$, if there exist $\widetilde{\mu}_1, \widetilde{\mu}_2 \in \mathbb{R}^d$ that are both (α, τ, ϕ, T)-certifiable means with respect to $\mathcal{Y}$, then $\|\widetilde{\mu}_1 - \widetilde{\mu}_2\|_2 \le O(1)$.*

As a corollary of Lemma 5.13, we have the following result.

Corollary 5.14. *Suppose that $\mathcal{Y}$ is a dataset with every y_i bounded in ℓ_2 norm by $R \cdot d^{100}$ that has an $(\alpha, \tau, \phi, 0.7\eta^* n)$-certifiable mean, and let $\hat{\mu} = \mathcal{L}[\mu']$ where $\mathcal{L}$ is an $(\alpha, \tau, \phi, 0.7\eta^* n)$-certificate. Also, suppose $\phi \le \alpha/\sqrt{d}$. Then, the set of $(\alpha, \tau, \phi, 0.8\eta^* n)$-certifiable means contains all $\widetilde{\mu}$ such that $\|\widetilde{\mu} - \hat{\mu}\|_\infty \le \phi$, and any $(\alpha, \tau, \phi, \eta^* n)$-certifiable mean $\widetilde{\mu}$ must satisfy $\|\widetilde{\mu} - \hat{\mu}\| \le O(1)$.*

Proof. If $\mathcal{L}$ is an $(\alpha, \tau, \phi, 0.7\eta^* n)$-certificate, it is also an $(\alpha, \tau, \phi, 0.8\eta^* n)$-certificate. This means every $\widetilde{\mu}$ such that $\|\hat{\mu} - \widetilde{\mu}\|_\infty \le \phi$ is $(\alpha, \tau, 0.8\eta^* n)$-certifiable. To see why, note that for a $(\alpha, \tau, 0.8\eta^* n)$-certificate $\mathcal{L}$ of $\mathcal{Y}$, Constraint 3 (which is the only constraint that deals with $\widetilde{\mu}$, which we recall is not indeterminate) just requires that $\mathcal{L}[\mu'_i] - \widetilde{\mu}_i \in [-\phi - \tau \cdot T, \phi + \tau \cdot T]$. So, any such $\widetilde{\mu}$ is an $(\alpha, \tau, 0.8\eta^* n)$-certifiable covariance.

The second part is immediate by Lemma 5.13. □

Therefore, by setting $\phi := \alpha/\sqrt{d}$, the set of $(\alpha, \tau, \phi, \eta^* n)$-certifiable means has volume at most $O(1/\sqrt{d})^d$, since the volume of a unit sphere is $O(1/\sqrt{d})^d$. The set of $(\alpha, \tau, \phi, 0.8\eta^* n)$-certifiable means has volume at least $\phi^d \ge \Omega(1/\sqrt{d})^d$. So, the ratio $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}) \le O(1/\alpha)^d$.

We now prove Theorem 5.2, by applying Theorem 4.2. First, note that we may truncate the samples so that no $y_i \in \mathcal{Y}$ has norm more than $R \cdot d^{100}$. Since we are promised $\|\mu\| \le R$, the probability that any uncorrupted sample has this norm is at most $e^{-d^{100}}$. We will set η^* to be a sufficiently small constant (such as 0.01). We just showed, using Corollary 5.14, that for all $\mathcal{Y}$ such that $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) \le 0.7\eta^* n$, $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}) \le O(1/\alpha)^{d^2}$. So, as long as $n \ge O\left(\frac{\log(1/\delta)+d\log(1/\alpha)}{\varepsilon}\right)$, the algorithm of Theorem 4.2 is (ε, δ)-differentially private. In addition, we have already verified all of the conditions, so the algorithm is accurate as long as we additionally have $n \ge \widetilde{O}((d + \log(1/\beta))/\eta^2)$ and

$$n \ge O\left(\max_{\eta' : \eta \le \eta' \le \eta^*} \frac{\log(V_{\eta'}(\mathcal{Y})/V_{\eta}(\mathcal{Y})) + \log(1/(\beta \cdot \eta'))}{\varepsilon \cdot \eta'}\right).$$

By our volume bounds, this means it suffices for

$$\begin{aligned} n &\ge \widetilde{O}\left(\frac{d + \log(1/\beta)}{\alpha^2}\right) + O\left(\max_{\eta \le \eta' \le \eta^*} \frac{d\log(1/\eta) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right) + O\left(\frac{\log(1/\delta) + d\log(1/\alpha)}{\varepsilon}\right) \\ &= \widetilde{O}\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{d + \log(1/\beta)}{\varepsilon \cdot \alpha} + \frac{\log(1/\delta)}{\varepsilon}\right). \end{aligned}$$

This concludes the proof of Theorem 5.2.

6 Preconditioning the Gaussian

6.1 Main Theorems

Our goal is to obtain polynomial time algorithms for private covariance estimation of a unknown Gaussian, with optimal sample complexity. Before achieving this, an important step is preconditioning the Gaussian so that the samples come from a near-isotropic Gaussian. This requires approximately learning the covariance up to spectral distance, which we focus on in this section.

We prove both a pure-DP and approx-DP result in this section, showing that one can privately (and robustly) learn the covariance of a Gaussian up to spectral distance using roughly d^2 samples. In addition, in the approx-DP setting, our sample complexity has no dependence on the parameter K, which describes the ratio between a priori upper and lower bounds on the true covariance matrix, though the runtime depends on $\log K$.

Theorem 6.1 (Private Preconditioning of a Gaussian, Pure-DP)**.** *Let $\Sigma \in \mathbb{R}^{d\times d}$ be such that $K^{-1}I \preccurlyeq \Sigma \preccurlyeq K \cdot I$. Then, there exists an ε-differentially private algorithm that takes n i.i.d. samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ and with probability $1 - \beta$ outputs $\widetilde{\Sigma}$ such that $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \le \alpha$, for*

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

Here $\widetilde{O}$ is hiding factors. Moreover, this algorithm runs in time $\mathrm{poly}(n, d)$, and succeeds with the same accuracy even if $\eta = \tilde{\Omega}(\alpha)$ fraction of the points are adversarially corrupted.

Theorem 6.2 (Private Preconditioning of a Gaussian, Approx-DP)**.** *Let $\Sigma \in \mathbb{R}^{d\times d}$ be such that $K^{-1}I \preccurlyeq \Sigma \preccurlyeq K \cdot I$. Then, there exists an (ε, δ)-differentially private algorithm that takes n i.i.d. samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ and with probability $1-\beta$ outputs $\widetilde{\Sigma}$ such that $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \le \alpha$, where*

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

Here $\widetilde{O}$ is hiding factors. Moreover, this algorithm runs in time $\mathrm{poly}(n, d, \log K)$, and succeeds with the same accuracy even if $\eta = \tilde{\Omega}(\alpha)$ fraction of the points are adversarially corrupted.

6.2 Resilience of Moments

Similar to the mean estimation case, we will also require higher-order moment bounds, and stability conditions that imply the top roughly η fraction of samples in any "covariance" direction cannot be too large.

Lemma 6.3. *Let $\{x_i\} \sim \mathcal{N}(\mathbf{0}, I)$ and $n \ge \widetilde{O}((d^2 + \log^2(1/\beta))/\eta^2)$. Then, with probability $1-\beta$, the following all hold for all symmetric $P \in \mathbb{R}^{d\times d}$ with $\|P\|_F = 1$ simultaneously, for some $\alpha = \tilde{O}(\eta)$.*

1. $\left|\frac{1}{n}\sum_{i=1}^n \langle (x_i x_i^\top - I)/\sqrt{2}, P\rangle\right| \le \alpha$.
2. $\left|\frac{1}{n}\sum_{i=1}^n \langle (x_i x_i^\top - I)/\sqrt{2}, P\rangle^2 - 1\right| \le \alpha$.
3. *For any real values $a_1, \ldots, a_n \in [0,1]$ such that $\sum_{i=1}^n a_i \le \eta \cdot n$, $\left|\frac{1}{n}\sum_{i=1}^n a_i \langle (x_i x_i^\top - I)/\sqrt{2}, P\rangle\right| \le \alpha$ and $\left|\frac{1}{n}\sum_{i=1}^n a_i \langle (x_i x_i^\top - I)/\sqrt{2}, P\rangle^2\right| \le \alpha$.*
4. $\frac{1}{n}\sum_{i=1}^n \left|\langle (x_i x_i^\top - I)/\sqrt{2}, P\rangle\right| \le O(1)$.

To our knowledge, such a result is not known with this number of samples. The best-known result we know of can obtain the same bounds but requires $\widetilde{O}(d^2 \log^5(1/\beta)/\eta^2)$ samples [DKK+16], which means the number of samples required is $d^{2+\Omega(1)}$ if we want exponentially small failure probability. We prove Lemma 6.3 in Appendix D.

Remark. As in the mean estimation case, Lemma 6.3 will be the only conditions we will require about the samples we draw. (Or if $x_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then $\{\Sigma^{-1/2}x_i\}$ are resilient.)

6.3 Robust Algorithm

Suppose x_i's are samples from $\mathcal{N}(\mathbf{0}, \Sigma)$. Let y_i's be an arbitrarily η-corruption of x_i's. Consider the following pseudo-expectation program, where y_i's are the input points and the domain is the

degree-12 pseudo-expectations with $\{w_i\}, \{x_i\}$ as indeterminates.

$$\begin{aligned}
&\text{find } \tilde{\mathbf{E}} \\
&\text{such that } \tilde{\mathbf{E}} \text{ satisfies } w_i^2 = w_i, \\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } \sum w_i \geq (1-\eta)n, \\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } w_i x_i' = w_i y_i, \\
&\qquad \tilde{\mathbf{E}}(2 + \widetilde{O}(\eta)) \cdot (v^\top \Sigma' v)^2 - \frac{1}{n}\sum (\langle v, x'\rangle^2 - v^\top \Sigma' v)^2 \text{ has a degree 4-SoS proof of} \\
&\qquad\qquad \text{nonnegativity in } v \in \mathbb{R}^d, \text{ where } \Sigma' = \frac{1}{n}\sum_{i=1}^n (x_i')(x_i')^\top.
\end{aligned}$$

To explain the last condition further, note that $\tilde{\mathbf{E}}\left[(2 + \widetilde{O}(\eta)) \cdot (v^\top \Sigma' v)^2 - \frac{1}{n}\sum (\langle v, x'\rangle^2 - v^\top \Sigma' v)^2\right]$ is a degree 4 polynomial in $v \in \mathbb{R}^d$: the claim is that this polynomial has a degree-4 sum of squares certificate of being nonnegative.

It can be proven that if n is as in Lemma 6.3, with probability $1-\beta$ over the choice of x_i's, if we output $\tilde{\mathbf{E}}\Sigma'$, then $\|\Sigma^{-1/2}(\tilde{\mathbf{E}}\Sigma')\Sigma^{-1/2} - I\|_{op} = \widetilde{O}(\eta)$.

6.4 Score Function and its Properties

Our goal is to use Theorem 4.1, so we relax the pseudo-expectation from the robust algorithm to a linear operator that behaves as an approximate pseudoexpectation.

Definition 6.4 (Certifiable Covariance)**.** Let $\alpha, \tau, T \in \mathbb{R}^{\geq 0}$, $y_1, \ldots y_n \in \mathbb{R}^d$ and let $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}$ be PSD. We call the point $\widetilde{\Sigma}$ an (α, τ, T)-certifiable covariance for y_i's if and only if there exists a linear functional $\mathcal{L}$ over the set of polynomials in indeterminates $\{w_i\}, \{x'_{i,j}\}, \{M_{\{j,j'\},\{k,k'\}}\}$ of degree at most 12 such that

1. $\mathcal{L}1 = 1$
2. for every polynomial p, where $\|\mathcal{R}(p)\|_2 \leq 1$
   (a) $\mathcal{L}p^2 \geq -\tau \cdot T$,
   (b) $\forall i, \mathcal{L}(w_i^2 - w_i)p^2 \in [-\tau \cdot T, \tau \cdot T]$,
   (c) $\mathcal{L}(\sum w_i - n + T)p^2 \geq -5\tau \cdot T \cdot n$,
   (d) $\forall i, \mathcal{L}w_i(x_i' - y_i)p^2 \in [-\tau \cdot T, \tau \cdot T]$,
3. $\mathcal{L}[\frac{1}{n}\sum_i \left(\langle v, x_i'\rangle^2 - v^\top \Sigma' v\right)^2 + (v^{\otimes 2})^\top M^\top M v^{\otimes 2} - (2+\alpha)(v^\top \Sigma' v)^2]$, as a degree-4 polynomial in $v = (v_1, \ldots, v_d)$, has all coefficients between $[-\tau \cdot T, \tau \cdot T]$, where $\Sigma' := \frac{1}{n}\sum (x_i')(x_i')^\top$.
4. $\mathcal{L}[(1+\alpha)\Sigma' - \widetilde{\Sigma}] \succcurlyeq -\tau \cdot T \cdot I$, and $\mathcal{L}[\widetilde{\Sigma} - (1-\alpha)\Sigma'] \succcurlyeq -\tau \cdot T \cdot I$, where $\mathcal{L}$ applied to a matrix is applied entrywise,
5. $(\frac{1}{2K} - \tau \cdot T) \cdot I \preccurlyeq \mathcal{L}[\Sigma'] \preccurlyeq (2K + \tau \cdot T) \cdot I$.

We also require $\|\mathcal{R}(\mathcal{L})\|_2 \le R' + T \cdot \tau$ for some sufficiently large $R' = \mathrm{poly}(n, d, K)$. As in the mean estimation case, this requirement is only needed for computability purposes. We will also say that $\mathcal{L}$ is an (α, τ, T)-certificate for $\mathcal{Y}$.

Note that one may think of $\mathcal{L}$ as an approximate pseudo-expectation, and it is clear that $\mathcal{L}$ generalizes pseudo-expectations. In addition, for each constraint 2a) to 2d) we implicitly assume a bound on the degree of p so that $\mathcal{L}$ is applied to a polynomial of degree at most 12.

For our purposes, we will end up setting $\tau = 1/(K \cdot n \cdot d)^{O(1)}$, for a large enough $O(1)$.

Now we use this definition to define a score function.

Definition 6.5 (Score Function)**.** Let $\alpha, \tau, T \in \mathbb{R}^{\ge 0}$, $y_1, \dots, y_n \in \mathbb{R}^d$ (with $\mathcal{Y} := \{y_1, \dots, y_n\}$), and $\widetilde{\Sigma} \in \mathbb{R}^{d \times d}$ be a symmetric matrix. We define the score function $\mathcal{S} : \mathbb{R}^{d \times d} \to \mathbb{R}$ as

$$\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}; \alpha, \tau) = \min_{T} \text{ such that } \widetilde{\Sigma} \text{ is a } (\alpha, \tau, T) \text{ certifiable covariance for } y_i\text{'s}\,.$$

In the rest of this subsection we will prove the following properties for this score function. This will allow us to use Theorem 4.1.

1. Score has sensitivity 1.
2. Score is quasi-convex as a function of $\widetilde{\Sigma}$.
3. All points $\widetilde{\Sigma}$ that have score at most $\eta \cdot n$ have spectral distance at most $\tilde{O}(\eta)$ away from Σ. (Robustness for volume/accuracy purposes).
4. The volume of points that have score at most $\eta \cdot n$ is sufficiently large, and the volume of points with score at most $\eta' \cdot n$ for $\eta' > \eta$ is not too large.
5. Score is efficiently computable.
6. We can approximately minimize score efficiently.

6.4.1 Existence of Low-Scoring Σ'

Before verifying the desired conditions of our score functions, we prove that for data points drawn from $\mathcal{N}(\mathbf{0}, \Sigma)$, with high probability some Σ' which is close to Σ has low score. This will be important both for sensitivity and for volume bounds. While such results are already known in the literature [KS17] for *certifiable* fourth moment bounds, which we will need to verify Condition 3 of Definition 6.4, the previous result requires $n = \tilde{O}(d^2 \log^2(1/\beta)/\alpha^2)$, as opposed to our goal of $n = \tilde{O}((d^2 + \log^2(1/\beta))/\alpha^2)$. As a result, we reprove some known results to establish a low-scoring Σ', but with better failure probability bounds.

First, we note the following basic proposition which is immediate by Cauchy-Schwarz.

Proposition 6.6. *For any two matrices $A, B \in \mathbb{R}^{d \times d}$, $|\operatorname{Tr}(AB)| \le \|A\|_F \cdot \|B\|_F$.*

The following proposition is also well-known.

Proposition 6.7. *For any two matrices $A, B \in \mathbb{R}^{d \times d}$, $\|AB\|_F \le \|A\|_{op} \cdot \|B\|_F$, $\|B\|_{op} \cdot \|A\|_F \le \|A\|_F \cdot \|B\|_F$.*

Proposition 6.8. *Let $M \in \mathbb{R}^{d\times d}$ be a real symmetric matrix, and let $J \in \mathbb{R}^{d\times d}$ be any real-valued matrix (possibly not symmetric) such that $\|JJ^\top - I\|_{op} \le \alpha$. Then, $\|J^\top MJ\|_F^2 = (1 \pm 3\alpha)\cdot\|M\|_F^2$.*

Proof. Start by writing

$$\|J^\top MJ\|_F^2 = \operatorname{Tr}((J^\top MJ)(J^\top MJ)^\top) = \operatorname{Tr}(J^\top MJJ^\top MJ) = \operatorname{Tr}(MJJ^\top MJJ^\top).$$

Now, write $JJ^\top = I + H$ for some symmetric matrix H such that $\|H\|_{op} \le \alpha$. Therefore,

$$\begin{aligned}\operatorname{Tr}(MJJ^\top MJJ^\top) &= \operatorname{Tr}(M(I+H)M(I+H))\\ &= \operatorname{Tr}(M^2) + \operatorname{Tr}(MHM) + \operatorname{Tr}(MMH) + \operatorname{Tr}(MHMH)\\ &= \operatorname{Tr}(M^2) + 2\operatorname{Tr}(MMH) + \operatorname{Tr}((MH)^2).\end{aligned}$$

Now, by Propositions 6.6 and 6.7, we have that $|\operatorname{Tr}((MH)^2)| \le \|MH\|_F^2 \le \|M\|_F^2\cdot\|H\|_{op}^2$. In addition, $|\operatorname{Tr}(MMH)| \le \|M\|_F\cdot\|MH\|_F \le \|M\|_F^2\cdot\|H\|_{op}$. Since $\|H\|_{op}\le\alpha$, this implies that $\|J^\top MJ\|_F^2 = \operatorname{Tr}(M^2) \pm 3\alpha\cdot\|M\|_F^2 = (1\pm 3\alpha)\cdot\|M\|_F^2$. □

Lemma 6.9. *Suppose $n \ge \widetilde{O}\left(\frac{d^2+\log^2(1/\beta)}{\eta^2}\right)$, and let $\alpha = \widetilde{O}(\eta)$. Let $\mathcal{X} = \{x_1,\dots,x_n\} \sim \mathcal{N}(\mathbf{0},\Sigma)$, and let $\mathcal{Y} = \{y_1,\dots,y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for any $\tau \ge 0$ and $T = \eta\cdot n$, with probability at least $1-\beta$, there exists a Σ' such that every $\widetilde{\Sigma}$ of spectral distance at most α from Σ' (i.e., $\|(\Sigma')^{-1/2}\widetilde{\Sigma}(\Sigma')^{-1/2} - I\|_{op} \le \alpha$) is an (α,τ,T)-certifiable covariance for $\mathcal{Y}$.*

Proof. As in the case for mean estimation, we use the fact that our linear operators generalize pseudo-expectations, which in turn generalize expectations over a single point. Again, we set $w_i = 1$ if $y_i = x_i$ and 0 otherwise, and $x_i' = x_i$ for all i. For $T = \eta n$, it is clear that Constraints 1 and 2a-2d are all satisfied in Definition 6.4.

To verify Constraint 3 in Definition 6.4, first note that $\Sigma^{-1/2}x_1,\dots,\Sigma^{-1/2}x_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0},I)$. Now, by part 2 of Lemma 6.3, where we replace α with $\alpha/4$, we have $\frac{1}{n}\sum_{i=1}^n\langle\Sigma^{-1/2}x_ix_i^\top\Sigma^{-1/2} - I, P\rangle^2 \le (2+\alpha/2)\cdot\|P\|_F^2$ with probability at least $1-\beta$, for all $d\times d$ symmetric matrices P. We can write

$$\begin{aligned}\langle\Sigma^{-1/2}x_ix_i^\top\Sigma^{-1/2} - I, P\rangle &= \operatorname{Tr}[(\Sigma^{-1/2}x_ix_i^\top\Sigma^{-1/2} - I)\cdot P]\\ &= \operatorname{Tr}[x_ix_i^\top\cdot\Sigma^{-1/2}P\Sigma^{-1/2} - P]\\ &= \operatorname{Tr}[(x_ix_i^\top - \Sigma)\cdot(\Sigma^{-1/2}P\Sigma^{-1/2})]\\ &= \langle x_ix_i^\top - \Sigma, \Sigma^{-1/2}P\Sigma^{-1/2}\rangle.\end{aligned}$$

So, by replacing P with $\Sigma^{1/2}P\Sigma^{1/2}$, we have that for all symmetric matrices P,

$$\frac{1}{n}\sum_{i=1}^n\langle x_ix_i^\top - \Sigma, P\rangle^2 \le (2+\alpha/2)\cdot\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2.$$

Now, note that the empirical covariance $\Sigma' = \frac{1}{n}\sum_{i=1}^n x_ix_i^\top$ of the uncorrupted samples satisfies $\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F \le \alpha/100$ with probability at least $1-\beta$, by Condition 1 of Lemma 6.3 (replacing α with $\alpha/200$). Therefore, by setting $J = \Sigma^{-1/2}(\Sigma')^{1/2}$, we have $\|JJ^\top - I\|_F \le \alpha/100$, which means

$\|(\Sigma')^{1/2}P(\Sigma')^{1/2}\|_F^2 = \|J^\top\Sigma^{1/2}P\Sigma^{1/2}J\|_F^2 \geq (1-3\alpha/100)\cdot\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2$ by Proposition 6.8. In addition, since Σ' is the empirical average of $x_ix_i^\top$, this means for any symmetric P,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\langle x_ix_i^\top-\Sigma',P\rangle^2 &\leq \frac{1}{n}\sum_{i=1}^n\langle x_ix_i^\top-\Sigma,P\rangle^2\\ &\leq \frac{2+\alpha/2}{(1-3\alpha/100)}\cdot\|(\Sigma')^{1/2}P(\Sigma')^{1/2}\|_F^2\\ &\leq (2+\alpha)\cdot\|(\Sigma')^{1/2}P(\Sigma')^{1/2}\|_F^2.\end{aligned}$$

For fixed x_i (and thus fixed Σ'), note that for a symmetric matrix P, $\langle x_ix_i^\top-\Sigma',P\rangle$ is a linear functional mapping P to $\mathbb{R}$, and $(\Sigma')^{1/2}P(\Sigma')^{1/2}$ is a linear map sending symmetric matrices P to symmetric matrices. For a symmetric matrix $P\in\mathbb{R}^{d\times d}$, let $P^\flat\in\mathbb{R}^{d^2}$ be the vector $\{P_{ij}\}_{i,j\leq d}$, and let $(P^\flat)'\in\mathbb{R}^{d(d+1)/2}$ be the vector $\{P_{ij}\}_{i\leq j}$. So, if we consider the embedding $P\to(P^\flat)'$, there exist vectors $v_1,\ldots,v_n\in\mathbb{R}^{d(d+1)/2}$ (corresponding to taking inner product with $x_ix_i^\top-\Sigma'$) and a $\frac{d(d+1)}{2}\times\frac{d(d+1)}{2}$ matrix J (corresponding to left- and right- multiplication by $(\Sigma')^{-1/2}$), such that $\frac{1}{n}\sum_{i=1}^n\langle v_i,(P^\flat)'\rangle^2\leq(2+\alpha)\cdot\|J\cdot(P^\flat)'\|_2^2$. Therefore, there is some other matrix J' such that $\frac{1}{n}\sum_{i=1}^n\langle v_i,(P^\flat)'\rangle^2+\|J'\cdot(P^\flat)'\|_2^2=(2+\alpha)\cdot\|J\cdot(P^\flat)'\|_2^2$, meaning that

$$\frac{1}{2}\sum_{i=1}^n\langle x_ix_i^\top-\Sigma',P\rangle^2+\|J'\cdot(P^\flat)'\|_2^2=(2+\alpha)\cdot\|(\Sigma')^{1/2}P(\Sigma')^{1/2}\|_F^2.$$

We can convert $J'\in\mathbb{R}^{d(d+1)/2\times d(d+1)/2}$ into a matrix $M\in\mathbb{R}^{d(d+1)/2\times d^2}$, by replacing any column in J' corresponding to entry (i,j) for $i<j$ with two copies for (i,j) and (j,i), each divided by 2. Importantly, $J'\cdot(P^\flat)'=M\cdot P^\flat$. Therefore, for any $P=vv^\top$, since $P^\flat=\{v_iv_j\}_{i,j\leq n}=v^{\otimes 2}$ and $(P^\flat)'=\{v_iv_j\}_{i\leq j}$, there exists a matrix $M\in\mathbb{R}^{d(d+1)/2\times d^2}$ such that

$$\frac{1}{n}\sum_{i=1}^n\left(\langle v,x_i\rangle^2-v^\top\Sigma'v\right)^2+(v^{\otimes 2})^\top M^\top Mv^{\otimes 2}=(2+\alpha)(v^\top\Sigma'v)^2.$$

While M is lacking in rows (it should have d^2 rows and columns), we can simply add additional 0 rows.

Now, we have such a Σ' so that the first 3 constraints are satisfied, and moreover, Σ' has spectral distance at most $\alpha/100$ from Σ, which means Constraint 5 is also satisfied since $\frac{1}{K}\cdot I\preccurlyeq\Sigma\preccurlyeq K\cdot I$. We can choose any $\widetilde{\Sigma}$ between $(1-\alpha)\Sigma'$ and $(1+\alpha)\Sigma'$, since then $(1+\alpha)\Sigma'-\widetilde{\Sigma}$ and $\widetilde{\Sigma}-(1-\alpha)\Sigma'$ are both PSD, so Constraint 4 is satisfied.

Finally, we remark that every w_i, $x_{i,j}$, and $M_{\{j,k\},\{j',k'\}}$ is bounded by $\mathrm{poly}(n,d,K)$. Therefore, the corresponding linear operator $\mathcal{L}$ satisfies $\|\mathcal{R}(\mathcal{L})\|_2\leq(Knd)^{O(1)}$. □

6.4.2 Sensitivity

The proof of sensitivity is similar to the mean estimation case. We again have an upper bound of n on the value of the score function. This time we can essentially use Lemma 6.9.

Lemma 6.10 (score function upper bound)**.** *The value of the score function $\mathcal{S}$ defined in Definition 6.5 is less than or equal to n.*

Proof. We use the fact that our linear operators generalize pseudo-expectations, which generalize expectations over a single point mass. In Lemma 6.9, we showed that for $\mathcal{X} = \{x_1, \dots, x_n\} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, we can set $x_i' = x_i$, and choose $\Sigma' = \frac{1}{n}\sum x_i x_i^\top$ and M to satisfy all of the constraints (where $\widetilde{\Sigma} = \Sigma$), with probability at least $1 - \beta$. So, there exists a set $\mathcal{X}$ that satisfies the constraints, which means for a general set of data points $\mathcal{Y} = \{y_1, \dots, y_n\}$, the score is at most n, since we can set $w_i = 0$ and $x_i' = x_i$ for all i. For $T = n$, it is clear that all constraints are satisfied. □

Lemma 6.11 (sensitivity)**.** *The score function $\mathcal{S}$ as defined in Definition 6.5 has sensitivity 1 with respect to its first input.*

Proof. Suppose that $\mathcal{Y}$, $\mathcal{Y}'$ are two neighboring datasets, and $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}$. Moreover, assume $\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) = T$. If we show that $\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}') \le \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) = T + 1$, by symmetry we are done.

Without loss of generality assume $\mathcal{Y}$ and $\mathcal{Y}'$ differ on index j. In order to construct $\mathcal{L}'$, for any monomial p, let

$$\mathcal{L}'p = \begin{cases} 0 & \text{if } p \text{ has a } w_j \text{ factor,} \\ \mathcal{L}p & \text{otherwise} \end{cases}.$$

To verify the constraints, Constraints 1 and 2a-2d are identical to in the mean estimation case (where checking Constraint 2c applies Lemma 6.10). Also, $\|\mathcal{R}(\mathcal{L}')\|_2 \le \|\mathcal{R}(\mathcal{L})\|_2$ clearly holds. So, we just need to verify Constraints 3, 4, and 5 in Definition 6.4.

