Efficient and Differentially Private Statistical Estimation via a Sum-of-Squares Exponential Mechanism

by

Mahbod Majid

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

© Mahbod Majid 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As machine learning is applied to more privacy-sensitive data, it is becoming increasingly crucial to develop algorithms that maintain privacy. However, even the most basic high-dimensional statistical estimation tasks were not fully understood under differential privacy, specifically, there were no known efficient algorithms for mean estimation using the optimal number of samples under pure differential privacy.

We propose a new method for designing efficient and information-theoretically optimal algorithms for statistical estimation tasks that preserve privacy, using a combination of the Sum-of-Squares hierarchy and the exponential mechanism. The Sum-of-Squares hierarchy, a convex programming method, has been used to design efficient algorithms in robust statistics. The exponential mechanism, which has been widely used in differential privacy, is often used to design information-theoretically optimal algorithms, but can be inefficient. By combining these two approaches, we are able to create efficient algorithms that are also information-theoretically optimal. We apply this approach to mean estimation for heavy-tailed distributions and learning Gaussian distributions and achieve optimal results. We also show that this approach can be applied to other problems captured by the Sum-of-Squares hierarchy through a meta-theorem. Additionally, our algorithms highlight the strong connection between robustness and privacy.

We establish information-theoretical lower bounds to show the statistical optimality of our approaches. Technically we use packing lower bounds; however, the novelty of our lower bounds is in capturing the high probability setting.

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Gautam Kamath, for his unwavering support and guidance. Gautam was extremely generous with his time, he always made himself available to meet and discuss research, regardless of the time. Beyond his role as an advisor, Gautam also played an important role as a friend, offering a listening ear, words of encouragement, and support. I am honored to have had the opportunity to work with such an exceptional advisor, mentor, and friend.

I am grateful to Sam Hopkins for his exceptional collaboration and mentorship. Sam played a significant role in shaping my development as a researcher and was incredibly supportive, consistently engaged, and always willing to delve into technical details for extended periods of time.

Additionally, I would like to thank my previous mentors, Anthony Man-Cho So, and Andrej Bogdanov for introducing me to mathematics and theoretical computer science research. My special thanks to David Steurer for hosting and mentoring me during my undergraduate studies and introducing me to the sum of squares paradigm. David's guidance and mentorship have been instrumental in my education and career development.

I am very grateful to have Sam Hopkins and Shyam Narayanan as collaborators, and I look forward to continuing to work with you in the future.

Thank you to Hongyang Zhang and Xi He for serving on my thesis committee.

I am grateful for the invaluable assistance and support provided by the dedicated administrative team at Waterloo, including Kimberley Beckman, Marie Kahkejian, Carly McLeod, and Denise Shantz. Their hard work and dedication have greatly contributed to making my experience at Waterloo smooth and enjoyable.

I would like to thank my many friends and colleagues here at Waterloo who made my time here so memorable: Aseem, Alex, Christian, Argyris (Morteza), Mahtab, Matt, Mohammadhossein, Niki, Sara, Sabrineh, and Vahid. Thank you for all the thoughtful discussions and shenanigans.

I would like to express my sincere appreciation to Amin, Amirreza, and Argyris for the valuable time we spent discussing academia and various aspects of life.

I am grateful for the friendship, support and companionship of many individuals in the area including Amin, Amirmojtaba, Elly, Fatemeh, Peyman, Saba, Shayan, and particularly Olena. They have become an important part of my life and I am thankful to have them as friends.

Finally, I would like to thank my family, for their continuing and unyielding support.

The work in this thesis was supported by grants and fellowships from the Natural Sciences and Engineering Research Council of Canada, Waterloo Cybersecurity and Privacy Institute, and the University of Waterloo.

Table of Contents

List of Tables vii

1 Introduction 1

1.1 Organization and Bibliographic Information . . . 2

2 Mean Estimation with Bounded Second Moments 3

2.1 Introduction . . . 3

2.1.1 Related Work . . . 7

2.2 Techniques . . . 8

2.2.1 The SoS Exponential Mechanism . . . 8

2.2.2 Private Mean Estimation . . . 10

2.3 Preliminaries . . . 15

2.3.1 SoS Proofs and Pseudoexpectations . . . 15

2.3.2 Privacy . . . 16

2.4 Meta-Theorem on SoS Exponential Mechanism . . . 17

2.4.1 Proof of Theorem 2.4.5 . . . 20

2.5 Coarse Estimation . . . 21

2.5.1 First Algorithm: Coordinate-Wise . . . 22

2.5.2 Second Algorithm: SoS Exponential Mechanism . . . 26

2.5.3 Robustness . . . 34

2.6 Fine Estimation . . . 34

2.6.1 Preliminaries . . . 39

2.6.2 Main Algorithm and Theorem . . . 41

2.6.3 Halt Estimation . . . 42

2.6.4 Distance Estimation . . . 43

2.6.5 Gradient Estimation . . . 47

2.6.6 Gradient Descent . . . 59

2.6.7 Robustness . . . 60
2.7 Lower Bounds . . . 60
2.8 Deferred Proofs . . . 63
2.9 Lemmata . . . 65

3 Robustness Implies Privacy: Learning Gaussian Distributions 68
3.1 Introduction . . . 68
3.1.1 Results . . . 70
3.1.2 Related Work . . . 72
3.2 Techniques . . . 74
3.2.1 Black-Box Reduction from Privacy to Robustness . . . 74
3.2.2 Algorithms . . . 77
3.3 Preliminaries . . . 80
3.4 A General Private Sampling Algorithm . . . 82
3.5 Estimating the Mean of a Gaussian . . . 84
3.5.1 Main Theorem . . . 84
3.5.2 Resilience of First and Second Moments . . . 84
3.5.3 Robust Algorithm . . . 85
3.5.4 Score Function and its Properties . . . 86
3.5.5 Proof of Theorem 3.5.1 . . . 91
3.5.6 The approx-DP setting . . . 92
3.6 Preconditioning the Gaussian . . . 93
3.6.1 Main Theorems . . . 93
3.6.2 Resilience of Moments . . . 94
3.6.3 Robust Algorithm . . . 95
3.6.4 Score Function and its Properties . . . 95
3.6.5 Proof of Theorem 3.6.1 and 3.6.2 . . . 98
3.7 Learning a Gaussian in Total Variation Distance . . . 99
3.7.1 Robust Algorithm . . . 99
3.7.2 Score Function and its Properties . . . 100
3.7.3 Proof of Theorem 3.7.1 . . . 102
3.7.4 Proof of Theorems 3.1.3 and 3.1.4 . . . 102
3.8 Computing Score Functions . . . 103

References 111

List of Tables

2.1 Comparing private mean estimation algorithms for distributions with bounded covariance. (Some papers contain several algorithms.) Sample complexity column ignores logarithmic factors in d. "-" indicates that we are not aware of any analysis in the literature. 7

3.1 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. 72

3.2 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. 73

Chapter 1

Introduction

The last decade has seen a surge in the prevalence and practice of machine learning. In many cases, machine learning methods are applied to sensitive datasets. However, the issue with classical approaches in machine learning is that they lead to privacy violations of users' data. Differential Privacy (DP) has been widely adopted in practice (by Google, Apple, Microsoft, US Census Bureau, etc.) to face this problem and is considered as a standard for privacy.

In these practical applications, private algorithms are often used to provably guarantee the privacy of users whose data is used in complex machine learning tasks. However, one persistent challenge is that the additional constraint of differential privacy often incurs significant drops in accuracy. One may naturally wonder, are these accuracy losses unavoidable? Can we develop time and data-efficient methods for differentially private data analysis? What are the limits of what we can compute privately?

Our end goal is to develop effective algorithms for highly complex statistical and machine learning tasks. Sophisticated machine learning algorithms used in practice are built layer-by-layer from simpler statistical inference procedures. Therefore, to gain deeper insight into these settings, we must first understand the more fundamental problems upon which they rely. Unfortunately, even the most basic statistical primitives are not well understood under the constraint of DP. For example, simply estimating the mean of a Gaussian distribution runs into the curse of dimensionality, requiring a running time exponential in the dimension. As a result, we seek to understand the algorithmic mathematics underlying statistical problems in DP and answer the following questions:

- How can we design efficient private algorithms for statistical tasks?
- What are the limits of what can be efficiently computed privately?

We present algorithms with provable run-time and statistical guarantees and prove information theoretical lower bounds to answer these questions.

1.1 Organization and Bibliographic Information

Most contents of this thesis have appeared previously as other publications, which we briefly outline.

Chapter 2 introduces the *sum of squares exponential mechanism* and explores its application to mean estimation for heavy-tailed distributions. This is based on the paper "Efficient Mean Estimation with Pure Differential Privacy via a Sum-of-Squares Exponential Mechanism," which is joint work with Samuel B. Hopkins, and Gautam Kamath, and appeared in the Proceedings of the 54th Annual ACM Symposium on Theory of Computing [51].

Chapter 3 introduces a more sophisticated version of the sum of squares exponential mechanism which is applicable to a greater family of sum of squares programs, and provides a black-box reduction from privacy to robustness. This is based on the paper "Robustness Implies Privacy in Statistical Estimation," which is a joint work with Samuel B. Hopkins, Gautam Kamath and Shyam Narayanan, and is currently available as a preprint [52].

Chapter 2

Mean Estimation with Bounded Second Moments

2.1 Introduction

Mean estimation is perhaps the most elementary statistical task: given samples from a probability distribution D, estimate its expected value. In this chapter, we study mean estimation in d dimensions subject to *differential privacy (DP)* [39], a rigorous notion of data privacy. Privacy is of natural concern in high-dimensional statistics, where data may be sensitive and standard estimators like the empirical mean may leak information about individuals in a dataset (see, e.g., privacy attacks in [35, 49, 22, 41] and the survey [40]). On the other hand, privacy and statistical estimation seem largely compatible: in the large-sample limit, good statistical estimators will have vanishing dependence on each individual sample anyway. Indeed, even though the empirical mean (the natural benchmark for accuracy in mean estimation) is not differentially private, estimators are known which match its accuracy guarantees while also providing differential privacy [20, 64]. Namely, an accurate and private estimate of the mean can be obtained using $n = O(d)$ samples. However, known estimators achieving this sample complexity require exponential running time.

If one instead focuses on polynomial-time algorithms, existing estimators all face significant drawbacks: either they require $n \geq d^{1+\Omega(1)}$ samples, or they may leak private information with probability 2^{-d^c} for small $c > 0$ [48, 58, 64]. (In privacy language, they satisfy only *concentrated* or *approximate* differential privacy.) Since a major goal of privacy is to make strong assurances on leakage of sensitive information, best practices disallow private data leakage even with "cryptographically" small probability $2^{-\operatorname{poly}(d)}$. One way to satisfy this stringent requirement is to provide *pure* differential privacy, which disallows substantial leakage of private information with any nonzero probability. Thus, the main question in this chapter is:

Is there a polynomial-time pure DP algorithm for mean estimation using $n = O(d)$ samples?

Our main result answers this question affirmatively, up to logarithmic factors. To state our main result, we define differential privacy:

Definition 2.1.1 ((Pure) differential privacy)**.** *For $\varepsilon > 0$, a (randomized) algorithm A which takes n inputs $X_1, \dots, X_n$ is ε-differentially private (ε-DP) if for every pair of inputs $X_1, \dots, X_n$ and $X'_1, \dots, X'_n$ such that $X_i = X'_i$ for all except a single index $i \in [n]$ and for all possible events $S \subseteq \mathrm{Range}(A)$,*

$$\mathbb{P}(A(X_1, \dots, X_n) \in S) \leq e^{\varepsilon} \cdot \mathbb{P}(A(X'_1, \dots, X'_n) \in S)\,.$$

Theorem 2.1.2 (Main Theorem — Combination of Theorem 2.5.1, Theorem 2.6.1)**.** *For every $n, d \in \mathbb{N}$ and $R, \alpha, \varepsilon, \beta > 0$ there is a polynomial-time ε-DP algorithm such that for every distribution D on $\mathbb{R}^d$ such that $\|\mathbf{E}_{X\sim D} X\| \leq R$ and $\mathrm{Cov}_{X\sim D}(X) \preceq I$, given $X_1, \dots, X_n \sim D$, with probability $1-\beta$ the algorithm outputs $\hat{\mu}$ such that $\|\hat{\mu} - \mathbf{E}_{X\sim D} X\| \leq \alpha$, so long as*

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

Furthermore, if an η-fraction of the samples $X_1, \dots, X_n$ are adversarially corrupted, the algorithm maintains the same guarantee, at the cost that now $\|\hat{\mu} - \mathbf{E}\, X\| \leq \alpha + O(\sqrt{\eta})$.

The sample complexity of our algorithm is nearly linear in d, thereby answering our main question up to logarithmic factors. Beyond this core goal, we highlight that our algorithm is *also* robust to adversarial contaminations and enjoys sub-Gaussian confidence intervals when $n \gg \frac{d \log R + \min(d, \log R) \log(1/\beta)}{\varepsilon}$, both sought-after features in recent non-private algorithms [29, 73, 50, 24]. The sample complexity of our algorithm is nearly optimal, with the exception of the term $(\log(1/\beta) \cdot \min(d, \log R))/\varepsilon$ – see Theorem 2.7.2 for corresponding lower bounds. (Put differently, information-theoretically it is possible to achieve sub-Gaussian confidence intervals under the slightly milder assumption $n \gg \frac{d \log R + \log(1/\beta)}{\varepsilon}$.)

Beyond clip-and-noise Obtaining the guarantees of our algorithm requires going beyond existing techniques for private mean estimation, which we briefly review. A common technique in differential privacy is to "just add noise." More precisely, suppose that $f(X_1, \dots, X_n)$ is a *bounded sensitivity* function of n inputs, meaning that replacing a single input X_i with an arbitrary X'_i changes the value of f by at most Δ (in an appropriate choice of norm). Then, $f(X_1, \dots, X_n) + Z$ will be private, where Z is an appropriate random variable whose magnitude depends on Δ.

Existing polynomial-time algorithms for private mean estimation all take this approach. First, to limit the sensitivity of the empirical mean, the algorithms clip the samples $X_1, \dots, X_n \in \mathbb{R}^d$ to lie in an ℓ_2 ball. Considering the case $\mathrm{Cov}(X) \approx I$, we will have $\|X - \mathbf{E}\, X\| \approx \sqrt{d}$, so using a ball of radius at least $\sqrt{d}$ is unavoidable without introducing too much error in the clipping phase. Then, the algorithms output $\overline{\mu} + Z$, where $\overline{\mu}$ is the empirical mean of the clipped samples.

If one takes the coordinates of Z to be draws from a Laplace distribution, the resulting algorithm will satisfy ε-DP, but the relatively heavy tails of the Laplace distribution impose a cost that n must be at least $d^{1.5}/\varepsilon$ to obtain nontrivial guarantees. On the other hand, if Z is Gaussian, the resulting algorithm appears to tolerate $n \approx d/\varepsilon$ samples, but it no longer satisfies *pure* DP. Instead, to guarantee privacy loss at most ε with probability at

least δ over the internal randomness in the algorithm, $n \geq d\sqrt{\log(1/\delta)}/\varepsilon$ is required (i.e. the algorithm satisfies *concentrated* DP), meaning that for "cryptographically small" δ, this algorithm, too, has super-linear sample complexity.

One might naturally suspect, therefore, that strong privacy guarantees like this simply require $\Omega(d^{1.5})$ samples. Indeed, if nastier distributions than we consider here (violating bounded covariance) are allowed, these two bounds of $\Omega(d^{1.5})$ and $\Omega(d)$ are known to be tight for pure and concentrated DP, respectively [48, 22, 58], and similar separations were previously conjectured even for bounded-covariance distributions [20].

However, [64] (building on related techniques of [20]) show that in exponential time one can go beyond the clip-and-noise approach to mean estimation. In particular, a tournament-based approach gives a pure-DP algorithm using $O(d/\varepsilon)$ samples. We obtain nearly-matching guarantees in polynomial time – ours is the first polynomial-time private algorithm to go beyond the clip-and-noise approach.

Sum-of-Squares Proofs to Private Algorithms *Proofs to Algorithms* has become a powerful algorithm-design technique in computational high-dimensional statistics. Roughly speaking, the proofs to algorithms technique shows that statistical estimation problems which can be solved to a given accuracy by some (not necessarily polynomial-time) estimator can actually be solved in polynomial time with the same accuracy *if the* analysis *of that estimator can be captured by a certain restricted but powerful formal proof system, known as the SoS proof system.* This insight has had major consequences in robust and heavy-tailed statistics, clustering, learning latent variable models, and beyond. (For instance, [54, 71, 70, 85], and the survey [87].)

We show that this approach also applies to private algorithm design. Our techniques give a generic method to turn (potentially) exponential-time instances of the workhorse *exponential mechanism*, whose breadth of applicability is hard to overstate [84], into polynomial time algorithms, when their analyses are captured by the same SoS proof system. This gives us the following *proofs-to-algorithms* principle for private algorithm design, which we anticipate will be widely applicable:

> **Proofs to Private Algorithms** (see Theorem 2.4.5): Instances of the exponential mechanism with low-degree SoS proofs of bounded sensitivity and utility automatically yield computationally-efficient private algorithms.

(See Section 3.2 for a description of the exponential mechanism and more on SoS proofs.) We are able to capture this principle as a formal meta-theorem, Theorem 2.4.5 – since its statement requires a few more technical definitions than we are ready to state, we defer it for now.

Convex Programming and Private Algorithms Like nearly all applications of the Sum of Squares method, algorithms which result from our proofs-to-private-algorithms approach ultimately use convex programs – semidefinite programs, in our case. Our techniques give a substantially novel approach to convex programming in private algorithms. Prior works largely do this in one of two ways. Some privatize the input *before* applying a

convex program to it (e.g., privatizing a graph and then post-processing with convex programming [18, 42]). This can be limiting, as the convex program cannot itself be helpful in achieving privacy, only in some post-processing. Other algorithms use a private solver for the convex program itself [56, 77]. This can be quite technically challenging: few DP solvers for convex programs are known, and in particular, we are not aware of any generic DP algorithm for solving semidefinite programs.