However, note that these three constraints do not involve w_j at all, so in fact their evaluation is the same regardless of $\mathcal{L}$ or $\mathcal{L}'$. The only difference is we are allowing the values $\mathcal{L}[\cdot]$ to have a greater range, which makes it easier. □

6.4.3 Quasi-convexity

Lemma 6.12 (quasi-convexity)**.** *The score function $\mathcal{S}$ as defined in Definition 6.5 is quasi-convex in its second input, $\widetilde{\Sigma}$.*

Proof. Suppose $\mathcal{S}(\widetilde{\Sigma}_1, \mathcal{Y}) = T_1, \mathcal{S}(\widetilde{\Sigma}_2, \mathcal{Y}) = T_2$, and suppose there exists $\mathcal{L}_1$ and $\mathcal{L}_2$ that satisfy the constraints in Definition 6.4 with $\widetilde{\Sigma}_1, T_1$,and $\widetilde{\Sigma}_2, T_2$ respectively. If we can construct a functional $\mathcal{L}_3$ such that the constraints in Definition 6.4, are satisfied with $\widetilde{\Sigma}_3 = \lambda\widetilde{\Sigma}_1 + (1-\lambda)\widetilde{\Sigma}_2$, and $T_3 = \max\{T_1, T_2\}$, we are done. Let $\mathcal{L}_3 = \lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2$. As in the mean estimation case, all of the constraints in Definition 6.4 will be satisfied trivially except for Constraints 2c and 5, and Constraint 2c is the same as in the mean estimation case. So, the same verification implies that this constraint is also satisfied. Constraint 5 is also straightforward, since if $(\frac{1}{2K} - \tau\cdot T_1)\cdot I \preccurlyeq \mathcal{L}_1[\Sigma'] \preccurlyeq (2K + \tau\cdot T_1)\cdot I$ and $(\frac{1}{2K} - \tau\cdot T_2)\cdot I \preccurlyeq \mathcal{L}_2[\Sigma'] \preccurlyeq (2K + \tau\cdot T_2)\cdot I$, then $(\frac{1}{2K} - \tau\cdot\max\{T_1, T_2\})\cdot I \preccurlyeq \lambda\cdot\mathcal{L}_1[\Sigma'] + (1-\lambda)\cdot\mathcal{L}_2[\Sigma'] \preccurlyeq (2K + \tau\cdot\max\{T_1, T_2\})\cdot I$. □

6.4.4 Accuracy

We show that any point $\widetilde{\Sigma}$ of low score with respect to i.i.d. samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ must be close to Σ in spectral distance, i.e., $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \le O(\alpha)$.

Lemma 6.13. *Let $\alpha = \widetilde{O}(\eta)$ and suppose α, η are bounded by a sufficiently small constant. Let $n \geq \widetilde{O}\left(\frac{d^2+\log^2(1/\beta)}{\alpha^2}\right)$, and $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, for $K^{-1}I \preccurlyeq \Sigma \preccurlyeq K \cdot I$. Also, suppose $\tau \ll (ndK/\varepsilon)^{-O(1)}$.*

Then, for any $\alpha^* \leq \alpha$, with probability at least $1-\beta$, every symmetric matrix $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}$ that is (α^*, τ, T)-certifiable for $\mathcal{X}$ with $T = \eta n$ must satisfy $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \leq O(\alpha)$.

As in the mean estimation case, the proof follows the same approach as [KMZ22], so we defer this to Appendix B.

6.4.5 Volume of Good Points

Lemma 6.14. *Let $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, and let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for every integer $T \in [\eta \cdot n, \eta^* \cdot n]$ for some fixed constant $\eta^* < 1$, with probability at least $1-\beta$, every (α, τ, T)-certifiable covariance with respect to $\mathcal{Y}$ has spectral distance at most $\widetilde{O}(T/n)$ from Σ.*

Proof. Since the score function has sensitivity at most 1 (Lemma 6.11), this means that any (α, τ, T)-certifiable mean with respect to $\mathcal{Y}$ is an $(\alpha, \tau, T+\eta n)$-certifiable mean with respect to $\mathcal{X}$.

Now, define $\eta' := \frac{T+\eta n}{n} = O(\frac{T}{n})$. In this case, by setting $\alpha' = \widetilde{O}(\eta')$ and since $\alpha = \widetilde{O}(\eta) \leq \alpha'$, we have that by Lemma 6.13 that any $(\alpha, \tau, T+\eta n)$-certifiable covariance $\widetilde{\Sigma}$ must satisfy $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \leq O(\alpha') \leq \widetilde{O}(T/n)$. □

So, for any $\eta' \in [\eta, \eta^*]$, the any $\widetilde{\Sigma}$ with score at most $\eta' \cdot n$ must be of the form $(1-\alpha') \cdot \Sigma + 2\alpha' \cdot R$ where $0 \preccurlyeq R \preccurlyeq \Sigma$ and $\alpha' = \widetilde{O}(\eta')$. So, if we define V_Σ to the set of PSD matrices spectrally bounded by Σ (where we think of symmetric matrices as vectors in $\mathbb{R}^{d(d+1)/2}$), the set of $(\alpha, \tau, \eta' n)$-certifiable covariance matrices has volume at most $e^{O(d^2)} \cdot V_\Sigma$, meaning $V_{\eta'} \leq O(1)^{d^2} \cdot V_\Sigma$ for $\eta' \in [\eta, \eta^*]$. In addition, by Constraint 5 of Definition 6.4, we know that $\mathcal{L}[\Sigma']$ is always spectrally bounded between $\frac{1}{4K} \cdot I$ and $4K \cdot I$, and so $\widetilde{\Sigma}$ is spectrally bounded between 0 and $8K \cdot I \preccurlyeq 8K^2 \cdot \Sigma$. Thus, for any $\eta' \in [\eta^*, 1]$, $V_{\eta'} \leq O(K^2)^{d^2} \cdot V_\Sigma = e^{O(d^2 \log K)} \cdot V_\Sigma$.

Finally, by Lemma 6.9, with probability at least $1-\beta$ every $\widetilde{\Sigma}$ within spectral distance α of $\frac{1}{n}\sum x_i x_i^\top$ (where $\{x_i\}$ are the uncorrupted points) have score at most $\eta \cdot n$. Since $n \geq \widetilde{O}((d^2 + \log^2(1/\beta))/\eta^2)$, $\frac{1}{n}\sum x_i x_i^\top$ has spectral distance at most $\alpha/10$ away from Σ, which means every $\widetilde{\Sigma}$ within spectral distance $\alpha/10$ of Σ has score at most $\eta \cdot n$. So, $V_\eta \geq V_\Sigma \cdot (\alpha/10)^{d^2}$.

6.4.6 Efficient Computability

As in the mean estimation case, we apply Theorem C.6 in Appendix C. This time, there are constraints where we wish to spectrally bound $\mathcal{L}$ applied to a matrix. However, this constraint is also captured by Theorem C.6. So, we have efficient computability.

6.4.7 Efficient Finding of Low-Scoring Point

To verify that the "robust algorithm finds low-scoring point", we simply remove the constraint that $\mathcal{L}[(1+\alpha)\Sigma' - \widetilde{\Sigma}] \succcurlyeq -\tau \cdot T \cdot I$, and $\mathcal{L}[\widetilde{\Sigma} - (1-\alpha)\Sigma'] \succcurlyeq -\tau \cdot T \cdot I$. We can apply Theorem C.6 in the same way to find some linear operator $\mathcal{L}$ with score at most $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) + 1$. Then, we can

compute $\mathcal{L}[\Sigma']$ and set $r \le \tau$, and obtain that every matrix $\widetilde{\Sigma}$ such that $\|\widetilde{\Sigma} - \mathcal{L}[\Sigma']\|_F \le r$ has score at most $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) + 1$.

6.5 Proof of Theorem 6.1

We apply Theorem 4.1, using the score function defined in Definition 6.5 and thinking of the candidate parameters Σ as lying in $\mathbb{R}^{d(d+1)/2}$. Indeed, for $r = \alpha/K^{O(1)}$ and $R = K^{O(1)}$, we have verified all conditions, as long as $n \ge \widetilde{O}((d^2 + \log^2(1/\beta))/\eta^2)$. Therefore, we have an ε-DP algorithm running in time $\mathrm{poly}(n, d, \log \frac{K}{\alpha})$ that finds a candidate covariance $\widetilde{\Sigma}$ of score at most $2\eta n$, as long as

$$n \ge O\left(\max_{\eta' : \eta \le \eta' \le 1} \frac{\log(V_{\eta'}(\mathcal{Y})/V_{\eta}(\mathcal{Y})) + \log(1/(\beta \cdot \eta'))}{\varepsilon \cdot \eta'}\right).$$

By our volume bounds, this means it suffices for

$$\begin{aligned} n &\ge O\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2}\right) + O\left(\max_{\eta \le \eta' \le \eta^*} \frac{d^2 \log(1/\eta) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'} + \max_{\eta^* \le \eta' \le 1} \frac{d^2 \log(K/\eta) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right) \\ &= \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2} + \frac{d^2 + \log(1/\beta)}{\varepsilon \cdot \alpha} + \frac{d^2 \log K}{\varepsilon}\right). \end{aligned}$$

Hence, our algorithm, using this many samples, can find a point $\widetilde{\Sigma}$ of score at most $2\eta n$ with respect to $\mathcal{Y}$, which means it has score at most $3\eta n$ with respect to the uncorrupted samples $\mathcal{X}$. Finally, by replacing η with 3η and applying Lemma 6.13, we have that any point $\widetilde{\Sigma}$ with score at most $3\eta n$ with respect to $\mathcal{X}$ is within $O(\alpha)$ spectral distance of Σ. This completes the proof.

6.6 The approx-DP setting

In this subsection, we prove Theorem 6.2. In this setting, the score function is identical, but we can afford fewer samples as we apply the algorithm of Theorem 4.2 instead of Theorem 4.1. The main additional thing we must check is that for *any* dataset $\mathcal{Y}$, if $\mathcal{S}(\Sigma, \mathcal{Y}) \le 0.7\eta^* n$ for some Σ, then the volume ratio $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y})$ is not too high.

Before proving our main result of this subsection, we must first establish the following lemma, which is important for ensuring privacy. We defer the proof to Appendix B.

Lemma 6.15. *Fix η^* to be a sufficiently small constant, and $T = \eta^* n$. Then, for a dataset $\mathcal{Y}$ with every y_i bounded in ℓ_2 norm by $K \cdot d^{100}$, if there exist linear operators $\mathcal{L}_1, \mathcal{L}_2$ that are both (α, τ, T)-certificates for $\mathcal{Y}$, then $\widetilde{\Sigma}_1 \preccurlyeq O(1) \cdot \widetilde{\Sigma}_2$ and $\widetilde{\Sigma}_2 \preccurlyeq O(1) \cdot \widetilde{\Sigma}_1$.*

As a corollary of Lemma 6.15, we have the following result.

Corollary 6.16. *Suppose that $\mathcal{Y}$ is a dataset with every y_i bounded in ℓ_2 norm by $K \cdot d^{100}$ that has an $(\alpha, \tau, 0.7\eta^* n)$-certifiable covariance, and let $\hat{\Sigma} = \mathcal{L}[\Sigma']$ where $\mathcal{L}$ is an $(\alpha, \tau, 0.7\eta^* n)$-certificate. Then, the set of $(\alpha, \tau, 0.8\eta^* n)$-certifiable covariance matrices $\widetilde{\Sigma}$ contains all matrices spectrally bounded between $(1-\alpha)\hat{\Sigma}$ and $(1+\alpha)\hat{\Sigma}$, and the set of $(\alpha, \tau, \eta^* n)$-certifiable covariance matrices is spectrally bounded between $\frac{1}{C} \cdot \hat{\Sigma}$ and $C \cdot \hat{\Sigma}$ for some constant $C = O(1)$.*

Proof. If $\mathcal{L}$ is an $(\alpha, \tau, 0.7\eta^* n)$-certificate, it is also an $(\alpha, \tau, 0.8\eta^* n)$-certificate. This means every $\widetilde{\Sigma}$ such that $(1-\alpha)\hat{\Sigma} \preccurlyeq \widetilde{\Sigma} \preccurlyeq (1+\alpha)\hat{\Sigma}$ is $(\alpha, \tau, 0.8\eta^* n)$-certifiable. To see why, note that for a $(\alpha, \tau, 0.8\eta^* n)$-certificate $\mathcal{L}$ of $\mathcal{Y}$, Constraint 4 (which is the only constraint that deals with $\widetilde{\Sigma}$, which we recall is not indeterminate) just requires that $\mathcal{L}[(1+\alpha)\Sigma' - \widetilde{\Sigma}] \geqslant -\tau \cdot T \cdot I$ and $\mathcal{L}[\widetilde{\Sigma} - (1-\alpha)\Sigma'] \geqslant -\tau \cdot T \cdot I$. So, any $\widetilde{\Sigma}$ spectrally bounded between $(1-\alpha)\hat{\Sigma}$ and $(1+\alpha)\hat{\Sigma}$ is an $(\alpha, \tau, 0.8\eta^* n)$-certifiable covariance.

The second part is immediate by Lemma 6.15. □

Therefore, if we let $V_{\hat{\Sigma}}$ represent the volume of PSD matrices spectrally bounded above by $\hat{\Sigma}$ (where we think of symmetric matrices as vectors in $\mathbb{R}^{d(d+1)/2}$), the set of $(\alpha, \tau, \eta^* n)$-certifiable covariance matrices has volume at most $O(1)^{d^2} \cdot V_{\hat{\Sigma}}$ and the set of $(\alpha, \tau, 0.8\eta^* n)$-certifiable covariance matrices has volume at least $\alpha^{d^2} \cdot V_{\hat{\Sigma}}$. So, the ratio $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}) \leq O(1/\alpha)^{d^2}$.

We now prove Theorem 6.2, by applying Theorem 4.2. First, note that we may truncate the samples so that no $y_i \in \mathcal{Y}$ has norm more than $K \cdot d^{100}$. Since we are promised $\|\Sigma\|_{op} \leq K$, the probability that any uncorrupted sample has this norm is at most $e^{-d^{100}}$. We will set η^* to be a sufficiently small constant (such as 0.01). We just showed, using Corollary 6.16, that for all $\mathcal{Y}$ such that $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) \leq 0.7\eta^* n$, $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}) \leq O(1/\alpha)^{d^2}$. So, as long as $n \geq O\left(\frac{\log(1/\delta) + d^2 \log(1/\alpha)}{\varepsilon}\right)$, the algorithm of Theorem 4.2 is (ε, δ)-differentially private. In addition, we have already verified all of the conditions, so the algorithm is accurate as long as $n \geq \widetilde{O}((d^2 + \log^2(1/\beta))/\eta^2)$ and

$$n \geq O\left(\max_{\eta' : \eta \leq \eta' \leq \eta^*} \frac{\log(V_{\eta'}(\mathcal{Y})/V_{\eta}(\mathcal{Y})) + \log(1/(\beta \cdot \eta'))}{\varepsilon \cdot \eta'}\right).$$

By our volume bounds, this means it suffices for

$$\begin{aligned} n &\geq \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2}\right) + O\left(\frac{\log(1/\delta) + d^2 \log(1/\alpha)}{\varepsilon}\right) + O\left(\max_{\eta \leq \eta' \leq \eta^*} \frac{d^2 \log(1/\eta) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right) \\ &= \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2} + \frac{d^2 + \log(1/\beta)}{\varepsilon \cdot \alpha} + \frac{\log(1/\delta)}{\varepsilon}\right). \end{aligned}$$

This concludes the proof of Theorem 6.2.

7 Learning a Gaussian in Total Variation Distance

The main result we prove in this section is is to privately learn the covariance Σ of a Gaussian up to low Frobenius norm error, if we are promised all eigenvalues of Σ are between $(1-\alpha)$ and $(1+\alpha)$.

Theorem 7.1 (Privately Learning a Preconditioned Gaussian)**.** *Let $\Sigma \in \mathbb{R}^{d \times d}$ where $(1-\alpha) \cdot I \preccurlyeq \Sigma \preccurlyeq (1+\alpha) \cdot I$. There exists an ε-differentially private algorithm that takes n i.i.d. samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ and with probability $1-\beta$ outputs $\widetilde{\Sigma}$ such that $\|\widetilde{\Sigma} - \Sigma\|_F \leq O(\alpha)$, where*

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

Moreover, this algorithm runs in time $\mathrm{poly}(n, d)$, and succeeds with the same accuracy even if $\eta = \tilde{\Omega}(\alpha)$ fraction of the points are adversarially corrupted.

By combining Theorem 7.1 with Theorem 5.1 and Theorem 6.1 (or Theorem 5.2 and Theorem 6.2), we will be able to prove our main results on privately learning Gaussians up to low total variation distance, namely Theorems 1.3 and 1.4. We prove these theorems in Section 7.4.

7.1 Robust Algorithm

Suppose $\{x_i\}$ is a set of samples from $\mathcal{N}(\mathbf{0}, \Sigma)$, where $(1-\alpha)\cdot I \preccurlyeq \Sigma \preccurlyeq (1+\alpha)\cdot I$. Let $\{y_i\}$ be an arbitrary η-corruption of $\{x_i\}$. Consider the following pseudo-expectation program, where $\{y_i\}$ are the input points and the domain is the degree-12 pseudo-expectations with $\{w_i\}, \{x_i\}, \{M_{\{j,j'\},\{k,k'\}}\}$ as indeterminates.

$$
\begin{aligned}
&\text{find } \tilde{\mathbf{E}} \\
&\text{such that } \tilde{\mathbf{E}} \text{ satisfies } w_i^2 = w_i, \\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } \sum w_i \geq (1-\eta)n, \\
&\qquad \tilde{\mathbf{E}} \text{ satisfies } w_i x_i' = w_i y_i, \\
&\qquad \tilde{\mathbf{E}} \left[\frac{1}{n} \sum_{i=1}^{n} (x_i' \otimes x_i' - S')(x_i' \otimes x_i' - S')^{\mathsf{T}} + MM^{\mathsf{T}} \right] = (2 + \widetilde{O}(\eta))I, \text{where } S' = \frac{1}{n} \sum_i x_i' \otimes x_i'.
\end{aligned}
$$

We use Σ' to represent $\frac{1}{n}\sum (x_i')(x_i')^{\mathsf{T}}$: note that S' is the flattening of Σ'. It can be proven that if n is as in Lemma 6.3, with probability $1-\beta$ over the choice of x_i's, if we output $\tilde{\mathbf{E}}\Sigma'$, then $\|\tilde{\mathbf{E}}\Sigma' - \Sigma\|_F = \widetilde{O}(\eta)$.

7.2 Score Function and its Properties

Again, we need design a suitable score function based on the robust algorithm, this time for learning covariance in up to low Frobenius norm error.

Definition 7.2 (Certifiable Covariance)**.** Let $\alpha, \tau, \phi, T \in \mathbb{R}^{\geq 0}$, $y_1, \ldots y_n \in \mathbb{R}^d$ (with $\mathcal{Y} := \{y_1, \ldots, y_n\}$), and $\widetilde{\Sigma} \in \mathbb{R}^d$. We call the point $\widetilde{\Sigma}$ an (α, τ, ϕ, T)-*certifiable covariance* for $\mathcal{Y}$ if and only if there exists a linear functional $\mathcal{L}$ over the set of polynomials in indeterminates $\{w_i\}, \{x'_{i,j}\}, \{M_{j,k}\}$ of degree at most 6 such that

1. $\mathcal{L}1 = 1$
2. for every polynomial p, where $\|\mathcal{R}(p)\|_2 \leq 1$:
   (a) $\mathcal{L}p^2 \geq -\tau \cdot T$,
   (b) $\forall i, \mathcal{L}(w_i^2 - w_i)p^2 \in [-\tau \cdot T, \tau \cdot T]$,
   (c) $\mathcal{L}(\sum w_i - n + T)p^2 \geq -5\tau \cdot T \cdot n$,
   (d) $\forall i, j, \mathcal{L}w_i(x'_{i,j} - y_{i,j})p^2 \in [-\tau \cdot T, \tau \cdot T]$,
3. $\forall j, k \ : \ \mathcal{L}\left(\frac{1}{n}\sum_i \left[((x_i')^{\otimes 2} - S')((x_i')^{\otimes 2} - S')^{\mathsf{T}} + MM^{\mathsf{T}} - (2+\alpha)I_{d^2}\right]_{\{j,j'\},\{k,k'\}}\right) \in [-\tau \cdot T, \tau \cdot T]$, where $x_i' = \{x'_{i,j}\}_{1 \leq j \leq d}$, and $S' = \mathbf{E}_i \, x_i'^{\otimes 2}$.

4. $\forall j,k \in [d], \mathcal{L}\Sigma'_{j,k} - \widetilde{\Sigma}_{j,k} \in [-\phi - \tau \cdot T, \phi + \tau \cdot T]$

We also require $\|\mathcal{R}(\mathcal{L})\|_2 \leq R' + T \cdot \tau$ for some sufficiently large $R' = \mathrm{poly}(n,d)$. As in the mean estimation case, this requirement is only needed for computability purposes. We will also say that $\mathcal{L}$ is an (α, τ, ϕ, T)-certificate for $\mathcal{Y}$.

Again, we may think of $\mathcal{L}$ as an approximate pseudo-expectation. In addition, for each constraint 2a) to 2d) we implicitly assume a bound on the degree of p so that $\mathcal{L}$ is applied to a polynomial of degree at most 12.

For our purposes, we will end up setting $\tau = 1/(n \cdot d)^{O(1)}$, for a large enough $O(1)$.

Now we use this definition to define a score function.

Definition 7.3 (Score Function)**.** Let $\alpha, \tau, \phi, T \in \mathbb{R}^{\geq 0}$, $y_1, \ldots y_n \in \mathbb{R}^d$ (with $\mathcal{Y} = \{y_1, \ldots, y_n\}$) and $\widetilde{\Sigma} \in \mathbb{R}^{d \times d}$. We define the score function $\mathcal{S} : \mathbb{R}^{d \times d} \to \mathbb{R}$ as

$$\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}; \alpha, \tau, \phi) = \min_T \text{ such that } \widetilde{\Sigma} \text{ is a } (\alpha, \tau, \phi, T) \text{ certifiable covariance for } \mathcal{Y} = \{y_1, \ldots, y_n\}.$$

In the rest of this section we will prove the following properties for this score function. This will allow us to use Theorem 4.1.

1. Score has sensitivity 1.
2. Score is quasi-convex as a function of $\widetilde{\Sigma}$.
3. All points $\widetilde{\Sigma}$ that have score at most $\eta \cdot n$ have $\|\widetilde{\Sigma} - \Sigma\|_F \leq \widetilde{O}(\eta)$. (Robustness for volume/accuracy purposes).
4. The volume of points that have score at most $\eta \cdot n$ is sufficiently large, and the volume of points with score at most $\eta' \cdot n$ for $\eta' > \eta$ is not too large.
5. Score is efficiently computable.
6. We can approximately minimize score efficiently.

Checking these constraints will, for the most part, be identical to the cases for mean estimation and covariance estimation in spectral distance. So for the sake of brevity, we omit any details that are essentially identical to these cases.

7.2.1 Existence of Low-Scoring Σ'

As in the case of covariance estimation, we must show that for data points drawn from $\mathcal{N}(\mathbf{0}, \Sigma)$, that some Σ' close to Σ has low score. In this setting, it actually turns out to be easier, because the dataset has already been well-conditioned and since the robust algorithm/score function are slightly easier to work with.

Lemma 7.4. *Suppose that $n \geq \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\eta^2}\right)$ and $\alpha = \widetilde{O}(\eta)$. Let $\mathcal{X} = \{x_1, \ldots, x_n\} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, where $\|\Sigma - I\|_{op} \leq \alpha$, and let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, with probability at least $1 - \beta$, for $\Sigma' = \frac{1}{n}\sum x_i x_i^\top$, every $\widetilde{\Sigma}$ such that $\|\widetilde{\Sigma} - \Sigma\|_F \leq \phi$ is $(\alpha, \tau, \phi, \eta n)$-certifiable.*

Proof. Again, we use the fact that $\mathcal{L}$ generalizes pseudoexpectations, which generalize expectations over a single data point. We will set $w_i = 1$ if $x_i = y_i$ and 0 otherwise, and $x'_i = x_i$ for all i. By part 2 of Lemma 6.3, we know that for all $d \times d$ symmetric matrices P with $\|P\|_F = 1$, if $x_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$, then $\left|\frac{1}{n}\langle x_i x_i^\top - I, P\rangle^2 - 2\right| \leq O(\alpha)$. But in our case, $x_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, but by the same argument as Lemma 6.9, we have

$$\frac{1}{n}\sum_{i=1}^{n}\langle x_i x_i^\top - \Sigma, P\rangle^2 \leq (2 + O(\alpha)) \cdot \|\Sigma^{1/2} P \Sigma^{1/2}\|_F^2$$

for any symmetric matrix P. Also, by Proposition 6.8,

$$\|\Sigma^{1/2} P \Sigma^{1/2}\|_F \leq (1 + 3\alpha) \cdot \|P\|_F.$$

So, by setting Σ' to be the empirical average of $x_i x_i^\top$, this means

$$\frac{1}{n}\sum_{i=1}^{n}\langle x_i x_i^\top - \Sigma', P\rangle^2 \leq \frac{1}{n}\sum_{i=1}^{n}\langle x_i x_i^\top - \Sigma, P\rangle^2 \leq (2 + O(\alpha)) \cdot \|\Sigma^{1/2} P \Sigma^{1/2}\|_F^2 \leq (2 + O(\alpha)) \cdot \|P\|_F^2$$

for any symmetric matrix P. Note, this is also true for non-symmetric matrices, because if P is nonsymmetric, then $\frac{P+P^\top}{2}$ has smaller Frobenius norm but $\langle x_i x_i^\top - \Sigma', P\rangle = \langle x_i x_i^\top - \Sigma', \frac{P+P^\top}{2}\rangle$.

By flattening and defining $S' = (\Sigma')^\flat$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\langle x_i^{\otimes 2} - S', v\rangle^2 \leq 2 + O(\alpha)$$

for all unit vectors $v \in \mathbb{R}^{d^2}$. Hence, $\frac{1}{n}\sum_{i=1}^{n}(x_i^{\otimes 2} - S')(x_i^{\otimes 2} - S')^\top$ has all eigenvalues at most $2 + O(\alpha)$, and thus we can find some positive semidefinite $MM^\top$ such that $\frac{1}{n}\sum_{i=1}^{n}(x_i^{\otimes 2} - S')(x_i^{\otimes 2} - S')^\top + MM^\top = 2 + O(\alpha)$.

If $\Sigma'_{j,k} - \widetilde{\Sigma}_{j,k} \in [-\phi, \phi]$ for all j, k, then $\widetilde{\Sigma}$ has score at most $\eta \cdot n$. So, every covariance matrix $\widetilde{\Sigma}$ with $\|\widetilde{\Sigma} - \Sigma'\|_F \leq \phi$ has score at least $\eta \cdot n$.

Finally, we remark that every w_i, $x_{i,j}$, and $M_{\{j,k\},\{j',k'\}}$ is bounded by $\mathrm{poly}(n, d)$. Therefore, the corresponding linear operator $\mathcal{L}$ satisfies $\|\mathcal{R}(\mathcal{L})\|_2 \leq (nd)^{O(1)}$. □

7.2.2 Sensitivity

Before proving sensitivity we need to prove the following upper bound on the value of the score function.

Lemma 7.5 (score function upper bound)**.** *The value of the score function $\mathcal{S}$ defined in Definition 7.3 is less than or equal to n.*

Proof. The proof is identical to Lemma 6.10: we again set $w_i = 0$ and $x'_i = x_i$ for all i, with $T = n$. □

Lemma 7.6 (sensitivity)**.** *The score function $\mathcal{S}$ as defined in Definition 7.3 has sensitivity 1 with respect to its first input.*

Proof. The proof is nearly identical to Lemma 6.11. Suppose that $\mathcal{Y}$, $\mathcal{Y}'$ are two neighboring datasets, and $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}$. Moreover, assume $\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) = T$. If we show that $\mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}') \le \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) = T+1$, by symmetry we are done.

The only constraints that are different in our setting from Lemma 6.11 are Constraints 3 and 4 in Definition 7.2. However, note that these three constraints do not involve w_j at all, so in fact their evaluation is the same regardless of $\mathcal{L}$ or $\mathcal{L}'$. The only difference is we are allowing the values $\mathcal{L}[\cdot]$ to have a greater range, which makes it easier. □

7.2.3 Quasi-convexity

Lemma 7.7 (quasi-convexity)**.** *The score function $\mathcal{S}$ as defined in Definition 7.3 is quasi-convex in its second input, $\widetilde{\Sigma}$.*

Proof. Again, all of the constraints are satisfied trivially except 2c), and the same proof as in Lemma 5.9 and Lemma 6.12 works for this case. □

7.2.4 Accuracy

We now show accuracy, meaning that any point $\widetilde{\Sigma}$ of low score with respect to i.i.d. samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ must be close to Σ. Because of our sensitivity bound, this will also imply a similar result for corrupted samples. Like for Lemma 5.10 and 6.13, we defer the proof to Appendix B.

Lemma 7.8. *Let $\alpha = \widetilde{O}(\eta)$ and suppose α, η are bounded by a sufficiently small constant. Let $n \ge \frac{d^2+\log^2(1/\beta)}{\alpha^2}$, and $\mathcal{X} = \{x_1, \dots, x_n\} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, for $\Sigma \in \mathbb{R}^{d\times d}$ with $(1-\alpha)I \preccurlyeq \Sigma \preccurlyeq (1+\alpha)I$.*

Then, for any $\alpha^* \le \alpha$, and assuming $\tau \ll 1/(nd)^{O(1)}$, with probability at least $1-\beta$, any covariance matrix $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}$ that is $(\alpha^*, \tau, \phi, T)$-certifiable for $\mathcal{X}$ with $T = \eta n$ and $\phi \le \alpha/d$ must satisfy $\|\widetilde{\Sigma} - \Sigma\|_F \le O(\alpha)$.

7.2.5 Volume of Good Points

Finally, we use our accuracy bounds to get an upper bound for the volumes of V_η. We already can obtain a lower bound from Lemma 7.4.