Instead, our algorithms *hide convex programs behind the exponential mechanism.* Namely, we use convex programs as score functions. (Again, see Section 3.2 for an explanation of this terminology.) Our key insight is: *when convex programs are married to the exponential mechanism in the correct way, the resulting exponential mechanisms can be implemented in polynomial time using private log-concave sampling algorithms.* This builds directly on [16], who use a similar approach but for much simpler un-constrained convex optimization problems resulting from empirical risk minimization problems. The analysis of this interplay of exponential mechanism and convex programming is completely generic – meaning it has nothing to do with the particular setting of mean estimation – and is captured in the proof of our (meta)-Theorem 2.4.5.

Robust Statistics and Privacy Robust statistics, the study of statistics in the presence of corrupted samples, has enjoyed a recent renaissance [30]. Robustness and privacy are at least spiritual cousins – both demand that a statistical estimator not change its behavior "too much" when one or several samples are replaced arbitrarily. However, the formal requirements for robustness and privacy are rather different. The output of a good robust mean estimator should not move far in Euclidean distance when 1% of the samples it is given are corrupted by a malicious adversary. Privacy, by contrast, demands that the *distribution* of outputs of an algorithm not shift too much when any single sample is replaced by another.

In spite of this formal difference, [38] was able to use robust estimators in one dimension to construct private estimators. This suggests that the flurry of recent algorithms for high-dimensional robust statistics with strong provable guarantees should yield high-dimensional private estimators.

Our work makes good on this promise. In particular, our algorithm for private mean estimation ultimately employs a well-studied Sum-of-Squares SDP from robust mean estimation, and our analysis applies several SoS proofs originally formulated to analyze that SDP in robust statistics. Thus, our proofs-to-private-algorithms approach gives a lens through which algorithms from robust statistics can yield private algorithms.

Lastly, there is an even more direct connection between our particular result and recent developments in robust statistics. The robustness renaissance was kicked off by [27, 73], which gave the first polynomial-time algorithms for robustly learning a Gaussian where corrupted samples can be tolerated with dimension-independent error. Prior polynomial-time algorithms for the same problem have guarantees no better than what can be obtained by naive sample-clipping, and as a result incur error scaling as $\eta \cdot \mathrm{poly}(d)$ when an η-fraction of samples are corrupted. Just as [27, 73] gave the first algorithms with guarantees going beyond naive sample clipping in the robust setting, our work is the first to go beyond clip-and-noise in the private setting.

Algorithm	Sample Complexity	Privacy	Sub-Gaussian Rates	Robust	Poly Time
Sample Mean	d	None	No	No	Yes
[64]-A	d	Pure	-	Yes	No
[64]-B	d	Concentrated	No	No	Yes
[64]-C	$d^{1.5}$	Pure	No	No	Yes
[77]-A	d	Approximate	-	Yes	No
[77]-B	$d^{1.5}$	Concentrated	No	Yes	Yes
Theorem 2.1.2	d	Pure	$n \gg \frac{d \log R + \min(d, \log R) \log(1/\beta)}{\varepsilon}$	Yes	Yes

Table 2.1: Comparing private mean estimation algorithms for distributions with bounded covariance. (Some papers contain several algorithms.) Sample complexity column ignores logarithmic factors in d. "-" indicates that we are not aware of any analysis in the literature.

2.1.1 Related Work

A mature body of work focuses on private mean estimation. The most relevant results restrict the underlying distribution to satisfy a moment bound (including sub-Gaussianity), examples include [15, 67, 20, 21, 58, 23, 64, 96, 36, 17, 1, 19, 59, 57, 61, 5]. All prior study of the multivariate case either focuses on concentrated or approximate DP, or provides computationally-inefficient algorithms for pure DP. We are the first to give an efficient $O(d)$-sample algorithm for multivariate mean estimation under pure DP. This matches the sample complexity of the best known algorithms under concentrated DP, while simultaneously strengthening the privacy guarantee.

Other prior work studies private mean estimation of arbitrary (bounded) distributions [22, 93, 41]. Interestingly, in this case, the optimal sample complexity under pure and concentrated DP are separated by a factor of $\sqrt{d}$, leading to the naive Laplace mechanism being effectively optimal. It appears that our bounded moments assumption induces a qualitatively different structure, which eliminates the benefits of relaxing to concentrated DP. Pinpointing the precise conditions under which a separation occurs remains an interesting direction for future work.

Other works study private statistical estimation in different and more general settings, including mixtures of Gaussians [63, 2], graphical models [98], discrete distributions [26], and median estimation [11, 95]. Some recent directions involve guaranteeing user-level privacy [80, 75], or a combination of local and central DP for different users [12]. See [65] for further coverage of DP statistical estimation.

Several other works employ sampling-based methods to efficiently implement the exponential mechanism [66, 16, 4, 74]. [66, 4, 74] construct sophisticated sampling algorithms by hand to sample from a non-log-concave distributions; by contrast, we use convex programming to construct our score functions, ensuring that we always stay within the realm of log-concave sampling algorithms. [16] constructs the private log-concave sampler we use in our work, but they employ it only to sample from comparatively simple log-concave distributions arising from empirical risk minimization of convex loss functions. A recent work of [83] provides a faster algorithm for private log-concave sampling.

Another recent line of work studies sub-Gaussian confidence intervals for mean estimation of heavy-tailed distributions (i.e., assuming only bounded covariance). Lugosi and

Mendelson [82] proposed an inefficient algorithm, with an SoS-based algorithm coming in [50]; see also the survey [81]. Some works considered efficient algorithms for simultaneously robust and heavy-tailed mean estimation [25, 86, 32]. A recent result shows that the core solution concepts for robust and heavy-tailed mean estimation can be considered equivalent [55]. Our work demonstrates that the line of efficient estimators inspired by [50] is simultaneously effective for robust, heavy-tailed, and private mean estimation.

Relatively limited work simultaneously considers privacy and the other constraints of robustness and heavy-tailed estimation. The main result is [77] which considers robust and private mean estimation. Our result can be seen as improving on the running time or sample complexity of their two algorithms, and the privacy guarantee of both. While other prior works focus on heavy-tailed distributions [15, 64, 96, 59], these do not address the primary goal in the non-private setting, which is to achieve sub-Gaussian rates with respect to the error probability. Instead, they generally prove guarantees with constant probability of success, which can be boosted at the cost of a multiplicative (rather than additive) logarithmic factor which is inverse in the failure probability. Note that some of the previous (inefficient) cover- and median-based approaches to private estimation [20, 64, 77, 19, 1] have varying levels of robustness and sub-Gaussian rates for heavy-tailed settings. However, none of their results give computationally efficient estimators for pure DP. A simultaneous and independent work of [78] demonstrates an interesting connection between resilience [92] and private estimation. They exploit this connection to design robust and private algorithms for a variety of settings, including mean estimation, covariance estimation, PCA, and more. However, their focus is on providing inefficient algorithms under the constraint of approximate DP, while our goal is to give a framework for efficient algorithms under pure DP. For a summarized comparison of our work with recent private algorithms, see Table 2.1.

2.2 Techniques

2.2.1 The SoS Exponential Mechanism

Before we turn to our algorithm for mean estimation, we offer a little more detail on the proofs-to-private-algorithms approach and its core component, which we call the SoS exponential mechanism. We begin with a review of the exponential mechanism itself.

The Exponential Mechanism Consider the general problem of privately selecting one among a set of candidates $\mathcal{C}$ given a dataset X, where the quality of a candidate $x \in \mathcal{C}$ depends on the dataset X. The candidates $\mathcal{C}$ could represent many different things depending on the context. In a statistical setting one may often think of $\mathcal{C}$ as a class of probability distributions, X as a list of samples from one of those distributions, and the goal is to select the distribution from which X came (up to small error).

To apply the exponential mechanism for this problem, one first finds a *score function* $s(X, x)$ which assigns a (real-valued) score to each dataset-candidate pair, ideally such that x's which are "good" for a given X receive high scores. Given X and a privacy parameter $\varepsilon > 0$, the exponential mechanism will sample a random x with probabilities

$\mathbb{P}(x) \propto \exp(\varepsilon \cdot s(X, x))$. The output is ε-differentially private so long as the score function satisfies a *bounded sensitivity* property: for any pair of neighboring datasets X, X'[1] and for all $x \in \mathcal{C}$, one has $|s(X, x) - s(X', x)| \leq 1$.

Beyond privacy, one also wants the resulting x to be useful. While this can also mean different things depending on the context, we will take "utility" to mean that x is close to some good $x^*(X)$. To prove that this happens for the exponential mechanism defined above, one shows:

1. High-scoring x's are good: if $s(X, x) \geq 0$, then $\|x - x^*\| \leq \alpha$, for a small $\alpha > 0$. (The choice of 0 is without loss of generality; the mechanism is invariant under additive shifts of s.)

2. Not too few high-scoring x's: $\frac{|\{x \in \mathcal{C} : s(X,x) \geq t\}|}{|\mathcal{C}|} \geq \frac{1}{r}$. (If $\mathcal{C}$ is an infinite set one can replace $|\cdot|$ with some measure of volume.)

Then a simple argument shows that x such that $\|x - x^*\| \leq \alpha$ is selected with probability at least $1 - r\exp(-\varepsilon t)$. In other words, the mechanism selects a good x so long as $t \gg \frac{\log r}{\varepsilon}$.

The exponential mechanism can lead to computationally inefficient algorithms for two basic reasons. First, in high-dimensional statistics it is often natural to use score functions which seem hard to compute – the Tukey depth, for just one example [77, 19]. Second, as we will face when we turn to mean estimation, even with score functions that are easy to compute, sampling x with $\mathbb{P}(x) \propto \exp(\varepsilon \cdot s(X, x))$ may not be computationally tractable.

Convex programs as score functions The SoS exponential mechanism can address both of these sources of intractability for instances of the exponential mechanism where the *proofs* of bounded sensitivity are captured in a powerful formal proof system, the *Sum of Squares proof system* (SoS). To de-mystify this a little without yet delving into the details of SoS proofs, we can see a high-level picture of the algorithms which use the SoS exponential mechanism. Ultimately, these algorithms fit into the exponential mechanism framework, but with special score functions.

To wit, suppose that we can arrange for the score function $s(X, x)$ to take the form of a linear optimization problem over a convex set $\mathcal{K}(X)$ with an additional linear constraint involving x:

$$s(X, x) = \max_{y \in \mathcal{K}(X)} \langle c, y \rangle \text{ such that } Ay = x$$

for some matrix A and vector c. As long as $\mathcal{K}(X)$ admits a computationally efficient separation oracle, $s(X, x)$ will be polynomial-time computable – this already addresses the first source of potential intractability in the exponential mechanism. A simple argument shows even more: for each X, $s(X, x)$ is actually concave in x. Thus, the distribution $\mathbb{P}(x) \propto \exp(\varepsilon \cdot s(X, x))$ is log-concave, so we can (usually) sample from it in polynomial time! This sampling step itself has to be done privately: luckily for us, log-concave sampling is so well understood that private methods for polynomial-time log-concave sampling are known [16].

[1] We say X, X' are neighboring if they differ on the presence/absence of just one individual.

The restriction that the score function take the form of an optimization problem is not a major one: many useful score functions in high-dimensional statistics, such as the Tukey depth, are naturally expressible in this way. But how to find a score function which is simultaneously a *convex* optimization problem and maintains both bounded sensitivity and utility? The SoS method gives us a way to construct such a score function automatically, starting from any score function which is expressible as a (potentially non-convex) optimization problem, and for which the *proofs* of bounded sensitivity and utility are expressible in the SoS proof system.

2.2.2 Private Mean Estimation

We proceed with a high-level description of our algorithm for private mean estimation and its analysis. One could obtain our algorithm as an instance of proofs-to-private-algorithms, by:

- Constructing a certain simple-to-analyze exponential mechanism-based mean estimator.
- Writing the simple analysis as a series of SoS proofs and applying (meta)-Theorem 2.4.5.

However, so that our algorithm can be understood without tackling the SoS exponential mechanism in full generality, we will now give a more concrete description of the algorithm and its analysis. We give a high level version of this description here, and in the main body of this chpater we actually provide a full analysis of the algorithm without appealing to Theorem 2.4.5.

Let us recall the setup for our problem. There is a random variable X on $\mathbb{R}^d$ with $\mathrm{Cov}(X) \preceq I$ and $\|\mathbf{E}\, X\| \leq R$. The goal is to estimate $\mathbf{E}\, X$ from i.i.d. copies $X_1, \ldots, X_n$, subject to ε-DP. The promise $\|\mathbf{E}\, X\| \leq R$ is information-theoretically necessary for pure differential privacy [67, 20].

Our algorithm has strong guarantees in the presence of adversarially-corrupted samples, and obtains sub-Gaussian confidence intervals given heavy-tailed samples. In fact, this is a side-effect of the fact that our algorithm uses well-studied SDPs from the robust statistics setting, viewing them through the SoS exponential mechanism lens to obtain privacy. In particular, folklore adaptations of the analyses in [24, 71] directly show that our algorithm is robust, so to avoid a proliferation of notation we give the proof in the non-robust setting – see Remarks 2.6.40 and 2.5.26.

Like prior algorithms for private mean estimation, our algorithm has two phases.

1. Reduce R to $\mathrm{poly}(d)$, roughly by finding a large ball containing a large number of samples.
2. Estimate $\mathbf{E}\, X$ under the assumption $\|\mathbf{E}\, X\| \leq \mathrm{poly}(d)$.

We start with the second step, which captures much of our conceptual contribution – even under the assumption $R \leq \sqrt{d}$, prior algorithms could not achieve pure differential privacy. Then we discuss the first step, where prior algorithms require $\Omega(d(\log R + \log(1/\beta)))$

samples. We give an algorithm with sample complexity $\log R(d+\log(1/\beta))$. When $\log R \ll d$, our algorithm improves over prior work. (Information-theoretically, $d \log R + \log(1/\beta)$ is possible.)

Private mean estimation when $\|\mathbf{E}\, X\| \le \mathrm{poly}(d)$

Let us write $\mu = \mathbf{E}\, X$. For simplicity, for now we focus on the case that our goal is to find $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le O(1)$ using $O(d)$ samples. (We can reduce from the α-error case to this one by placing the samples in buckets containing around $1/\alpha^2$ samples and taking sample means within each bucket.) We will also think of ε as a small constant. A merit of our approach is that it allows powerful techniques from robust statistics to be used for private algorithm design: in this case, our algorithm draws heavily on a robust mean estimation algorithm due to [24], using SoS exponential mechanism to privatize its use of convex programming.

Iterative refinement/gradient descent Our algorithm will iteratively refine an initial estimate of the mean (without loss of generality, the origin), producing a series of estimates $\hat{\mu}_0 = 0, \hat{\mu}_1, \ldots, \hat{\mu}_T$. In each step t, we will privately find a unit vector v such that $\langle v, \mu - \hat{\mu}_t \rangle \ge \Omega(1) \cdot \|\mu - \hat{\mu}_t\|$. Then we can replace $\hat{\mu}_t$ with $\hat{\mu}_{t+1} = \hat{\mu}_t + r \cdot v$ for some appropriate step size $r > 0$. By standard reasoning this means that $O(\log d)$ steps suffice to obtain $\|\hat{\mu}_T - \mu\| \le O(1)$.

This gradient-descent approach introduces only logarithmic overheads into our sample complexity and running time, so now we turn to the heart of our algorithm: privately finding v.

Finding private gradients Given samples $X_1, \ldots, X_n \in \mathbb{R}^d$ and $\hat{\mu}_t$, we would like to privately select a unit vector v such that $\langle v, \hat{\mu}_t - \mu \rangle \ge \Omega(1) \cdot \|\hat{\mu}_t - \hat{\mu}\|$. We will use the exponential mechanism, for which we need to define a score function. For this, we are inspired by recent work in robust and heavy-tailed statistics, where the following has become a standard fact [82]:

Fact 2.2.1 (Directions with many outliers are good, informal)**.** *Suppose that $\|\hat{\mu}_t - \mu\| \gg 1$. Then, with high probability over $X_1, \ldots, X_n$, so long as $n \gg d$, there are at least $0.9n$ samples X_i such that $\langle X_i - \hat{\mu}_t, \frac{\mu - \hat{\mu}_t}{\|\mu - \hat{\mu}_t\|} \rangle \ge \Omega(1) \cdot \|\hat{\mu}_t - \mu\|$. Furthermore, for every unit vector v such that*

$$|\{i \,:\, \langle X_i - \hat{\mu}_t, v \rangle \ge \Omega(1) \cdot \|\hat{\mu}_t - \mu\|\}| \ge 0.8n\,,$$

we have $\langle v, \mu - \hat{\mu}_t \rangle \ge \Omega(1) \cdot \|\mu - \hat{\mu}_t\|$.

Fact 2.2.1 makes a good choice of score function clear: we should use

$$s(X, v) = |\{i \,:\, \langle X_i - \hat{\mu}_t, v \rangle \ge \Omega(1) \cdot \|\hat{\mu}_t - \mu\|\}|\,.$$

Utility of this score function is captured by Fact 2.2.1, and bounded sensitivity is clear by construction. (In fact, it is exactly this bounded sensitivity property which has already

made Fact 2.2.1 so important in robust and heavy-tailed statistics.) We do not necessarily know the value of $\|\hat{\mu}_t - \mu\|$, but getting a private estimate of this quantity is not too difficult – we privatize a procedure due to [24].

A straightforward analysis shows that, since the volume of the d-dimensional unit ball is roughly $\exp(d)$, exponential mechanism with this score function will ε-privately select a good v so long as $n \gg d/\varepsilon$, which is exactly the sample complexity we expect for estimating μ to error $O(1)$.

Finding private gradients in polynomial time Of course, the key problem is that sampling with $\mathbb{P}(v) \propto \exp(\varepsilon \cdot s(X, v))$ may not be possible in polynomial time. Indeed, a seemingly-easier problem, finding the highest-scoring v, seems closely related to computing Tukey depth, which is NP-hard.

However, some hope comes again from robust/heavy-tailed statistics, where approximation algorithms (often based on semidefinite programming) for the problem of finding the highest scoring v have become an invaluable tool. Our starting point is the by-now standard construction of such a semidefinite relaxation, which we briefly review. First, we write this optimization problem as a degree-2 polynomial optimization problem in variables $v_1, \dots, v_d$ and $b_1, \dots, b_n$, where the latter are constrained so as to be 0/1 indicators of $\langle X_i - \hat{\mu}_t, v\rangle \geq r$ for some threshold value r:

$$\max \sum_{b=1}^{n} b_i \text{ such that } \|v\|^2 \leq 1, b_i^2 = b_i, \text{ and } b_i \langle X_i - \hat{\mu}_t, v\rangle \geq b_i \cdot r \text{ for all } i\,.$$

The degree-2 SoS relaxation of this optimization problem optimizes over (degree 2) *pseudoexpectations*, which are linear functionals $\tilde{\mathbf{E}} : \mathbb{R}[b, v]_{\leq 2} \to \mathbb{R}$ defined on degree at most 2 polynomials in b, v which are normalized and positive: $\tilde{\mathbf{E}}\, 1 = 1$ and $\tilde{\mathbf{E}}\, p(b, v)^2 \geq 0$ for all linear p. Concretely:

$$\max_{\tilde{\mathbf{E}}} \tilde{\mathbf{E}} \sum_{i \leq n} b_i \text{ s.t. } \tilde{\mathbf{E}}\, \|v\|^2 \leq 1, \tilde{\mathbf{E}}\, b_i^2 = \tilde{\mathbf{E}}\, b_i, \text{ and } \tilde{\mathbf{E}}\, b_i \langle X_i - \hat{\mu}_t, v\rangle \geq r \cdot \tilde{\mathbf{E}}\, b_i \text{ for all } i. \quad (2.1)$$

This optimization problem can be solved via semidefinite programming – the resulting (equivalent) SDP optimizes over $(1 + n + d) \times (1 + n + d)$ block matrices

$$\max \operatorname{Tr} B \text{ such that } \begin{pmatrix} 1 & \tilde{b}^\top & \tilde{v}^\top \\ \tilde{b} & B & W^\top \\ \tilde{v} & W & V \end{pmatrix} \succeq 0, \operatorname{Tr} V \leq 1, B_{ii} = \tilde{b}_i, \text{and } \langle X_i - \hat{\mu}_t, W_i\rangle \geq r \cdot B_{ii}$$

Here, V is a proxy for the rank-one matrix $vv^\top$ and similarly for B and $bb^\top$. This SDP is known to be a good approximation to the problem of finding the maximum-score v.

We prove the following fact. (As an aside, this is where SoS proofs enter the picture: establishing facts such as the below for SoS SDP relaxations in general requires constructing SoS proofs.)

Fact 2.2.2 (Bounded Sensitivity and Utility for (2.1), informal)**.** ***Bounded Sensitivity:*** *Changing a single sample in optimization problem (2.1) can change the objective value by at most* 1. ***Utility:*** *With high probability over $X_1, \dots, X_n$, if $\|\hat{\mu}_t - \mu\| \gg 1$, then any feasible $\tilde{\mathbf{E}}$ in (2.1) with objective value at least $0.8n$ satisfies $\langle \tilde{\mathbf{E}}\, v, \mu - \hat{\mu}_t\rangle \geq \Omega(1) \cdot \|\mu - \hat{\mu}_t\|$.*

While this proof largely adapts similar arguments in the robust statistics literature, there is a key technical innovation. Prior algorithms employing the SDP described above use a nontrivial rounding step to extract a good vector v from the $d \times d$ PSD matrix V. However, for reasons we discuss below, to use the SoS exponential mechanism, it is important that v can be read directly off of the SDP (more precisely, that the rounding algorithm used is linear). This means we need a stronger rounding procedure than that used in prior works – we are able to show that $\tilde{v} = \tilde{\mathbf{E}}\, v$, a simple linear function of $\tilde{\mathbf{E}}$, is a good choice.

Sampling with convex programs The "utility" part Fact 2.2.2 solves the problem of finding a high-scoring v in polynomial time, via semidefinite programming. But we want to find such a high-scoring v privately.

The key idea is to use the SDP to construct a score function: *convexity of the set of feasible solutions and linearity of the objective function now imply log-concavity of the resulting sampling problem*! This observation, while simple, is remarkably powerful: it allows us to employ a well-studied SDP from robust statistics nearly out-of-the-box to obtain strong privacy guarantees.

Ultimately, we employ the score function:

$$s(X, v_0) = \max_{\tilde{\mathbf{E}}} \tilde{\mathbf{E}} \sum_{i \le n} b_i \text{ s.t. } \tilde{\mathbf{E}}\, \|v\|^2 \le 1, \tilde{\mathbf{E}}\, b_i^2 = \tilde{\mathbf{E}}\, b_i, \text{ and } \tilde{\mathbf{E}}\, b_i \langle X_i - \hat{\mu}_t, v \rangle \ge r \cdot \tilde{\mathbf{E}}\, b_i \text{ for all } i \text{ and } \tilde{\mathbf{E}}\, v = v_0\,,$$

together with the private sampling algorithm of [16], to sample from $\mathbb{P}(v_0) \propto \exp(\varepsilon \cdot s(X, v_0))$.

We make a few remarks on the precise way that $s(X, v_0)$ depends on v_0; that is, via the constraint $\tilde{\mathbf{E}}\, v = v_0$, since this choice is not accidental. First of all, it is important that the constraints of the optimization problem defining $s(X, v_0)$ depend *linearly* on v_0; otherwise we might not retain log-concavity of the resulting sampling problem. We can do this only because the rounding algorithm described in Fact 2.2.2, which proves that $\tilde{\mathbf{E}}\, v$ itself is useful, is a simple linear function of $\tilde{\mathbf{E}}$.

We also note an important interplay between the "lifted" nature of the SDP and the utility analysis of the exponential mechanism. On the one hand, the power of the SDP comes from lifting from the $d+n$ variables v, b to $(d+n+1)^2$ variables, and solving a convex problem in the lifted space. On the other hand, for the exponential mechanism to satisfy utility with just $O(d)$ samples, it is important that we use it to sample in d dimensions (where, in particular, there is a $2^{O(d)}$-sized cover) rather than, say, d^2 dimensions. This tension is resolved by hiding the additional variables inside the optimization problem which defines s, so that exponential mechanism still samples from a d-dimensional distribution, but we can still use the power of SoS and semidefinite programming.

Lipschitzness The sampling algorithm of [16] (like other algorithms for sampling from log-concave probability distributions) runs in polynomial time in the ambient dimension, *so long as the distribution has Lipschitz log-probabilities*. Natural approaches to force $s(X, v_0)$ to be Lipschitz, which generally take the form of randomized smoothing, risk violating pure DP, because an algorithm computing a randomized smoothing of s will have some small probability of quietly failing, at which point privacy is at risk.

To ensure that the resulting algorithm runs in polynomial time, therefore, we have to show that $s(X, v_0)$ is Lipschitz with respect to v_0. To establish Lipschitzness, we make a two-step argument:

1. We show that dual solutions to the SDP defining $s(X, v_0)$ have optimal dual certificates which are not too large in norm. (At most, say, $\mathrm{poly}(d, n)$.)
2. We show that any such dual solution to the SDP for $s(X, v_0)$ which certifies an upper bound of c can be adapted to a dual solution for $s(X, v_0 + \Delta)$ which certifies an upper bound of $c + \|\Delta\| \mathrm{poly}(d, n)$, for any small perturbation vector Δ.

Together, these imply that $s(X, v_0)$ is $\mathrm{poly}(d, n)$-Lipschitz with respect to v_0. Since the Lipschitz constant only arises in our algorithm's running time, this suffices for our purposes. This concludes our overview of the second phase of our private mean estimation algorithm.

Coarse estimation: from R to $\mathrm{poly}(d)$

We now describe our algorithm to privately localize $\mu = \mathbf{E}\, X$ to a ball of radius $\mathrm{poly}(d)$, beginning only with the promise that $\|\mathbf{E}\, X\| \le R$ for some large number R. The goal is to do so using as few samples as possible. Existing efficient algorithms [58] can perform this task with probability $1 - \beta$ using $O(d(\log R + \log(1/\beta))/\varepsilon)$ samples (we present this analysis to capture the dependence on $\log(1/\beta)$). If we used this algorithm, we would obtain sub-Gaussian confidence intervals only when $n \gg \frac{d \log R + d \log(1/\beta)}{\varepsilon}$. However, if we do not worry about running time, the same task can be accomplished using the exponential mechanism using only $O(d \log R/\varepsilon + \log(1/\beta)/\varepsilon)$ samples – much fewer for $\beta \ll 1$, which is important for constructing confidence intervals.

While we do not quite obtain optimal complexity, we are able to improve on existing algorithms in the regime $\log R \ll d$; our algorithm requires $\tilde{O}(d \log R/\varepsilon + \log R \log(1/\beta)/\varepsilon)$ samples.

Our algorithm again follows the proofs-to-private-algorithms approach. We start with the following basic instantiation of the exponential mechanism. We are given samples $X_1, \ldots, X_n$. With probability at least $1 - \beta$ over the choice of $X_1, \ldots, X_n$, as $n \gg \log(1/\beta)$, any ball of radius $\mathrm{poly}(d)$ containing $0.9n$ of the samples will have center which has distance at most $\mathrm{poly}(d)$ to μ. So, we would like to use the exponential mechanism with score function $s(X, x) = |\{i : \|X_i - x\| \le \mathrm{poly}(d)\}|$ to select a point from the ball of radius R centered at the origin.

As before, it is not clear how to perform the sampling task that this would require in polynomial time. So, we replace the score function with a convex relaxation, in this case built from the "degree-4" SoS relaxation of of the following polynomial optimization problem:

$$\max_{b,x} \sum_{i=1}^{n} b_i \text{ s.t. } \|x\|^2 \le R^2, b_i^2 = b_i, \text{ and } b_i \|X_i - x\|^2 \le \mathrm{poly}(d) \cdot b_i \text{ for all } i\,.$$

(The degree-4 SoS relaxation shows up here because the optimization problem involves polynomials of degree 3, and SoS relaxations are defined only for even degrees.) That is,

we use the score function

$$s(X, x_0) = \max_{\tilde{\mathbf{E}}} \tilde{\mathbf{E}} \sum_{i=1}^{n} b_i \text{ s.t. } \tilde{\mathbf{E}} \text{ satisfies } \|x\|^2 \le R^2, b_i^2 = b_i, b_i\|X_i - x\|^2 \le b_i(R/10)^2,$$
$$\text{and } \tilde{\mathbf{E}}\, x = x_0$$

(for the definition of "satisfies", see Section 2.3).

Once we have decided to use this particular SoS relaxation, the outlines of the algorithm and its analysis are largely similar to the second phase of the algorithm, but with different SoS proofs plugged in. In particular, we prove an analogue of Fact 2.2.2 for this setting, and establish Lipschitzness of the SDP, so that we can use the same strategy as in second phase.

The key step is to show that a high-scoring x_0 has distance at most $R/2$ to any x' which has distance $\mathrm{poly}(d)$ to $0.8n$ samples. This statement turns out to have a relatively simple SoS proof. This shows that SoS exponential mechanism manages to reduce R to $R/2$. Iterating this $\log R$ times completes the algorithm.

2.3 Preliminaries

2.3.1 SoS Proofs and Pseudoexpectations

We give a brief overview of SoS; for details see [14].

SoS Proofs We first informally review the SoS proof system. Let $p(x), p_1(x), \ldots, p_m(x)$ be multivariate polynomials in indeterminates $x_1, \ldots, x_n$. The SoS proof system can prove statements of the form:

$$\text{For all } x \in \mathbb{R}^n, \text{ if } p_1(x) \ge 0, \ldots, p_m(x) \ge 0, \text{ then } p(x) \ge 0.$$

Polynomials are highly expressive, so a wide range of mathematical statements can be encoded in the above form. An SoS proof of such a statement is a family of polynomials $\{q_S(x) : S \subseteq [m]\}$ such that each $q_S(x) = \sum_{i=1}^{r} (q_S^{(i)}(x))^2$ is a sum of squares, and $p(x) = \sum_{S \subseteq [m]} q_S(x) \cdot \prod_{i \in S} p_i(x)$. The proof has degree $D \in \mathbb{N}$ if each term in this sum is a polynomial of degree at most D. We write:

$$p_1 \ge 0, \ldots, p_m \ge 0 \vdash_D p \ge 0\,.$$

Additionally, we will need one non-standard definition: for any given $j \in [m]$, we say that an SoS proof is degree D_j *with respect to* p_j if every term $q_S(x) \cdot \prod_{i \in S} p_i(x)$ in the proof such that $j \in S$ has degree at most D'. We will sometimes write $p_1 \ge 0, \ldots p_m \ge_0 \vdash_{D, \deg_{p_j} = D_j} p \ge 0$.

SoS proofs are dual solutions to semidefinite programs arising from the Sum of Squares method, where pseudoexpectations are primal solutions. As in many applications of convex programming, to prove facts about primal solutions, the main technique is to construct duals. In particular, to show that the SoS SDPs we use for exponential mechanism score functions satisfy bounded sensitivity and privacy, it suffices to construct SoS proofs witnessing these facts.

Pseudoexpectations For even $d \in \mathbb{N}$, a degree-d pseudoexpectation in indeterminates $x = x_1, \ldots, x_n$ is a linear operator $\tilde{\mathbf{E}} : \mathbb{R}[x_1, \ldots, x_n]_{\leq d} \to \mathbb{R}$ which satisfies $\tilde{\mathbf{E}}\, 1 = 1$ an $\tilde{\mathbf{E}}\, p^2 \geq 0$ for every degree-$d/2$ polynomial p. We say that $\tilde{\mathbf{E}}$ *satisfies* an inequality $p(x) \geq 0$ if for every q such that $\deg(p \cdot q^2) \leq d$ we have $\tilde{\mathbf{E}}\, pq^2 \geq 0$.

Archimedean systems and duality We say that a system of polynomial inequalities $p_1(x) \geq 0, \ldots, p_m \geq 0$ is Archimedean if for some real $M > 0$ it contains the inequality $\|x\|^2 \leq M$. SoS proofs and pseudoexpectations satisfy a natural duality for Archimedean systems, which we use often. Namely: for every Archimedian system $p_1, \ldots, p_m$ and every polynomial f and every degree d, exactly one of the following holds.

1. For every $\varepsilon > 0$ there is an SoS proof $p_1 \geq 0, \ldots, p_m \geq 0 \vdash_d f \geq -\varepsilon$.
2. There is a degree-d pseudoexpectation satisfying $p_1 \geq 0, \ldots, p_m \geq 0$ but $\tilde{\mathbf{E}}\, f < 0$.

2.3.2 Privacy

We have already seen the definition of (pure) differential privacy. Our approach relies heavily upon the exponential mechanism of [84], we employ a volume-based version which appears in [66]. The proof is standard but we include it for completeness.

Theorem 2.3.1 (volume-based exponential mechanism [84, 66])**.** *The exponential mechanism M_E on inputs $X, \mathcal{H} \subset \mathbb{R}^d$, s, selects and outputs some object $h \in \mathcal{H}$, where the probability a particular h is selected is proportional to $\exp(\frac{\varepsilon s(X,h)}{2\Delta})$. Let $\mathcal{H}^* \subseteq \mathcal{H}$ be a set such that, $\mathrm{OPT}(X) \leq \inf_{h \in \mathcal{H}^*} s(X,h)$ be a lower bound for the score attained by the objects in $\mathcal{H}^*$ with respect to the dataset X. Moreover, let $\mathrm{vol}(S)$ denote the Lebesgue measure of S in $\mathbb{R}^d$. Then*

$$\mathbb{P}\left[s(M_E(X)) \leq \mathrm{OPT}(X) - \frac{2\Delta}{\varepsilon}\left(\ln\left(\frac{\mathrm{vol}(\mathcal{H})}{\mathrm{vol}(\mathcal{H}^*)} + t\right)\right)\right] \leq \exp(-t).$$

Proof. We follow the same argument as the standard exponential mechanism analysis.

$$\begin{aligned}
\mathbb{P}[s(M_E(X)) \leq c] &= \frac{\int_{h : s(X,h) \leq c, h \in \mathcal{H}} \exp\left(\frac{\varepsilon s(X,h)}{2\Delta}\right) \mathrm{d}h}{\int_{h' : h' \in \mathcal{H}} \exp\left(\frac{\varepsilon s(X,h')}{2\Delta}\right) \mathrm{d}h'} \\
&\leq \frac{\mathrm{vol}(\mathcal{H}) \exp\left(\frac{\varepsilon c}{2\Delta}\right)}{\mathrm{vol}(\mathcal{H}^*) \exp\left(\frac{\varepsilon\, \mathrm{OPT}(X)}{2\Delta}\right)} \\
&= \frac{\mathrm{vol}(\mathcal{H})}{\mathrm{vol}(\mathcal{H}^*)} \exp\left(\frac{\varepsilon(c - \mathrm{OPT}(X))}{2\Delta}\right).
\end{aligned}$$

From this inequality, the theorem statement can be obtained by substituting in the prescribed value for c. It remains to explain the first inequality. The numerator can be upper bounded since we are taking the integral at most over $\mathcal{H}$, and the value of the integral at each point is less than $\exp(\varepsilon c/2\Delta)$. Similarly, the denominator can be lower bounded by considering only the points in $\mathcal{H}^*$, all of which have a lower bound of $\exp(\varepsilon\, \mathrm{OPT}(X)/2\Delta)$. □

We will also extensively use the following result of [16].

Theorem 2.3.2 (Lemma 6.5 of [16]). *For every ε and $d \in \mathbb{N}$, there is an ε-DP algorithm A with the following guarantees. Given access to an evaluation oracle for a concave, L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and to membership and projection oracles for a convex set $\mathcal{C} \subseteq \mathbb{R}^d$, the algorithm produces a sample from a distribution D such that for every (measurable) $S \subseteq \mathcal{C}$,*

$$e^{-\varepsilon}\,\mathbb{P}(A \in S) \leq \frac{\int_S \exp(f)}{\int_{\mathcal{C}} \exp(f)} \leq e^{\varepsilon}\,\mathbb{P}(A \in S)\,.$$

The algorithm runs in time $\mathrm{poly}(d, L\,\mathrm{diam}(\mathcal{C}), 1/\varepsilon, \log \mathrm{diam}(C))$, *making at most that many queries to the oracles.*

[16] actually provides the running time bound $\mathrm{poly}(d, L, \mathrm{diam}(\mathcal{C}), 1/\varepsilon)$, but we can deduce the more precise bound in the theorem statement by a simple scaling argument. We note that recent work of [83] provides an algorithm which improves the running time's dependence on $1/\varepsilon$ from polynomial to polylogarithmic, as well as reducing the polynomial dependence on d for convex sets $\mathcal{C}$ that contain a ball of radius r, at an additional cost of $\mathrm{poly}\log(1/r)$ (Remark 2.5 of [83]). As our main focus is on providing polynomial time algorithms and not on designing the fastest algorithms, we do not further employ their improved methods.