Lemma 7.9. *Let $\mathcal{X} = \{x_1, \dots, x_n\} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ (where $(1-\alpha)I \preccurlyeq \Sigma \preccurlyeq (1+\alpha)I$), and let $\mathcal{Y} = \{y_1, \dots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for every integer $T \in [\eta \cdot n, \eta^* \cdot n]$ for some fixed constant $\eta^* < 1$, with probability at least $1-\beta$, every (α, τ, ϕ, T)-certifiable covariance $\widetilde{\Sigma}$ with respect to $\mathcal{Y}$, for $\phi = \alpha/d$, satisfies $\|\Sigma - \widetilde{\Sigma}\|_F \le \widetilde{O}(T/n)$.*

Proof. Since the score function has sensitivity at most 1 (Lemma 7.6), this means that any (α, τ, ϕ, T)-certifiable covariance with respect to $\mathcal{Y}$ is an $(\alpha, \tau, \phi, T+\eta n)$-certifiable covariance with respect to $\mathcal{X}$.

Now, define $\eta' := \frac{T+\eta n}{n} = O(\frac{T}{n})$. In this case, by setting $\alpha' = \widetilde{O}(\eta')$ and since $\alpha = \widetilde{O}(\eta) \le \alpha'$, we have that by Lemma 7.8 that any $(\alpha, \tau, \phi, T+\eta n)$-certifiable covariance $\widetilde{\Sigma}$ must satisfy $\|\widetilde{\Sigma} - \Sigma\|_F \le O(\alpha') \le \widetilde{O}(T/n)$. □

We think of the set of potential covariances as lying in $\mathbb{R}^{d(d+1)/2}$, by taking the upper-diagonal entries. In addition, we know that the covariance has all eigenvalues between $1-\alpha$ and $1+\alpha$, so in $\mathbb{R}^{d(d+1)/2}$, they all lie in a ℓ_2-norm ball of radius $O(d)$ around the origin. If we set $\phi = \alpha/d$ and $\tau \ll 1/(nd)^{O(1)}$, this means the volume of (α, τ, ϕ, T)-certifiable covariances for $T = \eta n$ is at least $(\alpha/d)^{d(d+1)/2}$. However, for any $T = \eta' n$ for $\eta \le \eta' \le \eta^*$, the volume of (α, τ, ϕ, T)-certifiable covariances is at most $(\tilde{O}(\eta'))^{d(d+1)/2}$ times the volume of a $\frac{d(d+1)}{2}$-dimensional sphere, which is $(\tilde{O}(\eta')/d)^{d(d+1)/2}$. Finally, for $T = \eta' n$ with $\eta' > \eta^*$, the volume of Θ, the set of all candidate covariances $\widetilde{\Sigma}$, is at most $d^{O(d^2)}$.

7.2.6 Efficient Computability

As in the mean estimation case, we apply Theorem C.6 in Appendix C: the proof is identical to verify "efficient computability".

7.2.7 Efficient Finding of Low-Scoring Point

To verify that the "robust algorithm finds low-scoring point", we remove the constraint that $\mathcal{L}[\Sigma'_{j,k} - \widetilde{\Sigma}_{j,k}] \in [-\phi - \tau \cdot T, \phi + \tau \cdot T]$. We can apply Theorem C.6 in the same way to find some linear operator $\mathcal{L}$ with score at most $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) + 1$. Then, we can compute $\mathcal{L}[\Sigma']$ and set $r \le \tau$, and obtain that every matrix $\widetilde{\Sigma}$ with $\|\widetilde{\Sigma} - \mathcal{L}[\Sigma']\|_F \le r$ has score at most $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) + 1$.

7.3 Proof of Theorem 7.1

We apply Theorem 4.1, using the score function defined in Definition 7.3. Indeed, for $r = \phi = \alpha/d$, we have verified all conditions, as long as $n \ge \widetilde{O}((d^2 + \log^2(1/\beta))/\alpha^2)$. Therefore, we have an ε-DP algorithm running in time $\mathrm{poly}(n, d, \log \frac{R}{\alpha})$ that finds a candidate covariance $\widetilde{\Sigma}$ of score at most $2\eta n$, as long as

$$n \ge O\left(\max_{\eta' : \eta \le \eta' \le 1} \frac{\log(V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y})) + \log(1/(\beta \cdot \eta'))}{\varepsilon \cdot \eta'}\right).$$

Using Lemmas 7.4 and 7.9, and by the commentary after Lemma 7.9, we have that for $\eta' \le \eta^*$ for some $\eta^* = \Omega(1)$, if we set $\phi = \alpha/d$, then $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) \le (\widetilde{O}(\eta')/\alpha)^{d(d+1)/2} \le (O(1/\eta))^{d(d+1)/2}$. For $\eta' > \eta^*$, we have that $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) \le d^{O(d^2)}$. So overall, it suffices for

$$\begin{aligned} n &\ge O\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2}\right) + O\left(\max_{\eta \le \eta' \le \eta^*} \frac{d^2 \log(1/\eta) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'} + \max_{\eta^* \le \eta' \le 1} \frac{d^2 \log d + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right) \\ &= \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\alpha^2} + \frac{d^2 + \log(1/\beta)}{\varepsilon \cdot \alpha}\right). \end{aligned}$$

Hence, our algorithm, using this many samples, is ε-DP, and can find a point $\widetilde{\Sigma}$ of score at most $2\eta n$ with respect to $\mathcal{Y}$. So, by replacing η with 2η and applying Lemma 7.8, we have that any point $\widetilde{\Sigma}$ with score at most $2\eta n$ with respect to $\mathcal{Y}$ must have $\|\widetilde{\Sigma} - \Sigma\|_F \le O(\alpha)$. This completes the proof.

7.4 Proof of Theorems 1.3 and 1.4

By combining Theorems 6.1, 7.1, and 5.1, we are able to prove Theorem 1.3.

Proof of Theorem 1.3. Let the corrupted samples be $y_1, \dots, y_n$, and let the uncorrupted samples be $x_1, \dots, x_n$.

We may assume without loss of generality that $\alpha = \widetilde{O}(\eta)$ (either by raising α or η appropriately). Via the standard method of pairing samples and subtracting them, we may first assume that the mean is $\mathbf{0}$, and we will attempt to learn covariance. By Theorem 6.1, we can thus privately learn a $\widetilde{\Sigma}_1$ such that $\|\Sigma^{-1/2}\widetilde{\Sigma}_1\Sigma^{-1/2} - I\|_{op} \le \alpha$, given samples $y_1, \dots, y_n$.

Next, we learn Σ up to Mahalanobis distance rather than just spectral distance. Let $\hat{y}_i = \widetilde{\Sigma}_1^{-1/2} y_i$, and let $\hat{x}_i = \widetilde{\Sigma}_1^{-1/2} x_i$ and $x_i^* = \Sigma^{-1/2} x_i$. Note that $x_i^* \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$, and $\hat{x}_i = J \cdot x_i^*$ for $J = \widetilde{\Sigma}_1^{-1/2}\Sigma^{1/2}$. However, J may be adversarially dependent on the data points, as we chose $\widetilde{\Sigma}_1$ based on the samples $\mathcal{Y}$[9]. Nevertheless, we may still apply Theorem 7.1, because it will turn out that the $\{\hat{x}_i\}$ samples will have the desired resilience conditions for *every* choice of J with $\|JJ^\top - I\|_{op} \le \alpha$.

Indeed, note that $\langle \hat{x}_i\hat{x}_i^\top - JJ^\top, P\rangle = \langle Jx_i^*(x_i^*)^\top J^\top - JJ^\top, P\rangle = \langle (x_i^*)(x_i^*)^\top - I, J^\top P J\rangle$ for all J, and $\|J^\top P J\|_F = (1 \pm 3\alpha)\cdot\|P\|_F$ by Proposition 6.8, since $JJ^\top = \widetilde{\Sigma}_1^{-1/2}\Sigma\widetilde{\Sigma}_1^{-1/2}$. Thus, assuming $\{x_i^*\}$ satisfy the resilience properties (Lemma 6.3), $\frac{1}{n}\sum\langle \hat{x}_i\hat{x}_i^\top - JJ^\top, P\rangle^2 \le (2 + O(\alpha))\cdot\|P\|_F^2$ for all symmetric matrices P. This is sufficient to ensure Lemma 7.4 holds, if we replace $\mathcal{X}$ with $\{\hat{x}_1, \dots, \hat{x}_n\}$ and Σ with $JJ^\top = \widetilde{\Sigma}_1^{-1/2}\Sigma\widetilde{\Sigma}_1^{-1/2}$. Likewise, Lemma 7.8 will also work in the same way, replacing each x_i with $\hat{x}_i$, and replacing Σ with $JJ^\top$. The rest of the conditions also clearly hold (as they either do not depend on the dataset or follow from Lemmas 7.4 and 7.8). Therefore, we can apply Theorem 7.1 to privately and robustly find $\widetilde{\Sigma}_2$ such that $\|\widetilde{\Sigma}_2 - JJ^\top\|_F \le O(\alpha)$, by applying the algorithm on $\hat{y}_1, \dots, \hat{y}_n$. Since both Σ_2 and $JJ^\top$ are spectrally bounded between $1 \pm \alpha$, this implies $\|I - \widetilde{\Sigma}_2^{-1/2} JJ^\top \widetilde{\Sigma}_2^{-1/2}\|_F \le O(\alpha)$, which means $\|I - J^\top\widetilde{\Sigma}_2^{-1}J\|_F \le \alpha$. Note, however, that we can write this as $\|I - \Sigma^{1/2}(\widetilde{\Sigma}_1^{-1/2}\widetilde{\Sigma}_2^{-1}\widetilde{\Sigma}_1^{-1/2})\Sigma^{1/2}\|_F \le \alpha$, which implies that Σ and $\widetilde{\Sigma}_1^{1/2}\widetilde{\Sigma}_2\widetilde{\Sigma}_1^{1/2}$ are α-close in Mahalanobis distance. So, we can output $\hat{\Sigma} = \widetilde{\Sigma}_1^{1/2}\widetilde{\Sigma}_2\widetilde{\Sigma}_1^{1/2}$.

Finally, we must decide on $\hat{\mu}$. To do so, we return to our original samples $y_1, \dots, y_n$ (where we did not do sample pairing and subtraction), and redefine $\hat{y}_i = \hat{\Sigma}^{-1/2} y_i$, $\hat{x}_i = \hat{\Sigma}^{-1/2} x_i$. Also, redefine $x_i^* = \Sigma^{-1/2} x_i$. Now, $x_i^* \sim \mathcal{N}(\Sigma^{-1/2}\mu, I)$, and $\hat{x}_i = J \cdot x_i^*$ for some new choice of $J = \hat{\Sigma}^{-1/2}\Sigma^{1/2}$, and note $\|JJ^\top - I\|_F \le \alpha$, but J may be adversarial. However, this is sufficient to satisfy all resilience conditions by the remark after Corollary 5.4. Hence, using Theorem 5.1 on the corrupted samples $\hat{y}_i$, we learn $\hat{\Sigma}^{-1/2}\mu$ up to ℓ_2 error $O(\alpha)$. Multiplying this by $\hat{\Sigma}^{1/2}$, we find $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_{\hat{\Sigma}} \le O(\alpha)$, which implies $d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \hat{\Sigma})) \le O(\alpha)$. But since Σ and $\hat{\Sigma}$ have Mahalanobis distance at most $O(\alpha)$, this means $d_{\mathrm{TV}}(\mathcal{N}(\mu, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \le \alpha$. So, by the Triangle inequality, we have $d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \le O(\alpha)$, which completes the proof.

The privacy factor and increases by a factor of 3 via basic composition of privacy, the failure probability also increases by a factor of 3, and the sample complexity is simply the maximum of the sample complexities required by Theorems 6.1, 7.1, and 5.1. □

The proof of Theorem 1.4 is very similar: this time, we combine Theorems 6.2, 7.1, and 5.2.

[9] One may attempt to remove this issue by using different samples for this step, but due to the adversarial nature of the strong contamination model, previous samples may affect how later samples are corrupted!

Proof of Theorem 1.4. The proof is identical to the proof of Theorem 1.3. First, we privately learn $\widetilde{\Sigma}_1$ such that $\|\Sigma^{-1/2}\widetilde{\Sigma}_1\Sigma^{-1/2} - I\|_{op} \leq \alpha$, using Theorem 6.2. We then replace each y_i with $\hat{y}_i = \widetilde{\Sigma}_1^{-1/2} y_i$, and via the same procedure we learn some $\hat{\Sigma}$ such that $\Sigma, \hat{\Sigma}$ are close in Mahalanobis distance. Finally, we redefine $\hat{y}_i = \hat{\Sigma}^{-1/2} y_i$, and learn $\hat{\mu}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq O(\alpha)$, using Theorem 5.2, in the same way as we applied Theorem 5.1, to prove Theorem 1.3.

The privacy factor and failure probability increase by a factor of 3, and the sample complexity is the maximum of the sample complexities required by Theorems 6.2, 7.1, and 5.2. □

Acknowledgements

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

[AAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 185–216. JMLR, Inc., 2021.

[AD20a] Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020.

[AD20b] Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020.

[AKT+22] Daniel Alabi, Pravesh Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust and optimal. *Personal communication (in submission)*, 2022.

[AL22] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1075–1076, 2022.

[AM20] Marco Avella-Medina. The role of robust statistics in private data analysis. *Chance*, 33(4):37–42, 2020.

[AM21] Marco Avella-Medina. Privacy-preserving parametric inference: a case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

[BGS+21] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019.

[BS19] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 181–191. Curran Associates, Inc., 2019.

[DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[DK22] Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2022.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664, Washington, DC, USA, 2016. IEEE Computer Society.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 999–1008. JMLR, Inc., 2017.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 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, pages 486–503, Berlin, Heidelberg, 2006. Springer.

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 73–84, Washington, DC, USA, 2017. IEEE Computer Society.

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 2745–2754, Philadelphia, PA, USA, 2019. SIAM.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[EMN22] Hossein Esfandiari, Vahab S Mirrokni, and Shyam Narayanan. Tight and robust private mean estimation with few users. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 16383–16412. JMLR, 2022.

[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.

[GH22] Kristian Georgiev and Samuel B Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

[GKM21] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[GKMN21] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. Robust and private learning of halfspaces. In *The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *AISTATS '21*, pages 1603–1611. PMLR, 2021.

[HKM22] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. 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, New York, NY, USA, 2022. ACM.

[HL18a] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[HL18b] Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '18, pages 1021–1034, New York, NY, USA, 2018. ACM.

[HL19] Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? In *Conference on Learning Theory*, pages 1649–1682. PMLR, 2019.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

[JLLV21] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh S. Vempala. Reducing isotropy and volume to KLS: an $o^*(n^3\psi^2)$ volume algorithm. In *53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC '21, pages 961–974. ACM, 2021.

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemama. *Discret. Comput. Geom.*, 13:541–559, 1995.

[KLSU19] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1853–1902, 2019.

[KMS22a] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22, 2022.

[KMS+22b] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 544–572, 2022.

[KMV22] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 723–777, 2022.

[KMZ22] Pravesh K. Kothari, Peter Manohar, and Brian Hu Zhang. Polynomial-time sum-of-squares can robustly estimate mean and covariance of gaussians optimally. In Sanjoy Dasgupta and Nika Haghtalab, editors, *International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France*, volume 167 of *Proceedings of Machine Learning Research*, pages 638–667. PMLR, 2022.

[KS17] Pravesh K. Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pages 2204–2235, 2020.

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[LBY22] Mengchu Li, Thomas B Berrett, and Yi Yu. On robustness and local differential privacy. *arXiv preprint arXiv:2201.00751*, 2022.

[LKKO21] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[LKO22] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1167–1246, 2022.

[LV04] László Lovász and Santosh S. Vempala. Hit-and-run from a corner. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, STOC '04, pages 310–314. ACM, 2004.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

[MV22] Oren Mangoubi and Nisheeth K. Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

[RC21] Kelly Ramsay and Shoja'eddin Chenouri. Differentially private depth functions and their associated medians. *arXiv preprint arXiv:2101.02800*, 2021.

[RJC22] Kelly Ramsay, Aukosh Jagannath, and Shoja'eddin Chenouri. Concentration of the exponential mechanism and differentially private multivariate medians. *arXiv preprint arXiv:2210.06459*, 2022.

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians (ICM 2018)*, pages 3389–3423. WORLD SCIENTIFIC, 2018.

[RW17] Prasad Raghavendra and Benjamin Weitz. On the bit complexity of sum-of-squares proofs. *arXiv preprint arXiv:1702.05139*, 2017.

[SCV18] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 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, pages 45:1–45:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[SM22] Aleksandra Slavkovic and Roberto Molinari. Perturbed M-estimation: A further investigation of robust statistics for differential privacy. In Alicia L. Carriquiry, Judith M. Tanur, and William F. Eddy, editors, *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*, pages 337–361. Springer, 2022.

[TCK+22] Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 21828–21863. JMLR, Inc., 2022.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

A Omitted proofs for Private Sampling

In this section, we prove Lemmas 4.4 and 4.5.

A.1 Preliminaries

In this subsection, we note a few miscellaneous results that will be very important in proving Theorems 4.1 and 4.2.

We need the following ellipsoid theorem, showing that convex bodies are contained in reasonably small ellipsoids but also contain reasonably large ellipsoids.

Theorem A.1. *[KLS95, Theorem 4.1] Let $K \subset \mathbb{R}^d$ be a convex body in isotropic position, meaning that if X is uniformly drawn from K, $\mathbb{E}[X] = \mathbf{0}$ and $Cov(X) = I$. Then, for B the unit ball of radius 1, $\sqrt{\frac{d+2}{d}} \cdot B \subset K \subset \sqrt{d(d+2)} \cdot B$.*

Next, we need the following basic proposition.

Proposition A.2. *Suppose that K, K' are convex bodies such that $B(\mathbf{0}, r) \subset K \subset K'$, and suppose that $K' \not\subset (1+\gamma_1)K$, where $(1+\gamma_1)K$ represents K dilated by a factor $1+\gamma_1$ around the origin. Then, $\mathrm{vol}(K') - \mathrm{vol}(K) \geq (\frac{\gamma_1 \cdot r}{6d})^d$.*

Proof. Since $K' \not\subset (1+\gamma_1)K$, there exists a vector v such that $v, (1+\frac{\gamma_1}{2})v$ are both contained in $K' \backslash K$. Now, let's consider the ball of radius ρ around $(1+\frac{\gamma_1}{4})v$ for some small value ρ. We will show that for ρ appropriately chosen, this ball is contained in K' but is disjoint from K.

For any such point, we can write it as $v + (\frac{\gamma_1}{4}v + w)$ for some w with $\|w\|_2 \leq \rho$. If this point were in K, then since $v \notin K$, this means by convexity $v - \lambda(\frac{\gamma_1}{4}v + w)$ is not in K for all $\lambda \geq 0$. By choosing $\lambda = \frac{4}{\gamma_1}$, we have that $-\frac{4}{\gamma_1} \cdot w$ is not in K. This is a contradiction if we choose $\rho \leq \frac{\gamma_1 \cdot r}{4}$, since this implies $-\frac{4}{\gamma_1} \cdot w$ has norm at most r so it must be in K. Thus, if $\rho \leq \frac{\gamma_1 \cdot r}{4}$, the ball of radius ρ around $(1+\frac{\gamma_1}{4})v$ is disjoint from K.

Next, we alternatively write the point as $(1+\frac{\gamma_1}{2})v - (\frac{\gamma_1}{4}v - w)$. To show it is in K', note that $(1+\frac{\gamma_1}{2})v$ is in K', so by convexity it suffices to show that $(1+\frac{\gamma_1}{2})v - \lambda(\frac{\gamma_1}{4}v - w)$ is in K' for some $\lambda \geq 1$. By setting $\lambda = (1+\frac{\gamma_1}{2}) \cdot \frac{4}{\gamma_1}$, it suffices to show that $(1+\frac{\gamma_1}{2}) \cdot \frac{4}{\gamma_1} \cdot w = \left(\frac{4}{\gamma_1} + 2\right) \cdot w$ is in K'. But this is similarly true as long as $\rho \leq r/(\frac{4}{\gamma_1} + 2)$, which holds as long as $\rho \leq \frac{\gamma_1 \cdot r}{6}$.

Therefore, $K' \backslash K$ contains a ball of radius $\frac{\gamma_1 \cdot r}{6}$, which has volume at least $(\frac{\gamma_1 \cdot r}{6d})^d$. □

A.2 Sampling from a well-rounded convex body with an imperfect oracle

In this subsection, our goal is to sample uniformly from a convex body, but we wish to do this even if we only can afford polynomial bit precision and do not have a perfect membership oracle. To do this, we will apply the well-known hit-and-run Markov chain, but with some minor adjustments to avoid the issue of requiring infinite precision arithmetic.

First, we describe the standard hit-and-run Markov chain assuming infinite precision arithmetic. Given a convex body K for which we have a membership oracle, the hit-and-run algorithm starts with a point $x_0 \in K$. At each step t, we move from $x_{t-1} \in K$ to $x_t \in K$ as follows. We first pick a vector v at random from the unit sphere. We then let x_t be uniformly chosen on the line segment $\{x_{t-1} + \lambda \cdot v\}_{\lambda \in \mathbb{R}} \cap K$, i.e., the line segment parallel to v that goes through x_{t-1}, but restricted by K since we cannot sample outside K.

The main result of the hit-and-run algorithm we apply is the following, due to Lovász and Vempala.

Theorem A.3. *[LV04] Let K be a d-dimensional convex body that contains the ball $B(\mathbf{0}, 1)$ and is contained in the ball $B(\mathbf{0}, D)$. Then, for a sufficiently large constant C and for any $0 < \gamma < \frac{1}{2}$, after $m \geq Cd^2D^2 \log \gamma^{-1}$ steps of hit-and-run starting from the origin (i.e., setting x_0 to be the origin), the distribution of the final point x_m has total variation distance at most γ from the uniform distribution over K.*

In our setting, we cannot directly use the hit-and-run algorithm for two reasons. The first reason is that we cannot pick a truly uniform direction and sample truly uniformly along that direction from a starting point. The second reason is that we don't have a perfect membership oracle. That being said, we will be able to make minor modifications to the algorithm and show that we output a distribution that is "close" to uniform on K.

We assume we are given two unknown convex bodies K_1, K_2, such that $B(\mathbf{0}, 1) \subset K_1 \subset K_2 \subset (1 + \gamma_1)K_1 \subset B(\mathbf{0}, D)$. One should think of D as polynomially large (we will later improve this to being exponentially large) and γ_1 as exponentially small. We also assume we have an (K_1, K_2)-membership oracle $\mathcal{O}$.

For some small parameter $\gamma > 0$, we define the hit-and-run algorithm with γ precision as follows. Given a point x_{t-1}, we select a random unit vector v and round the coordinates of v to multiples of γ. Next, we attempt to sample along the line $x_{t-1} + \lambda \cdot v$ for $\lambda \in \mathbb{R}$, restricted to K_1. To do this with our oracle $\mathcal{O}$, we perform a binary search to find a positive integer a_1 such that $\mathcal{O}$ accepts $x_{t-1} + a_1 \cdot \gamma \cdot v$ but rejects $x_{t-1} + (a_1 + 1) \cdot \gamma \cdot v$. Likewise we find a negative integer $-a_2$ such that oracle accepts $x_{t-1} - a_2 \cdot \gamma \cdot v$ but rejects $x_{t-1} - (a_2 + 1) \cdot \gamma \cdot v$. Finally, we compute $x_t := x_{t-1} + a \cdot \gamma \cdot v$, where a is an integer chosen uniformly at random betwen $-a_2$ and a_1, inclusive. We note that it may be possible to choose x_t that the oracle rejects, but we know that x_t is always in K_2.

Before analyzing the modified hit-and-run algorithm, we first show the following proposition.

Proposition A.4. *Let K be any convex body. Suppose that x is a point and $\gamma > 0$ is a parameter such that the ball $B(x, \gamma)$ is contained in K, and let Λ be an arbitrary line passing through x. Define L to be the length of $\Lambda \cap K$. Then, for any parameter $0 < \lambda < 1$, the length of points x' on $\Lambda \cap K$ such that $B(x', \lambda \cdot \gamma)$ is contained in K is at least $(1 - \lambda) \cdot L$.*

Proof. Let Λ' represent the segment of Λ that is contained in K, with endpoints y and z. Since the ball $B(x, \gamma)$ is contained in K, we can consider the convex hull of this ball and the points y and z. Note that the ball of radius $\lambda' \cdot r$ around $\lambda' x + (1-\lambda')y$ or around $\lambda' x + (1-\lambda')z$ is contained in this convex hull. So, all points x' on Λ' such that $B(x', \lambda \cdot \gamma)$ is not contained in K cannot be between $\lambda x + (1-\lambda)y$ and $\lambda x + (1-\lambda)z$, so the length of the interval of such points is at least $(1-\lambda) \cdot L$. □

Next, we show that the hit-and-run algorithm with γ precision, assuming γ is sufficiently small, always stays within K_1 up to a small margin of error.

Proposition A.5. *Let K_1, K_2 be convex bodies such that $B(\mathbf{0}, 1) \subset K_1 \subset K_2 \subset (1+\gamma_1)K_1 \subset B(\mathbf{0}, D)$. Consider running m steps of hit-and-run from the origin with γ_1 precision, with x_t being the point chosen after the t^{th} step for all $0 \le t \le m$. Then, for any $0 < \tau < 1$ such that $(\tau/2)^{m+1} \ge D \cdot \gamma_1$, we have that with probability at least $1 - O(m \cdot \tau)$, all the points x_t satisfy the $B(x_t, (\tau/2)^m) \subset K_1$.*

Proof. Suppose that after t steps of hit and run, the point selected is x_t. Suppose that $B(x_t, \gamma^{(t)})$ is contained in K_1, for some positive real $\gamma^{(t)}$, which also means $B(x_t, \gamma^{(t)}) \subset K_2$. Let Λ represent an arbitrary line through x_t. By making oracle calls to O using the binary search procedure, we obtain some line segment $\Lambda' \subset \Lambda$ that goes entirely through K_1 but is contained in K_2. Let L represent the length of the line segment we found, and L_1 represent the length of $\Lambda \cap K_1$. Also, let L_2 represent the length of $\Lambda \cap K_2$, so $L_1 \le L \le L_2$.

Recall that $B(x_t, \gamma^{(t)}) \subset K_2$, but note that for any point x' outside K_1, $B(x', \gamma_1 \cdot D) \not\subset K_2$. Therefore, by Proposition A.4, the value of $L_2 - L_1$ is at most $\frac{\gamma_1 \cdot D}{\gamma^{(t)}} \cdot L_2$, which assuming $\gamma^{(t)} \ge 2\gamma_1 D$ is at most $\frac{2\gamma_1 \cdot D}{\gamma^{(t)}} \cdot L$. In addition, the length of points x' in L_1 such that $B(x', \tau \cdot \gamma^{(t)}) \not\subset K_1$ is at most $\tau \cdot L_1 \le \tau \cdot L$. So, if we sample randomly from L even after discretizing by rounding coordinates to the nearest multiples of γ_1, the probability of selecting a point x' such that $B(x', \tau \cdot \gamma^{(t)} - \gamma_1) \not\subset K_1$ is at most $\tau + O\left(\frac{\gamma_1 \cdot D}{\gamma^{(t)}}\right)$.

For $t = 0$, we assume x_0 is the origin, so we can set $\gamma^{(0)} = 1$. In general, we fix some parameter τ, and let $\gamma^{(t+1)} := \tau \cdot \gamma^{(t)} - \gamma_1$. If $(\tau/2)^{m+1} \ge D \cdot \gamma_1$, then we will inductively have that $\gamma^{(t)} \ge (\tau/2)^t$ for all $0 \le t \le m$, and so $\frac{\gamma_1 \cdot D}{\gamma^{(t)}} \le \tau/2$. Therefore, by a union bound over all $0 \le t \le m$, we have that with probability at least $1 - O(m \cdot \tau)$, every x_t selected satisfies $B(x_t, \gamma^{(t)}) \subset K_1$, so $B(x_t, (\tau/2)^m) \subset K_1$. □

We show that the hit-and-run algorithm with limited precision outputs a distribution that is "close" to uniform on the convex body K_1. We will use the following formal definition of closeness.

Definition A.6. We define two distributions $\mathcal{D}, \mathcal{D}'$ over Euclidean space $\mathbb{R}^d$ to be (γ, γ')**-close** if there exists a coupling of $\mathcal{D}, \mathcal{D}'$ such that $\mathbb{P}_{(a,c)\sim(\mathcal{D},\mathcal{D}')}(\|a - c\|_2 \ge \gamma) \le \gamma'$.

Lemma A.7. *Given parameters D, γ_2, γ_3, there exists γ_1 such that $\log \gamma_1^{-1} = \mathrm{poly}(d, D, \log \gamma_2^{-1}, \log \gamma_3^{-1})$, and the following holds. If K_1, K_2 are convex bodies such that $B(\mathbf{0}, 1) \subset K_1 \subset K_2 \subset (1+\gamma_1)K_1 \subset B(\mathbf{0}, D)$, then after $m \ge O(d^2 D^2 \log \gamma_3^{-1})$ steps of hit-and-run starting from the origin with γ_1 precision, the final point is (γ_2, γ_3)-close to the uniform distribution over K_1.*

Proof. We create a coupling between running hit-and-run with perfect precision and running hit-and-run with γ_1 precision. After t steps, let x_t be the point we sampled for hit-and-run with

perfect precision, and let x'_t be the point we sampled for hit-and-run with γ_1 precision. We start with $x_0 = x'_0$ as the origin.