2.4 Meta-Theorem on SoS Exponential Mechanism

To describe the SoS exponential mechanism more completely, we prove a meta-theorem on its performance. This meta-theorem could be extended in various ways, but the version we give here captures the mechanism as it is used in our paper. The reader interested solely in our result on mean estimation may comfortably skip this section as we do not rely on it elsewhere, but the exposition and abstraction here may make it easier to follow.

Meta-theorem Setup Consider the following template to capture the exponential mechanism in the language of polynomials. Let $\mathcal{X}$ be a universe of possible datasets and $\mathcal{C} \subseteq \mathbb{R}^n$ be a set of candidates. Consider a score function of the following form. For each dataset X, suppose there is a family of polynomials $p^X, p_1^X, \ldots, p_m^X$ in variables $x_1, \ldots, x_n, y_1, \ldots, y_N$, and the score function $s(X, x)$ is given by

$$s(X, x) = \max_y p^X(x, y) \text{ such that } p_1^X(x, y) \geq 0, \ldots, p_m^X(x, y) \geq 0\,.$$

Example 2.4.1 (Tukey depth with margin). *To make things a bit less abstract, let us see that a score function closely related to the Tukey depth is expressible in this form. Suppose $X = X_1, \ldots, X_N \in \mathbb{R}^n$ is a dataset. Recall that the Tukey depth of a point $z \in \mathbb{R}^n$ is given by the maximum number of X_i's which lie on one side of a hyperplane through z.*

Later on, for a fixed $z \in \mathbb{R}^n$ with $\|z\| = 1$, we will want to run the exponential mechanism to select a direction v such that many X_i's lie above a hyperplane through z in

direction v, at distance at least distance $r > 0$ from that hyperplane, so our score function for v is the number of such X_i's. We can express this as follows:

$$s(X, v) = \max_{b_1,\ldots,b_N} \sum_{i=1}^{N} b_i \text{ s.t. for all } i,\ b_i^2 = b_i \text{ and } b_i\langle X_i - z, v\rangle \geq b_i \cdot r\,. \tag{2.2}$$

Note that for fixed z and v it is easy to compute the number of X_i's lying above the resulting hyperplane, but finding the maximizing v already appears to be a hard problem. (Of course, we want to accomplish only a related task: sampling *from a distribution on high-scoring v's.) Indeed, approximation algorithms for finding such v play a key role in recent algorithmic advances in robust and heavy-tailed statistics statistics.*

Utility and Bounded Sensitivity in the Language of Polynomials Utility in this framework takes exactly the same form as in the usual exponential mechanism – to demonstrate utility, one would show that high-scoring x's are good, and there are not too many low-scoring x's. The first of these statements is naturally expressible in the SoS proof system, and it turns out that the latter will not need to be expressed as an SoS proof. There is just one technical subtlety: the requirement that $x \in \mathcal{C}$, if it is used in the proof of utility, must be captured by the polynomials $p_1(x, y) \geq 0, \ldots, p_m(x, y) \geq 0$, as the SoS proof system has no other way of natively using the hypothesis that $x \in \mathcal{C}$. (We illustrate this in the example below.)

To formulate bounded sensitivity within SoS will take a little additional work – SoS will require for there to be a certain kind of witness to bounded sensitivity. This is not a major restriction, as natural proofs of bounded sensitivity for optimization-based score functions typically yield such witnesses anyway.

Making this concrete, a natural way to show that a score function like the above satisfies bounded sensitivity is to relate feasible solutions to the optimization problem for X to those for a neighboring dataset X', without losing too much in the objective value. Let us suppose that for each neighboring pair X, X' there is a transformation $y'(y)$ such that for all x, if y is feasible for X (i.e. $p_1^X(x, y) \geq 0, \ldots, p_m^X(x, y) \geq 0$) then $y'(y)$ is feasible for X' (i.e. $p_1^{X'}(x, y'(y)) \geq 0, \ldots, p_m^{X'}(x, y'(y)) \geq 0$). If additionally $p^X(x, y) - p^{X'}(x, y'(y)) \leq 1$, this transformation (together with the corresponding one mapping y's to ys) witnesses bounded sensitivity for s. For technical reasons, our meta-theorem imposes the restriction that y' is a linear function of y, but this still suffices for our algorithms.

Example 2.4.2 (Continuation of Example 2.4.1)**.** *First addressing utility: it turns out that the proof of utility for v having high score according to (2.2) will rely on $\|v\| \leq 1$. So we will have to strengthen our system of polynomials to include this constraint. As a technicality, we will also shift our objective function so that having positive score is good enough for utility.*

$$\max_{b,v} \sum_{i=1}^{N} b_i - 0.9N \text{ such that } b_i^2 = b_i, b_i\langle X_i - z, v\rangle \geq b_i \cdot r, \|v\|^2 \leq 1\,. \tag{2.3}$$

Now, to establish bounded sensitivity, consider two neighboring datasets $X = X_1, \ldots, X_N$ and $X' = X_1', X_2, \ldots, X_N$, where X and X' differ on the first vector. If (b, v) is a feasible

solution to (2.3) for X with objective value t, then we can replace it with $(0, b_2, \ldots, b_N, v)$ to get a feasible to solution for X' with objective value at least $t - 1$. The meta-theorem requires there to be an SoS proof of this fact; this (easy) SoS proof was first established in [50].

Robustly Satisfiable Polynomials Before stating our meta-theorem, we need one more technical definition, capturing a certain well-conditioned-ness property of a polynomial optimization problem. Ultimately, this condition will imply a Lipschitz property of semidefinite relaxations of that optimization problem. This Lipschitz property will be used, in turn, to bound the running time of MCMC-based samplers for probability densities using those semidefinite relaxations as log-probabilities.

The details of the following definition (Definition 2.4.3) may appear opaque, but they are not too important – a good intuitive interpretation is that the polynomial optimization problem

$$\max_{x,y} p(x,y) \text{ s.t. } p_1(x,y) \geq 0, \ldots, p_m(x,y) \geq 0$$

has a robust space of feasible solutions. Roughly, this means that for any x there is a small ball B around x such that for any $x' \in B$ there is a feasible solution (x', y) whose objective value isn't too large. (The actual condition we use is slightly weaker than this.) This type of condition is common in meta-theorems involving the SoS proof system, to rule out the use of pathological $p, p_1, \ldots, p_m$, and it is typically not too difficult to establish – see, e.g. [53, 97].

Definition 2.4.3 (Robustly satisfiable polynomial systems)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$ and let $p, p_1, \ldots, p_m$ be polynomials in $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_N$. Let $\eta > 0$. Consider a family of optimization problems, one for each $x \in \mathcal{C}$, given by*

$$\max_{y} p(x,y) \text{ s.t. } p_1(x,y) \geq 0, \ldots, p_m(x,y) \geq 0\,.$$

We say this family is η-robustly satisfiable if, for each $x \in \mathcal{C}$, each x' in the ball of radius η around x can satisfy the constraints. That is, for each x' such that $\|x' - x\| \leq \eta$, there exists y such that $p_1(x', y) \geq 0, \ldots, p_m(x', y) \geq 0$.

Note that in our algorithms, η will factor only into running times and not sample complexity or accuracy guarantees, so rather coarse bounds on η, perhaps loose by polynomial factors, suffice for our purposes.

Example 2.4.4 (Continuation of Examples 2.4.1, 2.4.2)**.** *Let us imagine now that $\mathcal{C}$ is a ball of radius 0.9 centered at the origin, to see a proof sketch of η-robust satisfiability for (2.3). For each $v \in \mathcal{C}$ and every vector Δ with $\|\Delta\| \leq 0.1$, the constraints of (2.3) are satisfied by $(v + \Delta, 0)$. So, (2.3) is η-well-conditioned for $\eta = 0.1$, with respect to $\mathcal{C}$.*

With this setup in hand, we can state our meta-theorem.

Theorem 2.4.5 (Meta-Theorem on SoS Exponential Mechanism)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a compact, convex set and $\mathcal{X}$ a universe of possible datasets, equipped with a "neighbors" relation.*

Suppose that for every dataset X there exists an Archimedean and η-robustly satisfiable system of polynomial inequalities $\mathcal{P}^X(x,y) = \{p_1^X(x,y) \geq 0, \ldots, p_N^X(x,y) \geq 0\}$ and a polynomial $p^X(x,y)$, all of degree at most D, in indeterminates $x_1, \ldots, x_n, y_1, \ldots, y_N$ such that for every neighboring dataset X' there is a linear function $y'(y)$ such that bounded sensitivity has an SoS proof:

$$\forall j,\ \mathcal{P}^X(x,y) \vdash_{\deg(p_j^{X'})} p_j^{X'}(x,y'(y)) \geq 0 \text{ and } \mathcal{P}^X(x,y) \vdash_D p^X(x,y) - p^{X'}(x,y') \leq 1\,.$$

Suppose also that for every X, there are SoS proofs $\mathcal{P}^X(x,y) \vdash_D p(x,y) \leq 1/\eta$ and $\mathcal{P}^X(x,y) \vdash_D -p(x,y) \leq 1/\eta$. Furthermore, suppose that the polynomials $\mathcal{P}^X$ and p^X, and the polynomials used in the above SoS proofs, all have coefficients expressible in at most B bits.

Then for every $\varepsilon > 0$ and $D \in \mathbb{N}$ there exists an ε-differentially private algorithm which takes as input the polynomials $p^X, p_1^X, \ldots, p_m^X$ and $B, \eta > 0$, with the following guarantees:

Utility: *For every X, if there is an SoS proof of utility for X which is degree-$\deg(p)$ with respect to p, i.e.,*

$$\mathcal{P}^X(x,y) \cup \{p(x,y) \geq 0\} \vdash_{D,\deg_p=\deg(p)} \|x - x^*(X)\|^2 \leq \alpha^2$$

for some vector $x^*(X) \in \mathbb{R}^n$ and $\alpha > 0$, where the coefficients of all polynomials involved in the proof are expressible with B bits, and if

$$\frac{\mathrm{vol}(\mathcal{C})}{\mathrm{vol}(\{x \in \mathcal{C} \,:\, \exists y \text{ s.t. } \mathcal{P}^X(x,y) \text{ and } p^X(x,y) \geq t\})} \leq r\,,$$

then the algorithm outputs x such that $\|x - x^*(X)\| \leq \alpha + 2^{-B}$ with probability at least $1 - r\exp(-\Omega(\varepsilon t))$.

Running time: *The algorithm runs in time*

$$\mathrm{poly}\left(n^D, N^D, m^D, \frac{1}{\varepsilon}, \frac{1}{\eta}, \mathrm{diam}(\mathcal{C}), B\right)\,,$$

making at most this many calls to membership and projection oracles for $\mathcal{C}$.

2.4.1 Proof of Theorem 2.4.5

We describe the algorithm we use to prove Theorem 2.4.5; then we assemble the lemmas we need for the analysis.

SoSExponentialMechanism Input: polynomials $p^X, p_1^X, \ldots, p_m^X$, $D \in \mathbb{N}$, $\eta > 0$, $B \in \mathbb{N}$.

1. For $x_0 \in \mathcal{C}$, let

$$s(X, x_0) = \max_{\tilde{\mathbf{E}}} \tilde{\mathbf{E}}\, p^X(x,y) \text{ such that } \deg \tilde{\mathbf{E}} = D, \tilde{\mathbf{E}} \text{ satisfies } \mathcal{P}^X, \text{ and } \tilde{\mathbf{E}}\, x = x_0$$

   where the optimization is over $\tilde{\mathbf{E}}$ in indeterminates x, y.

2. Let $B' = B + T(\operatorname{diam}\mathcal{C}, 1/\eta, 1/\varepsilon, d)$, where T is a sufficiently-large polynomial in the running time of the log-concave private sampler of [16], when run with finite-precision arithmetic.

3. Run the log-concave private sampling algorithm of [16] (Lemma 6.5) with score function s, Lipschitz parameter $\operatorname{poly}(1/\eta)$, and privacy parameter $\varepsilon/4$. Whenever the sampling algorithm makes a call to $s(X, x_0)$, solve the underlying SDP to $\operatorname{poly}(B')$ bits of precision.

Theorem 2.4.5 is immediate from the following lemmas, all of which we establish in the next section.

Lemma 2.4.6 (High-scoring x_0 is found in polynomial time)**.** *Given the setup of Theorem 2.4.5, for all X, if*

$$\frac{\operatorname{vol}(\mathcal{C})}{\operatorname{vol}(\{x \in \mathcal{C} \,:\, \exists y \text{ s.t. } \mathcal{P}^X(x,y) \text{ and } p^X(x,y) \geq t\})} \leq r\,,$$

then with probability at least $1 - r\exp(-\varepsilon(t/2-1))$, the x_0 output in step (2) of SoSExponentialMechanism has $s(X, x_0) \geq 0$.

Lemma 2.4.7 (High-scoring x_0 is useful)**.** *Under the assumptions of Theorem 2.4.5, for all X, if there is an SoS proof of utility for X as described in Theorem 2.4.5, then for all x_0 such that $s(X, x_0) \geq 0$, $\|x^*(X) - x_0\| \leq \alpha + 2^{-B}$.*

Lemma 2.4.8 (Privacy)**.** *Under the assumptions of Theorem 2.4.5, SoSExponentialMechanism satisfies ε-DP.*

For a proof of these lemmas, we refer the reader to the full version of the paper [51].

2.5 Coarse Estimation

Our overall algorithm for private mean estimation proceeds in two phases. At start of the algorithm, we assume that we have a point μ_0 such that $\|\mu - \mu_0\|_2 \leq R$. Equivalently, by re-centering, we assume that $\|\mu\|_2 \leq R$. Recall that packing lower bounds for differential privacy generally necessitate an assumption of this nature [48].

In the first phase of the algorithm we perform coarse mean estimation, improving our initial estimate of the mean from R to $\mathcal{O}(\sqrt{d})$ error in ℓ_2-norm. In the second phase, we perform fine mean estimation, further improving our estimate from $\mathcal{O}(\sqrt{d})$ to α error. The overall algorithm will follow by combining these two phases. In this section, we describe our algorithm for coarse mean estimation, quantified in Theorem 2.5.1.

Theorem 2.5.1 (Combination of Theorem 2.5.5 and Theorem 2.5.10)**.** *Suppose $\mathcal{D}$ is a d-dimensional distribution with mean μ and covariance Σ, such that $\|\mu\|_2 \leq R$, $\|\Sigma\|_2 \leq 1$. Then there exists a polynomial-time ε-DP algorithm that takes n i.i.d. samples from $\mathcal{D}$ and outputs y such that $\|y - \mu\|_2 \leq \mathcal{O}(\sqrt{d})$ with probability $1 - \beta$, and sample complexity*

$$n = \tilde{\mathcal{O}}\left(\frac{d\log R + \min\left(d, \log R\right)\log(1/\beta)}{\varepsilon}\right).$$

To demystify the sample complexity in the above theorem, note that in order to prove this theorem, we present two algorithms with different sample complexities. The above sample complexity is the minimum of the two.

Our first algorithm uses a rather simple approach: privately estimate the mean in each coordinate up to constant error (using, e.g., the exponential mechanism, though one could also employ private histograms as in [67]), and use composition over the d coordinates. The sample complexity of this approach is $\mathcal{O}(d(\log R + \log 1/\beta + \log d)/\varepsilon)$ (see Theorem 2.5.5). Similar methods have been previously employed for multivariate mean estimation (see, e.g., [58]), but we provide an analysis for completeness, and include a particular focus on the error probability β.

We observe that this coordinate-wise approach incurs a cost of $d \log(1/\beta)$, whereas the primary goal for heavy-tailed mean estimation in the non-private setting is to decouple these two factors. To this end, we give a second algorithm, which uses our SoS exponential mechanism paradigm. This algorithm has sample complexity $\mathcal{O}(\log R(d + \log 1/\beta + \log\log R)/\varepsilon)$ (see Theorem 2.5.10). In the settings where $\log R \ll d$, this algorithm yields a better sample complexity than the simpler approach. An interesting question is to design an efficient algorithm for this setting with sample complexity $\tilde{\mathcal{O}}((d \log R + \log(1/\beta))/\varepsilon)$; an appropriate application of the exponential mechanism provides an inefficient algorithm with this sample complexity. See Remark 2.5.15 for more discussion around shaving the additional $\log R$ factor.

2.5.1 First Algorithm: Coordinate-Wise

Main Steps

As the name suggests, our coordinate-wise algorithm works by running a one-dimensional algorithm on each coordinate of the samples from a d-dimensional distribution. The one-dimensional algorithm exploits the fact that the true mean is within constant distance of $0.99n$ of the samples with high probability. We use the exponential mechanism to find any point that satisfies the same property. By triangle inequality, this new point will be close to the true mean.

In the following we describe the main theorem of this subsection and the main steps of proving it.

Lemma 2.5.2. *Suppose $X_1, \dots, X_n$ are i.i.d. samples from a one-dimensional distribution $\mathcal{D}$ with mean μ and variance bounded above by 1. Furthermore, suppose $n \geq 4\log(1/\beta)/\alpha^2$, and $R^* \geq \sqrt{2/\alpha}$. Then with probability $1-\beta$ there exists a set $G \subseteq [n]$ such that $|G| \geq (1-\alpha)n$, and*

$$\forall i \in G : |X_i - \mu| \leq R^* .$$

Proof. This lemma is a standard application of Chebyshev's inequality and Hoeffding's bound. Suppose the X_i's are i.i.d. samples from $\mathcal{D}$ and suppose $\mathcal{D}$ has variance σ^2. Then by Chebyshev's inequality we know that

$$\mathbb{P}[|X_i - \mu| \geq R^*] \leq \frac{\sigma^2}{R^{*2}} ,$$

Let $I_i = \mathbb{1}[|X_i - \mu| \geq R^*]$, then I_i's are independent random variables between 0 and 1. Therefore, we may use Hoeffiding's bound. Let $S_n = \sum_{i=1}^n I_i$, then

$$\mathbb{P}\left[S_n \geq \frac{\sigma^2}{R^{*2}} n + \sqrt{n \log 1/\beta}\right] \leq \beta\,.$$

Since $R^* \geq \sqrt{2/\alpha}$, $n \geq 4\log(1/\beta)/\alpha^2$, $\sigma^2 \leq 1$, each of these terms are bounded by $\alpha n/2$, implying that $|G| \geq (1-\alpha)n$ as desired. □

We will prove the following theorem using a combination of the arguments from Section 2.5.1 to Section 2.5.1.

Theorem 2.5.3 (COARSE-1D-ESTIMATE)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}$, $R > R^* \geq 1$, $0 \leq \alpha < 1/3$. Suppose there exists a point $x_0 \in [-R, R]$ such that $|\{i \mid |X_i - x_0| \leq R^*\}| \geq (1-\alpha)n$. Then there exists an ε-DP algorithm and a universal constant C_0 such that if $n \geq C_0 \left(\log R + \log(1/\beta)\right)/\varepsilon\alpha$, the algorithm returns y such that $|y - x_0| \leq \mathcal{O}\left(R^*\right)$, with probability $(1-\beta)$, and runtime $\mathcal{O}\left(n^2 \log R\right)$.*

By combining Theorem 2.5.3 and Lemma 2.5.2, we arrive at the following theorem for univariate coarse mean estimation.

Theorem 2.5.4 (one-dimensional coarse mean estimation)**.** *Suppose $\mathcal{D}$ is a one-dimensional distribution with mean μ, and covariance σ^2, such that $|\mu| \leq R$ and $\sigma^2 \leq 1$. Then there exists an ε-DP algorithm that takes n i.i.d. samples from $\mathcal{D}$ and outputs y such that $|y - \mu| \leq \mathcal{O}(1)$ with probability $1 - \beta$, sample complexity*

$$n = \mathcal{O}\left(\frac{\log R + \log(1/\beta)}{\varepsilon}\right),$$

and runtime $\mathrm{poly}(n, \log R, 1/\varepsilon)$.

To extend this to the multivariate setting, we apply the algorithm from Theorem 2.5.4 with privacy budget ε/d and failure probability β/d to each coordinate (d times in total), and using composition of differential privacy.

Theorem 2.5.5 (simple high dimensional coarse mean estimation)**.** *Suppose $\mathcal{D}$ is a d-dimensional distribution with mean μ and covariance Σ, such that $\|\mu\|_2 \leq R$ and $\|\Sigma\|_2 \leq 1$. Then there exists an ε-DP algorithm that takes n i.i.d. samples from $\mathcal{D}$ and outputs y such that $\|y - \mu\|_2 \leq \mathcal{O}(\sqrt{d})$ with probability $1 - \beta$, sample complexity*

$$n = \mathcal{O}\left(\frac{d \log R + d \log(1/\beta) + d \log d}{\varepsilon}\right),$$

and runtime $\mathrm{poly}\left(n, d, \log R, 1/\varepsilon\right)$.

In the rest of this subsection we prove Theorem 2.5.3. To do so, we propose an algorithm and provide privacy, accuracy and runtime analysis for it.

Algorithm

Our algorithm (Algorithm 1) is based on the exponential mechanism. We construct a cover over $[-R, R]$, and run the exponential mechanism over it to capture a large number of X_i's in a ball of radius $\mathcal{O}(R^*)$. This quantity is formalized using the score function as defined in Definition 2.5.6.

Definition 2.5.6 (Score function)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}, R^* \geq 1$. For any point $y \in \mathbb{R}$ we define*

$$\textsc{Score}(y) = |\{i \mid |X_i - y| \leq 2R^*\}| \ .$$

Algorithm 1 COARSE-1D-ESTIMATE

Input: Parameters $R > R^* \geq 1$, dataset $X_1, \ldots, X_n \in \mathbb{R}$,

Operation:
1. Let $I = \{R^* k \mid \lfloor -R/R^* \rfloor \leq k \leq \lceil R/R^* \rceil, k \in \mathbb{Z}\}$.
2. Run the exponential mechanism over the points in I with score function SCORE (Definition 2.5.6), privacy budget ε, and sensitivity 1, to obtain $\widehat{y}$.

Output: $\widehat{y}$

Privacy analysis

Since SCORE has sensitivity 1, the algorithm is ε-DP by privacy of the exponential mechanism.

Accuracy analysis

Lemma 2.5.7. *Suppose $0 \leq \alpha < 1/3$, $n \geq 2\left(\log\left(2\lceil R/R^* \rceil + 1\right) + \log(1/\beta)\right)/\varepsilon\alpha$, and that there exists some point $x_0 \in [-R, R]$ such that $|\{i \mid |X_i - x_0| \leq R^*\}| \geq (1-\alpha)n$. Then the exponential mechanism in Algorithm 1 will return $\widehat{y}$ such that*

$$|\widehat{y} - x_0| \leq 3R^* \ ,$$

with probability $1 - \beta$.

Proof. We show that there exists a point in $I = \{R^* k \mid \lfloor -R/R^* \rfloor \leq k \leq \lceil R/R^* \rceil, k \in \mathbb{Z}\}$ that has SCORE at least $(1-\alpha)n$. Let G be the set of points which are close to the point x_0, that is, $G = \{i \mid |X_i - x_0| \leq R^*\}$ and $|G| \geq (1-\alpha)n$. Let y^* be the point in I that is closest to x_0. Since $x_0 \in [-R, R]$, $|x_0 - y^*| \leq R^*$. Then

$$\forall i \in G \ : \ |y^* - X_i| \leq |y^* - x_0| + |x_0 - X_i| \leq R^* + R^* \leq 2R^* \ .$$

Therefore y^* has $\textsc{Score}$ at least $(1-\alpha)\,n$. By the guarantees of the standard exponential mechanism, the exponential mechanism when instantiated with the $\textsc{Score}$ function over I will return $\widehat{y}$ such that

$$\textsc{Score}(\widehat{y}) \geq (1-\alpha)n - \frac{2}{\varepsilon}\left(\log\left(2\lceil R/R^*\rceil + 1\right) + \log(1/\beta)\right)\,,$$

with probability $1-\beta$. Since $\alpha n \geq \frac{2}{\varepsilon}\left(\log\left(2\lceil R/R^*\rceil + 1\right) + \log(1/\beta)\right)$, the exponential mechanism will return $\widehat{y}$ such that

$$\textsc{Score}(\widehat{y}) \geq (1-2\alpha)n\,,$$

with probability $1-\beta$. Let $T = \{i \mid |X_i - \widehat{y}| \leq 2R^*\}$, then $|T| \geq (1-2\alpha)n$, with probability $1-\beta$. Moreover $G = \{i \mid |X_i - x_0| \leq 2R^*\}$ has at least $(1-\alpha)n$ members. Therefore since $\alpha < 1/3$, we have that $T \cap G \neq \emptyset$, and therefore there exists some $j \in T \cap G$. Hence $|X_j - x_0| \leq R^*$ and $|X_j - \widehat{y}| \leq 2R^*$. Therefore $|y - x_0| \leq 3R^*$ with probability $1-\beta$, as desired. □

Runtime analysis: lazy exponential mechanism

If the exponential mechanism in Algorithm 1 is naïvely implemented it will take $\Theta(Rn)$ time. This is a pseudo-polynomial time algorithm, but we would prefer a truly polynomial time algorithm, with running time polynomial in $\log R$. In the following theorem, we describe how this application of the exponential mechanism can be implemented in $\mathcal{O}(n^2 \log R)$ time. This style of "lazy exponential mechanism" may be folklore, but we include the argument for completeness.

Theorem 2.5.8 (lazy exponential mechanism)**.** *Suppose $C \in \mathbb{N}$, $I = \{b_1, \ldots, b_m\}$ is a set of bins, and $n \in \mathbb{N}$ objects $O = \{o_1, \ldots, o_n\}$ are given. Moreover suppose that each item can belong to at most C bins, and that for each object o, the set of the bins it belongs to can be computed in time $\mathcal{O}(C \log m)$. For a bin b_i, let*

$$\textsc{Score}(b_i) = |\{j \mid o_j \text{ belongs to } b_i\}|\,.$$

Then the exponential mechanism over I with score function $\textsc{Score}$ *can be implemented in time $\mathcal{O}(Cn^2 \log m)$.*

Proof. Let $J = \{b_i \mid \textsc{Score}(b_i) > 0\}$. We can construct J and compute the score function for all of its members by iterating over O and finding which bins each of them belongs to. This will take time $\mathcal{O}(Cn \log m)$. Now since all of the other bins are empty, the probability that the exponential mechanism with sensitivity Δ and privacy budget ε selects $b_j \in J$ has to be

$$\mathbb{P}[b_j] := \frac{\exp(\frac{\varepsilon \textsc{Score}(b_j)}{2\Delta})}{\sum_{k, b_k \in J} \exp(\frac{\varepsilon \textsc{Score}(b_k)}{2\Delta}) + m - |J|}\,.$$

Therefore the probability of choosing each member of J can be computed in time $\mathcal{O}(n \log m)$. Moreover, let $\perp$ be a special symbol representing the event that the exponential mechanism selects a b_j such that $b_j \notin J$. It is easy to see that

$$\mathbb{P}[\perp] := \frac{m - |J|}{\sum_{k, b_k \in J} \exp(\frac{\varepsilon \textsc{Score}(b_k)}{2\Delta}) + m - |J|}\,.$$

We now draw an element from the set $\{b_j \mid j \in J\} \cup \{\perp\}$ where the probabilities are as established above. If the sampling algorithm picks a $b_j \in J$ we can return that, and we are done. Otherwise, if $\perp$ is picked, we can sample an index from $i \in \{1, \ldots, m - |J|\}$ uniformly at random in time $\mathcal{O}(\log m)$. Now we aim to construct a uniformly random member from $I - J$ from this random index i. To do we just need to consider $K = \{b_i, \ldots, b_{i+|J|}\}$. By the intermediate value theorem, there exists a member $b_{i+k} \in K$ such that k equals the number of the members of J which have index smaller than or equal to $i + k$. For the smallest k that has such property, we return b_{j+k}, it is easy to see that this scheme returns a uniformly random member from $I - J$. Moreover, the overall algorithm takes time at most $\mathcal{O}(Cn^2 \log m)$. Note that in this proof, the $\log m$ factors come from the bit complexity of indices of the members of I. □

Corollary 2.5.9. *The exponential mechanism in Algorithm 1 can be implemented in time* $\mathcal{O}(n^2 \log R)$.

Proof. Each X_i can be counted in the SCORE of at most 5 members of I. Furthermore, for each X_j, to find the members of I for which it is counted in, we just need to find the points in $[X_i - 2R^*, X_i + 2R^*]$ which are of the form kR^*, for some $k \in \mathbb{Z}$. This can be done in time $\mathcal{O}(\log R)$. The statement thus holds by an application of Theorem 2.5.8. □

2.5.2 Second Algorithm: SoS Exponential Mechanism

Main Steps

We proceed to describe our second coarse-estimation algorithm, based on our SoS Exponential Mechanism framework. Our coverage is for a concrete problem (as opposed to the abstract setting in our meta-theorem in Section 2.4) and conceptually simpler than our fine-estimation algorithm (Section 2.6), so it may be a good starting point for a reader to understand the framework.

Consider the following simple but inefficient method for performing coarse estimation. Let the score function for a point v be the number of datapoints within a ball of radius $\text{poly}(d)$ centered at v. One could run the exponential mechanism with this score function, with the set of candidate points being the ball of radius R. Compared to the coordinate-wise algorithm described in Section 2.5.1, this approach works directly in the high-dimensional space.

Of course, this algorithm is not polynomial time, as the naive method of implementing it would involve computing the score function for all points in a cover of size exponential in d. Instead, appealing to our SoS Exponential mechanism framework, we first replace the score function with a similar SoS optimization problem. The resulting algorithm and optimization problem are both described in Section 2.5.2. Note that for technical reasons, we actually run the exponential mechanism $\mathcal{O}(\log R)$ times, making a constant factor improvement in the radius each time. This iterative process is what gives rise to a sub-optimal $\mathcal{O}(\log R \log(1/\beta)/\varepsilon)$ term in the sample complexity, To prove accuracy of the overall procedure, we first show that the solution to the SoS optimization problem will give a point with high utility (Lemma 2.5.14 in Section 2.5.2). Then, assuming the instantiations of the exponential mechanism are all "successful" (that is, produce points which are of

comparable utility to the one which optimizes the score function), it is not hard to show that the final point output will satisfy our desired accuracy guarantee (Corollary 2.5.17 in Section 2.5.2).

At this point, there are three components of the exponential mechanism left to argue: privacy, accuracy, and computational efficiency. Privacy is fairly straightforward: we prove that the score function has bounded sensitivity (Lemma 2.5.18 in Section 2.5.2), which implies privacy (Corollary 2.5.19) by combining the exponential mechanism's privacy guarantee with basic composition. Utility follows the standard recipe for the volume-based exponential mechanism (Theorem 2.3.1): in addition to the aforementioned sensitivity bound, we require that points with high utility are sufficiently dense (Lemma 2.5.22 in Section 2.5.2). Combining the two gives the desired accuracy guarantee in Lemma 2.5.24 in Section 2.5.2. Finally, to ensure that the algorithm is polynomial time, we employ the sampling-based exponential mechanism of [16] (Theorem 2.3.2). We again use the fact that the score function has bounded sensitivity, but also Lipschitzness (Lemma 2.5.21 in Section 2.5.2), and concavity (Corollary 2.5.23 in Section 2.5.2). Combining these properties allow us to conclude the exponential mechanism can be implemented in polynomial time (Lemma 2.5.25 in Section 2.5.2).

These allow us to conclude the following result for coarse mean estimation.

Theorem 2.5.10 (high dimensional coarse mean estimation via SoS exponential mechanism)**.** *Suppose $\mathcal{D}$ is a d-dimensional distribution with mean μ and covariance Σ, where $\|\mu\|_2 \le R$, and $\|\Sigma\|_2 \le 1$. Then there exists a polynomial time ε-DP algorithm* COARSE-ESTIMATE *that takes*

$$n = \mathcal{O}\left(\frac{\log R\,(d + \log(1/\beta) + \log\log R)}{\varepsilon}\right)$$

i.i.d. samples from $\mathcal{D}$ and outputs y, such that $\|y - \mu\|_2 \le \mathcal{O}(\sqrt{d})$ with probability $1 - \beta$.

Before we begin the main analysis, we establish a deterministic condition on the dataset under which our algorithm succeeds. This condition holds with high probability if the dataset is sampled from an appropriate distribution. Specifically, the following is a multivariate analogue of Lemma 2.5.2, which states that most samples from a distribution with bounded covariance will be within a bounded distance of the mean. The proof is similar to before.

Lemma 2.5.11. *Suppose $X_1, \dots, X_n$ are i.i.d. samples from a d-dimensional distribution $\mathcal{D}$, with mean μ and covariance Σ, where $\|\Sigma\|_2 \le 1$. Furthermore suppose $n \ge 4\log(1/\beta)/\alpha^2$ and $R^* \ge \sqrt{2d/\alpha}$. Then with probability $1 - \beta$, there exists a set G such that $|G| \ge (1-\alpha)n$, and*

$$\forall i \in G : \|X_i - \mu\|_2 \le R^*\,.$$

With this in place, it suffices to prove the following theorem, which is our main technical result of this section.

Theorem 2.5.12. *For every $\varepsilon > 0$ and $n, d \in \mathbb{N}$ there is an ε-differentially private polynomial-time algorithm* COARSE-ESTIMATE$(R, R^*, X_1, \dots, X_n)$ *with the following guarantees. There exists a universal constant C_0 that given $X_1, \dots, X_n \in \mathbb{R}^d$, $R, R^* > 0$,*

COARSE-ESTIMATE *outputs* $y \in \mathbb{R}^d$ *such that if there exists* $x \in \mathbb{R}^d$ *with* $\|x\|_2 \le R$ *and such that* $|\{X_i : \|X_i - x\|_2 \le R^*\}| \ge (1-\alpha)n$, $R \ge 1000R^* > 1$, $n \ge C_0 \frac{\log R}{\alpha\varepsilon}(d + \log(1/\beta) + \log\log R)$, $0 < \alpha < 10^{-4}$, *then* $\|y - x\|_2 \le O(R^*)$, *with probability at least* $1 - \beta$.

In the rest of this section (from Section 2.5.2 to Section 2.5.2) we prove Theorem 2.5.12. To do so, we present an algorithm and provide privacy, accuracy, and runtime analysis for it.

Algorithm

We describe the score function employed in the exponential mechanism. Our algorithm repeatedly invokes the exponential mechanism with a geometrically decreasing radius.

Definition 2.5.13 (COARSE-SDP)**.** *Let* $X_1, \dots, X_n, y \in \mathbb{R}^d$, *and* $R > r > 1$. *Let* COARSE-SDP *be the following convex program, where the maximization is over degree-4 pseudo-distributions in indeterminates* $v_1, \dots, v_d, b_1, \dots, b_n$.

$$\begin{aligned} \max_{\tilde{\mathbf{E}}} \ & \tilde{\mathbf{E}} \sum_{i=1}^{n} b_i \ \textit{s.t.} \\ & \tilde{\mathbf{E}} \ \textit{satisfies} \ b_i^2 = b_i \quad \forall i \\ & \tilde{\mathbf{E}} \ \textit{satisfies} \ \|v\|_2^2 \le (2R + R/1000)^2 \\ & \tilde{\mathbf{E}} \ \textit{satisfies} \ b_i \|X_i - v\|_2^2 \le b_i \cdot r^2 \quad \forall i \\ & \tilde{\mathbf{E}}\, v = y\,. \end{aligned}$$

Algorithm 2 COARSE-ESTIMATE

Input: Parameters $R \ge 1000R^* \ge 1$, dataset $X_1, \dots, X_n \in \mathbb{R}^d$

Operation: Let $m = \lfloor \log_5(R/1000R^*) \rfloor$. For $0 \le t < m$ do:

1. Let $R_t = R(0.2)^t$.
2. Run the exponential mechanism with score function COARSE-SDP ($R \leftarrow R_t, r \leftarrow R_t/100$), privacy budget ε/m, and sensitivity 1 over the ball of radius $R_t + R_t/1000$ in $\mathbb{R}^d$ to obtain y_t.
3. For every j let $X_j = X_j - y_t$.