Let Λ be the random line drawn through x_t, and let Λ' be the rounded random line drawn through x'_t. We will couple the lines so that with $1-\gamma_1$ probability, the lines are essentially parallel up to γ_1 error. Let's write $\Lambda = \{x_t + \lambda \cdot v_t\}_{\lambda \in \mathbb{R}}$ and $\Lambda' = \{x'_t + \lambda \cdot v'_t\}_{\lambda \in \mathbb{R}}$, where v_t, v'_t are unit vectors with $\|v_t - v'_t\|_2 \le \sqrt{d} \cdot \gamma_1$ with probability at least $1-\gamma_1$. If $B(x_t, (\tau/2)^m) \subset K_1$, then the probability that a random point x' on $\Lambda \cap K_1$ satisfies $B(x', (\tau/2)^{m+1}) \not\subset K_1$ is at most $\tau + O\left(\frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$, by the argument of Proposition A.5. Likewise, if $B(x'_t, (\tau/2)^m) \subset K_1$, then the probability that a random point x' on Λ' in the segment selected by step $t+1$ of the algorithm satisfies $B(x', (\tau/2)^{m+1}) \not\subset K_1$ is at most $\tau + O\left(\frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$.

Now, suppose that $B(x_t, (\tau/2)^m), B(x'_t, (\tau/2)^m) \subset K_1$, and $\|x_t - x'_t\|_2 \le \tau^{(t)}$ for some parameter $\tau^{(t)} \le (\tau/2)^m - 2D \cdot \gamma_1$. Then, if we selected λ uniformly such that $x_t + \lambda v_t \in K_1$, then $x'_t + \lambda v'_t$ has distance at most $\tau^{(t)} + 2D \cdot \gamma_1$ from $x_t + \lambda v_t$ (even after rounding off λ to the nearest multiple of γ_1). This means that with probability at most $\tau + O\left(\frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$, $x'_t + \lambda v'_t \in K_1$. Likewise, if we selected λ' according to the distribution from hit-and-run on x'_t with γ_1-precision, then $x_t + \lambda' v_t \in K_1$ with probability at most $\tau + O\left(\frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$ (even after replacing λ' with a uniform real in $[\lambda' - \gamma_1/2, \lambda' + \gamma_1/2]$ to "un-round" it). So, by keeping λ the same (up to rounding) whenever possible (which can happen with at most $O\left(\tau + \frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$ failure probability, we have that $\|x_{t+1} - x'_{t+1}\|_2 \le \tau^{(t)} + 2D \cdot \gamma_1$ if $B(x_t, (\tau/2)^m), B(x'_t, (\tau/2)^m) \subset K_1$, $\|x_t - x'_t\|_2 \le \tau^{(t)}$, and $\tau^{(t)} \le (\tau/2)^m - 2D \cdot \gamma_1$.

To finish the proof, we set $\tau^{(t)} = 2D\gamma_1 \cdot t$. We assume that $\tau^{(t)} \le (\tau/2)^m - 2D\gamma_1$, so it suffices for $4Dm\gamma_1 \le (\tau/2)^m$. Let $\mathcal{E}_t$ be the event that $B(x_t, (\tau/2)^m), B(x'_t, (\tau/2)^m) \subset K_1$, and $\|x_t - x'_t\|_2 \le \tau^{(t)}$. Then, if $\mathcal{E}_t$ holds, the probability that $\|x_{t+1} - x'_{t+1}\|_2 \le \tau^{(t+1)}$ does not hold is at most $O\left(\tau + \frac{\gamma_1 \cdot D}{(\tau/2)^m}\right)$. In addition, the probability that $B(x_{t+1}, (\tau/2)^m), B(x'_{t+1}, (\tau/2)^m) \subset K_1$ does not hold for any choice of $t+1$ is at most $O(m \cdot \tau)$ if $(\tau/2)^{m+1} \ge D \cdot \gamma_1$, by Proposition A.5. So, $\mathbb{P}(\mathcal{E}_t \setminus \mathcal{E}_{t+1}) \le O(m \cdot \tau + \tau + \frac{\gamma_1 \cdot D}{(\tau/2)^m})$, which means that the probability that $\mathcal{E}_m$ doesn't hold is at most $O\left(m^2 \cdot \tau + \frac{m \cdot \gamma_1 \cdot D}{(\tau/2)^m}\right)$, as long as $4Dm\gamma_1 \le (\tau/2)^m$. Assuming $\mathcal{E}_m$, we have that $\|x_m - x'_m\|_2 \le 2Dm\gamma_1$, and by Theorem A.3, if $m \ge Cd^2D^2 \log \gamma^{-1}$ then the distribution of x_m is γ-far from uniform over K_1.

To summarize, we have that there exists a coupling of x'_m (which is the random walk after m steps of hit-and-run with γ_1-precision) with a uniform distribution x over K_1 such that $\mathbb{P}((\|x - x'_m\|_2 \le 2Dm\gamma_1) \le O\left(\gamma + m^2\tau + \frac{m\gamma_1 \cdot D}{(\tau/2)^m}\right)$, as long as $4Dm\gamma_1 \le (\tau/2)^m$, $D\gamma_1 \le (\tau/2)^{m+1}$, and $m \ge Cd^2D^2 \log \gamma^{-1}$. Given some small parameters γ_2, γ_3, we set $\gamma = c\gamma_3$, $m = Cd^2D^2 \log \gamma^{-1}$, and $\tau = \frac{\gamma}{m^2}$ for some small constant c. Finally, we set $\gamma_1 = \min\left(\frac{\gamma_2}{2Dm}, \frac{\gamma(\tau/2)^m}{mD}, \frac{(\tau/2)^{m+1}}{4Dm}\right)$ so that the conditions are satisfied and $\mathbb{P}(\|x - x'_m\|_2 \ge \gamma_2) \le O(\gamma) \le \gamma_3$. □

Next, we must show that, rather than having $K_1, K_2 \subset B(\mathbf{0}, D)$ for some polynomially sized D, we can have $K_1, K_2 \subset B(\mathbf{0}, R)$ for R exponentially large. In other words, one can avoid issues when the convex body is poorly conditioned.

Lemma A.8. *Let γ_2, γ_3 be as in Lemma A.7, and let γ_1 be defined as in the end of Lemma A.7, assuming $D := 2d^3$. For some $r < 1 < R$, Let K_1, K_2 be convex bodies with a (K_1, K_2)-membership oracle, such that $B(\mathbf{0}, r) \subset$*

$K_1 \subset K_2 \subset B(\mathbf{0}, R)$, *and* $\mathrm{vol}(K_2) - \mathrm{vol}(K_1) \le \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$. *Then, there exists a* $\mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_2^{-1}, \log \gamma_3^{-1})$*-time algorithm that can find an affine transformation* $\mathbb{A}$ *such that* $\mathbb{A}(K_1)$ *is contained in* $B(\mathbf{0}, 2d^3)$ *but contains* $B(\mathbf{0}, 1)$*.*

Proof. The proof is an ellipsoid method, modified to deal with the fact that we do not have a perfect membership oracle and that we do not have a separation oracle. We will keep track of an interior ellipsoid $E_1 \subset K_1$ and an exterior ellipsoid $E_2 \supset K_2$, and keep trying to either grow E_1 or shrink E_2. The way we do this will be inspired by some recent work on sampling and volume computation of convex bodies [JLLV21].

Given current interior ellipsoid E_1 and exterior ellipsoid E_2, let $\mathbb{A}$ be some affine transformation that sends E_1 to the ball $B(\mathbf{0}, 1) = \mathbb{A}E_1$, and where the largest axis of $\mathbb{A}E_2$ is parallel to the first coordinate direction. (Note that $\mathbb{A}E_2$ may not have center as the origin.) At every step, we will only increase the volume of E_1 and decrease the volume of E_2, and since E_1 started out as $B(\mathbf{0}, r)$, the affine transformation $\mathbb{A}$ multiplies the volume by at most $\left(\frac{1}{r}\right)^d$. Therefore, $\mathrm{vol}(\mathbb{A}K_2) - \mathrm{vol}(\mathbb{A}K_1) \le \left(\frac{\gamma_1}{6d}\right)^d$.

We may assume that this largest axis of $\mathbb{A}E_2$ has length at least $D := 2d^3$, or else we are already done. Define $K_1' := \mathbb{A}K_1 \cap B(\mathbf{0}, D)$ and $K_2' := \mathbb{A}K_2 \cap B(\mathbf{0}, D)$. Note that $\mathrm{vol}(K_2') - \mathrm{vol}(K_1') \le \mathrm{vol}(\mathbb{A}K_2) - \mathrm{vol}(\mathbb{A}K_1) \le \left(\frac{\gamma_1}{6d}\right)^d$, and $B(\mathbf{0}, 1) \subset K_1'$, so by Proposition A.2, $\mathbb{A}K_2 \subset (1+\gamma_1)\mathbb{A}K_1$ and $K_2' \subset (1+\gamma_1)K_1'$. Given (K_1, K_2)-membership oracle access, it is simple to obtain (K_1', K_2')-membership oracle access. Therefore, by Lemma A.7, we can produce a sample from a distribution that is (γ_2, γ_3)-close to uniform over K_1' in $\mathrm{poly}(d, D, \log \gamma_2^{-1}, \log \gamma_3^{-1})$ time.

Assuming without loss of generality that $\gamma_2, \gamma_3 \le d^{-100}$, we can repeat the sampling $\mathrm{poly}(d, D) = d^{O(1)}$ times and approximately learn the mean μ_1 and covariance Σ_1 of the uniform distribution with respect to K_1', up to ℓ_2 norm (resp., Frobenius norm) error 1 by using the empirical mean $\hat{\mu}_1$ and empirical covariance $\hat{\Sigma}_1$ as our estimates.

First, suppose that one of our (approximate) samples from K_1' was a point x with ℓ_2 norm at least $25d$. Then, if we define $y = (1-\gamma_1)x$, then $y \in \mathbb{A}K_1$ and $\|y\|_2 \ge 20d$. If we rotate the space $\mathbb{R}^d$ and assume $y = (y_1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ for $y_1 \ge 20d$, then the ellipse $E = \left\{z : (\frac{z_1 - 10}{10})^2 + \sum_{i=2}^d \left(\frac{z_i}{(1-1/d)}\right)^2 \le 1\right\}$ is contained in the convex hull of $B(\mathbf{0}, 1)$ and y. The volume ratio $\mathrm{vol}(E)/\mathrm{vol}(B(\mathbf{0}, 1))$ is $10 \cdot (1 - \frac{1}{d})^{d-1} \ge \frac{10}{e} \ge 2$, so we can replace $\mathbb{A}E_1 = B(\mathbf{0}, 1)$ with the larger ellipsoid $E \subset \mathbb{A}K_1$.

Alternatively, every sample we drew has ℓ_2 norm at most $25d$, which means that the empirical covariance $\hat{\Sigma}_1$ has operator norm at most $25d$. Thus, Σ_1 has operator norm at most $30d$, meaning $x^\top \Sigma_1 x \le 30d$ for all $\|x\|_2 = 1$. Now, by Theorem A.1, $K_1' \subset \left\{x : (x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) \le d(d+2)\right\}$. So if $x \subset K_1'$, then $(x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) \le d(d+2)$, and since the minimum eigenvalue of Σ_1^{-1} is at least $\frac{1}{25d+1}$, this means that $\|x - \mu_1\|^2 \le d(d+2)(25d+1)$ for all $x \in K_1'$. Since the origin is in K_1', this implies that every point in K_1' has norm bounded by $O(d^{3/2})$.

Recall that the original convex bodies K_1, K_2 are known to be in E_2, and $\mathbb{A}E_2$ has largest axis parallel to the first coordinate direction. If the major radius of $\mathbb{A}E_2$ is some F, then we claim that all points in $\mathbb{A}K_1$ or $\mathbb{A}K_2$ have first coordinate bounded in magnitude by $O\left(\frac{F}{d^{3/2}}\right)$. To see why, for any $x \in \mathbb{A}K_1$, $x \cdot \frac{D}{F}$ is in K_1 by convexity. Moreover, since $\|x\|_2 \le F$, this means that $\|x \cdot \frac{D}{F}\|_2 \le D$ so $x \cdot \frac{D}{F} \in K_1'$. Therefore, we actually have $\left\|x \cdot \frac{D}{F}\right\| \le O(d^{3/2})$, which means that $\|x\| \le O(F \cdot d^{3/2}/D) = O(F/d^{3/2})$. This implies that $|x_1|$ is at most $O(F/d^{3/2})$ for all $x \in \mathbb{A}K_1$: this therefore is also true for all $x \in \mathbb{A}K_2$. The intersection of the ellipsoid $\mathbb{A}E_2$ with the set of points

with first coordinate at most $O(F/d^{3/2})$ is contained in the ellipsoid E which shrinks the first axis of $\mathbb{A}E_2$ by a factor of 10 and grows all other directions by $1 + \frac{1}{d}$. So, we can replace $\mathbb{A}E_2$ with another ellipsoid $E \supset \mathbb{A}K_2$ with volume at most $\frac{e}{10} \leq 0.5$ times the volume of $\mathbb{A}E_2$.

Therefore, unless $2d^3 \cdot \mathbb{A}E_1 \supset \mathbb{A}E_2$, we can find either a new larger E_1 or a new smaller E_2 in polynomial time. Each time this takes $\mathrm{poly}(d, D, \log \gamma_2^{-1}, \log \gamma_3^{-1}) = \mathrm{poly}(d, \log \gamma_2^{-1}, \log \gamma_3^{-1})$ time. However, the volume ratio of the original ellipsoids $B(\mathbf{0}, r)$ and $B(\mathbf{0}, R)$ is $(R/r)^d$, so we can only repeat this process at most $O(d \log \frac{R}{r})$ times. □

We combine Lemma A.7 and Lemma A.8 to obtain the following corollary.

Corollary A.9. *For any parameters $r, \gamma_2, \gamma_3 < 1 < R$, there exists some γ_1 such that $\log \gamma_1^{-1} = \mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_2^{-1}, \log \gamma_3^{-1})$ and the following holds. If K_1, K_2 are convex bodies such that $B(\mathbf{0}, r) \subset K_1 \subset K_2 \subset B(\mathbf{0}, R)$ and $\mathrm{vol}(K_2) - \mathrm{vol}(K_1) \leq \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$, then there is a $\mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_2^{-1}, \log \gamma_3^{-1})$-time algorithm that can sample from a distribution that is (γ_2, γ_3)-close to uniform on K_1.*

Proof. First, use Lemma A.8 to find an affine transformation $\mathbb{A}$ such that $\mathbb{A}$ applied to K_1 contains $B(\mathbf{0}, 1)$ but is contained in $B(\mathbf{0}, 2d^3)$. Then, if we define $\gamma_2' = \frac{\gamma_2}{(R/r) \cdot d^3}$, we can produce a sample that is (γ_2', γ_3)-close to uniform on $\mathbb{A}K_1$. Finally, undo the affine transformation and the sample will still be (γ_2, γ_3)-close to uniform. □

Unfortunately, being (γ_2, γ_3)-close to uniform does not necessarily ensure privacy. This is because one may extract information about the data based on minor perturbations of the generated sample. To fix this, we convert this version of closeness to pointwise closeness to uniform on a fine grid of points.

Lemma A.10. ***(Convex body sampling, Lemma 4.4)*** *Fix any parameters $\gamma_6 \leq d^{-100}$ and $r < 1 < R$. Let K_1, K_2 be convex bodies such that $B(\mathbf{0}, r) \subset K_1 \subset K_2 \subset B(\mathbf{0}, R)$, and $\mathrm{vol}(K_2) - \mathrm{vol}(K_1) \leq \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$, for γ_1 that will be defined in terms of γ_6. Suppose we have a (K_1, K_2)-membership oracle $\boldsymbol{O}$. Then, in $\mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_6^{-1})$ time and queries to $\boldsymbol{O}$, we can output a point z that is $(1 \pm \gamma_6)$-pointwise close to uniform on the set of points in $\mathbb{R}^d$ with all coordinates integer multiples of γ_5 that are accepted by $\boldsymbol{O}$, for $\gamma_5 = \frac{r \cdot \gamma_6}{d^3}$.*

Proof. First, we will define parameters γ_1 through γ_5 based on r, R, and γ_6. Define $\gamma_4 := \frac{r}{d^2}$ and $\gamma_5 := \frac{\gamma_4 \cdot \gamma_6}{d}$. Next, define $\gamma_2 := \frac{\gamma_5 \gamma_6}{d^2}$ and $\gamma_3 := \left(\frac{\gamma_5}{2R}\right)^d \cdot \frac{\gamma_6}{d}$. Finally, define γ_1 to be the value for γ_1 that appears when applying Corollary A.9 with γ_2 and γ_3.

Let $K_1' = (1 + \frac{1}{d})K_1$ and $K_2' = (1 + \frac{1}{d})K_2$. Let $\mathcal{D}_1$ be the uniform distribution over K_1'. By applying Corollary A.9 on (K_1, K_2) and then scaling the point by $1 + \frac{1}{d}$, we obtain a point $c \sim \mathcal{D}_2$, where $\mathcal{D}_2$ is (γ_2, γ_3)-close to $\mathcal{D}_1$.

Our algorithm works as follows. First, replace c with $c + y$, where each coordinate y_i was uniformly chosen from $[-\gamma_4, \gamma_4]$ with precision γ_1. Then, round each coordinate of $c + y$ to the nearest multiple of γ_5 to get a point z. Finally, we run a rejection sampling algorithm by checking whether the (K_1, K_2)-membership oracle accepts z. If so, we return z. If not, we restart the procedure until we accept some z. It will be simple to see that each step of the rejection sampling algorithm succeeds with probability $(1 - \Omega(\frac{1}{d}))^d \geq \Omega(1)$ because z will be in K_1 with this probability, so we can stop the rejection sampling after $O(\log \gamma_3^{-1})$ steps, to incur additional additive error γ_3.

We now analyze the accuracy. Let $\mathcal{D}_3$ be a distribution so that we have a coupling between $\mathcal{D}_1, \mathcal{D}_3, \mathcal{D}_2$ such that if $(a, b, c) \sim \mathcal{D}_1, \mathcal{D}_3, \mathcal{D}_2$, then $\|a-b\|_2 \le \gamma_2$ with probability 1, and $\mathbb{P}(b \neq c) \le \gamma_3$. Now, for any point z with all coordinates multiples of γ_5 such that the (K_1, K_2)-membership oracle accepts z, we compute the probability of sampling z. In order to sample z, we must have sampled c and y so that $c + y$ rounds to z. This probability is the same as the probability that $b + y$ rounds to z, up to additive error γ_3. If we condition on choosing a such that $\|a - z\|_\infty \le \gamma_4 - (\gamma_1 + \gamma_2 + \gamma_5)$, then $\|b - z\|_\infty \le \gamma_4 - (\gamma_1 + \gamma_5)$, so if each coordinate y_i were chosen uniformly from $[-\gamma_4, \gamma_4]$ with perfect precision, the probability that $b + y$ rounds to z will exactly be $(\gamma_5/(2\gamma_4))^d$. Due to precision issues, the actual probability that $b + y$ rounds to z is $((\gamma_5 \pm O(\gamma_1))/(2\gamma_4))^d$. Likewise, if we choose a such that $\|a - z\|_\infty \ge \gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)$, then $\|b - z\|_\infty \ge \gamma_4 + (\gamma_1 + \gamma_5)$, so we will never select $b + y$ to round to z. Finally, if $\gamma_4 - (\gamma_1 + \gamma_2 + \gamma_5) \le \|a - z\|_\infty \le \gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)$, then the probability that $b + y$ rounds to z is between 0 and $((\gamma_5 + O(\gamma_1))/(2\gamma_4))^d$.

Since a is truly uniform from $K_1' = (1 + \frac{1}{d})K_1$, we claim that the probability of selecting an a with $\|a - z\|_\infty \le \gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)$ is $(2(\gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)))^d / \mathrm{vol}(K_1')$. For this to be true, we need every point a with $\|a - z\|_\infty \le \gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)$ to be in K_1'. Since O accepts z, this means $z \in K_2 \subset (1 + \gamma_1)K_1$, so every point within ℓ_2 distance $(\frac{1}{d} - \gamma_1) \cdot r$ of z is contained in $(1 + \frac{1}{d})K_1 = K_1'$. So, it suffices for $\sqrt{d} \cdot (\gamma_1 + \gamma_2 + \gamma_4 + \gamma_5) \le (\frac{1}{d} - \gamma_1) \cdot r$. Likewise, the probability of selecting an a with $\|a - z\|_\infty \le \gamma_4 - \gamma_5 - \gamma_2$ is $(2(\gamma_4 - \gamma_5 - \gamma_2))^d / \mathrm{vol}(K_1')$.

So, the overall probability that $b + y$ rounds to z is at least $(2(\gamma_4 - (\gamma_1 + \gamma_2 + \gamma_5)))^d / \mathrm{vol}(K_1') \cdot ((\gamma_5 - O(\gamma_1))/(2\gamma_4))^d$ and at most $(2(\gamma_4 + (\gamma_1 + \gamma_2 + \gamma_5)))^d / \mathrm{vol}(K_1') \cdot ((\gamma_5 + O(\gamma_1))/(2\gamma_4))^d$. Assuming that $d \cdot \gamma_1, \gamma_2 \ll \gamma_5$ and $d \cdot \gamma_5 \ll \gamma_4$, these bounds equal $\frac{\gamma_5^d}{\mathrm{vol}(K_1')} \cdot \left(1 \pm O\left(\frac{d \cdot \gamma_5}{\gamma_4} + \frac{d \cdot \gamma_1}{\gamma_5}\right)\right)$. We also need that $\gamma_4 \ll \frac{r}{d\sqrt{d}}$, so that $\sqrt{d} \cdot (\gamma_1 + \gamma_2 + \gamma_4 + \gamma_5) \le (\frac{1}{d} - \gamma_1) \cdot r$. Finally, we had an additive error of γ_3 due to the coupling of the points b and c, as well as another γ_3 for the rejection algorithm failing. So, the final probability of choosing some point z with all coordinates integer multiples of γ_5 that is accepted by the (K_1, K_2)-membership oracle is $\left(\frac{d}{d+1}\right)^d \cdot \frac{\gamma_5^d}{\mathrm{vol}(K_1)} \cdot \left(1 \pm O\left(\frac{d \cdot \gamma_5}{\gamma_4} + \frac{d \cdot \gamma_1}{\gamma_5}\right)\right) \pm O(\gamma_3)$, where we used the fact that $\mathrm{vol}(K_1') = \mathrm{vol}(K_1) \cdot \left(1 + \frac{1}{d}\right)^d$.

Based on how we set $\gamma_1, \ldots, \gamma_5$, all the conditions hold, and we can simplify the probability as $\left(\frac{d}{d+1}\right)^d \cdot \frac{\gamma_5^d}{\mathrm{vol}(K_1)} \cdot \left(1 \pm O\left(\frac{\gamma_6}{d}\right)\right) \pm \left(\frac{\gamma_5}{2R}\right)^d \cdot \gamma_6$. However, since $\mathrm{vol}(K_1) \le (2R)^d$ and $\left(\frac{d}{d+1}\right)^d \ge \frac{1}{e}$, in total this equals $\left(\frac{d}{d+1}\right)^d \cdot \frac{\gamma_5^d}{\mathrm{vol}(K_1)} \cdot \left(1 \pm O\left(\frac{\gamma_6}{d}\right)\right)$. So, our sampling algorithm is pointwise accurate up to a $1 \pm O\left(\frac{\gamma_6}{d}\right)$ factor. □

Finally, we show that our sampling algorithm can also allow us to approximately compute the volume of points accepted by the oracle O. More accurately, we can approximate the number of points in the grid of precision γ_5 that are accepted by O.

Lemma A.11. ***(Volume sampling, Lemma 4.5)*** *Let all notation be as in Lemma 4.4. Fix any $\varepsilon < 0.5$, and set $\gamma_6 \le \frac{\varepsilon}{d \log \frac{R}{r}}$ and $\gamma_1, \ldots, \gamma_5$ in terms of γ_6 as in Lemma 4.4. Then, for any $\gamma < 1$, in $\mathrm{poly}(d, \log \frac{R}{r}, \frac{1}{\varepsilon}, \log \gamma^{-1})$ time and oracle accesses, we can approximate the number of points in $\mathbb{R}^d$ with all coordinates integer multiples of γ_5 that are accepted by O, up to a $1 \pm \varepsilon$ multiplicative factor, with failure probability γ.*

Proof. For some $\rho \in [r, R]$, let $K_1^{(\rho)} = K_1 \cap B(\mathbf{0}, \rho)$ and $K_2^{(\rho)} = K_2 \cap B(\mathbf{0}, \rho)$. Clearly, $B(\mathbf{0}, r) \subset K_1^{(\rho)} \subset$

$K_2^{(\rho)} \subset B(\mathbf{0}, R)$, and $\mathrm{vol}(K_2^{(\rho)}) - \mathrm{vol}(K_1^{(\rho)}) \le \left(\frac{\gamma_1 \cdot r}{6d}\right)^d$. Also, let $S^{(\rho)}$ be the set of points in $K_2^{(\rho)}$ with all coordinates multiples of γ_5 that are accepted by the oracle, and let $N^{(\rho)} := |S^{(\rho)}|$.

Since $\gamma_2 \le \gamma_5 \le \frac{r}{d^3}$, we have that $\mathrm{vol}(K_1^{(\rho)}) = (1 \pm o(1)) \cdot N^{(\rho)} \cdot (\gamma_5)^d$. To see why, suppose x is a point that, after rounding each coordinate to the nearest multiple of γ_5, is in $(1 - \frac{\gamma_5\sqrt{d}}{r}) \cdot K_1^{(\rho)}$. Then, since x moved by at most $\gamma_5\sqrt{d}$ in absolute value, and since $B(\mathbf{0}, r) \subset K_1^{(\rho)}$, x must be in $K_1^{(\rho)}$ and so is accepted by the oracle. Therefore, $(\gamma_5)^d \cdot N^{(\rho)} \ge (1 - \frac{\gamma_5\sqrt{d}}{r})^d \cdot \mathrm{vol}(K_1^{(\rho)})$ which means $\mathrm{vol}(K_1^{(\rho)}) \le (1 + o(1)) \cdot N^{(\rho)} \cdot (\gamma_5)^d$. For the other direction, any point that is in $K_2^{(\rho)} \subset (1 + \gamma_2)K_1^{(\rho)}$, if we change each coordinate by up to γ_5, is still in $(1 + \gamma_2) \cdot (1 + \frac{\gamma_5\sqrt{d}}{r}) \cdot K_1^{(\rho)}$. Therefore $(\gamma_5)^d \cdot N^{(\rho)} \le \mathrm{vol}(K_1^{(\rho)}) \cdot (1 + o(1))$.

Now, if $\rho' \le (1 + \frac{1}{d})\rho$, note that $K_2^{(\rho')} \subset (1 + \frac{1}{d})K_2^{(\rho)}$. Therefore, this means $\mathrm{vol}(K_1^{(\rho')}) \le (e + o(1)) \cdot \mathrm{vol}(K_1^{(\rho)})$, which means that $N^{(\rho')} \le (e + o(1)) \cdot N^{(\rho)}$. Given this, by Lemma 4.4, we can generate $1 \pm \gamma_6$-pointwise random samples from $S^{(\rho')}$ and check the fraction of them that are in $S^{(\rho)}$ by determining for each sample if its ℓ_2 norm is at most ρ. By Hoeffding's inequality, for any $\varepsilon' < 1$ we can compute $\frac{N^{(\rho)}}{N^{(\rho')}}$ with failure probability γ up to an additive error of $\pm O(\varepsilon' + \gamma_6)$ in $O((\varepsilon')^{-2} \log \gamma^{-1})$ random samples, and since $1 \le \frac{N^{(\rho')}}{N^{(\rho)}} \le e + o(1)$, this also implies we can compute the ratio up to a multiplicative factor of $1 \pm O(\varepsilon')$, assuming $\gamma_6 \le \varepsilon'$.

Now, consider $r = \rho_0, \rho_1, \rho_2, \ldots, \rho_M = R$, where $\frac{\rho_1}{\rho_0} \le 1 + \frac{1}{d}$. We can let $M = O(d \log \frac{R}{r})$. Then, by setting $\varepsilon' = \frac{\varepsilon}{M}$ we can compute $\frac{N^{(\rho_{t+1})}}{N^{(\rho_t)}}$ up to multiplicative error $e^{O(\varepsilon/M)}$ in $\mathrm{poly}\left(d, \log \frac{R}{r}, \log \gamma_6^{-1}, \log \gamma_1^{-1}, \frac{1}{\varepsilon}\right)$ time, with failure probability $\frac{\gamma_1}{M}$. By multiplying all of our estimates to form a telescoping product, we can compute $\frac{N^{(R)}}{N^{(r)}}$ up to a multiplicative factor $e^{\pm O(\varepsilon)}$ with failure probability γ_1. Our goal is precisely to compute $N^{(R)}$, so it suffices to compute $N^{(r)}$. But since $B(\mathbf{0}, r) \subset K_1$, this is just the number of points with all coordinates integral multiples of γ_5 that are in $B(\mathbf{0}, r)$. By the argument of the above paragraph, this is just $\gamma_5^{-d} \cdot \mathrm{vol}(B(\mathbf{0}, r)) \cdot \left(1 \pm \frac{\gamma_5\sqrt{d}}{r}\right)^d = \left(\frac{r}{\gamma_5}\right)^d \cdot e^{\pm\gamma_5 \cdot d^{3/2}/r} \cdot \mathrm{vol}(B(\mathbf{0}, 1))$. Since $\gamma_5 = \frac{r \cdot \gamma_6}{d^3}$, $\gamma_5 \cdot d^{3/2}/r \le \gamma_6 \le \varepsilon$. Therefore, since the volume of a d-dimensional sphere has an explicit representation, we can compute $N^{(R)}$ up to multiplicative error $e^{\pm O(\varepsilon)}$ in time $\mathrm{poly}(d, \log \frac{R}{r}, \log \gamma_6^{-1}, \frac{1}{\varepsilon}, \log \gamma^{-1}) = \mathrm{poly}(d, \log \frac{R}{r}, \frac{1}{\varepsilon}, \log \gamma^{-1})$ time. □

B Sum-of-squares proofs

In this section, we prove sum-of-squares proofs that are crucial in establishing accuracy of our algorithms, as well as privacy in the approx-DP setting. These include both results both when the data points are sampled from a Gaussian, and for worst-case results. Due to precision issues when solving a semidefinite program, our bounds must hold with respect to not only all pseudo-expectations but also with respect to linear operators that are "approximate pseudoexpectations". The exponentially-small numerical errors this introduces are manageable by observing that the coefficients in the SoS proofs we use to analyze these approximate pseudoexpectations are at most some fixed polynomial in the bit-representation of the input; see e.g. the discussion in [HL18b].