Output: $\widehat{y} = \sum_{i=0}^{m-1} y_i$

SoS Analysis

We prove that a point with a high score function value will also have high utility by containing a large fraction of the dataset within a bounded radius.

Lemma 2.5.14. *Let $X_1, \dots, X_n \in \mathbb{R}^d$, $R > r > R^* > 1$, and $\alpha \in [0, 1/3]$ and let $\tilde{\mathbf{E}}$ be a feasible solution to* COARSE-SDP *with $\tilde{\mathbf{E}} \sum_{i \le n} b_i \ge (1-2\alpha)n$. For any $x_0 \in \mathbb{R}^d$ such that $|\{i : \|X_i - x_0\|_2 \le R^*\}| \ge (1-\alpha)n$ and $\|x_0\|_2 \le R$,*

$$\|y - x_0\|_2 < 10(r + \sqrt{\alpha}R)\,.$$

Proof. Let $G = \{i : \|X_i - x_0\| \le R^*\}$ be the *good* set of indices. Since $\sum_{i \in G} \tilde{\mathbf{E}}[b_i] + \sum_{i \notin G} \tilde{\mathbf{E}}[b_i] = \sum_{i=1}^n \tilde{\mathbf{E}}[b_i] \ge (1-2\alpha)n$, and $|G| \ge (1-\alpha)n$,

$$\sum_{i \in G} \tilde{\mathbf{E}}[b_i] \ge (1-2\alpha)n - (n - |G|) \ge (1-3\alpha)n\,.$$

We have

$$|G| \cdot \|x_0 - v\|^2 = \sum_{i \in G} b_i \cdot \|x_0 - v\|^2 + \sum_{i \in G} (1-b_i) \cdot \|x_0 - v\|^2\,.$$

We will bound the two terms separately. For the first, for any $i \in G$, by SoS triangle inequality (Fact 2.9.7) we have

$$\tilde{\mathbf{E}}\, b_i \|x_0 - v\|^2 \le 2\,\tilde{\mathbf{E}}\, b_i (\|x_0 - X_i\|^2 + \|X_i - v\|^2) \le 4r^2 \cdot \tilde{\mathbf{E}}\, b_i\,.$$

(Here we used that $\tilde{\mathbf{E}}$ satisfies $b_i \ge 0$.) For the second term, again using triangle inequality, we have

$$\tilde{\mathbf{E}}(1-b_i)\|x_0 - v\|^2 \le 2\,\tilde{\mathbf{E}}(1-b_i)(\|v\|_2^2 + \|x_0\|_2^2) \le 2\,\tilde{\mathbf{E}}(1-b_i)(R^2 + (2R + R/1000)^2) \le 11R^2\,\tilde{\mathbf{E}}(1-b_i)\,.$$

(Here we used that $\tilde{\mathbf{E}}$ satisfies $1 - b_i \ge 0$.) So overall,

$$|G| \cdot \tilde{\mathbf{E}}\, \|x_0 - v\|^2 \le 4r^2 \cdot \tilde{\mathbf{E}} \sum_{i \in G} b_i + 11R^2 \cdot \tilde{\mathbf{E}} \sum_{i \in G} (1 - b_i)\,.$$

Now $\sum_{i \in G} \tilde{\mathbf{E}}\, b_i \ge (1-3\alpha)n$, so $\tilde{\mathbf{E}} \sum_{i \in G} (1-b_i) \le 3\alpha n$, and hence

$$\tilde{\mathbf{E}}\, \|x_0 - v|^2 \le 11 \cdot \frac{nr^2 + 3\alpha n R^2}{|G|} \le 66(r^2 + \alpha R^2) < 100(r^2 + \alpha R^2)\,.$$

Since $\|\tilde{\mathbf{E}}\, v - x_0\| \le \sqrt{\tilde{\mathbf{E}}\, \|v - x_0\|^2}$, we are done. □

Remark 2.5.15. *In the proof above, if instead of $\tilde{\mathbf{E}} \sum_{i=1}^n b_i \ge (1-2\alpha)n$, we knew that $\tilde{\mathbf{E}}$ satisfies $\sum b_i \ge (1-2\alpha)n$, then we could have bounded the the second term by a small constant factor of $\alpha n\, \tilde{\mathbf{E}} \|x_0 - v\|_2^2$. If we could have done this, we would have avoided the large R term that we get in the end. Note that the R term is the only reason why in this solution we require to run the algorithm for $\mathcal{O}(\log R)$ iterations, which results in the suboptimal dependence of the complexity on $\log R$. Making sure that $\tilde{\mathbf{E}}$ satisfies $\sum b_i \ge (1-2\alpha)n$ and avoiding the unnecessary $\log R$ term in the sample complexity can be achieved by using a more sophisticated technique and sampling scheme presented in Chapter 3, and [52].*

Accuracy Analysis Assuming Exponential Mechanism's Success

Here we show that if each iteration of Algorithm 2 produces a high utility solution, the resulting estimates will improve in accuracy at a rate exponential in the number of iterations.

Proposition 2.5.16. *Suppose* $|\{i : \|X_i - x_0\|_2 \le R^*\}| \ge (1-\alpha)n$, $\alpha \le 10^{-4}$, $\|x_0\|_2 \le R$, *and assume the exponential mechanism in Algorithm 2 is successful in every step, i.e.,* y_t *has score larger than* $(1-2\alpha)n$ *for every* t. *Then for each* t,

$$\|\sum_{i=0}^{t-1} y_i - x_0\|_2 \le R_t\,.$$

Proof. We prove this proposition through induction. For the base case $t = 0$, the sum $\sum_{i=0}^{t-1} y_i$ is empty and we are left with $\|x_0\| \le R_0 = R$, which holds by assumption. For the inductive case, we suppose the statement holds for $t-1$, and prove that it holds for t. The inductive hypothesis is that $\|\sum_{i=0}^{t-2} y_i - x_0\|_2 \le R_{t-1}$. We may now use Lemma 2.5.14, where x_0 and R in the lemma statement are replaced by $\sum_{i=0}^{t-2} y_i - x_0$ and R_{t-1}, respectively. This gives us

$$\|\sum_{i=0}^{t-1} y_t - x_0\|_2 \le 10(R_{t-1}/100 + R_{t-1}/100) \le R_{t-1}/5 = R_t\,,$$

as desired. □

As an immediate corollary, we can bound the resulting error after Algorithm 2 is complete.

Corollary 2.5.17. *Under the same assumptions as Proposition 2.5.16 and noting that that* $m = \log_5(R/1000R^*)$, *Algorithm 2 outputs* y, *such that*

$$\|y - x_0\|_2 \le 1000R^*\,.$$

Sensitivity

We now show that the score function has bounded sensitivity.

Lemma 2.5.18 (COARSE-SDP sensitivity)**.** COARSE-SDP *has sensitivity* 1 *with respect to the* X_i*'s.*

Proof. Without loss of generality suppose X, X' are two datasets that differ in the first index. Suppose for inputs $(X_1, X_2, \dots X_n)$, $\tilde{\mathbf{E}}$ is the degree-4 pseudo-expectation that maximizes the objective function of COARSE-SDP. In order to prove that COARSE-SDP has sensitivity 1, we prove that there exists a feasible solution $\tilde{\mathbf{E}}'$ to the COARSE-SDP convex program, on inputs $(X_1', X_2, \dots, X_n)$, such that

$$\tilde{\mathbf{E}}'[\sum b_i] \ge \tilde{\mathbf{E}}[\sum b_i] - 1\,.$$

If we do so, sensitivity follows by symmetry. To construct $\tilde{\mathbf{E}}'$, we describe it by the pseudo-moments up to the fourth degree. For every degree-4 monomial f in b, v that does not contain b_1, let

$$\tilde{\mathbf{E}}'[f(b,v)] = \tilde{\mathbf{E}}[f(b,v)]\,,$$

and for every degree-4 monomial f in b, v, that contains b_1, let

$$\tilde{\mathbf{E}}'[f(b,v)] = 0\,.$$

Then $\tilde{\mathbf{E}}'$ satisfies the constraints in COARSE-SDP. Furthermore, $\tilde{\mathbf{E}}'[\sum b_i] \geq \tilde{\mathbf{E}}[\sum b_i] - 1$. It remains to show that $\tilde{\mathbf{E}}'$ is a pseudo-expectation. This is because $\tilde{\mathbf{E}}'[((1,b,v)^{\otimes 2})((1,b,v)^{\otimes 2})^T]$ is PSD, which is a direct result of Lemma 2.9.1. □

Combining this with the privacy guarantees of the exponential mechanism and basic composition, we see that the overall algorithm is ε-DP.

Corollary 2.5.19 (privacy analysis)**.** *The exponential mechanism in Algorithm 2 is ε/m-DP. Hence Algorithm 2 is ε-DP by composition.*

Lipschitzness

To prove that the score function is Lipschitz, we rely upon the following lemma, which is a direct result of the duality of pseudo-distributions and SoS proofs (see, e.g., [14]).

Lemma 2.5.20. *Suppose $R, r, X_1, \ldots, X_n$ are fixed. Let $f(y)$ denote the maximum value* COARSE-SDP *can obtain for y. Then for every $y \in \mathbb{R}^d, \omega > 0$, there exists $\alpha_i \in \mathbb{R}[b,v]_{\leq 2}$, a degree-2 SoS polynomial β, $\gamma_i \geq 0$, $\lambda_i \in \mathbb{R}$, and degree-4 SoS polynomial σ such that*

$$\begin{aligned} f(y) + \omega - \sum_{i=1}^n b_i = \sum_{i=1}^n \alpha_i(b,v)(b_i^2 - b_i) + \beta(b,v)((2R + R/1000)^2 - \|v\|_2^2) \\ + \sum_{i=1}^n \gamma_i b_i (r^2 - \|X_i - v\|_2^2) + \sum_{i=1}^d \lambda_i (v_i - y_i) + \sigma(b,v)\,. \end{aligned}$$

Note that the above equality is a polynomial identity in b and v.

Note that the constraints on b and v in the optimization problem imply boundedness (Lemma 2.8.3) and therefore our problem satisfies the duality conditions. We use this lemma to prove the following theorem.

Lemma 2.5.21 (COARSE-SDP lipschitzness)**.** *In the convex optimization problem* COARSE-SDP*, suppose $R, r, X_1, \ldots, X_n$ are fixed. Let $f(y)$ denote the maximum value* COARSE-SDP *can obtain for y. Then $f(y)$ is L-Lipschitz over the ball of radius $R + R/1000$ in $\mathbb{R}^d$, where $L = n\sqrt{d}/R$.*

Proof. Let

$$\begin{aligned} g(\alpha, \beta, \gamma, \sigma; y) = \sum_{i=1}^n b_i + \sum_{i=1}^n \alpha_i(b,v)(b_i^2 - b_i) + \beta(b,v)((2R + R/1000)^2 - \|v\|_2^2) \\ + \sum_{i=1}^n \gamma_i b_i (r^2 - \|X_i - v\|_2^2) + \sum_{i=1}^d \lambda_i (v_i - y_i) + \sigma(b,v)\,. \end{aligned}$$

Note that g is a polynomial in b and v. Then by Lemma 2.5.20 we know that for any $y \in \mathbb{R}^d, \omega > 0$, there exists $\alpha, \beta, \gamma, \lambda, \sigma$ such that

$$f(y) + \omega = g(\alpha, \beta, \gamma, \lambda, \sigma; y)\,, \tag{2.4}$$

is a polynomial identity in b and v. First, we prove that the λ_i's are bounded. To do so we plug different values for b and v in Equation (2.4). Let $b = 0$, and $v = y + te_j$, such that $\|y + te_j\|_2 = 2R + R/1000$, and $t > 0$. Such t exists and is larger than R due to $\|y\|_2 \le R + R/1000$, and the triangle inequality. Plugging this b and v into Equation (2.4), gives us

$$f(y) + \omega \ge \lambda_j t\,.$$

Since $t \ge R$, $f(y) + \omega \ge R\lambda_j$. Similarly, by plugging $b = 0, v = y - te_j$, we obtain $f(y) + \omega \ge -R\lambda_j$, and therefore, $|\lambda_j| \le (f(y) + \omega)/R$. Since $f(y) \le n$, we obtain $|\lambda_i| \le (n + \omega)/R$, and hence

$$\|\lambda\|_2 \le \sqrt{d}(n + \omega)/R\,.$$

Now that we have bounded λ, we can prove that f is Lipschitz. To do so, suppose $y' = y + \delta u$, where $\|u\|_2 = 1$. If for y' we can construct a dual solution that obtains value $f(y) + L\delta + h(\omega)$, where $\lim_{\omega \to 0} h(\omega) = 0$, we have proven that $f(y') \le f(y) + L\delta$, and Lipschitzness will follow by symmetry. To construct this dual solution, let $(\alpha', \beta', \gamma', \lambda', \sigma') = (\alpha, \beta, \gamma, \lambda, \sigma)$. Then

$$\begin{aligned} g(\alpha', \beta', \gamma', \lambda'\sigma'; y') &= g(\alpha, \beta, \gamma, \lambda, \sigma; y) + \delta\langle \lambda, u\rangle \\ &\le f(y) + \omega + \delta\|\lambda\|_2 \\ &\le f(y) + \omega + \delta\sqrt{d}(n + \omega)/R \\ &\le f(y) + \delta\sqrt{d}n/R + \omega(1 + \delta\sqrt{d}/R)\,, \end{aligned}$$

as desired. Therefore f is $L = n\sqrt{d}/R$-Lipschitz. □

Volume

We show that the set of points with high score function is sufficiently dense. This essentially follows from the fact that every point which is close to the optimal solution also has high utility, and by an argument about the volume of ℓ_2-balls.

Lemma 2.5.22 (COARSE-SDP volume)**.** *Let $R > 1000R^* > 1$. Suppose there exists a point $x_0 \in \mathbb{R}^d$ such that $|\{i : \|X_i - x_0\|_2 \le R^*\}| \ge (1 - \alpha)n$ and $\|x_0\|_2 \le R$. Let $\mathbb{B}$ denote the ball of radius $R + R/1000$ in $\mathbb{R}^d$. Then there exists a set $\mathcal{H}^* \subseteq \mathbb{B}$, such that for every point y in $\mathcal{H}^*$, the value of* COARSE-SDP *with $(R \leftarrow R, r \leftarrow R/100)$ is larger than $(1 - \alpha)n$ and*

$$\log \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathcal{H}^*)} \le d \log(2000)\,.$$

Proof. Let $\mathcal{H}^*$ be the ball of radius $R/1000$ centered at x_0. Then $\mathcal{H}^* \subseteq \mathbb{B}$, and

$$\log \frac{\mathrm{vol}\,\mathbb{B}}{\mathrm{vol}(\mathcal{H}^*)} = \log \left(\frac{R + R/1000}{R/1000}\right)^d \le d \log(2000)\,.$$

It remains to prove that every point in $\mathcal{H}^*$ has high COARSE-SDP value. Suppose y is a point in $\mathcal{H}^*$. We know that COARSE-SDP is a relaxation of the polynomial optimization problem; therefore, if we propose a solution to the polynomial optimization problem with value $(1-\alpha)n$, we have proven that COARSE-SDP has value at least $(1-\alpha)n$. Hence, it remains to show that there exists a set of size at least $(1-\alpha)n$ from the X_i's such that $\|X_i - y\|_2 \leq R/100$. Consider the set $G = \{i \,:\, \|X_i - x_0\|_2 \leq R^*\}$. We know that $|G| \geq (1-\alpha)n$. Moreover,

$$\forall i \in G: \; \|X_i - y\|_2 \leq \|X_i - x_0\|_2 + \|x_0 - y\|_2 \leq R^* + R/1000 \leq 2R/1000 \leq R/100\,,$$

as desired. □

Concavity

It is not hard to see that the score function is concave.

Corollary 2.5.23 (COARSE-SDP concavity)**.** *Let $f(y)$ denote the value of* COARSE-SDP *for y. Then f is concave.*

Proof. Direct result of Lemma 2.9.2. □

Accuracy Analysis of the Exponential Mechanism

Given that we have established bounded sensitivity (Lemma 2.5.18) and that the set of good solutions is sufficiently dense (Lemma 2.5.22), we can apply the volume-based exponential mechanism (Theorem 2.3.1).

Lemma 2.5.24 (guarantee of the exponential mechanism for COARSE-SDP)**.** *Let $f(y)$ denote the value of* COARSE-SDP *for y. The exponential mechanism in Algorithm 2 with score function f, and privacy budget ε/m returns y such that*

$$f(y) \geq (1-\alpha)n - \frac{2m}{\varepsilon}\left(d\log(2000) + \log(m/\beta)\right)\,,$$

with probability $1-\beta/m$. Specifically, if

$$n \geq \frac{2m}{\alpha\varepsilon}\left(d\log\left(2000\right) + \log(m/\beta)\right)\,,$$

then

$$f(y) \geq (1-2\alpha)\,,$$

with probability $1-\beta/m$. We say the exponential mechanism has been successful if the above inequality holds.

Runtime Analysis of the Exponential Mechanism

At this point, we have established bounded sensitivity (Lemma 2.5.18), Lipschitzness (Lemma 2.5.21), and concavity (Corollary 2.5.23). These properties suffice to apply the efficient sampling algorithm of Theorem 2.3.2.

Lemma 2.5.25 (runtime of the exponential mechanism)**.** *There exists a sampling scheme for the exponential mechanism in Algorithm 2 that takes time* $\mathrm{poly}(n, d, \log R, 1/\varepsilon)$.

Note that, since the function is $L = n\sqrt{d}/R$-Lipschitz and since we are running the algorithm over the ball of radius $R + R/1000$ (i.e., $\mathrm{diam}(\mathcal{C}) \le R + R/1000$), we have that $L\,\mathrm{diam}(\mathcal{C}) = O(n\sqrt{d})$ thus making the running time polynomial in n and d, but not in R.

2.5.3 Robustness

To conclude, we note that both of our coarse estimation algorithms are robust to contamination.