In Appendix B.1, we recall the sum-of-squares results from [KMZ22], and use these to establish accuracy for Gaussian data. Namely, we show that a low-scoring point with respect to samples

drawn from a Gaussian (or more generally for samples with the required resilience samples) must be a good estimate for the mean/covariance of the Gaussian. Next, we prove two sum-of-squares results showing that any set of data points, no matter how corrupted, cannot have a very large volume of potential means (or covariances) which all have low scores. This differs from accuracy results proven in prior work, which assume that a large fraction of the points come from some distribution. This establishes a "worst-case accuracy" result, which is crucial to establishing privacy in our approx-DP algorithms. We prove a result for covariance estimation in Appendix B.2 and a result for mean estimation in Appendix B.3.

B.1 Proofs of Accuracy Lemmas

In this subsection, we prove the accuracy results for mean and covariance estimation (Lemmas 5.10, 6.13, and 7.8).

The main sum-of-squares result that we apply is the following lemma due to Kothari, Manohar, and Zhang.

Lemma B.1. *[KMZ22, Lemma 4.1, restated] Let $x_1, \dots, x_n \in \mathbb{R}^d$, and let $\mu_0 = \frac{1}{n}\sum_{i=1}^n x_i$ be the sample mean. Let $V(\mu, v)$ for $v \in \mathbb{R}^d$ be a degree at most 2 polynomial in μ, that is always nonnegative for all $\mu \in \mathbb{R}^d$ and v in some fixed subset $\mathcal{S} \subset \mathbb{R}^d$. Suppose that for all vectors $a \in [0,1]^n$ with $\sum_{i=1}^n a_i \geq (1-\eta)n$, and for all $v \in \mathcal{S}$, we have*

$$\left|\frac{1}{n}\sum_{i=1}^n a_i \langle x_i - \mu_0, v\rangle\right| \leq \widetilde{O}(\eta)\cdot\sqrt{V(\mu_0, v)} \quad \text{and} \quad \left|\frac{1}{n}\sum_{i=1}^n a_i[\langle x_i - \mu_0, v\rangle^2 - V(\mu_0, v)]\right| \leq \widetilde{O}(\eta)\cdot V(\mu_0, v).$$

Let $\tilde{\mathbf{E}}$ be a degree-6 pseudoexpectation on $\{x_i'\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ such that

1. *$\forall i \in [n]$, $\tilde{\mathbf{E}}$ satisfies $w_i^2 = w_i$,*
2. *$\forall i \in [n]$, $\tilde{\mathbf{E}}$ satisfies $w_i x_i' = w_i x_i$,*
3. *$\tilde{\mathbf{E}}$ satisfies $\sum_{i=1}^n w_i \geq (1-\eta)n$,*
4. *For all $v \in \mathcal{S}$, $\tilde{\mathbf{E}}\left[\frac{1}{n}\sum_{i=1}^n \langle x_i' - \mu', v\rangle^2\right] \leq (1 + \widetilde{O}(\eta))\cdot \tilde{\mathbf{E}}[V(\mu', v)]$, where $\mu' := \frac{1}{n}\sum_{i=1}^n x_i'$.*

Then, for every unit vector $v \in \mathcal{S}$, the following two inequalities hold:

$$\tilde{\mathbf{E}}\left[\langle \mu' - \mu_0, v\rangle^2\right] \leq O(\eta)\cdot(\tilde{\mathbf{E}}[V(\mu', v)] + V(\mu_0, v)).$$
$$|\langle \tilde{\mathbf{E}}[\mu'] - \mu_0, v\rangle| \leq \widetilde{O}(\eta)\cdot\sqrt{V(\mu_0, v) + \tilde{\mathbf{E}}[V(\mu', v)]} + \sqrt{\widetilde{O}(\eta)\cdot(\tilde{\mathbf{E}}[V(\mu', v)] - V(\mu_0, v))}.$$

We first prove Lemma 5.10.

Proof of Lemma 5.10. Since $\phi \leq \alpha/\sqrt{d}$, it suffices to show that any $(\alpha^*, \tau, \phi, T)$-certificate $\mathcal{L}$ for X satisfies $\|\mathcal{L}[\mu'] - \mu\| \leq O(\alpha)$. If we assume $\tau = 0$ and $\alpha = \widetilde{O}(\eta)$, then in fact $\mathcal{L}$ is a degree-6 pseudoexpectation that precisely satisfies the four required conditions of Lemma B.1, if we set $V(\mu', v) := 1$ and define $\mathcal{S}$ to be the set of unit vectors in $\mathbb{R}^d$. In addition, by Corollary 5.4, the required conditions on x_i hold up to replacing α with 2α and the sample mean μ_0 with the true

mean μ. However, Corollary 5.4 implies that $\|\mu - \mu_0\|_2 \le \alpha$, so $|\langle x_i - \mu_0, v\rangle - \langle x_i - \mu, v\rangle| \le \alpha$ and $|\langle x_i - \mu_0, v\rangle^2 - \langle x_i - \mu, v\rangle^2| \le \alpha(1 + |\langle x_i - \mu, v\rangle|)$. But $\frac{1}{n}\sum_{i=1}^n |\langle x_i - \mu, v\rangle| \le O(1)$. Together this means that the conditions of Lemma B.1 hold up to replacing α with $O(\alpha)$.

Therefore, for any unit vector v, $|\langle \mathcal{L}[\mu'] - \mu, v\rangle| \le \widetilde{O}(\eta) \le O(\alpha)$ by Lemma B.1, as desired.

While our proof was for exact pseudoexpectations since we set $\tau = 0$, as mentioned in [KMZ22], the proof also extends to approximate pseudoexpectations for small τ, since the coefficients at each step in the sum-of-squares proof are polynomially bounded (see, e.g., the discussion in [HL18b] or [KMZ22]). Here, we must make the assumption that every x_i has magnitude bounded by $(ndR)^{O(1)}$, which holds automatically assuming the resilience properties. □

Next, we prove Lemma 6.13.

Proof of Lemma 6.13. Given samples $x_1, \dots, x_n$ and d-dimensional indeterminates $x'_1, \dots, x'_n$, we define the indeterminates $z'_1, \dots, z'_n$ as $z'_i := (x'_i)(x'_i)^\top$ (note that each z'_i is a $d \times d$-dimensional matrix), and $\Sigma' := \frac{1}{n}\sum z'_i$. We also define $z_i := x_i x_i^\top$ and $\Sigma_0 := \frac{1}{n}\sum z_i$.

We will apply Lemma B.1, but replacing d with d^2, x_i with z_i, x'_i with z'_i, μ_0 with Σ_0, and μ' with Σ'. We also define $\mathcal{S}$ to be the subset of vectors of the form $vv^\top$ where v is a d-dimensional unit vector. (Note that $vv^\top$ is d^2-dimensional and has ℓ_2 norm 1 when flattened). Finally, for $\Sigma, M \in \mathbb{R}^{d\times d}$, we define $V(\Sigma, M) := 2 \cdot \langle \Sigma, M\rangle^2$.

Now, for any (α^*, τ, T)-certificate $\mathcal{L}$ with $\alpha^* \le \alpha$, it suffices to show that for any unit vector $v \in \mathbb{R}^d$, $(1 - O(\alpha))v^\top \Sigma v \le \mathcal{L}[v^\top \Sigma' v] \le (1 + O(\alpha))v^\top \Sigma v$. This would imply that $(1 - O(\alpha))\Sigma \preccurlyeq \mathcal{L}[\Sigma'] \preccurlyeq (1 + O(\alpha))\Sigma$, which means for $\tau \ll 1/\mathrm{poly}(n, d, K)$, $(1 - O(\alpha))\Sigma \preccurlyeq \widetilde{\Sigma} \preccurlyeq (1 + O(\alpha))\Sigma$.

We start by assuming $\tau = 0$, so $\mathcal{L}$ is actually a degree-12 pseudoexpectation. Then, $\mathcal{L}$ satisfies $w_i^2 = w_i$, $w_i(x'_i)(x'_i)^\top = w_i x_i x_i^\top$, and $\sum w_i \ge (1 - \eta)n$. In addition, since $\mathcal{L}\left[\|(v^{\otimes 2})^\top M^\top M v^{\otimes 2}\|_2^2\right] = \mathcal{L}\left[\|Mv^{\otimes 2}\|_2^2\right] \ge 0$, this means

$$\mathcal{L}\left[\frac{1}{n}\sum_{i=1}^n \left(\langle x'_i, v\rangle^2 - v^\top \Sigma' v\right)^2\right] \le (2 + \widetilde{O}(\eta)) \cdot \mathcal{L}\left[(v^\top \Sigma' v)^2\right].$$

But note that $V(\mu', v)$ is precisely replaced with $2 \cdot \langle \Sigma', vv^\top\rangle^2 = 2(v^\top \Sigma' v)^2$. In addition, $\langle x'_i, v\rangle^2 - v^\top \Sigma' v = \langle z'_i - \Sigma', vv^\top\rangle$. Hence, the 4 conditions in Lemma B.1 are satisfied.

In addition, to apply Lemma B.1 we need to verify the desired conditions for z_i. By the resilience properties (Lemma 6.3) of $\{\Sigma^{-1/2}x_i\}$, we have that $(1-\alpha)\|v\|_2^2 \le \frac{1}{n}\sum v^\top \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} v \le (1 + \alpha)\|v\|_2^2$, which means by replacing v with $\Sigma^{1/2}v$, we have

$$(1 - \alpha)(v^\top \Sigma v) \le v^\top \Sigma_0 v \le (1 + \alpha)v^\top \Sigma v. \tag{6}$$

Now, for any unit vector v and $a_1, \dots, a_n \in [0, 1]$ with $\sum a_i \ge (1-\eta)n$, $\left|\frac{1}{n}\sum_{i=1}^n a_i(\langle v, \Sigma^{-1/2}x_i\rangle^2 - 1)\right| = \left|\frac{1}{n}\sum_{i=1}^n a_i\langle (\Sigma^{-1/2}x_i)(\Sigma^{-1/2}x_i)^\top - I, vv^\top\rangle\right| \le \widetilde{O}(\eta)$ if $\{\Sigma^{-1/2}x_i\}$ satisfy the resilience properties. By scaling, for general vectors v, $\left|\frac{1}{n}\sum_{i=1}^n a_i(\langle v, \Sigma^{-1/2}x_i\rangle^2 - \|v\|_2^2)\right| \le \widetilde{O}(\eta) \cdot \|v\|_2^2$, which means by replacing v with $\Sigma^{1/2}v$, we have $\left|\frac{1}{n}\sum_{i=1}^n a_i(\langle v, x_i\rangle^2 - v^\top \Sigma v)\right| \le \widetilde{O}(\eta) \cdot v^\top \Sigma v$. By (6), this implies

$$\left|\frac{1}{n}\sum_{i=1}^n a_i\langle vv^\top, z_i - \Sigma_0\rangle\right| \le \widetilde{O}(\eta) \cdot \langle vv^\top, \Sigma_0\rangle.$$

Next, note that

$$\begin{aligned}\langle z_i - \Sigma_0, vv^\top\rangle^2 &= \langle z_i - \Sigma, vv^\top\rangle^2 + 2\langle \Sigma - \Sigma_0, vv^\top\rangle \cdot \langle z_i - \Sigma, vv^\top\rangle + \langle \Sigma - \Sigma_0, v^\top\rangle^2 \\ &= \langle z_i - \Sigma, vv^\top\rangle^2 \pm O(\alpha) \cdot (v^\top \Sigma v) \cdot |\langle z_i - \Sigma, vv^\top\rangle| \pm O(\alpha^2) \cdot (v^\top \Sigma v)^2.\end{aligned}$$

We can rewrite $\langle z_i - \Sigma, vv^\top\rangle = \langle \Sigma^{-1/2} x_i, \Sigma^{1/2} v\rangle^2 - \|\Sigma^{1/2} v\|_2^2$. This means by applying Lemma 6.3 with $P = (\Sigma^{1/2} v)(\Sigma^{1/2} v)^\top$, we have that

$$\frac{1}{n}\sum_{i=1}^n \langle z_i - \Sigma, vv^\top\rangle^2 = (2 \pm O(\alpha)) \cdot \|\Sigma^{1/2} v\|_2^4 = (2 \pm O(\alpha)) \cdot (v^\top \Sigma v)^2.$$

and

$$\frac{1}{n}\sum_{i=1}^n \left|\langle z_i - \Sigma, vv^\top\rangle\right| \le O(1) \cdot \|\Sigma^{1/2} v\|_2^2 = O(1) \cdot (v^\top \Sigma v).$$

Together, this implies that

$$\left|\frac{1}{n}\sum_{i=1}^n a_i \langle vv^\top, z_i - \Sigma_0\rangle^2\right| = (2 \pm O(\alpha)) \cdot (v^\top \Sigma v)^2 = (1 \pm O(\alpha)) \cdot V(\Sigma_0, vv^\top).$$

Since $\sum a_i \ge (1-\eta)n$, this completes the verification of the conditions.

Now, we may apply Lemma B.1. We first have that

$$\mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] \le O(\eta) \cdot \left(\mathcal{L}[(v^\top \Sigma' v)^2] + \mathcal{L}[(v^\top \Sigma_0 v)^2]\right).$$

By Cauchy-Schwarz, we know that

$$\mathcal{L}[(v^\top \Sigma' v)^2] \le 2 \cdot \left(\mathcal{L}[(v^\top \Sigma_0 v)^2] + \mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2]\right),$$

which means

$$\mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] \le O(\eta) \cdot \left(\mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] + \mathcal{L}[(v^\top \Sigma_0 v)^2]\right)$$

and therefore,

$$\mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] \le O(\eta) \cdot (v^\top \Sigma_0 v)^2$$

since v, Σ_0 are fixed determinates. This also implies that $\mathcal{L}[(v^\top \Sigma' v)^2] \le O(1) \cdot (v^\top \Sigma_0 v)^2$.

Let $A := v^\top \Sigma_0 v$, and $B := \mathcal{L}[v^\top(\Sigma' - \Sigma_0)v]$. Then, $V(\Sigma_0, vv^\top) = 2A^2$, $\mathcal{L}[V(\Sigma', vv^\top)] = O(A^2)$, and $\mathcal{L}[V(\Sigma', vv^\top)] - V(\Sigma_0, vv^\top) = 2 \cdot \mathcal{L}[(v^\top \Sigma' v)^2 - (v^\top \Sigma_0 v)^2] = 2 \cdot \mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] + 4A \cdot B$. In addition, we know that $\mathcal{L}[(v^\top(\Sigma' - \Sigma_0)v)^2] \le O(\eta) \cdot A^2$. Hence, Lemma B.1 implies that

$$|B| \le \widetilde{O}(\eta) \cdot A + \sqrt{\widetilde{O}(\eta) \cdot \widetilde{O}(\eta) \cdot A^2 + \widetilde{O}(\eta) \cdot A \cdot B} \le \widetilde{O}(\eta) \cdot A + \sqrt{\widetilde{O}(\eta)} \cdot \sqrt{A \cdot |B|}.$$

This means that $|B| \le \widetilde{O}(\eta) \cdot A$, which means that $\mathcal{L}[v^\top \Sigma' v] = (1 \pm \widetilde{O}(\eta)) \cdot v^\top \Sigma_0 v = (1 \pm O(\alpha)) \cdot v^\top \Sigma v$, where the last equation follows by (6).

This completes the proof for true pseudoexpectations. Again, the proof extends to approximate pseudoexpectations, since the coefficients at each step in the sum-of-squares proof are polynomially bounded. □

Finally, we prove Lemma 7.8.

Proof of Lemma 7.8. As in Lemma 6.13, we apply Lemma B.1 with some replacements. This time, we replace d with d^2, x_i with $z_i = x_i^{\otimes 2}$, x_i' with $z_i' = (x_i')^{\otimes 2}$, μ_0 with $S_0 = \frac{1}{n}\sum_i z_i$, and μ' with $S' = \frac{1}{n}\sum_i z_i'$. In addition, the set $\mathcal{S} \subset \mathbb{R}^{d^2}$ will represent all vectors P of norm 1 such that the $d \times d$ matrix M such that $M^\flat = P$ is symmetric. Finally, we define $V(S, P) := 2$.

For any $(\alpha^*, \tau, \phi, T)$-certificate $\mathcal{L}$ with $\phi \le \alpha/d$ and $\alpha^* \le \alpha$, it suffices to show that $\|\mathcal{L}[\Sigma'] - \Sigma\|_F \le \alpha$, since $\mathcal{L}[\Sigma']_{j,k} - \widetilde{\Sigma}_{j,k} \le O(\alpha/d)$ for all indices $j, k \le d$.

We again assume $\tau = 0$, so $\mathcal{L}$ is actually a degree-12 pseudoexpectation. Then, $\mathcal{L}$ satisfies $w_i^2 = w_i$, $\sum w_i \ge (1-\eta)n$, and $w_i(x_i')^{\otimes 2} = w_i x_i^{\otimes 2}$. Next, for any $P \in \mathcal{S}$, $\langle z_i' - S', P\rangle^2 = P^\top((x_i')^{\otimes 2} - S')((x_i')^{\otimes 2} - S')^\top P$, and we are assuming $\mathcal{L}[((x_i')^{\otimes 2} - S')((x_i')^{\otimes 2} - S')^\top] \preccurlyeq (2+\alpha)\cdot\mathcal{L}[I] = (2+\alpha)\cdot I$, where I refers to the $d^2 \times d^2$-identity matrix. Hence, $\mathcal{L}[\langle z_i' - S', P\rangle^2] \le 2 + \alpha \le (1 + \widetilde{O}(\eta))\cdot\mathcal{L}[V(S', P)]$ since $V \equiv 2$, so the 4 conditions in Lemma B.1 are satisfied.

Next, we must verify the desired conditions for z_i. Note that $\langle x_i^{\otimes 2} - S_0, P\rangle = \langle x_i x_i^\top - \Sigma_0, P^\sharp\rangle$ (where $P^\sharp$ is the symmetric matrix that flattens to P). Also, note that $\langle x_i x_i^\top - \Sigma, P^\sharp\rangle = \langle \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - I, \Sigma^{1/2} P^\sharp \Sigma^{1/2}\rangle$. Writing $Q = \Sigma^{1/2} P^\sharp \Sigma^{1/2}$, by Proposition 6.8 we have that $\|Q\|_F = 1 \pm O(\alpha)$. This implies, using the resilience of $\{\Sigma^{-1/2} x_i\}$ (Lemma 6.3) that

$$\left|\frac{1}{n}\sum_{i=1}^n a_i \langle x_i x_i^\top - \Sigma, P^\sharp\rangle\right| = \left|\frac{1}{n}\sum_{i=1}^n a_i \langle \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - I, Q\rangle\right| \le O(\alpha),$$
$$\frac{1}{n}\sum_{i=1}^n a_i \langle x_i x_i^\top - \Sigma, P^\sharp\rangle^2 = \frac{1}{n}\sum_{i=1}^n a_i \langle \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - I, Q\rangle^2 = 2 \pm O(\alpha).$$

In addition, note that $\|\Sigma - \Sigma_0\|_F \le \alpha$ due to the resilience guarantees (Lemma 6.3), which means $\langle x_i x_i^\top - \Sigma_0, P^\sharp\rangle = \langle x_i x_i^\top - \Sigma, P^\sharp\rangle \pm \alpha$. In addition, Lemma 6.3 implies that $\frac{1}{n}\sum \left|\langle x_i x_i^\top - \Sigma, P^\sharp\rangle\right| = \frac{1}{n}\sum \left|\langle \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - I, Q\rangle\right| \le O(1)$. This immediately implies that

$$\left|\frac{1}{n}\sum_{i=1}^n a_i \langle z_i - S_0, P\rangle\right| = \left|\frac{1}{n}\sum_{i=1}^n a_i \langle x_i x_i^\top - \Sigma_0, P^\sharp\rangle\right| \le O(\alpha),$$
$$\frac{1}{n}\sum_{i=1}^n a_i \langle z_i - S_0, P\rangle^2 = \frac{1}{n}\sum_{i=1}^n a_i \langle x_i x_i^\top - \Sigma_0, P^\sharp\rangle^2 = 2 \pm O(\alpha).$$

Since $V \equiv 2$, this immediately implies we can apply Lemma B.1. Doing so, we obtain $|\langle \mathcal{L}[S'] - S_0, P\rangle| = |\langle \mathcal{L}[\Sigma'] - \Sigma_0, P^\sharp\rangle| \le \widetilde{O}(\eta)$ for all symmetric $P^\sharp$ with $\|P^\sharp\|_F = 1$. Hence, $\|\mathcal{L}[\Sigma'] - \Sigma_0\|_F \le O(\alpha)$, which means $\|\mathcal{L}[\Sigma'] - \Sigma\|_F \le O(\alpha)$ as well.

This completes the proof for true pseudoexpectations. Again, the proof extends to approximate pseudoexpectations, since the coefficients at each step in the sum-of-squares proof are polynomially bounded. □

B.2 SoS bounds for arbitrary samples: Covariance estimation

In this subsection, we prove Lemma 6.15, which is our worst-case robustness result for covariance estimation. First, we establish a 1-dimensional Sum-of-Squares result that will be crucial in proving Lemma 6.15.

Lemma B.2. *Let $z_1, \dots, z_n$ be a set of n reals, such that the 95th percentile of the z_i^2 values is 1. Suppose that there exists a degree-6 pseudoexpectation $\tilde{\mathbf{E}}$ on the variables $\{w_i\}, \{z_i'\}$ such that:*

1. *$\forall i$, $\tilde{\mathbf{E}}$ satisfies $w_i^2 - w_i = 0$,*
2. *$\tilde{\mathbf{E}}$ satisfies $\sum w_i - 0.99n \geq 0$,*
3. *$\forall i$, $\tilde{\mathbf{E}}$ satisfies $w_i(z_i' - z_i) = 0$,*
4. *$\tilde{\mathbf{E}}\left[\frac{1}{n}\sum_i((z_i')^2 - \sigma')^2\right] \leq 3 \cdot \tilde{\mathbf{E}}\left[(\sigma')^2\right]$, where we define $\sigma' := \frac{1}{n}\sum (z_i')^2$.*

Then, $\tilde{\mathbf{E}}[\sigma'] = \Theta(1)$. Moreover, if the 95th percentile of z_i^2 is less than 1, we still have $\tilde{\mathbf{E}}[\sigma'] \leq O(1)$.

Proof. First, let's show that $\tilde{\mathbf{E}}[\sigma'] \geq \Omega(1)$. To prove this, note that by Constraint 1, $\tilde{\mathbf{E}}$ satisfies $w_i = w_i^2 \geq 0$ and $(1 - w_i) = (1 - w_i)^2 \geq 0$. So,

$$\begin{aligned}
\tilde{\mathbf{E}}[\sigma'] &= \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{E}}[w_i(z_i')^2 + (1 - w_i)(z_i')^2] && \text{(Definition of } \sigma') \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{E}}[w_i(z_i')^2] && \text{(Positivity of } 1 - w_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{E}}[w_i z_i^2] && \text{(Constraint 3)} \\
&= \frac{1}{n}\sum_{i=1}^{n} z_i^2 \tilde{\mathbf{E}}[w_i] && \text{(Linearity)}
\end{aligned}$$

Since $\tilde{\mathbf{E}}[w_i]$ is bounded between 0 and 1 (as $\tilde{\mathbf{E}}w_i = \tilde{\mathbf{E}}w_i^2$ and $\tilde{\mathbf{E}}[1 - w_i] = \tilde{\mathbf{E}}[(1 - w_i)^2]$), and since $\sum \tilde{\mathbf{E}}[w_i] \geq 0.99n$, the minimum possible value of $\sum z_i^2 \tilde{\mathbf{E}}[w_i]$ is the sum of z_i^2 over the $0.99n$ smallest values of z_i^2. Since the 95th percentile of the z_i^2 values is 1, this means $\sum z_i^2 \tilde{\mathbf{E}}[w_i] \geq 0.04n$. Thus, $\tilde{\mathbf{E}}[\sigma'] \geq 0.04$.

Next, we must show that $\tilde{\mathbf{E}}[\sigma'] \leq O(1)$, if the 95th percentile of the z_i^2 values is at most 1. To do this, we consider restricting $\tilde{\mathbf{E}}$ to the (at least) $0.95n$ indices S where $z_i^2 \leq 1$ (note that z_i are fixed real numbers, not variables). More formally, we define $\tilde{\mathbf{E}}'$ to be a pseudoexpectation where on any monomial p, $\tilde{\mathbf{E}}'p = 0$ if p has a positive power of some w_i for $i \notin S$, and $\tilde{\mathbf{E}}'p = \tilde{\mathbf{E}}p$ otherwise. It is clear that $\tilde{\mathbf{E}}'$ is still a degree-6 pseudoexpectation, since $\tilde{\mathbf{E}}'1 = \tilde{\mathbf{E}}1 = 1$, and $\tilde{\mathbf{E}}'[p^2] = \tilde{\mathbf{E}}[(p')^2]$ where p' is the polynomial that removes all monomials containing some w_i for $i \notin S$. In addition, if we replace $\tilde{\mathbf{E}}$ with $\tilde{\mathbf{E}}'$, Constraint 4 is unchanged. Constraints 1 and 3 are unchanged for $i \in S$, and trivially hold for $i \notin S$. Finally, since $\tilde{\mathbf{E}}$ *satisfies* $w_i \leq 1$ for all i (since $\tilde{\mathbf{E}}'$ satisfies $1 - w_i = (1 - w_i)^2 \geq 0$), we thus have that $\tilde{\mathbf{E}}$ satisfies $\sum_{i \in S} w_i + \sum_{i \notin S} 1 - 0.99n \geq 0$, which means $\sum_{i \in S} w_i \geq 0.94n$. So, $\tilde{\mathbf{E}}'$ satisfies $\sum w_i - 0.94n \geq 0$. Overall, by replacing $\tilde{\mathbf{E}}$ with $\tilde{\mathbf{E}}'$, we have the constraints are unchanged except 2, and the goal of showing $\tilde{\mathbf{E}}'[\sigma'] \leq O(1)$ is sufficient.

We also remark that we can rewrite Constraint 4 (now with $\tilde{\mathbf{E}}'$) as

$$\tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^{n}(z_i')^4\right] \leq 4 \cdot \tilde{\mathbf{E}}'\left[(\sigma')^2\right] = 4 \cdot \tilde{\mathbf{E}}'\left[\left(\frac{1}{n}\sum_{i=1}^{n}(z_i')^2\right)^2\right]. \tag{7}$$

Now, note that

$$\frac{1}{n}\cdot \underbrace{\tilde{\mathbf{E}}'\left[\sum_{i=1}^{n}(1-w_i)(z_i')^4\right]}_{A} \le \tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^{n}(z_i')^4\right] \qquad \text{(Constraint 1)}$$

$$\le 4\cdot\tilde{\mathbf{E}}'\left[\left(\frac{1}{n}\sum_{i=1}^{n}(z_i')^2\right)^2\right] \qquad \text{(Equation (7))}$$

$$= 4\cdot\tilde{\mathbf{E}}'\left[\left(\frac{1}{n}\sum_{i=1}^{n}(1-w_i)(z_i')^2+\frac{1}{n}\sum_{i=1}^{n}w_i(z_i')^2\right)^2\right]$$

$$\le 8\cdot\tilde{\mathbf{E}}'\left[\left(\frac{1}{n}\sum_{i=1}^{n}(1-w_i)(z_i')^2\right)^2+\left(\frac{1}{n}\sum_{i=1}^{n}w_i(z_i')^2\right)^2\right] \qquad \text{(Cauchy-Schwarz)}$$

$$= \frac{8}{n^2}\cdot\Bigg(\underbrace{\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}(1-w_i)(z_i')^2\right)^2\right]}_{B}+\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}w_iz_i^2\right)^2\right]\Bigg) \qquad \text{(Constraint 3).}$$

Note that $\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}w_iz_i^2\right)^2\right]=\sum_{i,j\in S}\tilde{\mathbf{E}}[w_iw_j]z_i^2z_j^2$. In addition, for any i,j, $\tilde{\mathbf{E}}[w_iw_j]\le\frac{1}{2}\left(\tilde{\mathbf{E}}[w_i^2]+\tilde{\mathbf{E}}[w_j^2]\right)\le 1$. So, since $0\le z_i^2\le 1$ for all $i\in S$, we have that $\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}w_iz_i^2\right)^2\right]\le n^2$. Therefore,

$$\frac{1}{n}\cdot A\le\frac{8}{n^2}\cdot B+8. \tag{8}$$

Also,

$$0.06n\cdot A=\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}(1-w_i)(z_i')^4\right)\cdot 0.06n\right] \qquad \text{(Definition of } A\text{)}$$

$$=\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}(1-w_i)^2(z_i')^4\right)\cdot 0.06n\right] \qquad \text{(Constraint 1)}$$

$$\ge\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}(1-w_i)^2(z_i')^4\right)\cdot\left(\sum_{i=1}^{n}(1-w_i)^2\right)\right] \qquad \text{(Constraints 1 and 2)}$$

$$\ge\tilde{\mathbf{E}}'\left[\left(\sum_{i=1}^{n}(1-w_i)^2(z_i')^2\right)^2\right] \qquad \text{(Cauchy-Schwarz),}$$

Therefore, $0.06n\cdot A\ge B$, but (8) tells us that $n\cdot A\le 8(B+n^2)$. So, $B\le 0.06n\cdot A\le 0.48\cdot(B+n^2)$, which means $B\le n^2$. Therefore, by Cauchy-Schwarz, $\tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^{n}(1-w_i)(z_i')^2\right]^2\le\tilde{\mathbf{E}}'\left[\left(\frac{1}{n}\sum_{i=1}^{n}(1-w_i)(z_i')^2\right)^2\right]=\frac{1}{n^2}\cdot B\le 1$, which means $\tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^{n}(1-w_i)(z_i')^2\right]\le 1$. In addition, we

know that $\tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^n w_i(z_i')^2\right] \le \tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^n w_i z_i^2\right] \le 1$. So overall, since σ' has no coefficients with w_i, we obtain

$$\tilde{\mathbf{E}}[\sigma'] = \tilde{\mathbf{E}}'[\sigma'] = \tilde{\mathbf{E}}'\left[\frac{1}{n}\sum (z_i')^2\right] = \tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^n (1-w_i)(z_i')^2\right] + \tilde{\mathbf{E}}'\left[\frac{1}{n}\sum_{i=1}^n w_i(z_i')^2\right] \le 2. \qquad \square$$

Proof of Lemma 6.15. Our main goal will be to show that $\mathcal{L}_1[\Sigma'], \mathcal{L}_2[\Sigma']$ are close in spectral distance. To do so, we show that for any unit vector v, $v^\top \hat{\Sigma}_1 v$ and $v^\top \hat{\Sigma}_2 v$ are equal up to an $O(1)$ multiplicative factor. This will imply that $\mathcal{L}_1[\Sigma'] \preccurlyeq O(1)\cdot\mathcal{L}_2[\Sigma']$ and $\mathcal{L}_2[\Sigma'] \preccurlyeq O(1)\cdot\mathcal{L}_1[\Sigma']$.