Remark 2.5.26 (Robustness of Coarse Estimation)**.** *For both coarse estimation algorithms (i.e., the coordinate-wise approach, Algorithm 1, and the SoS exponential mechanism approach, Algorithm 2), the* Score *function has sensitivity* 1*. Moreover, there exists a constant* η_0 *(say* $\eta_0 \le 0.01$*), such that if* Score *is perturbed by* ηn*,* $\eta \le \eta_0$*, all of our arguments still hold. Therefore, there exists a universal constant* η_0 *for which our algorithm has the same guarantees if* η*-fraction of the samples are corrupted, where* $\eta \le \eta_0$*.*

2.6 Fine Estimation

From Section 2.5, we know how to find an estimate of the mean up to distance $\mathcal{O}(\sqrt{d})$. In this section we prove the following theorem. Putting the following theorem and our theorem for coarse estimation (Theorem 2.5.1) together, gives us an efficient algorithm for privately estimating the mean of a d-variate distribution with bounded mean and covariance up to error α.

Theorem 2.6.1 (high dimensional fine mean estimation via SoS exponential mechanism)**.** *Suppose* $\mathcal{D}$ *is a* d*-dimensional distribution with mean* μ *and covariance* Σ*, where* $\|\Sigma\|_2 \le 1$*. Moreover, suppose that an initial estimate* $\widetilde{\mu}_0$ *is given such that* $\|\mu - \widetilde{\mu}_0\|_2 \le \mathcal{O}(\sqrt{d})$*. Then there exists a polynomial time* ε*-DP algorithm* fine-estimation *that takes* n *samples from* $\mathcal{D}$ *and outputs* $\widetilde{\mu}^*$ *such that* $\|\widetilde{\mu}^* - \mu\|_2 \le \alpha$*, with probability* $1 - \beta$*, and sample complexity*

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

More specifically, $\tilde{\mathcal{O}}$ *is hiding the following lower order logarithmic factors:* $\log d$*, and* $\log(1/\alpha)$

Proof. Applying Theorem 2.6.12 with $m = 1/\alpha^2$, and noting that Assumption 2.6.9 holds with probability $1 - \beta$, due to Lemma 2.6.8. □

Suppose $\mathcal{D}$ is a d-dimensional distribution with mean μ, and covariance Σ such that $\|\Sigma\|_2 \leq 1$. Our goal is to privately estimate the mean of this distribution up to accuracy α with probability $1-\beta$. Moreover, suppose that we have access to an initial estimate of the mean such that $\|\mu - \widetilde{\mu}_0\|_2 \leq \mathcal{O}(\sqrt{d})$. We build our algorithm on top of the recent work in efficient high-dimensional mean estimation where the high-dimensional median of means technique is used [24, 50, 82]. As opposed to non-biased estimators that have $\mathcal{O}(\frac{d}{\beta\alpha^2})$ sample complexity, these estimators achieve the $\mathcal{O}(\frac{d+\log(1/\beta)}{\alpha^2})$, or the *sub-Gaussian rate* sample complexity.

In this technique, we take $n = mk$ samples, and divide them into k bins $\{\mathcal{B}\}_{i=1}^{k}$ each having m elements. Then we compute the means of each of these bins and name them Z_i's. A key observation then is that if $k \geq \max\{d, \Omega(\log(1/\beta))\}$, then with probability $1-\beta$ over the randomness of the samples, the true mean, μ is $(0.99, \mathcal{O}(\sqrt{1/m}))$-central.

Definition 2.6.2 ((ρ, r)-centrality [82, 50])**.** *Suppose $Z_1, \ldots, Z_k \in \mathbb{R}^d$. A point $x \in \mathbb{R}^d$ is (ρ, r)-central, if and only if for all unit vectors v we have*

$$|\{i \mid \langle Z_i - x, v\rangle \leq r\}| \geq \rho k$$

In other words, with high probability, for any direction v, if we center Z_i's at the point μ, and project them onto the direction v, a large fraction of the projections will fall in a radius of $\mathcal{O}(\sqrt{1/m})$ of the mean. This can be viewed as μ being a median in every direction.

Now note that if we can find some other point $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \leq \mathcal{O}(\sqrt{1/m})$, by setting $m = 1/\alpha^2$ we may obtain the sub-Gaussian rate as desired. Suppose there exists another point $\widetilde{\mu}$ which is also $(0.1, \mathcal{O}(\sqrt{1/m}))$-central, then by choosing v along the direction of $\mu - \widetilde{\mu}$, and applying the pigeonhole principle and the triangle inequality over the projections of Z_i's in this direction, we may conclude that $\|\widetilde{\mu} - \mu\|_2 \leq \mathcal{O}(\sqrt{1/m})$. Another useful observation is that if $\|\widetilde{\mu} - \mu\|_2 \leq \mathcal{O}\left(\sqrt{\frac{1}{m}}\right)$, then $\widetilde{\mu}$ is a central point. Therefore, we arrive at the following conclusion: assuming μ is central; then a point $\widetilde{\mu}$ is central, if and only if it is close to μ. This is the key observation that helps us characterize the points that are close to μ. One can think of this centrality property as the "median" part of the "median of means" approach.

Therefore, answering the following question, gives us an algorithm for estimating the mean with sub-Gaussian rates.

> **Main Problem** Suppose $Z_1, \ldots, Z_k \in \mathbb{R}^d$ are given, and there exists a point $\mu \in \mathbb{R}^d$ which is a central point for $Z_1, \ldots, Z_k$. How can we find a central point, both privately and efficiently?

Suppose we did not care about either privacy or efficiency. Similar to [24], our goal is to start off from an estimate $\widetilde{\mu}$ and move towards a center μ. If we were close to μ, from the discussion above, $\widetilde{\mu}$ would have been a center. Therefore, one way to check whether or not we have reached our destination is by checking if $\widetilde{\mu}$ is central. What if $\widetilde{\mu}$ were far from the center μ — which direction would move us towards μ? If $\widetilde{\mu}$ is far from the center, then $\widetilde{\mu}$ is not central and moreover, there exists some direction v such that

$$|\{i \mid \langle Z_i - \widetilde{\mu}, v\rangle \geq 0.99\|\mu - \widetilde{\mu}\|_2\}| \geq 0.9k\,.$$

This direction v certifies that $\widetilde{\mu}$ is not central. Now a key observation is that if v certifies non-centrality, then v points towards μ, i.e.

$$\left\langle v, \frac{\mu - \widetilde{\mu}}{\|\mu - \widetilde{\mu}\|_2} \right\rangle \geq \Omega(1) .$$

Now, if we move in the direction of v (which certifies non-centrality) with step-size proportional to the distance we are from μ (i.e., $\|\widetilde{\mu} - \mu\|_2$), we can move towards μ at a linear rate. Doing this for a number of steps which is logarithmic in the initial distance gives us a point which is central with respect to Z_i's and has distance $\mathcal{O}(\sqrt{1/m})$ to μ.

In order to employ this approach both efficiently and privately we need to execute the following tasks efficiently and privately:

1. Check whether or not the current estimate $\widetilde{\mu}$ is central.
2. Find an estimate of the distance to the central point or $\|\widetilde{\mu} - \mu\|_2$.
3. Find a direction v that certifies the non-centrality of $\widetilde{\mu}$.

Even setting the concern of privacy aside for the moment, it is not clear how to efficiently check whether or not the current estimate $\widetilde{\mu}$ is central. One approach involves constructing a net over the unit sphere $\mathcal{S}^{d-1}$ and counting the number of Z_i's that fall far for each direction in the unit sphere. However, doing so would take exponential time as the number of points on the net grows exponentially in d. In order to perform this efficiently, [24] uses the standard SDP relaxation of the following quadratic program, which finds the direction that has most of Z_i's at distance at least r from the current estimate $\widetilde{\mu}$. Note that a point x is (ρ, r)-central if and only if the following quadratic program has a low value.

Definition 2.6.3 (quadratic optimization problem [50, 24])**.** *Let* $Z_1, \ldots, Z_k, \widetilde{\mu} \in \mathbb{R}^d$, $r > 0$. *Let* $\mathrm{QUAD}(\widetilde{\mu}, r, Z)$ *be the following quadratic program.*

$$\begin{aligned} \mathrm{QUAD}(\widetilde{\mu}, r, Z) := \max_{v,b} \quad & \sum_{i=1}^{k} b_i^2 \\ s.t. \quad & b_i^2 = b_i \quad \forall i \\ & \sum_{i=1}^{d} v_i^2 = 1 \\ & b_i^2 r \leq \langle Z_i - \widetilde{\mu}, b_i v \rangle \quad \forall i \end{aligned}$$

The following is the standard SDP relaxation of the above quadratic program.

Definition 2.6.4 (SDP Relaxation [50, 24])**.** *Let* $Z_1, \ldots, Z_k, \widetilde{\mu} \in \mathbb{R}^d$, $r > 0$. *Let* $\mathrm{SDP}(\widetilde{\mu}, r, Z)$

be the following semi-definite program.

$$\begin{aligned}\mathrm{SDP}(\widetilde{\mu}, r, Z) := \max_{v,b,B,U,W} \quad & \mathrm{Tr}\,(B) \\ s.t. \quad & \begin{bmatrix} 1 & b^\mathsf{T} & v^\mathsf{T} \\ b & B & W \\ v & W^\mathsf{T} & V \end{bmatrix} \succcurlyeq 0 \\ & B_{ii} = b_i \quad \forall i \\ & \mathrm{Tr}(V) = 1 \quad \forall i \\ & B_{ii} \cdot r \leq \langle Z_i - \widetilde{\mu}, W_i \rangle \quad \forall i\end{aligned}$$

Now based on the above SDP relaxation, we may define a new notion of centrality.

Definition 2.6.5 ((ρ, r)-SDP-centrality [50])**.** *A point is (ρ, r)-SDP-central if and only if the SDP relaxation's value for that point is less than $1 - \rho$.*

The main advantage SDP-centrality has over the previous notion of centrality is that it can be computed efficiently. It can be proven that SDP-centrality has the same good properties that the previous definition of centrality provides:

1. The true mean μ is $(0.999, \mathcal{O}(\sqrt{1/m}))$-SDP-central with high probability (Lemma 2.6.8).
2. Under the assumption that μ is $(0.999, \mathcal{O}(\sqrt{1/m}))$-SDP-central, we have that $\widetilde{\mu}$ is $(0.1, \mathcal{O}(\sqrt{1/m}))$-SDP-central if and only if $\|\mu - \widetilde{\mu}\|_2 \leq \mathcal{O}(\sqrt{1/m})$ (a corollary of the proof of Theorem 2.6.13, Lemma 4 of [24])
3. The solution to the SDP relaxation can certify the non-SDP-centrality of points which are far from a SDP-center. Moreover, there exists a rounding scheme that obtains a direction v from this solution such that $\langle v, \frac{\mu - \widetilde{\mu}}{\|\mu - \widetilde{\mu}\|} \rangle \geq \Omega(1)$.[2]

These properties of the SDP relaxation, together with its computational efficiency, yield a computationally efficient algorithm for estimating the mean up to $\mathcal{O}(\sqrt{1/m})$ error.

Now that we have reviewed [24]'s method, we describe how our *private* algorithm works. One naive approach to ensuring privacy would be to add Laplace noise to the direction after the rounding step. However, it is not clear how to prove sensitivity bounds in this case to guarantee pure differential privacy. In order to privately find the direction, we run the exponential mechanism over the unit ball.[3] If the score function of the exponential mechanism for a given vector y is

$$\textsc{Score}(y) = \sum_{i=1}^{k} \mathbb{1}[\langle Z_i - \widetilde{\mu}, y \rangle \geq r]$$

and we assume that the exponential mechanism always returns the y with the highest score, the exponential mechanism would return a direction y such that $\langle y, \mu - \widetilde{\mu} \rangle \geq \Omega(1)$.

[2] [24] shows that taking the top eigenvector of V in Definition 2.6.4 is one such rounding scheme. In Theorem 2.6.23, we show that the simple rounding scheme of taking v itself provides the same guarantees.

[3] To be precise, in our final algorithm this is not exactly the unit ball. We run the exponential mechanism over a ball of radius $1 - \Theta(1)$. This helps us ensure Definition 2.4.3 or Lipschitzness.

Moreover, the sensitivity of the score function is only 1. However, naïve implementations of the exponential mechanism would not be computationally efficient, running in time exponential in d. To alleviate this issue, we recall the efficient sampling algorithm of [16]. Their results say that if a score function is concave and Lipschitz, and is defined over a bounded and convex domain, then appropriate techniques from the literature on sampling log-concave distributions allow us to efficiently implement the exponential mechanism. Meanwhile, this score function may not even be continuous.

To define a score function $\textsc{Score}(y)$ that has the mentioned properties, we use the SDP relaxation mentioned in Definition 2.6.4, and intersect it with a new condition $v = y$ to define SDP-VAL (see Definition 2.6.7). In order to show that sampling from this SCORE function can be done efficiently, we show that it is concave (Corollary 2.6.29) and Lipschitz (Lemma 2.6.28). Moreover, in order to show that with high probability the exponential mechanism returns a high scoring vector, we show that the score function has sensitivity 1 and that the volume of the high scoring points is large (Corollary 2.6.26, Section 2.6.5). It remains to show that given a feasible solution $X(1, b, v, B, V)$ to the SDP program defined in Definition 2.6.4 with high value, $\langle v, \|\mu - \widetilde{\mu}\|_2 \rangle \geq \Omega(1)$. This guarantees that if the exponential mechanism returns a high-scoring vector, then that vector points towards μ. Note that previous works use a slightly different rounding scheme. In [24], they show that the top eigenvector of the sub-matrix V of $X(1, b, v, B, V)$ has the said property; however we show that a simpler rounding scheme (alluded to above) also gives our desired guarantees (Theorem 2.6.23). For further details on gradient estimation, see Section 2.6.5.

It remains to address the other two components of the algorithm: determining when to stop, and estimating the distance. For the former, the problem would be trivial without privacy: one could simply check the value of the SDP with $r = \Theta(\sqrt{1/m})$. If the value is low, we would be close, otherwise, we would have been far. Our private approach is similar, but we also add noise to the value of the SDP with Laplace mechanism to ensure privacy. We prove the SDP value has sensitivity 1, in order to bound the magnitude of the Laplace noise. Further discussion on how to determine when to stop appears in Section 2.6.3.

Finally, for distance estimation sans privacy, a binary search suffices to select an r: find the maximum r such that for this choice of r, the SDP has value larger than $0.9k$. It can be proven that choosing r this way gives an approximation of the distance to the center [24]. However, it is not clear how to directly make this approach private. Therefore, we instead find some r for which the SDP has value between $0.9k$ and $0.95k$ via a private binary search. Note that if we are far enough from the center, and if we are given a good coarse estimate, both of these values should be feasible, which is ensured by the guarantee of the stopping step and having access to a coarse estimate. Since the value of SDP is a decreasing function in r and we are sufficiently far from the center, it can be proven that if the SDP has value between $0.9k$ and $0.95k$ for r, then r is an approximation of the distance to the center. Full discussion on the private binary search procedure and private distance estimation appears in Section 2.6.4.

Structure. First we define a new quadratic program and a new SDP that we use to define a score function for the exponential mechanism that helps us find the gradient in our algorithm. Then we review some useful lemmas and assumptions from [24] in Section 2.6.1. After that we propose the main algorithm and the state the guarantees it

provides in Section 2.6.2. As discussed above, our algorithm has three parts: estimating whether it should stop or not, estimating the distance, and estimating the direction to the true mean. We present our algorithms, guarantees and proofs, for these three tasks in Section 2.6.3, Section 2.6.4, and Section 2.6.5 respectively. In Section 2.6.6 we show that after sufficient number of rounds of the gradient descent our algorithm obtains a point close to the true mean. Finally, in Section 2.6.7, we explain why our algorithm is robust under constant fraction corruption in the samples.

2.6.1 Preliminaries

Definitions

In order to run the exponential mechanism we need to define a score function over the set of directions. In order to do so, we define the following programs based on the previous definitions (Definition 2.6.3, Definition 2.6.4). These programs are the intersection of the above programs with the new condition, $v = y$. These new programs can be interpreted as: for a given direction v, how many of the Z_i's are at distance at least r from the current estimate $\widetilde{\mu}$, in the direction v?

Definition 2.6.6 (QUAD-VAL)**.** *Let* $Z_1, \ldots, Z_k, \widetilde{\mu}, y \in \mathbb{R}^d$, $r > 0$. *Let* QUAD-VAL$(y; \widetilde{\mu}, r, Z)$ *be the following quadratic program.*

$$\begin{aligned} \text{QUAD-VAL}(y; \widetilde{\mu}, r, Z) := \max_{v,b} \quad & \sum_{i=1}^{k} b_i^2 \\ s.t. \quad & b_i^2 = b_i \quad \forall i \\ & v_i = y_i \quad \forall i \\ & b_i^2 r \leq \langle Z_i - \widetilde{\mu}, b_i v \rangle \quad \forall i \end{aligned}$$

Unlike the previous SDP (Definition 2.6.4), the following SDP is *not* a standard SDP relaxation of the above quadratic program. This is because, in order to be able to run the exponential mechanism efficiently, we need to define our score function over the unit ball as opposed to the unit sphere; hence, $\|y\|_2$ may not be equal to 1. However, we show that SDP-VAL(y) is larger than QUAD-VAL(y), in Lemma 2.8.5.

Definition 2.6.7 (SDP-VAL)**.** *Let* $Z_1, \ldots, Z_k, \widetilde{\mu}, y \in \mathbb{R}^d$, $r > 0$. *Let* SDP-VAL$(y; \widetilde{\mu}, r, Z)$ *be the following semi-definite program.*

$$\begin{aligned} \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) := \max_{v,b,B,W,V} \quad & \text{Tr}\,(B) \\ s.t. \quad & \begin{bmatrix} 1 & b^{\mathsf{T}} & v^{\mathsf{T}} \\ b & B & W \\ v & W^{\mathsf{T}} & V \end{bmatrix} \succcurlyeq 0 \\ & v_i = y_i \quad \forall i \\ & B_{ii} = b_i \quad \forall i \\ & \text{Tr}(V) = 1 \\ & B_{ii} \cdot r \leq \langle Z_i - \widetilde{\mu}, W_i \rangle \quad \forall i \end{aligned}$$

Note that the above semi-definite program can also be viewed as the following convex program, where the maximization is over degree-2 pseudo-distributions in indeterminates $v_1, \dots, v_d, b_1, \dots b_k$. In our proofs, we use the two definitions interchangeably.

$$\begin{aligned} \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) := \max_{\tilde{\mathbf{E}}} \quad & \tilde{\mathbf{E}} \sum_{i=1}^{k} b_i \\ s.t. \quad & \tilde{\mathbf{E}}\, v_i = y_i \quad \forall i \\ & \tilde{\mathbf{E}} \text{ satisfies } b_i = b_i^2 \quad \forall i \\ & \tilde{\mathbf{E}} \text{ satisfies } \|v\|^2 = 1 \\ & \tilde{\mathbf{E}} \text{ satisfies } b_i^2 r \leq \langle Z_i - \widetilde{\mu}, b_i v \rangle \quad \forall i \end{aligned}$$

Useful Lemmas

In this subsection we state some useful lemmas regarding the solutions of the quadratic program and the SDP defined in Definition 2.6.3, and Definition 2.6.4. These lemmas are due to [24]. We state a more general version of them due technical reasons, however the proofs remain similar.