Assume first that $\mathcal{L}_1, \mathcal{L}_2$ are actual pseudoexpectations (i.e., if $\tau = 0$). We define $z_i := \langle x_i, v\rangle$ and $z_i' := \langle x_i', v\rangle$. If $\mathcal{L}_1, \mathcal{L}_2$ are (α, τ, T)-certificates for $\tau = 0$ and $T \le 0.01n$, then it is clear that $\mathcal{L}_1$ and $\mathcal{L}_2$ satisfy Constraints 1, 2, and 3 of Lemma B.2. To check Constraint 4, note that by Constraint 3 of Definition 6.4,

$$\begin{aligned}\mathcal{L}\left[\frac{1}{n}\sum_{i=1}^n ((z_i')^2 - \sigma')^2 - (2+\alpha)(\sigma')^2\right] &= \mathcal{L}\left[\frac{1}{n}\sum_{i=1}^n \left(\langle x_i', v\rangle^2 - v^\top\Sigma' v\right)^2 - (2+\alpha)\cdot(v^\top\Sigma' v)^2\right] \\ &= -\mathcal{L}\left[\|Mv^{\otimes 2}\|_2^2\right] \\ &\le 0,\end{aligned}$$

for either $\mathcal{L} = \mathcal{L}_1$ or $\mathcal{L} = \mathcal{L}_2$, where $\Sigma' := \frac{1}{n}\sum (x_i')(x_i')^\top$ and $\sigma' := \frac{1}{n}\sum (z_i')^2$.

Hence, both we can apply Lemma B.2 for both $\mathcal{L}_1$ and $\mathcal{L}_2$. If the 95th percentile of $\langle y_i, v\rangle^2$ is equal to 1, this implies that $\mathcal{L}_1[v^\top\Sigma' v], \mathcal{L}_2[v^\top\Sigma' v]$ are both $\Theta(1)$. If the 95th percentile of $\langle y_i, v\rangle^2$ is some value G, we may rescale and use linearity to say that $\mathcal{L}_1[v^\top\Sigma' v], \mathcal{L}_2[v^\top\Sigma' v]$ are both $\Theta(G^2)$.

Hence, this implies that $\mathcal{L}_1[\Sigma']$ and $\mathcal{L}_2[\Sigma']$ are within $O(1)$ spectral distance of each other, at least when $\tau = 0$. For general τ, we note that again the coefficients at each step in the sum-of-squares proof are bounded by $\mathrm{poly}(n, d, K)$. The only possible issue is the rescaling, if $G \gg (ndK)^{O(1)}$ or $G \ll (ndK)^{-O(1)}$. We avoid the former case because we are assuming that every sample is bounded by $\mathrm{poly}(n, d, K)$ in magnitude, using truncation. In the latter case, we use the fact that if the 95th percentile of z_i^2 is less than 1, then $\mathcal{L}[\sigma'] \le O(1)$ in Lemma B.2. In this case, by scaling by $\frac{1}{K^2}$, we have $\mathcal{L}[v^\top\Sigma' v] \le \frac{O(1)}{K^2}$, which violates Constraint 5 of Definition 6.4.

In summary, we have that $\mathcal{L}_1[\Sigma'] \preccurlyeq O(1)\cdot\mathcal{L}_2[\Sigma']$ and $\mathcal{L}_2[\Sigma'] \preccurlyeq O(1)\cdot\mathcal{L}_1[\Sigma']$, and both are spectrally bounded between $\frac{1}{4K}$ and $4K$. Since we have the requirements that $(1-\alpha)\mathcal{L}[\Sigma'] - \tau\cdot T\cdot I \preccurlyeq \widetilde{\Sigma} \preccurlyeq (1+\alpha)\mathcal{L}[\Sigma'] + \tau\cdot T\cdot I$, this implies that $\widetilde{\Sigma}_1 \preccurlyeq O(1)\cdot\widetilde{\Sigma}_2$ and $\widetilde{\Sigma}_2 \preccurlyeq O(1)\cdot\widetilde{\Sigma}_1$. $\square$

B.3 SoS bounds for arbitrary samples: Mean estimation

In this subsection, we prove Lemma 5.13, which is our worst-case robustness result for mean estimation. First, we establish a 1-dimensional Sum-of-Squares result that will be crucial in proving Lemma 5.13.

Lemma B.3. *Let $z_1, \dots, z_n$ be a set of n reals, such that at least $n/4$ of the z_i's are at least 20. Then, for any degree-6 pseudoexpectation $\tilde{\mathbf{E}}$ on the variables $\{w_i\}, \{z_i'\}$ such that*

1. *$\forall i$, $\tilde{\mathbf{E}}$ satisfies $w_i^2 - w_i = 0$,*

2. $\tilde{\mathbf{E}}$ *satisfies* $\sum w_i - 0.99n = 0$,

3. $\forall i$, $\tilde{\mathbf{E}}$ *satisfies* $w_i(z'_i - z_i) = 0$,

4. $\tilde{\mathbf{E}}[\mu'] = 0$, *where* $\mu' = \frac{1}{n}\sum z'_i$,

we must have that $\tilde{\mathbf{E}}\left[\frac{1}{n}\sum(z'_i - \mu')^2\right] \geq 2$.

Proof. Using the fact that $\tilde{\mathbf{E}}$ satisfies $w_i^2 = w_i$, we have that $(1 - w_i w_j)^2 = (1 - w_i w_j)$, which means $w_i w_j \leq 1$ is satisfied. In addition, $\tilde{\mathbf{E}}$ satisfies $w_i w_j = w_i^2 w_j^2 \geq 0$, and also satisfies $w_i w_j \geq w_i + w_j - 1$, since $w_i w_j - (w_i + w_j - 1) = (1 - w_i)(1 - w_j) = (1 - w_i)^2(1 - w_j)^2$.

This means

$$\begin{aligned}
\tilde{\mathbf{E}}\left[\sum_{i,j=1}^{n}(z'_i - z'_j)^2\right] &\geq \tilde{\mathbf{E}}\left[\sum_{i,j} w_i w_j (z'_i - z'_j)^2\right] \\
&\geq \tilde{\mathbf{E}}\left[\sum_{i,j} w_i w_j (z_i - z_j)^2\right] && \text{(Condition 3)} \\
&= \sum_{i,j}(z_i - z_j)^2 \cdot \tilde{\mathbf{E}}\left[w_i w_j\right] && \text{(Linearity)} \\
&\geq \sum_{i,j}(z_i - z_j)^2 \cdot \max(\tilde{\mathbf{E}}[w_i] + \tilde{\mathbf{E}}[w_j] - 1, 0). && (9)
\end{aligned}$$

Now, C_1 and C_2 be the 25th and 75th percentiles, respectively, of the elements z_i sorted in increasing order. We show that we may assume $C_2 - C_1 \leq 8$. Otherwise, there exists a set S of $0.25n$ elements z_i that are at least $C_1 + 8$, and a set T of $0.25n$ elements that are at most C_1. In this case, we can bound (9) as at least

$$\begin{aligned}
&2 \cdot \sum_{i \in S, j \in T}(z_i - z_j)^2 \cdot \max(\tilde{\mathbf{E}}[w_i] + \tilde{\mathbf{E}}[w_j] - 1, 0) \\
&\geq 2 \cdot \sum_{i \in S, j \in T} 8^2 \cdot (\tilde{\mathbf{E}}[w_i] + \tilde{\mathbf{E}}[w_j] - 1) \\
&= 2 \cdot \left[64(n/4) \cdot \sum_{i \in S}\tilde{\mathbf{E}}[w_i] + 64(n/4) \cdot \sum_{j \in T}\tilde{\mathbf{E}}[w_j] - 64(n/4)^2\right] \\
&\geq 4 \cdot 64(n/4)(n/4 - 0.01n) - 2 \cdot 64(n/4)^2 \\
&\geq 6n^2,
\end{aligned}$$

where the penultimate inequality uses the fact that $\tilde{\mathbf{E}}[w_i] \in [0, 1]$ and $\sum_{i=1}^{n}\tilde{\mathbf{E}}[w_i] \geq 0.99n$. Overall, this means $\tilde{\mathbf{E}}[\sum_{i,j}(z'_i - z'_j)^2] \geq 6n^2$. But, $\sum_{i,j=1}^{n}(z'_i - z'_j)^2 = 2n \cdot \sum_{i=1}^{n}(z'_i - \mu')^2$, which means $\tilde{\mathbf{E}}\left[\sum_{i=1}^{n}(z'_i - \mu')^2\right] \geq 3n$, as desired.

Hence, we may assume that the 25th and 75th percentiles are within 8 of each other. Re-define $S \subset [n]$ to be the set of indices of size $n/2$ between the 25th and 75th percentile. By our assumption

in the lemma that at least $n/4$ values are at least 20, $z_i \in [C-4, C+4]$ for all $i \in S$, for some $C \geq 16$. Note that

$$\tilde{\mathbf{E}}\left[\sum_{i\in S} w_i z_i'\right] = \tilde{\mathbf{E}}\left[\sum_{i\in S} w_i z_i\right] = \sum_{i\in S} z_i \tilde{\mathbf{E}}[w_i] \geq (C-4)\cdot \sum_{i\in S} \tilde{\mathbf{E}}[w_i] \geq (C-4)\cdot(0.49n), \tag{10}$$

but

$$\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} w_i z_i'\right)^2\right] = \tilde{\mathbf{E}}\left[\left(\sum_{i\in S} w_i z_i\right)^2\right] = \sum_{i,j\in S} z_i z_j \tilde{\mathbf{E}}[w_i w_j] \leq (C+4)^2\cdot(0.5n)^2. \tag{11}$$

In addition, if we assume $\tilde{\mathbf{E}}[\frac{1}{n}\sum_{i=1}^n (z_i'-\mu')^2] \leq 2$, then since S is fixed and has size $n/2$,

$$\begin{aligned}
\tilde{\mathbf{E}}\left[\sum_{i\in S}(z_i')^2\right] &\leq \tilde{\mathbf{E}}\left[\sum_{i\in S}(z_i')^2\right] + \frac{1}{|S|}\cdot\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i' - |S|\cdot\mu'\right)^2\right] \\
&= \tilde{\mathbf{E}}\left[\sum_{i\in S}(z_i'-\mu')^2\right] + \frac{1}{|S|}\cdot\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i'\right)^2\right] \\
&\leq \tilde{\mathbf{E}}\left[\sum_{i=1}^{n}(z_i'-\mu')^2\right] + \frac{1}{|S|}\cdot\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i'\right)^2\right] \\
&\leq 2n + \frac{1}{|S|}\cdot\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i'\right)^2\right].
\end{aligned} \tag{12}$$

Making use of the fact that $\tilde{\mathbf{E}}$ satisfies $(1-w_i) = (1-w_i)^2$, we have

$$\begin{aligned}
\underbrace{\tilde{\mathbf{E}}\left[\left(\sum_{i\in S}(1-w_i)z_i'\right)^2\right]}_{A} &\leq \tilde{\mathbf{E}}\left[\left(\sum_{i\in S}(1-w_i)\right)\cdot\left(\sum_{i\in S}(1-w_i)(z_i')^2\right)\right] && \text{(Cauchy-Schwarz)} \\
&\leq 0.01n\cdot\tilde{\mathbf{E}}\left[\sum_{i\in S}(1-w_i)(z_i')^2\right] && \text{(Condition 2)} \\
&\leq 0.01n\cdot\tilde{\mathbf{E}}\left[\sum_{i\in S}(z_i')^2\right] && \text{(Condition 1)} \\
&\leq 0.01\cdot\left(\tilde{\mathbf{E}}\left[2\left(\sum_{i\in S} z_i'\right)^2\right] + 2n^2\right) && \text{(Equation (12))} \\
&\leq 0.01\cdot\left(4\underbrace{\tilde{\mathbf{E}}\left[\left(\sum_{i\in S}(1-w_i)z_i'\right)^2\right]}_{A} + 4\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} w_i z_i'\right)^2\right] + 2n^2\right). && \text{(Cauchy-Schwarz)}
\end{aligned}$$

Hence, we have that $A \le 0.05\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} w_i z_i'\right)^2\right] + 0.03n^2 \le 0.02C^2n^2$ for $C \ge 16$, using (11).

So, by Cauchy-Schwarz, we have that $\left|\tilde{\mathbf{E}}\left[\sum_{i\in S}(1-w_i)z_i'\right]\right| \le 0.15Cn$. But $\tilde{\mathbf{E}}\left[\sum_{i\in S} w_i z_i'\right] \ge (C-4)\cdot 0.49n \ge 0.35Cn$ by (10), which means $\tilde{\mathbf{E}}\left[\sum_{i\in S} z_i'\right] \ge 0.2Cn$.

However, if $\tilde{\mathbf{E}}\left[\sum_{i=1}^n z_i'\right] = 0$, then $\tilde{\mathbf{E}}\left[\sum_{i\in S} z_i' - \frac{1}{2}\sum_{i=1}^n z_i'\right] \ge 0.2Cn$. By Cauchy-Schwarz, this means $\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i' - \frac{1}{2}\sum_{i=1}^n z_i'\right)^2\right] \ge 0.04C^2n^2$. Since S is a fixed set of size $n/2$, defining $T := [n]\backslash S$, we have

$$\begin{aligned}
\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i' - \frac{1}{2}\sum_{i=1}^n z_i'\right)^2\right] &= \frac{1}{4}\cdot\tilde{\mathbf{E}}\left[\left(\sum_{i\in S} z_i' - \sum_{i\in T} z_i'\right)^2\right] \\
&= \frac{1}{n^2}\tilde{\mathbf{E}}\left[\left(\sum_{i\in S, j\in T}(z_i' - z_j')\right)^2\right] \\
&\le \frac{1}{4}\cdot\tilde{\mathbf{E}}\left[\sum_{i\in S, j\in T}(z_i' - z_j')^2\right]. \qquad \text{(Cauchy-Schwarz)}
\end{aligned}$$

This implies that $\tilde{\mathbf{E}}\left[\sum_{i\in S, j\in T}(z_i' - z_j')^2\right] \ge 0.16C^2n^2$, which means $2n\cdot\tilde{\mathbf{E}}\left[\sum_{i=1}^n (z_i' - \mu')^2\right] = \tilde{\mathbf{E}}\left[\sum_{i,j=1}^n (z_i' - z_j')^2\right] \ge 0.32C^2n^2$. So, $\tilde{\mathbf{E}}\left[\sum_{i=1}^n (z_i' - \mu')^2\right] \ge 0.16C^2n \ge 3n$. □

Proof of Lemma 5.13. Our main goal will be to show that $\hat{\mu}_1 := \mathcal{L}_1[\mu']$, $\hat{\mu}_2 := \mathcal{L}_2[\mu']$ are close in ℓ_2 distance. To do so, we show that for any unit vector v, $\langle \mathcal{L}_1[\mu'] - \mathcal{L}_2[\mu'], v\rangle \le O(1)$.

We first focus on $\mathcal{L}_1$: suppose $\mathcal{L}_1$ is an actual pseudoexpectation (i.e., if $\tau = 0$). We define $z_i := \langle x_i - \hat{\mu}_1, v\rangle$ and $z_i' := \langle x_i' - \hat{\mu}_1, v\rangle$. If $\mathcal{L}_1$ is an (α, τ, T)-certificate for $\tau = 0$ and $T \le 0.01n$, then it is clear that $\mathcal{L}_1$ satisfies Constraints 1, 2, and 3 of Lemma B.3. To check Constraint 4, note that $\mathcal{L}_1\left[\frac{1}{n}\sum z_i'\right] = \frac{1}{n}\sum \mathcal{L}_1[\langle x_i', v\rangle - \langle \hat{\mu}_1, v\rangle] = 0$.

Hence, by Lemma B.3, if the median of $\langle x_i - \hat{\mu}_1, v\rangle$ was greater than 20, then $\mathcal{L}_1\left[\frac{1}{n}\sum\langle x_i' - \mu', v\rangle^2\right] \ge 2$, where $\mu' := \frac{1}{n}\sum x_i'$. This, however, contradicts Condition 2e in Definition 5.5. For general τ, we note that again the coefficients at each step in the sum-of-squares proof are bounded by $\mathrm{poly}(n, d, K)$. So, this implies that if $\mathcal{L}_1$ is an $(\alpha^*, \tau, \phi, T)$-certifiable mean, then for every unit vector v, $\langle \hat{\mu}_1, v\rangle$ is at most 20 away from the median of $\langle x_i, v\rangle$. (This is true in both directions since we can replace v with $-v$).

Likewise, the same is true for $\mathcal{L}_2$, which means that $\left|\langle \hat{\mu}_1, v\rangle - \langle \hat{\mu}_2, v\rangle\right| \le 40$ for all vectors v. Therefore, $\|\hat{\mu}_1 - \hat{\mu}_2\|_2 \le 40$. Finally, we note that $\|\widetilde{\mu}_1 - \hat{\mu}_1\|_\infty, \|\widetilde{\mu}_2 - \hat{\mu}_2\|_\infty \le \phi + \tau\cdot T \le O(\alpha/\sqrt{d})$, so $\|\widetilde{\mu}_1 - \widetilde{\mu}_2\|_2 \le 42$. □

C Computing Score Functions

In this section we will describe how we can compute the value of the score functions efficiently.

In our problems, we usually have some family of properties $\{P_T\}$, parameterized by T. The higher values of T correspond to more lenient settings and the lower values of T correspond to more stringent settings. We are interested in how well (or poorly) a parameter θ satisfies these properties. We can easily define a score function to measure this. These score functions are later

used to run the exponential mechanism and design private algorithms. These score functions are defined in the following fashion.

$$\mathcal{S}(\theta) := \inf_T \text{ such that } \theta \text{ satisfies } P_T \ .$$

As mentioned, because P_T's are increasingly lenient, θ satisfies P_T for all $T > \mathcal{S}(\theta)$, and does not satisfy P_T for all $T < \mathcal{S}(\theta)$. In our problems we describe $\{P_T\}$ through systems of polynomial inequalities and the existence of linear functionals that approximately satisfy them. We define polynomial constraints $q_1 \geq 0, \ldots, q_k \geq 0$, which depend on T and θ, and if there exists a linear functional (an approximate pseudo-expectation) that *approximately satisfies* these polynomial constraints, we say that θ satisfies P_T. We first make some assumptions on these generic polynomial constraints and after that we will define approximate satisfiability formally in definition C.2.

Assumption C.1. We make the assumption that in problems that we deal with parameterized families of polynomials $\{Q_T\}_{T=0}^{T_{\max}}$ that are in the following form and may include the following different types of constraints.

1. Regular constraints: $q \geq 0$.
2. PSD constraints: $\forall h$, where $\|h\|_2 = 1$: $qh^2 \geq 0$.
3. T-constraint: Each Q_T, has exactly one constraint that depends on T. We call this constraint the "T-constraint".The other constraints do not depend on T, and are the same over all Q_T's. Let q_T denote this constraint. This constraint is also a PSD constraint and it appears only in the form of $\forall h : q_T h^2 \geq 0$. We also make the assumption that q_T depends linearly on T and

   $$\forall 0 \leq T, T' \leq T_{\max} : (q_T - q'_T) = (T - T')/(2T_{\max}).$$

   Note that this is a polynomial identity.
4. Matrix PSD constraints: $q \succcurlyeq 0$.

Definition C.2 (approximate satisfiability)**.** Suppose $R > 1$, and a parameterized family of polynomials $\{Q_T\}$ of up to degree d, over $\mathbb{R}^n$ are given as in assumption C.1. We say a linear functional $\mathcal{L}$ over the set of polynomials of degree at most d over R^n, τ-approximately satisfies Q_T and write $\mathcal{L} \vdash_\tau Q_T$ if and only if

1. $\mathcal{L}1 = 1$,
2. $\mathcal{L}h^2 \geq -\tau \cdot T$, for every polynomial h such that $2 \deg h \leq d$ and $\|h\|_2 \leq 1$.
3. $\mathcal{L}q \geq -\tau \cdot T$, for every polynomial $q \in Q_T$ that is a regular constraint.
4. $\mathcal{L}qh^2 \geq -\tau \cdot T$, for every polynomial $q \in Q_T$ that is a PSD constraint and every polynomial h such that $2 \deg h + \deg q \leq d$ and $\|h\|_2 = 1$.
5. $\mathcal{L}q \succcurlyeq -\tau \cdot T \cdot I$, for every polynomial $q \in Q_T$, that is a matrix PSD constraint.
6. $\|\mathcal{R}(\mathcal{L})\|_2 \leq R + \tau \cdot T$.

In addition, for any $\gamma > 0$ we write $\mathcal{L} \vDash_{\tau,\gamma} Q_T$ if the above conditions hold but replacing $\tau \cdot T$ with $\tau \cdot (T + \gamma)$. (Note that the constraint Q_T has *not* been replaced with $Q_{T+\gamma}$.)

Remark. In order to run the ellipsoid algorithm, we should have a full dimensional ball of positive volume. If we attempt to run the ellipsoid algorithm over the set of functionals with $\mathcal{L}1 = 1$, this is trivially not going to be the case. Therefore, instead we only consider the space of linear functionals excluding the $S = \emptyset$ index, which corresponds to the monomial 1.

Lemma C.3 (efficient functional search)**.** *Suppose $R > 1$, and Q_T is a set of polynomial constraints of up to degree d, over $\mathbb{R}^n$ as in Assumption C.1, with fixed parameter T. Let $\mathcal{R}(\mathcal{L})_{\overline{\emptyset}}$ denote the representation of a functional $\mathcal{L}$ for every multiset of size up to d, excluding the empty set index. Then, for any $r, \gamma > 0$, there exists an algorithm that runs in time $\mathrm{poly}(n^d, \mathrm{Size}(Q_T), \log(R'/r), \log(1/\gamma))$ that either*

1. *finds the representation of a linear functional $\mathcal{L}$ such that $\|\mathcal{R}(\mathcal{L})_{\overline{\emptyset}}\|_2 \le R'$, and $\mathcal{L} \vDash_{\tau,O(\gamma)} Q_T$; or,*
2. *shows that the volume of representations of functionals $\mathcal{L}$ such that $\|\mathcal{R}(\mathcal{L})_{\overline{\emptyset}}\|_2 \le R'$ and $\mathcal{L} \vDash_{\tau} Q_T$, when projected to the entries $S \neq \emptyset$, is less than the volume of a ball of radius r,*

where $R' = \sqrt{(R + \tau \cdot T)^2 - 1}$. Note that here $\mathcal{R}(\mathcal{L}) \in \mathbb{R}^{\binom{n}{\le d}}$, and $\mathcal{R}(\mathcal{L})_{\overline{\emptyset}} \in \mathbb{R}^{\binom{n}{\le d}-1}$, and the volume in the second case is measured with respect to $\mathbb{R}^{\binom{n}{\le d}-1}$.

In essence, we use reductions to semi-definite programs. For a textbook treatment of this approach see Chapter 3 of [FKP19].

Proof. Firstly note that under the assumption that $\mathcal{R}(\mathcal{L})_\emptyset = 1$, we have that $\|\mathcal{R}(\mathcal{L})\|_2 \le R$ is equivalent to $\|\mathcal{R}(\mathcal{L})_{\overline{\emptyset}}\|_2 \le R'$. Let

$$K = \{\mathcal{R}(\mathcal{L}) \mid \mathcal{L} \vDash_\tau Q_T\}, K_{\overline{\emptyset}} = \left\{\mathcal{R}(\mathcal{L})_{\overline{\emptyset}} \;\middle|\; \mathcal{L} \vDash_\tau Q_T\right\}.$$

It is easy to see that $K \subset \mathbb{R}^{\binom{n}{\le d}}$ is equal to $K_{\overline{\emptyset}} \subset \mathbb{R}^{\binom{n}{\le d}-1}$ with the adjustment that all of its members have the additional $\emptyset$ entry 1. We want to apply the ellipsoid algorithm over the ball of radius R in $\mathbb{R}^{\binom{n}{\le d}-1}$, if we show that

1. $K_{\overline{\emptyset}}$ is convex; and,
2. $K_{\overline{\emptyset}}$ admits an efficient (approximate) membership and separation oracle,

we are done and we obtain the desired guarantees via the ellipsoid algorithm.

Convexity. In order to show that $K_{\overline{\emptyset}}$ is convex, it suffices to show that K is convex. let $M_1, M_2 \in K$, we need to prove that $\forall \alpha \in [0,1]$, $M_3 = \alpha M_1 + (1-\alpha)M_2 \in K$. By triangle inequality it is easy to see that $\|M_3\|_2 \le \alpha\|M_2\|_2 + (1-\alpha)\|M_2\|_2 \le R$. Let $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ be the corresponding functionals of M_1, M_2, M_3. It suffices to show that $\mathcal{L}_3 \vDash Q_T$. Let's verify this.

1. $\mathcal{L}_3 1 = \alpha\mathcal{L}_1 1 + (1-\alpha)\mathcal{L}_2 1 = 1$.
2. $\mathcal{L}_3 q = \alpha\mathcal{L}_1 q + (1-\alpha)\mathcal{L}_2 q \ge -\tau \cdot T$, for every regular constraint $q \in Q_T$.

3. $\mathcal{L}_3 h^2 = \alpha \mathcal{L}_1 h^2 + (1-\alpha)\mathcal{L}_2 h^2 \geq -\alpha\tau \cdot T - (1-\alpha)\tau \cdot T = -\tau \cdot T$, for every polynomial h such that $2\deg(h) \leq d$.
4. $\mathcal{L}_3 q h^2 = \alpha \mathcal{L}_1 q h^2 + (1-\alpha)\mathcal{L}_2 q h^2 \geq -\alpha\tau \cdot T - (1-\alpha)\tau \cdot T = -\tau \cdot T$, for every PSD polynomial constraint $q \in Q_T$ and every polynomial h such that $\deg q + 2\deg(h) \leq d$.
5. $\mathcal{L}_3 q = \alpha \mathcal{L}_1 q + (1-\alpha)\mathcal{L}_2 q \succcurlyeq -\alpha\tau \cdot T \cdot I - (1-\alpha)\tau \cdot T \cdot I = -\tau \cdot T \cdot I$, for every matrix PSD polynomial constraint $q \in Q_T$ and every polynomial h such that $\deg q + 2\deg(h) \leq d$.
6. $\|\mathcal{R}(\mathcal{L}_3)\|_2 \leq \alpha \|\mathcal{R}(\mathcal{L}_1)\|_2 + (1-\alpha)\|\mathcal{R}(\mathcal{L}_2)\|_2 = R + \tau \cdot T$.

Therefore K is convex as desired.

Membership/Separation oracle. Suppose $M \in \mathbb{R}^{\binom{n}{\leq d}-1}$ is given. We need to verify $M \in K_{\overline{\emptyset}}$, or not. Let M' be equal to M with the additional entry $M'_{\emptyset} = 1$. Then it is easy to see that $M \in K_{\overline{\emptyset}}$, if and only if $M' \in K$. Suppose $\mathcal{L}$ is the linear functional with M' as its representation. We need to come up with membership/separation oracles for each of the constraints in Definition C.2.

Regular Constraints.

$$\mathcal{L}q \geq -\tau \cdot T, \text{ for every regular constraint } q \in Q_T.$$

In order to check this constraint we can just compute the value $\langle M', \mathcal{R}(q)\rangle$. If its value is greater than or equal to $-\tau \cdot T$, then that means $\mathcal{L}q \geq -\tau \cdot T$ is satisfied, and this constraint does not refute $\mathcal{L} \vDash_{\tau, \mathcal{O}(\gamma)}$, and we would be in the setting where $\mathcal{L} \vDash_{\tau,\mathcal{O}(\gamma)} Q_T$, if all of the other constraints hold as well.

If this is not the case then let $H \in \mathbb{R}^{\binom{n}{\leq d}}$ be as

$$H = \mathcal{R}(q).$$

Then $\langle M, H_{\overline{\emptyset}}\rangle = \langle M', H\rangle - H_{\emptyset} < -\tau \cdot T - H_{\emptyset}$. Moreover, for every $N \in K_{\overline{\emptyset}}$, we have that $\langle N, H_{\overline{\emptyset}}\rangle \geq -\tau \cdot T - H_{\emptyset}$. Therefore $H_{\emptyset}$ is a separating hyperplane. Therefore we have an efficient separation oracle as desired.

PSD Constraints. These constraints are in the following form.

$$\mathcal{L}qh^2 \geq -\tau \cdot T, \text{ for every polynomial } h \text{ where } \|\mathcal{R}(h)\|_2 \leq 1, \text{ and } \deg q + 2\deg h \leq d,$$
$$\text{and for every polynomial } q \text{ that is either } 1 \text{ or a PSD constraint in } Q_T.$$

Suppose $q = \langle a, v_d(x)\rangle$, $h = \langle b, v_d(x)\rangle$. Then,

$$\begin{aligned}
\mathcal{L}qh^2 &= \mathcal{L}\left(\sum_U \left(a_U x^U\right)\right) \cdot \left(\sum_V \left(b_V x^V\right)\right) \cdot \left(\sum_W \left(b_W x^W\right)\right) \\
&= \mathcal{L} \sum_{U,V,W} a_U b_V b_W \cdot x^{U+V+W} \\
&= \sum_{V,W} b_V b_W \left[\sum_U a_U \mathcal{L}(x^{U+V+W})\right].
\end{aligned}$$

Define the matrix $X \in \mathbb{R}^{\binom{n}{\leq (d-\deg q)/2} \times \binom{n}{\leq (d-\deg q)/2}}$ as

$$\begin{aligned} X_{V,W} &= \sum_U a_U \mathcal{L}(x^{U+V+W}) \\ &= \sum_U \mathcal{R}(q)_U \mathcal{R}(\mathcal{L})_{U+V+W}. \end{aligned}$$

Then

$$\mathcal{L}qh^2 = b^\mathsf{T} X b.$$

Our goal is to verify whether $\mathcal{L}qh^2$ is larger than $-\tau \cdot T$ for every h, where $\|\mathcal{R}(h)\|_2 \leq 1$ or not. This is equivalent to $b^\mathsf{T} X b$ being larger than $-\tau \cdot T$ for every b, where $\|b\|_2 \leq 1$. We can check this by looking at the spectral value decomposition of X. Suppose that the spectral decomposition of $X = PDP^\mathsf{T}$, where D is a diagonal matrix whose entries are the eigenvalues of X, and the rows of P are the corresponding eigenvectors. This decomposition can be computed in polynomial time using standard algorithms for obtaining eigenvalue decompositions. More accurately, for any $\gamma > 0$, we can learn the minimum eigenvalue up to error $\tau \cdot \gamma$ in time $\mathrm{poly}(n^d, \log \frac{1}{\tau \cdot \gamma})$. Then, if (our estimate of) the minimum eigenvalue is at least $-\tau \cdot (T + 3\gamma)$, this means that the constraint $\mathcal{L}qh^2 \geq \tau \cdot (T + 4\gamma)$ is satisfied, and this constraint does not refute $\mathcal{L} \vDash_{\tau, O(\gamma)}$, and we would be in the setting where $\mathcal{L} \vDash_{\tau, O(\gamma)} Q_T$, if all of the other constraints hold as well.

If this is not the case then we know that the minimum eigenvalue is less than $-\tau \cdot (T + 2\gamma)$, and we need to return a separating hyperplane that separates M and $K_{\overline{\varnothing}}$. Suppose the minimum eigenvalue of X is less than $-\tau \cdot (T + 2\gamma)$. Then we can find a vector c such that $c^\top X c < -\tau \cdot (T + \gamma)$. Let the vector $H \in \mathbb{R}^{\binom{n}{\leq d}}$ be as

$$H_S = \sum_{U \cup V = S} c_U \mathcal{R}(q)_V.$$

Note that we can compute this vector efficiently. Then we have that

$$\begin{aligned} \langle M', H \rangle &= \mathcal{L}q \langle c, v_{(d-\deg q)/2}(x) \rangle^2 \\ &= c^\mathsf{T} X c \\ &< -\tau \cdot (T + \gamma). \end{aligned}$$

Since $M'_\varnothing = 1$, we have that $\langle M, H_{\overline{\varnothing}} \rangle = \langle M', H \rangle - H_\varnothing < -\tau \cdot (T + \gamma) - H_\varnothing$. Now assume $N \in K_{\overline{\varnothing}}$. Similarly, we can show that $\langle N, H_{\overline{\varnothing}} \rangle \geq -\tau \cdot T - H_\varnothing$. Therefore $H_{\overline{\varnothing}}$ is a separating hyperplane. Therefore we have an efficient separation oracle as desired.

Matrix PSD Constraints.

$$\mathcal{L}q \succcurlyeq -\tau \cdot T, \text{ for every matrix PSD constraint } q \in Q_T.$$

Note that here q is a square matrix with polynomials as its entries. We use $q_{i,j}$ to denote the (i,j)-entry of this matrix, which is a polynomial. In order to check this constraint just define X as

$$X_{i,j} = \mathcal{L}q_{i,j} = \langle M, \mathcal{R}(q_{i,j}) \rangle.$$

In order to check the constraint $\mathcal{L}q \succcurlyeq -\tau \cdot T$, we can check the spectral value decomposition of X. Suppose that the spectral decomposition of $X = PDP^\mathsf{T}$, where D is a diagonal matrix whose entries are the eigenvalues of X, and the rows of P are the corresponding eigenvectors. This decomposition can be computed in polynomial time using standard algorithms for obtaining eigenvalue decompositions. More accurately, for any $\gamma > 0$, we can compute the minimum eigenvalue of to error $\tau \cdot \gamma$ in time $\mathrm{poly}(n^d, \log \frac{1}{\tau \cdot \gamma})$. Then, if our estimate of the eigenvalue is at least $-\tau \cdot (T + 3\gamma)$, this means that the constraint $\mathcal{L}q \succcurlyeq -\tau \cdot (T + 4\gamma) \cdot I$ is satisfied, and this constraint does not refute $\mathcal{L} \vdash_{\tau, O(\gamma) Q_T}$, and we would be in the setting where $\mathcal{L} \vdash_{\tau, O(\gamma)} Q_T$, if all of the other constraints hold as well.

If this is not the case then we know that the minimum eigenvalue is less than $-\tau \cdot (T + 2\gamma)$, and we need to return a separating hyperplane that separates M and $K_{\overline{\emptyset}}$. Suppose the minimum eigenvalue of X is less than $-\tau \cdot (T + 2\gamma)$. Then we can find a vector c such that $c^\mathsf{T} X c < -\tau \cdot (T + \gamma)$. Now consider $c^\mathsf{T} \mathcal{L} q c$, and assume c and q are constants and $\mathcal{L}$ is variable. We can write this as

$$c^\mathsf{T} \mathcal{L} q c = \sum_U H_U \mathcal{L}(x^U),$$

for some H_U's that depend only on q and c. Moreover, give q and c we can compute this H efficiently. Now since $c^\mathsf{T} X c < -\tau \cdot (T + \gamma)$, we have that

$$\langle \mathcal{R}(\mathcal{L}), H \rangle = \langle M', H \rangle < -\tau \cdot (T + \gamma),$$

and therefore $\langle M, H_{\overline{\emptyset}} \rangle = \langle M', H \rangle - H_\emptyset < -\tau \cdot (T + \gamma) - H_\emptyset$. Similarly, if $N \in K_{\overline{\emptyset}}$, we can show that $\langle N, H_{\overline{\emptyset}} \rangle \geq -\tau \cdot T - H_\emptyset$. Therefore $H_{\overline{\emptyset}}$ is a separating hyperplane. Therefore we have obtained an efficient separation oracle as desired.

Norm Bound Constraints.

$$\|\mathcal{R}(\mathcal{L})\|_2 \leq R + \tau \cdot T.$$

In order to check this constraint we just compute $\|M\|_2^2$. If its value is less than or equal to R'^2, then that means $\|\mathcal{R}(\mathcal{L})\|_2 \leq R + \tau \cdot T$ is satisfied. If this is not the case then let $H \in \mathbb{R}^{\binom{n}{\leq d}}$ be as $H = \mathcal{R}(L) = M'$. Note that $\|H\|_2 > R$, since $\|M\|_2 > R'$. Then $\langle M, H_{\overline{\emptyset}} \rangle = \langle M', H \rangle - H_\emptyset = \|H\|_2^2 - 1$. Moreover, for every $N \in K_{\overline{\emptyset}}$, we have that $\langle N, H_{\overline{\emptyset}} \rangle \leq R\|H\|_2 - 1$. Therefore $H_{\overline{\emptyset}}$ is a separating hyperplane. □

Lemma C.4 (robust satisfiability)**.** *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption C.1. Moreover, suppose that there exists some linear functional $\mathcal{L}_0$, such that $\mathcal{L}_0 \vdash Q_{T_0}$. Then there exists a set of linear functionals $\mathcal{F}$ such that*

$$\left\{ \mathcal{R}(\mathcal{L})_{\overline{\emptyset}} \;\middle|\; \mathcal{L} \in \mathcal{F} \right\}$$

contains a full-dimensional ball of radius $r = \mathrm{poly}(1/\mathrm{poly}(n^d), \tau, \gamma, 1/k, 1/\|R(Q_{T_0})\|_\infty)$, and for all $\mathcal{L} \in \mathcal{F}$, we have that $\mathcal{L} \vdash_\tau Q_{T_0 + \gamma}$. Here $\|\mathcal{R}(Q_{T_0})\|_\infty$ denotes the infinity norm over all coefficients that appear in Q_{T_0}..

Proof. Suppose $E \in \mathbb{R}^{\binom{n}{\leq d}}$ be such that $\|E_{\overline{\emptyset}}\|_2 \leq r$ and $E_\emptyset = 0$. Let $\mathcal{L}$ be the linear functional with the representation $\mathcal{R}(\mathcal{L}) = \mathcal{R}(\mathcal{L}_0) + E$. Our goal is to choose r, in a way that for every choice $E_{\overline{\emptyset}}$, where $\|E_{\overline{\emptyset}}\|_2 \leq r$, we can prove that $\mathcal{L} \vdash_\tau Q_{T_0 + \gamma}$.

1. $\mathcal{L}1 = \mathcal{L}_0 1 = 1$.

2. For every regular constraint q, we have that
$$\mathcal{L}q = \mathcal{L}_0 q + \langle E, \mathcal{R}(q)\rangle \geq -\tau \cdot T_0 - r\|q\|_\infty.$$

3. For all h such that $2\deg h \leq d$ and $\|h\|_2 \leq 1$ we have that
$$\mathcal{L}h^2 = \mathcal{L}_0 h^2 + \langle E, \mathcal{R}(h^2)\rangle \geq -\tau \cdot T_0 - r\,\mathrm{poly}(n^d).$$

4. For every PSD constraint q, excluding the T-constraint, and every polynomial h such that $2\deg h \leq d - \deg q$, and $\|h\|_2 \leq 1$ we have that
$$\begin{aligned}\mathcal{L}qh^2 &= \mathcal{L}_0 qh^2 + \langle E, \mathcal{R}(qh^2)\rangle \\ &\geq -\tau \cdot T_0 - r\|\mathcal{R}(q)\|_\infty \cdot \mathrm{poly}(n^d).\end{aligned}$$

5. Let $c = 1/2T_{\max}$. For the T-constraint $q_{T_0+\gamma}$, and every polynomial h such that $2\deg h \leq d - \deg q_{T_0+\gamma}$, and $\|h\|_2 \leq 1$, we have that
$$\begin{aligned}\mathcal{L}q_{T_0+\gamma}h^2 &= \mathcal{L}(q_{T_0} + c\gamma)h^2 \\ &= \mathcal{L}_0 q_{T_0}h^2 + c\gamma\mathcal{L}h^2 + \langle E, \mathcal{R}(q_{T_0}h^2)\rangle \\ &\geq -\tau \cdot T_0 - c\gamma\left(\tau \cdot T_0 + r\,\mathrm{poly}(n^d)\right) - r \cdot \|\mathcal{R}(q_{T_0})\|_\infty \cdot \mathrm{poly}(n^d)\end{aligned}$$

6. Let $\mathcal{E}$ be the corresponding linear functional for E. For every $k \times k$ matrix PSD constraint q, we have that
$$\begin{aligned}\|\mathcal{E}q\|_2 &\leq \sqrt{k}\|\mathcal{E}q\|_\infty \\ &= \sqrt{k}\max_{i,j}\left|\langle E, \mathcal{R}(q_{i,j})\rangle\right| \\ &\leq r \cdot \sqrt{k} \cdot \mathrm{poly}(n^d) \cdot \max_{i,j}\|\mathcal{R}(q_{i,j})\|_\infty.\end{aligned}$$

   Therefore
$$\mathcal{L}q \succcurlyeq \mathcal{L}_0 q + \mathcal{E}q \succcurlyeq -\tau \cdot T - r \cdot \sqrt{k} \cdot \mathrm{poly}(n^d) \cdot \max_{i,j}\|\mathcal{R}(q_{i,j})\|_\infty.$$

7. We have
$$\|\mathcal{R}(cL)\|_2 \leq \|\mathcal{R}(\mathcal{L}_0)\|_2 + \|E\|_2 \leq R + \tau \cdot T_0 + r.$$

Therefore it suffices to take r such that

1. $r \cdot \mathrm{poly}(n^d) \leq \tau\gamma$. In order to do this take $r \leq \tau\gamma/\mathrm{poly}(n^d)$.

2. $c\gamma r\,\mathrm{poly}(n^d) + r \cdot \|\mathcal{R}(q_{T_0})\|_\infty \cdot \mathrm{poly}(n^d) \leq \tau\gamma/2$. In order to do this take r to be
$$r \leq \frac{\tau}{4\,\mathrm{poly}(n^d)} \cdot \min\left(\frac{\gamma}{\left\|\mathcal{R}(q_{T_0})\right\|_\infty}, \frac{1}{T_{\max}}\right),$$

3. $r \cdot \sqrt{k} \cdot \mathrm{poly}(n^d) \cdot \max_{i,j} \|\mathcal{R}(q_{i,j})\|_\infty \le \tau \cdot \gamma$. In order for this to hold take r to be

$$r \le \frac{\tau\gamma}{\sqrt{k}\,\mathrm{poly}(n^d)\max_{i,j}\|\mathcal{R}(q_{i,j})\|_\infty}.$$

Therefore there exists a ball of radius $\mathrm{poly}(1/\mathrm{poly}(n^d), \tau, \gamma, 1/k, 1/\|R(Q_{T_0})\|_\infty)$ such that for every $\mathcal{R}(\mathcal{L})_{\overline{\varnothing}}$ in that ball we have that $\mathcal{L} \vDash_\tau Q_{T_0+\gamma}$, as desired.

□

Lemma C.5. *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption C.1. Suppose there exists some linear functional $\mathcal{L}$ such that $\mathcal{L} \vDash_{\tau,\gamma} Q_T$. Then if $\gamma \le T_{\max}/2$, we have that $\mathcal{L} \vDash_\tau Q_{T+4\gamma}$.*

Proof. All of the inequalities in $\mathcal{L} \vDash_\tau Q_{T+4\gamma}$ will be trivially satisfied because of $\mathcal{L} \vDash_{t+\gamma} Q_T$ except for the T-constraint. So we should prove the inequality for the T-constraint. Suppose h is a polynomial such that $\|h\|_2 \le 1$, and $2\deg h \le d - \deg q_T$. Then

$$\begin{aligned}
\mathcal{L}q_{T+4\gamma}h^2 &= \mathcal{L}q_T h^2 + \frac{4\gamma}{2T_{\max}}\mathcal{L}h^2 \\
&\ge -\tau\cdot(T+\gamma) - 2\gamma\tau\cdot(T+\gamma)/T_{\max} \\
&\ge -\tau\cdot(T+\gamma) - 3\gamma\tau \\
&\ge -\tau\cdot(T+4\gamma),
\end{aligned}$$

as desired.

□

Theorem C.6 (computability of score functions)**.** *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption C.1. Let*

$$T_0 = \inf_T \textit{ such that there exists } \mathcal{L} \textit{ such that } \mathcal{L} \vDash_\tau Q_T.$$

Then we can compute T_0 in time $\mathrm{poly}(n^d, \mathrm{Size}(Q_T), \log(R), \log(T_{\max}), \log(1/\gamma), \log(1/\tau))$ up to error $O(\gamma)$. Note that R is as in Definition C.2.

Proof. We apply binary search in order to estimate T_0. Suppose T is given, run the ellipsoid algorithm from Lemma C.3, either we can find some functional $\mathcal{L}$ such that $\mathcal{L} \vDash_{\tau,\gamma} Q_T$, or a proof that no ball of radius $r(\gamma)$ of functionals $\mathcal{L}$ that satisfy $\mathcal{L} \vDash_\tau Q_T$ exists. Note that $r(\gamma)$ here is as in Lemma C.4. If we are in the first case, by Lemma C.5 we know that $\mathcal{L} \vDash_\tau Q_{T+4\gamma}$. Therefore, $T + 4\gamma \ge T_0$, and we decrease the value of T. If we are in the second case, we must have $T < T_0 + \gamma$, since otherwise we know that by Lemma C.4 there should exists a ball of radius $r(\gamma)$. This gives us an efficient algorithm for approximating the score function. □

D High-Probability Bound for Stability of Covariance

D.1 Preliminaries

Lemma D.1. *[DKK+19, Corollary 4.8, rephrased] There exists $\alpha = O(\eta \log \frac{1}{\eta})$, such that for any $n \geq O\left(\frac{d^2+\log(1/\beta)}{\alpha^2}\right)$ and $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$, then with probability at least $1-\tau$, for all symmetric matrices $P \in \mathbb{R}^{d\times d}$ with Frobenius norm 1 and all $b \in [0,1]^n$ with $\mathbf{E}_i\, b_i \geq 1-\eta$,*

$$\frac{1}{n}\sum_{i=1}^{n} b_i \langle x_i x_i^\top - I, P\rangle \leq \alpha.$$

Lemma D.2 (Hanson-Wright Inequality)**.** *Let $x \sim \mathcal{N}(\mathbf{0}, I)$ be a d-dimensional Gaussian vector. Then, there exists a universal constant c such that for any symmetric matrix P and for all $t > 0$,*

$$\mathbb{P}\left[\left|\langle xx^\top - I, P\rangle\right| > t\right] \leq 2\exp\left(-c \cdot \min\left(\frac{t^2}{\|P\|_F^2}, \frac{t}{\|P\|_{op}}\right)\right).$$

Proposition D.3 (Theorem 4.5, [Ver18])**.** *Let A be a rectangular $m \times n$-dimensional matrix with each entry i.i.d. $\mathcal{N}(0,1)$. Then, there exists a constant C_0 such that for any $t \geq 0$, $\mathbb{P}(\|A\|_{op} > C_0(\sqrt{m}+\sqrt{n}+t)) \leq 2e^{-\Omega(t^2)}$.*

Proposition D.4. *For some fixed $1 \leq k \leq d$, let $\mathcal{P}$ be the set of symmetric $d \times d$ matrices with Frobenius norm at most 1 and all nonzero eigenvalues at least $\sqrt{1/k}$ in absolute value. Then, for any $0 < \gamma < 1/2$, $\mathcal{P}$ has a γ-net (in the Frobenius norm distance) of size $(1/\gamma)^{O(k\cdot d)}$.*

Proof. For such a $P \in \mathcal{P}$, note that P must have rank at most k. Therefore, we can write $P = UDU^\top$, where D is a diagonal matrix of Frobenius norm at most 1 and U is a $d \times k$-dimensional matrix with orthonormal columns. Let $\mathcal{T}$ be a $\gamma/10$-net of the d-dimensional unit sphere, of size $(1/\gamma)^{O(d)}$. Define $\mathcal{V} \in \mathbb{R}^{d\times k}$ to be the set of $d \times k$-matrices where each column is in $\mathcal{T}$. Then, every orthogonal $U \in \mathbb{R}^{d\times k}$ has a corresponding $V \in \mathcal{V}$ such that each corresponding column in U, V are unit vectors of distance at most $\gamma/10$. Therefore, there exists a set $\mathcal{W}$ of orthogonal matrices in $\mathbb{R}^{d\times k}$ such that every U has a corresponding W where $\|U - W\|_F \leq k \cdot \gamma/5$. $\mathcal{W}$ is created by choosing a single representative near each $V \in \mathcal{V}$, should one exist, which means $|\mathcal{W}| \leq (1/\gamma)^{O(d\cdot k)}$. Finally, let $\mathcal{T}'$ be a $\gamma/5$-net of the unit ball in k-dimensions, which corresponds to a $\gamma/5$-net $\mathcal{D}$ of diagonal matrices of Frobenius norm at most 1.

Now, we claim that the set of matrices $WD'W^\top$, for $W \in \mathcal{W}$ and $D' \in \mathcal{D}$, form a γ-net for the set of P. Indeed, for any $P = UDU^\top$, we associate U with W such that each column of U and of W differ by at most $\gamma/5$ in ℓ_2-distance, and D with D' such that $\|D - D'\|_F \leq \gamma/5$. We want to show that $\|UDU^\top - WD'W^\top\|_F \leq \gamma$.

Note we can bound $\|UDU^\top - WD'W^\top\|_F \leq \|UD(U-W)^\top\|_F + \|(U-W)DW^\top\|_F + \|W(D'-D)W^\top\|_F$, so it suffices to bound each of these terms by $\gamma/5$. Since U and W are orthogonal matrices, $\|UM\|_F = \|WM\|_F = \|M\|_F$ and $\|MU^\top\|_F = \|MW^\top\|_F = \|M\|_F$ for any matrix M (fitting the dimensions). Therefore, it suffices to show that $\|D(U-W)^\top\|_F$, $\|(U-W)D\|_F$, and $\|D'-D\|_F \leq \gamma/5$. Indeed, we already know $\|D'-D\|_F \leq \gamma/5$, and $\|D(U-W)^\top\|_F = \|(U-W)D\|_F$ since D is diagonal, so we just need to show $\|(U-W)D\|_F \leq \gamma/5$. To prove this, note that $U-W$ is

a $d \times k$-dimensional matrix with very column having ℓ_2 norm at most $\gamma/5$. When we multiply by D, this multiplies the ith column of $U - W$ by D_{ii}, the ith diagonal entry of D. Therefore, the ith column of $(U - W)D$ has ℓ_2 norm at most $\gamma/5 \cdot D_{ii}$. Therefore, the Frobenius norm of $(U - W)D$ is at most $\sqrt{\sum_{i=1}^{k} (\gamma/5)^2 \cdot D_{ii}^2} = \gamma/5 \cdot \sqrt{\sum_{i=1}^{k} D_{ii}^2} = \gamma/5$.

Finally, the size of this net is at most $|\mathcal{W}| \cdot |\mathcal{D}| \leq (1/\gamma)^{O(d \cdot k)} \cdot (1/\gamma)^{O(k)} = (1/\gamma)^{O(d \cdot k)}$. □

D.2 Main Probability Bound

Lemma D.5. *Let $n \geq \widetilde{O}\left(\frac{(d+\log(1/\delta))^2}{\eta^2}\right)$ and let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$. Then, with probability at least $1 - \delta$, for any $d \times d$ symmetric matrix P with Frobenius norm ≤ 1, and for any subset $S \subset [n]$ of size at most $\eta \cdot n$,*

$$\sum_{i \in S} \langle x_i x_i^\top - I, P \rangle^2 \leq O\left(\eta \log^2 \frac{1}{\eta}\right) \cdot n.$$

Proof. For simplicity, we may assume without loss of generality that $\delta = e^{-d}$. This is because if $\delta > e^{-d}$, we can decrease the failure probability to e^{-d}. Likewise, if $\delta < e^{-d}$, then $d < \log(1/\delta)$, so we may increase the dimension to $\log(1/\delta)$ by sampling additional random standard Gaussians for the rest of the coordinates of x_i, and then only proving the result for all P with all nonzero values supported on the first d rows and columns.

Let $\mathcal{R}$ be a $1/2$-net of the set of symmetric matrices with Frobenius norm at most 1, and suppose we successfully prove the lemma for all $P \in \mathcal{R}$. Then, for a general P, we can write $P = \sum_{i=0}^{\infty} 2^{-i} R_i$, for each $R_i \in \mathcal{R}$. Then, for any i,

$$\left\langle x_i x_i^\top - I, \sum_{i=0}^{\infty} 2^{-i} R_i \right\rangle^2 = \left(\sum_{i=0}^{\infty} 2^{-i} \cdot \langle x_i x_i^\top - I, R_i \rangle\right)^2 \leq \left(\sum_{i=0}^{\infty} 2^{-i}\right) \cdot \left(\sum_{i=0}^{\infty} 2^{-i} \langle x_i x_i^\top - I, R_i \rangle^2\right),$$

using the Cauchy-Schwarz inequality. Then, we can write

$$\sum_{i \in S} \langle x_i x_i^\top - I, P \rangle^2 \leq 2 \cdot \left(\sum_{i=0}^{\infty} 2^{-i} \cdot \sum_{i \in S} \langle x_i x_i^\top - I, R_i \rangle^2\right) \leq 4 \cdot \widetilde{O}(\eta) \cdot n.$$

So it suffices to show the theorem for the net $\mathcal{R}$.

Next, note that for a sufficiently large constant C_0,

$$\begin{aligned}
\sum_{i \in S} \langle x_i x_i^\top - I, P \rangle^2 &= \sum_{i \in S} \int_{t=0}^{\infty} \mathbb{I}\left[t \leq \langle x_i x_i^\top - I, P \rangle^2\right] dt \\
&= \int_{t=0}^{\infty} \#\left\{i \in S : \langle x_i x_i^\top - I, P \rangle^2 \geq t\right\} dt \\
&\leq (C_0 \log(1/\eta))^2 \cdot \eta n + \int_{t=(C_0 \log(1/\eta))^2}^{\infty} \#\left\{i : \langle x_i x_i^\top - I, P \rangle^2 \geq t\right\} \cdot (t \log^2 t) \cdot \frac{1}{t \log^2 t} dt \\
&\leq (C_0 \log(1/\eta))^2 \cdot \eta n + \max_{t \geq (C_0 \log(1/\eta))^2} \left(t \log^2 t \cdot \#\left\{i : \langle x_i x_i^\top - I, P \rangle^2 \geq t\right\}\right) \\
&\leq (C_0 \log(1/\eta))^2 \cdot \eta n + \max_{C \geq C_0 \log(1/\eta)} \left(C^2 \log^2 C \cdot \#\left\{i : \left|\langle x_i x_i^\top - I, P \rangle\right| \geq C\right\}\right).
\end{aligned}$$

The second-to-last line uses the fact that $\int_3^\infty \frac{1}{t \log^2 t} dt < 1$, and the last line is just a substitution $C = \sqrt{t}$.

Therefore, it will suffice to show that for all $C \geq C_0 \log(1/\eta)$, with probability at least $1 - e^{-d}/C$ the following holds. For all $P \in \mathcal{R}$, the number of $i \in [n]$ such that $|\langle x_i x_i^\top - I, P \rangle| \geq C$ is at most $n \cdot \eta/(C^2 \log^2 C)$. The probability bound is sufficient since it suffices to prove this for all C that is a power of 2, and the sum of e^{-d}/C over C a power of 2 is e^{-d}. □

So, to prove Lemma D.5, it suffices to prove the following lemma.

Lemma D.6. *Suppose $\frac{n}{\log^{20} n} \geq O\left(\frac{d^2}{\eta^2}\right)$ and let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$. Then, there exists a sufficiently large constant C_0 and a $1/2$-net $\mathcal{R}$ for $d \times d$ symmetric matrices with Frobenius norm at most 1, such that for any $C \geq C_0 \log(1/\eta)$, with probability at least $1 - e^{-d}/C$, for any $P \in \mathcal{R}$, the number of indices $i \in [n]$ such that $\langle x_i x_i^\top - I, P \rangle \geq C$ is at most $n \cdot \eta/(C^2 \log^2 C)$.*

Proof. First, assume that $C \leq \sqrt{\eta n/(\log^9 n \cdot d)}$. For $j \geq 1$, let $\mathcal{P}_j$ be a $\gamma_j := (1/(10j^2))$-net of the matrices in $\mathcal{P}$ with all nonzero eigenvalues in the range $[-2/\sqrt{2^j}, -1/\sqrt{2^j}] \cup [1/\sqrt{2^j}, 2/\sqrt{2^j}]$. Also, let $\mathcal{Q}_j$ be a $1/10$-net of the set of matrices in $\mathcal{P}$ with all eigenvalues below $1/\sqrt{2^j}$ in absolute value.

Now, for some fixed P, suppose we can write $P = P_1 + P_2 + \cdots + P_{\lceil \log_2 C^2 \rceil} + Q$, where each $P_j \in \mathcal{P}_j$ and $Q \in \mathcal{Q}_{\lceil \log_2 C^2 \rceil}$. Then, if the event that $\langle P, xx^\top - I \rangle \geq C$ holds, then we must have that either $\langle P_j, xx^\top - I \rangle \geq C/(4j^2)$ for some j or $\langle Q, xx^\top - I \rangle \geq C/2$. For any fixed choice of $\{P_j\}_{1 \leq j \leq \lceil \log_2 C^2 \rceil}$ and Q, the probability that this event occurs for each P_j is at most $\exp\left(-c_1 \min\left(\frac{C^2}{j^4}, \frac{C \cdot 2^{j/2}}{j^2}\right)\right) \leq \exp\left(-c_1 \cdot \frac{C \cdot 2^{j/2}}{j^4}\right)$, by the Hanson-Wright inequality and since $2^{j/2} \leq 2C$ for $j \leq \lceil \log_2 C^2 \rceil$. The probability that this event holds for Q, by Hanson-Wright, is at most $\exp\left(-c_1 \cdot \min\left(C^2, C \cdot 2^{\lceil \log_2 C^2 \rceil/2}\right)\right) \leq \exp\left(-c_1 \cdot C^2\right)$.

For a fixed $P = P_1 + P_2 + \cdots + P_{\lceil \log_2 C^2 \rceil} + Q$, and for $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I)$, we bound the probability of the event that $\langle P, x_i x_i^\top - I \rangle \geq C$ for at least $\eta \cdot n/(C^2 \cdot \log^2 C)$ different choices of $i \in [n]$. For simplicity we define $\bar{C} = C^2 \cdot (\log C)^2/\eta$. Now, if the event holds, then either some $\langle P_j, x_i x_i^\top - I \rangle \geq C/(4j^2)$ for $n/(4\bar{C}j^2)$ indices, or $\langle Q, x_i x_i^\top - I \rangle \geq C/2$ for $n/(2\bar{C})$ different choices of $i \in [n]$. For fixed $j \leq \lceil \log_2 C^2 \rceil$, the probability of this occurring for P_j is at most

$$\begin{aligned}
\binom{n}{n/(4\bar{C}j^2)} \cdot \exp\left(-c_1 \cdot \frac{C \cdot 2^{j/2}}{j^4} \cdot \frac{n}{4\bar{C}j^2}\right) &\leq O(\bar{C}j^2)^{n/(4\bar{C}j^2)} \cdot \exp\left(-c_1 \cdot \frac{C \cdot 2^{j/2}}{j^4} \cdot \frac{n}{4\bar{C}j^2}\right) \\
&\leq \exp\left(-c_2 \cdot \frac{n \cdot 2^{j/2} \cdot C}{\bar{C} \cdot j^6}\right) \\
&= \exp\left(-c_2 \cdot \frac{n \cdot 2^{j/2} \cdot \eta}{C(\log C)^2 \cdot j^6}\right).
\end{aligned}$$

where we used the fact that $\log(\bar{C}j^2) \leq O(\log(C/\eta))$, which is much smaller than $C \leq O(C \cdot 2^{j/2}/j^4)$ since $C \geq C_0 \log(1/\eta)$.

Likewise, the probability of this occurring for Q is at most

$$\begin{aligned}\binom{n}{n/(2\bar{C})} \cdot \exp\left(-c_1 \cdot C^2 \cdot \frac{n}{2\bar{C}}\right) &\le O(\bar{C})^{n/(2\bar{C})} \cdot \exp\left(-c_1 \cdot C^2 \cdot \frac{n}{2\bar{C}}\right) \\ &\le \exp\left(-c_2 \cdot C^2 \cdot \frac{n}{\bar{C}}\right) \\ &= \exp\left(-c_2 \cdot \frac{\eta \cdot n}{(\log C)^2}\right).\end{aligned}$$

Finally, recall that $|\mathcal{P}_j| \le O(j^2)2^{j \cdot d} = e^{O(\log j \cdot 2^j \cdot d)}$ and $|\mathcal{Q}_{\lceil \log_2 C^2 \rceil}| = e^{O(d^2)}$. So overall, the probability of there even existing such a P that can be written as $P_1 + \cdots + P_{\lceil \log_2 C^2 \rceil} + Q$ where each $P_j \in \mathcal{P}_j$ and $Q \in \mathcal{Q}_{\lceil \log_2 C^2 \rceil}$ is at most

$$\sum_{j=1}^{\lceil \log_2 C^2 \rceil} \exp\left(-c_2 \cdot \frac{n \cdot 2^{j/2} \cdot \eta}{C(\log C)^2 \cdot j^6}\right) \cdot \exp\left(C_1 \cdot \log j \cdot 2^j \cdot d\right) + \exp\left(-c_2 \cdot \frac{n \cdot \eta}{(\log C)^2}\right) \cdot \exp\left(C_1 \cdot d^2\right). \quad (13)$$

Since $C \le \sqrt{\eta n/(\log^9 n \cdot d)}$ and $2^j \le 2C^2$, this means $\frac{n \cdot 2^{j/2} \cdot \eta}{C(\log C)^2 \cdot j^6} \gg \log j \cdot 2^j \cdot d$. To see why, this is equivalent to $\eta \cdot \frac{n}{d} \gg C(\log C)^2 \cdot j^6 \log j \cdot 2^{j/2}$, and since $2^{j/2} \le 2C$ and $C \ll n$, this is implied by $\eta \cdot \frac{n}{d} \gg C^2(\log n)^9$. In addition, assuming that $n \gg (\log n)^2 \cdot d^2/\eta$, we also have that $\frac{n \cdot \eta}{(\log C)^2} \gg d^2$. Therefore, we can further bound (13) by

$$n \cdot \max_{j \le \lceil \log_2 C^2 \rceil} \exp\left(-c_3 \cdot \frac{n \cdot 2^{j/2} \cdot \eta}{C(\log C)^2 \cdot j^6}\right) + \exp\left(-c_3 \cdot \frac{n \cdot \eta}{(\log C)^2}\right) \le (n+1) \cdot \exp\left(-c_4 \frac{\eta \cdot n}{C(\log C)^2}\right) \le e^{-d}/C.$$

So, our probability bound is sufficient, but we need to make sure that the set $\mathcal{R}$ of matrices that can be written as $P_1 + \cdots + P_{\lceil \log_2 C^2 \rceil} + Q$ for $P_j \in \mathcal{P}_j$ and $Q \in \mathcal{Q}_{\lceil \log_2 C^2 \rceil}$ is a $1/2$-net. However, by looking at the singular value decomposition of any symmetric matrix $\tilde{P}$ with $\|P\|_F = 1$, we can write it as $\tilde{P}_1 + \cdots + \tilde{P}_{\lceil \log_2 C^2 \rceil} + \tilde{Q}$, where $\tilde{P}_j$ has all nonzero eigenvalues in $[-2/\sqrt{2^j}, -1/\sqrt{2^j}] \cup [1/\sqrt{2^j}, 2/\sqrt{2^j}]$ and $\tilde{Q}$ has all eigenvalues at most $1/\sqrt{2^{\lceil \log_2 C^2 \rceil}}$ in absolute value. In addition, each $\tilde{P}_j$ is within distance $1/(10j^2)$ of some $P_j \in \mathcal{P}_j$ and $\tilde{Q}$ is within distance $1/10$ of some $Q \in \mathcal{Q}_{\lceil \log_2 C^2 \rceil}$. So, by the triangle inequality, $\mathcal{R}$ is a $1/2$-net.

Next, suppose $C \ge \sqrt{\eta n/(\log^9 n \cdot d)}$, but $C \le \sqrt{\eta n/(\log^2 n)}$ so $\bar{C} \le n$. Then, for any fixed choice of $n/\bar{C} = \eta \cdot n/(C^2(\log C)^2)$ indices S, the probability that there exists $P \in \mathbb{R}^{d \times d}$ such that $\|P\|_F = 1$ and $\langle x_i x_i^\top - I, P \rangle \ge C$ for all $i \in S$ is at most

$$\begin{aligned}\mathbb{P}\left(\exists P : \sum_{i \in S} \langle x_i x_i^\top - I, P \rangle \ge C \cdot \frac{n}{\bar{C}}\right) &= \mathbb{P}\left(\left\|\sum_{i \in S} (x_i x_i^\top - I)\right\|_F \ge C \cdot \frac{n}{\bar{C}}\right) \\ &\le \mathbb{P}\left(\left\|\sum_{i \in S} x_i x_i^\top\right\|_F \ge \frac{n \cdot C}{\bar{C}} - \frac{n\sqrt{d}}{\bar{C}}\right) \\ &\le \mathbb{P}\left(\left\|\sum_{i \in S} x_i x_i^\top\right\|_F \ge \frac{n \cdot C}{2\bar{C}}\right).\end{aligned}$$

The second line is true because the Frobenius norm of I is $\sqrt{d}$ so the Frobenius norm of $|S| \cdot I$ is $\frac{n\sqrt{d}}{\bar{C}}$. The third line is true because $C \geq 2\sqrt{d}$ if $n \geq 4d^2 \log^9 n/\eta$.

So, for the final event to occur, it is equivalent for $\|AA^\top\|_F \geq \frac{n \cdot C}{2\bar{C}}$, where A is the $(n/\bar{C}) \times d$-dimensional matrix with each row of A being x_i for $i \in S$. Since A and therefore $AA^\top$ have rank at most $n/\bar{C}$, this requires $\|A\|_{op}^2 = \|AA^\top\|_{op} \geq \frac{n \cdot C}{2\bar{C}} \cdot \sqrt{\frac{\bar{C}}{n}} \geq \frac{\sqrt{n} \cdot C}{2\sqrt{C}} = \frac{\sqrt{\eta \cdot n}}{2 \log C}$.

Assuming $n \gg d^2 \log^2 n/\eta$, then $\frac{\sqrt{\eta \cdot n}}{2 \log C} \gg d$. Also, assuming $n \gg d^2 \log^{18} n/\eta$, then $\sqrt{\eta n} \gg d \log^9 n$, which means $\sqrt{\eta n} \ll \frac{\eta n}{\log^9 n \cdot d} \leq C^2 \cdot \log C$. This means that $\frac{n}{\bar{C}} = \frac{\eta \cdot n}{C^2 (\log C)^2} \ll \frac{\sqrt{\eta \cdot n}}{2 \log C}$. So, this means $\frac{\sqrt{\eta \cdot n}}{2 \log C} \gg d + \frac{n}{\bar{C}}$, which means that by Proposition D.3, the probability of $\|A\|_{op}^2 \geq \frac{\sqrt{\eta \cdot n}}{2 \log C}$ is at most $2 \exp\left(-\Omega\left(\frac{\sqrt{\eta \cdot n}}{\log C}\right)\right)$.

Therefore, for any fixed choice of n/C^2 indices, the probability that there exists $P \in \mathbb{R}^{d \times d}$ such that $\|P\|_F = 1$ and $\langle x_i x_i^\top - I, P \rangle \geq C$ for all $i \in S$ is at most $2 \exp\left(-\Omega\left(\frac{\sqrt{\eta \cdot n}}{\log C}\right)\right)$. There are at most $\binom{n}{n/\bar{C}} \leq e^{\log n \cdot n/\bar{C}} \leq e^{d \log^{10} n / \log^2 C}$. Note that $\frac{d \log^{10} n}{\log^2 C} \ll \frac{\sqrt{\eta \cdot n}}{\log C}$ for any $n \gg d \log^{20} n/\eta$. So, this means that the overall failure probability is at most $2 \exp\left(-\Omega\left(\frac{\sqrt{\eta \cdot n}}{\log C}\right)\right) \leq e^{-d}/C$.

The final case is if $C \geq \sqrt{\eta \cdot n/(\log^2 n)}$. In this case, the probability that even a single index has $\|x_i x_i^\top - I\|_F \geq C$ means $\|x_i x_i^\top\|_F \geq \frac{C}{2}$ which means $\|x_i\|_2^2 \geq \frac{C}{2}$. We can again apply Hanson-Wright to conclude that, since $C \gg d$, the probability that $\|x_i\|_2^2 \geq \frac{C}{2}$ is at most $2e^{-\Omega(-C)}$, which means the probability that this is true for even a single x_i is at most $2ne^{-\Omega(C)} \leq e^{-d}/C$. □

D.3 Proof of Lemma 6.3

First, we note the following corollary of Lemma D.5.

Corollary D.7. *With probability at least $1 - \beta$, every $d \times d$ symmetric matrix P with Frobenius norm exactly 1, $\frac{1}{n} \sum_{i=1}^n \langle x_i x_i^\top - I, P \rangle^2 = 2 \pm O\left(\eta \cdot \log^2 \frac{1}{\eta}\right)$.*

Proof. Suppose $x_1, \ldots, x_n$ has the property of Lemma D.5. Now, for a fixed P with $\|P\|_F = 1$, note that $\sum_{i=1}^n \langle x_i x_i^\top - I, P \rangle^2 \leq \sum_{i=1}^n \min\left(\langle x_i x_i^\top - I, P \rangle^2, C_0 \log^2 \frac{1}{\eta}\right) + O(n \cdot \eta \cdot \log^2 \frac{1}{\eta})$. This is because the number of indices i such that $\langle x_i x_i^\top - I, P \rangle^2 \geq C_0 \log^2 \frac{1}{\eta}$ is at most $O(\eta \cdot n)$ by Lemma D.5, and for those indices, we know that $\sum \langle x_i x_i^\top - I, P \rangle^2$ is at most $O(n \cdot \eta \cdot \log^2 \frac{1}{\eta})$.

Now, note that for any fixed P with $\|P\|_F = 1$, $\mathbb{E}_{x \sim \mathcal{N}(\mathbf{0}, I)} \langle xx^\top - I, P \rangle^2 = 2$. Indeed, this is simple to see if P is diagonal (using the fact that the fourth moment of a Gaussian is 3), and for general symmetric P we can diagonalize P and use the same diagonalization on each x_i, to show this is true. Therefore, since $\mathbb{P}(\langle xx^\top - I, P \rangle^2 \geq t^2) \leq 2e^{-\Omega(t)}$ by Hanson-Wright, this means if C_0 is sufficiently large, $\mathbb{E}_{x \sim \mathcal{N}(\mathbf{0}, I)} \min\left(\langle xx^\top - I, P \rangle^2, C_0 \log^2 \frac{1}{\eta}\right) \in [2 - \eta, 2]$. In addition, this variable is bounded between 0 and $C_0 \log^2 \frac{1}{\eta}$, so by Hoeffding's inequality, the probability that $\frac{1}{n} \cdot \sum_{i=1}^n \min\left(\langle xx^\top - I, P \rangle^2, C_0 \log^2 \frac{1}{\eta}\right)$ is not in the range $[2 - 2\eta, 2 + 2\eta]$ is at most $e^{-2n\eta^2 / \log^4 \frac{1}{\eta}}$.

We can union bound over a $1/n^2$-net of symmetric matrices with Frobenius norm 1, which has

size $e^{O(d^2 \log d)}$, to say that if $n \gg \frac{d^2}{\eta^2} \cdot \log n \log^4 \frac{1}{\eta}$, then with probability at least $e^{-n \cdot \eta^2 / \log^4 \frac{1}{\eta}}$, every P in the net satisfies $\frac{1}{n} \cdot \sum_{i=1}^n \min\left(\langle xx^\top - I, P\rangle^2, C_0 \log^2 \frac{1}{\eta}\right) \in [2-2\eta, 2+2\eta]$.

For a general P, write $P = P_0 + P'$, where P_0 is in the net and $\|P'\|_F \le 1/n^2$. Assuming the event of Lemma D.5, for every choice of P' and every choice of x_i, $\langle x_i x_i^\top - I, P'\rangle \le \frac{1}{n}$. Therefore, the difference between $\min\left(\langle xx^\top - I, P\rangle^2, C_0 \log^2 \frac{1}{\eta}\right)$ and $\min\left(\langle xx^\top - I, P_0\rangle^2, C_0 \log^2 \frac{1}{\eta}\right)$ is always at most $O\left(\frac{1}{n} \cdot \log^2 \frac{1}{\eta}\right) \le \eta$. So, for every symmetric matrix P with Frobenius norm 1, we have that $\frac{1}{n} \cdot \sum_{i=1}^n \min\left(\langle xx^\top - I, P\rangle^2, C_0 \log^2 \frac{1}{\eta}\right) \in [2-3\eta, 2+3\eta]$ with probability at least $1-\beta$, as long as $n \ge \widetilde{O}\left(\frac{(d+\log(1/\beta))^2}{\eta^2}\right)$.

Therefore, $\sum_{i=1}^n \langle x_i x_i^\top - I, P\rangle^2 = \left(2 \pm \widetilde{O}(\eta \cdot \log^2 \frac{1}{\eta})\right) \cdot n$, as desired. □

Proof of Lemma 6.3. Part 1 and the first half of Part 3 are immediate from Lemma D.1. The second half of Part 3 follows from Lemma D.5 and Part 2 follows from Corollary D.7. Finally, Part 4 follows from Lemma D.1 in the same way that Part 4 of Corollary 5.4 follows from Lemma 5.3. For instance, we can set $\eta = 0.01$ to obtain that for any subset S of size at most $0.01n$, $\frac{1}{n} \cdot \sum_{i \in S} c_i \langle x_i x_i^\top - I, P\rangle \le O(1)$ for any choice of $c_i \in \{-1, 1\}$, which means $\frac{1}{n} \cdot \sum_{i \in S} |\langle x_i x_i^\top - I, P\rangle| \le O(1)$. We can then partition $[n]$ into 100 such sets S. □

E Mean Estimation in ℓ_∞

In this section, we start by providing an algorithm for robust Gaussian mean estimation in ℓ_∞ distance (Proposition E.1). We then show that this robust algorithm allows us to derive a pure DP algorithm with better sample complexity than a black-box application of Lemma 2.1 (Proposition E.3).

Proposition E.1. *There is a robust estimator $\hat{\mu}_0 : (\mathbb{R}^d)^n \to \mathbb{R}$ such that for every $\mu \in \mathbb{R}^d$ and small-enough $\eta > 0$, with high probability over $x_1, \dots, x_n \sim \mathcal{N}(\mu, I)$, letting $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, given any η-corruption $y_1, \dots, y_n$ of $x_1, \dots, x_n$, $\|\hat{\mu}(y_1, \dots, y_n) - \overline{x}\|_2 \le O(\sqrt{\eta d (\log n)/n} + \eta\sqrt{\log n})$ and $\|\hat{\mu} - \overline{x}\|_\infty \le O(\eta\sqrt{\log n})$, as long as $n \gg d$.*

To prove the proposition, we establish a few facts about $x_1, \dots, x_n \sim \mathcal{N}(\mu, I)$.

Fact E.2. *The following all hold with high probability for $x_1, \dots, x_n \sim \mathcal{N}(\mu, I)$, with $\mu \in \mathbb{R}^d$, and letting $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, if $n \gg d$.*

1. *For a big-enough constant C and all $t \in \{C\sqrt{\log n}, 2C\sqrt{\log n}, 4C\sqrt{\log n}, \dots, d\}$*

$$\sup_{\|v\|=1} \sum_{i \le n} 1[\langle x_i - \overline{x}, v\rangle > t] \le O(d(\log d + \log\log n)/t^2),$$

2. $\|x_i - \overline{x}\| \le O(\sqrt{d} + \sqrt{\log n})$

3. *every coordinate $i \in [d]$ and $j, k \le n$ have $|(x_j)_i - (x_k)_i| \le O(\sqrt{\log nd})$.*

Proof. First, the following simultaneously occur with high probability by standard Gaussian concentration arguments:

- each x_i has $\|x_i\| \le \sqrt{d} + O(\sqrt{\log n})$, and
- $\|\overline{x}\| \le O(\sqrt{d/n})$.

Now, let S be a δ-net of the ℓ_2 unit sphere; we can take S to have $2^{O(d\log(1/\delta))}$ elements. For any $x_1, \dots, x_n$ and $t > 0$, let $n_t = \sup_{v\in S} \sum_{i\le n} 1[\langle x_i, v\rangle > t]$. We claim that

$$\sum_{i\le n} 1[\langle x_i - \overline{x}, v\rangle > t] \le n_{t-\delta\cdot\max_i \|x_i\| - \|\overline{x}\|}\,.$$

To see this, we can write $v = w + \Delta$, where $w \in S$ and $\|\Delta\| \le \delta$. Then $\langle x_i - \overline{x}, v\rangle = \langle x_i, w\rangle - \langle \overline{x}, v\rangle + \langle x_i, \Delta\rangle > t$ only if $\langle x_i, w\rangle > t + \langle \overline{x}, v\rangle - \langle x_i, \Delta\rangle \ge t - \|\overline{x}\| - \delta\|x_i\|$. If $\max_i \|x_i\| \le 1/(2\delta)$ and $n \gg d$, then we get $\sum_{i\le n} 1[\langle x_i, v\rangle > t] \le n_{t-1}$.

We just need to establish a high-probability upper bound on n_{t-1} for $\delta \ll 1/(\sqrt{d} + O(\sqrt{\log n}))$ and $t \in \{C\sqrt{\log n}, 2\sqrt{\log n}, \dots, d\}$. If $x_1, \dots, x_n \sim \mathcal{N}(\mathbf{0}, I)$, then for any fixed $v \in S$ and fixed t, we have

$$\mathbb{P}\left(\sum_{i\le n} 1[\langle x_i, v\rangle > t] > s\right) \le n^s \exp(-\Omega(st^2))\,.$$

via a union bound over n^s choices of s indices $i \in [n]$. If $t \ge C\sqrt{\log n}$ and $s = O(d\max(\log d, \log\log n)/t^2)$, we can take a union bound over the net S and get that for any fixed t, $\sum_{i\le n} 1[\langle x_i, v\rangle > t] \le O(d\max(\log d, \log\log n)/t^2)$ with probability at least $1 - e^{-\Omega(d)}$; the proof for (1) is finished by a union bound over $O(\log d)$ choices of t.

The proof for (2) is standard Gaussian concentration, and the proof for (3) is a union bound over $n^2 d$ pairs $(x_j)_i, (x_k)_i$. □

Proof of Proposition E.1. Define the estimator $\hat{\mu}$ as: given $y_1, \dots, y_n$, find any $x'_1, \dots, x'_n$ which (a) agree with the y_is on $(1-\eta)n$ vectors and (b) have both properties in Fact E.2, and output $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x'_i$. If no such set $\{x'_i\}$ exists, output $\emptyset$.

With high probability over $x_1, \dots, x_n \sim \mathcal{N}(\mu, I)$, by Fact E.2, such a set of x's exists, since the xs are such a set.

Let's bound $\|\overline{x} - \overline{x'}\|_2$. Let $B \subseteq [n]$, $|B| \le 2\eta n$, be the indices where $x_i \ne x'_i$. For any unit $v \in \mathbb{R}^d$,

$$\langle \overline{x} - \overline{x'}, v\rangle = \frac{1}{n}\sum_{i\le n}\langle x_i - x'_i, v\rangle = \frac{1}{n}\sum_{i\in B}\langle x_i - \overline{x}, v\rangle - \frac{1}{n}\sum_{i\in B}\langle x'_i - \overline{x'}, v\rangle + \frac{|B|}{n}\langle \overline{x} - \overline{x'}, v\rangle\,.$$

For each of the sums, we group terms in the average by their magnitudes. Terms smaller than $O(\sqrt{\log n})$ can only contribute $O(\eta\sqrt{\log n})$. At most ηn terms are smaller than $\sqrt{d(\log d + \log\log n)/(\eta n)}$; they contribute at most $\sqrt{\eta d(\log d + \log\log n)/n}$. So we have

$$\frac{1}{n}\sum_{i\in B}\langle x_i - \overline{x}, v\rangle \le O(\eta\sqrt{\log n}) + O\left(\sqrt{\frac{\eta d(\log d + \log\log n)}{n}}\right) + \sum_{i\,:\,|\langle x_i - \overline{x}, v\rangle| > \sqrt{d(\log d + \log\log n)/\eta n}} \langle x_i - \overline{x}, v\rangle\,.$$

The remaining terms on the RHS we can group by their magnitudes; for each $t = C2^j\sqrt{\log n}$ there are at most $O(d(\log d + \log\log n)/t^2)$ terms of magnitude t, so the total contribution to the average is also $O(\sqrt{\eta d(\log d + \log\log d)/n})$. The same argument applies symmetrically to $\frac{1}{n}\sum_{i\in B}\langle x_i' - \overline{x'}, v\rangle$; this proves our bound on $\|\overline{x} - \overline{x'}\|$.

We turn to the bound on $\|\overline{x} - \overline{x'}\|_\infty$. Fix a coordinate $j \in [d]$. Then we have

$$\overline{x}(j) - \overline{x'}(j) = \frac{1}{n}\sum_{i\in B} x_i(j) - x_i'(j) = \frac{1}{n}\sum_{i\in B} x_i(j) - \overline{x}(j) - (x_i'(j) - \overline{x'}) + \frac{|B|}{n}(\overline{x}(j) - \overline{x'}(j))$$

Each term in the average on the RHS is at most $O(\sqrt{\log nd})$, so we obtain

$$|\overline{x}(j) - \overline{x'}(j)| \le O(\eta\sqrt{\log nd})\,.$$

□

Proposition E.3. *There is an ε-DP estimator which takes n i.i.d. samples $y_1, \dots, y_n \sim \mathcal{N}(\mu, I)$, assuming $\|\mu\| \le R$, and with high probability produces $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_\infty \le \alpha$, as long as $n \ge \tilde{O}(\frac{d\log R}{\varepsilon} + \frac{d^{2/3}}{\alpha\varepsilon^{2/3}} + \frac{\sqrt{d}}{\alpha\varepsilon} + \frac{\log d}{\alpha^2})$.*

Proof. Before we describe the ε-DP estimator, we establish a few geometry statements. Define B to be the intersection between the ℓ_∞ ball of radius α and the ℓ_2 ball of radius $c\frac{\alpha\sqrt{d}}{\sqrt{\log d}}$, for some small constant c. Let W_d be the volume of the d-dimensional unit ℓ_2 ball. We claim that

$$\frac{1}{2}\cdot W_d \cdot \left(c\frac{\alpha\sqrt{d}}{\sqrt{\log d}}\right)^d \le \mathrm{vol}(B) \le W_d \cdot \left(c\frac{\alpha\sqrt{d}}{\sqrt{\log d}}\right)^d\,.$$

The upper bound is simply because B is contained in the ℓ_2 ball of radius $c\alpha\sqrt{d}/\sqrt{\log d}$. For the lower bound, note that, having taken c small enough, for a random z in the ℓ_2 ball of radius $c\alpha\sqrt{d}/\sqrt{\log d}$, we have $\mathbb{P}(\|z\|_\infty \le \alpha) \ge 1/2$, and hence $\mathbb{P}(z \in B) \ge 1/2$, so B contains at least half the volume of the ℓ_2 ball of this radius.

Now we describe the estimator $\hat{\mu}$. Let $\hat{\mu}_0$ be the robust estimator whose guarantees are described in Proposition E.1. Given a dataset $\mathcal{Y}$, we define

$$S(\widetilde{\mu}; \mathcal{Y}) = \min_{\mathcal{Y}'} d(\mathcal{Y}, \mathcal{Y}') \text{ such that } \hat{\mu}_0(\mathcal{Y}') - \widetilde{\mu} \in B\,.$$

In words, the score of $\widetilde{\mu}$ is the minimum distance from $\mathcal{Y}$ to a dataset $\mathcal{Y}'$ which causes the robust estimator $\hat{\mu}_0$ to output a point which is both ℓ_∞ and ℓ_2-close to $\widetilde{\mu}$. The estimator $\hat{\mu}$ is given by outputting a random draw from the exponential mechanism with score function $S(\cdot; \mathcal{Y})$, over the R-radius ℓ_2 ball.

Privacy holds by construction, so we just have to analyze accuracy. We claim that any $\tilde{\mu}$ with $S(\tilde{\mu}; \mathcal{Y}) \ll \alpha n/\sqrt{\log n}$ has $\|\tilde{\mu} - \overline{\mu}\|_\infty \le \alpha/2$, where $\overline{\mu} = \frac{1}{n}\sum_{i\le n} y_i$; indeed, this follows from the ℓ_∞ accuracy guarantee of $\hat{\mu}_0$. And, since $n \gg (\log d)/\alpha^2$, with high probability we have $\|\overline{\mu} - \mu\|_\infty \le \alpha/2$. So, we just need to show that the estimator outputs $\tilde{\mu}$ with score $\ll \alpha/\sqrt{\log n}$ with high probability.

First of all, there's a set $\tilde{\mu}$s of volume at least $\mathrm{vol}(B)$ with score 0 – the set B, centered at $\hat{\mu}_0(\mathcal{Y})$.

Now consider the set of $\widetilde{\mu}$ with score ηn for $\eta \geq \alpha/\sqrt{\log n}$. By the robustness guarantee of $\hat{\mu}_0$ and the definition of B, any $\widetilde{\mu}$ with score ηn has $\|\widetilde{\mu}-\overline{\mu}\|_2 \leq O(\max(\sqrt{\eta d \log n/n}, \eta\sqrt{\log n}) + c\alpha\sqrt{d}/\sqrt{\log d})$, so is contained in a ball around $\overline{\mu}$ of volume at most

$$O\left(\frac{\sqrt{\frac{\eta d \log n}{n}} + \eta\sqrt{\log n} + \frac{c\alpha\sqrt{d}}{\sqrt{\log d}}}{\frac{c\alpha\sqrt{d}}{\sqrt{\log d}}}\right)^d \cdot \mathrm{vol}(B) \leq \exp\left(O\left(\frac{d\sqrt{\eta}\log n}{\alpha\sqrt{n}} + \frac{\sqrt{d}\eta\log n}{\alpha}\right)\right) \cdot \mathrm{vol}(B)\,.$$

Following the same argument as in Lemma 2.1, the mechanism outputs $\tilde{\mu}$ with $S(\widetilde{\mu}; \boldsymbol{y}) \ll \alpha n/\sqrt{\log n}$ with high probability so long as for every $1/2 > \eta \geq \Omega(\alpha/\sqrt{\log n})$,

$$\frac{O\left(\frac{d\sqrt{\eta}\log n}{\alpha\sqrt{n}} + \frac{\sqrt{d}\eta\log n}{\alpha}\right) + \log(\eta n)}{\eta\varepsilon} \ll n\,.$$

This occurs so long as $n \gg \tilde{O}(\frac{d^{2/3}}{\alpha\varepsilon^{2/3}} + \frac{\sqrt{d}}{\alpha\varepsilon})$. □