The following lemma states that if we are at the true mean, then with high probability, the semidefinite program defined in Definition 2.6.4, cannot have a high score. In the quadratic program's language, this corresponds to a large fraction of Z_i's being close to the true mean when projected onto any given direction.

Lemma 2.6.8 (Lemma 7 in [24])**.** *Let $M \geq 20$, and $\mathcal{Y} = (Y_1, \dots, Y_k) \in \mathbb{R}^{k \times d}$ be k i.i.d. random vectors with mean μ and covariance Λ and let $\mathcal{S}$ denote the set of feasible solutions of the* SDP$(\mu, r, \mathcal{Y})$ *in Definition 2.6.4. Then, we have for $r \geq 5M(\sqrt{\operatorname{Tr}\Lambda/k} + \sqrt{\|\Lambda\|})$ and $k \geq 8M^2 \log 1/\beta$*

$$\max_{(v,b,B,U,W) \in \mathcal{S}} \operatorname{Tr} B \leq \frac{k}{M},$$

with probability at least $1 - \beta$. More specifically, for $r \geq 10M\sqrt{\|\Lambda\|_2}$ if $k \geq 8M^2 \log 1/\beta$, and $k \geq d$,

$$\max_{(v,b,B,U,W) \in \mathcal{S}} \operatorname{Tr} B \leq \frac{k}{M},$$

with probability at least $1 - \beta$.

Therefore, we may assume that with high probability, the following assumption holds. In [24], they state the above lemma and the following assumption with $M = 20$. In our proofs we require a larger value for M. For the sake of tidiness, we do not set a value for M in the following assumptions and we state them for general M. However, in our proofs we only require the following assumptions to hold for $M = 1000$.

Assumption 2.6.9 (Assumption 2 in [24])**.** *Suppose $r^* = 2\sqrt{k/n} = 2\sqrt{1/m}$. For the bucket means $Z = (Z_1, \dots, Z_k)$ let $\mathcal{S}_r$ denote the set of feasible solutions for* SDP(μ, r, Z)*. Then, for all $r \geq 5Mr^*$,*

$$\max_{v,b,B,W,V \in \mathcal{S}_r} \sum_{i=1}^{k} \operatorname{Tr} B \leq \frac{k}{M}$$

The above assumption is a strengthening of the following assumption.

Assumption 2.6.10 (Assumption 1 in [24]). *For the bucket means, $Z = (Z_1, \ldots, Z_k)$, we have:*

$$\forall v \in \mathbb{R}^d, \|v\|_2 = 1 \implies |\{i : \langle Z_i - \mu, v\rangle \geq 5Mr^*\}| \leq \frac{k}{M}$$

The following lemma is about the feasible solutions to the SDP problem that have high values.

Lemma 2.6.11 (Lemma 3 in [24]). *Let us assume Assumption 2.6.9. Let*

$$X = \begin{bmatrix} 1 & b^\mathsf{T} & v^\mathsf{T} \\ b & B & W \\ v & W^\mathsf{T} & V \end{bmatrix} \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$$

be a PSD matrix. Moreover, suppose that

$$B_{i,i} = b_i, \quad \mathrm{Tr}(V) = 1, \quad \mathrm{Tr}(B) \geq \psi k\,.$$

Then, there is a set of at least $(\psi - 1/M)k$ indices $\mathcal{T}$ such that for all $i \in \mathcal{T}$:

$$\langle Z_i - \mu, W_i\rangle < 5Mr^* B_{i,i}\,,$$

and a set of at least $(\psi - \psi')/(1-\psi')k$ indices $\mathcal{R}$ such that for all $j \in \mathcal{R}$, we have $B_{j,j} \geq \psi'$, where $0 \leq \psi' \leq \psi$.

2.6.2 Main Algorithm and Theorem

In this section we state our main algorithm and theorem. As discussed before, our main algorithm is a gradient descent style algorithm and consists of three parts: HALT-ESTIMATION, DISTANCE-ESTIMATION, GRADIENT-ESTIMATION. We discuss these algorithms and their guarantees in Section 2.6.3, Section 2.6.4, and Section 2.6.5 respectively. The following theorem is the guarantee that our algorithm provides, under the assumption that Assumption 2.6.9 holds, and an initial estimate is given. This theorem together with Lemma 2.6.8, gives us Theorem 2.6.1.

Theorem 2.6.12 (FINE-ESTIMATION). *Let $Z_1, \ldots, Z_k, \widetilde{\mu}_0, \mu \in \mathbb{R}^d$ $r^* = 2\sqrt{k/n} = 2\sqrt{1/m}$. Suppose $\|\widetilde{\mu}_0 - \mu_0\|_2 \leq C_0\sqrt{d}$ where C_0 is a universal constant, $M = 1000$, and Assumption 2.6.9 holds. Then there exists an efficient ε-DP algorithm FINE-ESTIMATION that returns $\widetilde{\mu}^*$ such that if $k \geq \tilde{\mathcal{O}}((d + \log(1/\beta))/\varepsilon)$, then $\|\widetilde{\mu}^* - \mu\|_2 \leq \mathcal{O}(r^*)$, with probability $1-\beta$. Note that $\tilde{O}$ is hiding the following lower order logarithmic factors: $\log d$, and $\log m$.*

Proof. Putting together Theorem 2.6.13 (HALT-ESTIMATION), Theorem 2.6.14 (DISTANCE-ESTIMATION), Theorem 2.6.22 (GRADIENT-ESTIMATION), and Theorem 2.6.39 (gradient descent), and setting $T = \log(dm) \geq \log(d/r^*)$. □

Algorithm 3 FINE-ESTIMATION

Input: Parameters: Data points $X_1, \ldots, X_n \in \mathbb{R}^d$, number of bins k, number of iterations T, initial estimate $\widetilde{\mu}_0$.

Operation: Step-size $\eta = 0.075$
Split the data in k bins $\{\mathcal{B}_i\}_{i=1}^k$, each having m members. For all $1 \leq i \leq k$, let $Z_i = \text{MEAN}(\mathcal{B}_i)$.
For $1 \leq t \leq T$ do:

1. Let $h_t = \text{HALT-ESTIMATION}\,(Z, \widetilde{\mu}_{t-1})$
2. If h_t, then return $\widetilde{\mu}_{t-1}$.
3. Let $d_t = \text{DISTANCE-ESTIMATION}\,(Z, \widetilde{\mu}_{t-1})$
4. Let $g_t = \text{GRADIENT-ESTIMATION}\,(Z, \widetilde{\mu}_{t-1}, d_t)$
5. Let $\widetilde{\mu}_t = \widetilde{\mu}_{t-1} + \eta d_t g_t$

Output: $\widetilde{\mu}_T$

2.6.3 Halt Estimation

If we did not care about privacy, in order to know whether we are close to μ or not, and we had a current estimate $\widetilde{\mu}$, we could just check if SDP has score smaller than $0.9k$ for a small enough distance choice or not. If that was the case, $\widetilde{\mu}$ would have been close enough to μ. In this section we privatize this condition by adding Laplace noise to the value of SDP, and argue that this algorithm provides privacy guarantees and high accuracy with high probability.

Algorithm 4 HALT-ESTIMATION

Input: Parameters: Data points $Z_1, \ldots, Z_k \in \mathbb{R}^d$, current estimate $\widetilde{\mu}$.

Operation: Let $M = 1000, N = 4M(5M+1) + 10M$
Let $H = \mathbb{1}[\text{SDP}(\widetilde{\mu}, Nr^*, Z) + \mathcal{L}(T/\varepsilon) \leq 0.91k]$, where $\mathcal{L}$ is the Laplace distribution.

Output: H

Theorem 2.6.13 (HALT-ESTIMATION)**.** *Let $Z_1, \ldots Z_k, \widetilde{\mu}, \in \mathbb{R}^d$, $r^* = 2\sqrt{k/n}$, $M \geq 1000$, $N = 4M(5M+1) + 10M$. Suppose $k \geq 100T\log(T/\beta)/\varepsilon$, and that Assumption 2.6.9 holds. There exists an ε/T-DP algorithm* HALT-ESTIMATION *that succeeds with probability $1 - \beta/T$. Moreover, in the regime where this algorithms succeeds we have*

$$H = 0 \implies \|\mu - \widetilde{\mu}\|_2 \geq M(5M+1)r^*\,,$$

and

$$H = 1 \implies \|\mu - \widetilde{\mu}\|_2 \leq Nr^* + 5Mr^* .$$

Proof. Since SDP has sensitivity 1 (Corollary 2.6.26), from the guarantees of the Laplace mechanism we know that Algorithm 4 is ε/T-DP and with probability $1 - \beta/T$, it succeeds i.e. $\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) + \mathcal{L}(T/\varepsilon) \in [\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) - T\log(T/\beta)/\varepsilon, \mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) + T\log(T/\beta)/\varepsilon]$. Since $k \geq 100T\log(T/\beta)$, with probability $1 - \beta/T$

$$\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) + \mathcal{L}(T/\varepsilon) \in [\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) - 0.1k, \mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) + 0.1k] .$$

Therefore if we are in the success regime, $H = 0$, means that

$$\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) \geq 0.90k ,$$

and $H = 1$, means that

$$\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) \leq 0.92k .$$

Now by Corollary 2.6.24, if we are in the success regime

$$H = 0 \implies \|\mu - \widetilde{\mu}\|_2 \geq \frac{0.7Nr^* - 5Mr^*}{2} \geq M(5M+1)r^* ,$$

where the last inequality comes from $N = 4M(5M+1) + 10M$. Now suppose that we are in the success regime and $H = 1$. Then $\mathrm{SDP}(\widetilde{\mu}, Nr^*, Z) \leq 0.92k$, which implies that $\mathrm{QUAD}(\widetilde{\mu}, Nr^*, Z) \leq 0.92k$, since SDP is a relaxation of QUAD. Therefore for any $v \in \mathbb{R}^d$, where $\|v\|_2 = 1$ we have that

$$|\{i \mid \langle Z_i - \widetilde{\mu}, v\rangle \leq Nr^*\}| \geq 0.08k .$$

On the other hand, from Assumption 2.6.10 (which is a result of Assumption 2.6.9), we know that for any $v' \in \mathbb{R}^d$, where $\|v'\|_2 = 1$ we have that

$$|\{i \mid \langle Z_i - \mu, v'\rangle \leq 5Mr^*\}| \geq (1 - \frac{1}{M})k .$$

Let $v = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$, $v' = -v$. Then $M \geq 1000$ and $0.08 + (1 - 1/M) > 1$, there exists some index j which is in the intersection of the above sets. Therefore for this j we have that

$$\langle Z_j - \mu + \mu - \widetilde{\mu}, v\rangle \leq Nr^* ,$$

therefore

$$\|\mu - \widetilde{\mu}\|_2 \leq Nr^* + \langle Z_i - \mu, -v\rangle \leq Nr^* + 5Mr^* ,$$

as desired. □

2.6.4 Distance Estimation

Suppose we have a current estimate $\widetilde{\mu}$, and μ is a central point. Furthermore suppose that $\widetilde{\mu}$ and μ are sufficiently far from each other. In this section our goal is to find an estimate of the distance $\|\widetilde{\mu} - \mu\|_2$. If we did not care about privacy, we could look for an approximation of the distance to a central point by finding the maximum r such that SDP obtains value at least $0.9k$ via a binary search up to arbitrary accuracy. However, it's not clear how to privatize this approach. Instead, we look for an r for which SDP obtains a value between $0.91k$ and $0.94k$, via a private binary search (Theorem 2.6.15, Lemma 2.6.16). If r has this property, then we can use the monotonicity of SDP in r to argue that r is an approximation of the distance the central point (Lemma 2.6.17, Lemma 2.6.18).

Algorithm

Algorithm 5 DISTANCE-ESTIMATION

Input: Parameters: Data points $Z_1, \ldots, Z_k \in \mathbb{R}^d$, current estimate $\widetilde{\mu}$.

Operation: Let $r^* = 2\sqrt{k/n}$, $S = \log(1.14C_0\sqrt{d}/r^*)$, $s = 0.92k$, $e = 0.93k$.
Do the following binary search for S rounds over $[0, 1.14C_0\sqrt{d}]$, with r_0 being the middle point

1. Let $u = \mathrm{SDP}(\widetilde{\mu}, r_t, Z) + N_i$, where $N_i \sim \mathcal{L}(TS/\varepsilon)$.
2. Let u is larger than ek move right, if it is less than sk move left, if it is neither, move in an arbitrary direction.

Output: r_T.

Main Theorem

Theorem 2.6.14 (DISTANCE-ESTIMATION)**.** *Let $Z_1, \ldots, Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d$. Suppose $r^* = 2\sqrt{k/n}$, $m = n/k$, $\|\mu - \widetilde{\mu}\|_2 \geq M(5M+1)r^*$, $k \geq 100ST\log(ST/\beta)/\varepsilon$, $M \geq 100$, and that Assumption 2.6.9 holds. Moreover, suppose that $\|\mu - \widetilde{\mu}\|_2 \leq C_0\sqrt{d}$, where C_0 is a universal constant. Then there exists a universal constant C_2 and a ε/T-dp algorithm* DISTANCE-ESTIMATION *using S rounds of binary search that returns $\widehat{d}$ such that*

$$0.99\|\mu - \widetilde{\mu}\|_2 \leq \widehat{d} \leq 1.15\|\mu - \widetilde{\mu}\|_2 ,$$

with probability $1-\beta/T$, where $S = C_2(\log m + \log d)$. We say that DISTANCE-ESTIMATION *has been successful if the above guarantee holds.*

Proof. See Section 2.6.4. □

Private Binary Search

The following theorem allows us to perform private binary search over a decreasing function, with low sensitivity.

Theorem 2.6.15 (Private Binary Search)**.** *Suppose $f\colon [0, D] \times \mathcal{X} \to [0, k]$ is a decreasing function and has sensitivity 1 with respect to its second component. Moreover two values $0 \leq s < e \leq 1$ are given. Let $r_s = \mathrm{argmax}_{f(r) \geq sk - \Delta} r$, $r_e = \arg\min_{f(r) \leq ek + \Delta} r$, then there exists an ε-DP algorithm using S rounds of binary search that returns x such that $x \in [r_e - a, r_s + a]$ with probability $1 - \beta$, where $\Delta = S\log(S/\beta)/\varepsilon$, and $S = \log(D/a)$.*

Proof. Suppose x_i is the binary search parameter at step i, do the following binary search up to S steps over the domain $[0, D]$ and in the end return x_S.

1. Let $f_i = f(x_i)$, and $f_i' = f_i + N_i$, where $N_i \sim \mathcal{L}(S/\varepsilon)$.
2. If f_i' is larger than ek move right, if it is less than sk move left, if it is neither, move in an arbitrary direction.

Now we need to prove privacy and accuracy guarantees for this algorithm.

Privacy The algorithm is ε-DP by basic composition. Note that each step is ε/S-DP.

Accuracy Note that for every $t \geq 0$

$$\mathbb{P}\left[|N_i| \geq tS/\varepsilon\right] = \exp(-t).$$

and by union bound

$$\mathbb{P}\left[\max |N_i| \geq tS/\varepsilon\right] \leq S\exp(-t).$$

Let $t = -\log(\beta/T)$, since $\Delta = S\log(S/\beta)/\varepsilon$,

$$\mathbb{P}\left[\max |N_i| \geq \Delta\right] \leq \beta.$$

Now with probability $1-\beta$, $\forall i, N_i \leq \Delta$. Suppose we are in this setting where all of the N_i's are smaller than Δ. We have three cases.

1. If $x_i \in (r_s, D]$, then $f_i \in [0, sk-\Delta)$, and $f_i' \in [-\Delta, sk)$, therefore, the binary search algorithm will choose left.
2. If $x_i \in [r_e, r_s]$, then $f_i \in [sk-\Delta, ek+\Delta]$, and $f_i' \in [sk-2\Delta, ek+\Delta]$, therefore the binary search algorithm might choose right or left (arbitrary direction).
3. If $x_i \in [0, r_e)$, then $f_i \in (ek+\Delta, k]$, and $f_i' \in (ek, k+\Delta]$, the binary search algorithm will choose right.

Therefore this algorithm is performing Binary Interval Search for the interval $[r_e, r_s]$ with accuracy parameter a. Therefore, by Lemma 2.9.3, with probability $1-\beta$, the algorithm returns a point x, such that $x \in [r_e - a, r_s + a]$. □

Proof of Theorem 2.6.14

In this section we prove Theorem 2.6.14. This theorem is proven by putting together Lemma 2.6.16, Lemma 2.6.17, and Lemma 2.6.18.

Lemma 2.6.16 (SDP Binary Search)**.** *Let $Z_1, \ldots, Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d$. Suppose $r^* = 2\sqrt{k/n}$, $m = n/k$, $\|\mu - \widetilde{\mu}\|_2 \leq C_0\sqrt{d}$, $\|\mu - \widetilde{\mu}\|_2 \geq M(5M+1)r^*$, $k \geq 100ST\log(ST/\beta)/\varepsilon$, $M \geq 100$, $C_1 = 1.14C_0$, and that Assumption 2.6.9 holds. Let $f(r) := \mathrm{SDP}(\widetilde{\mu}, r, Z)$, where $f : [0, C_1\sqrt{d}] \to \mathbb{R}$. Let $\rho_s = \mathrm{argmax}_{f(r)\geq 0.91k} r, \rho_e = \arg\min_{f(r)\leq 0.94k} r$,. Then there exists a universal constant C_2 and a ε/T-dp algorithm using S rounds of binary search that returns $\widehat{r}$ such that $\widehat{r} \in [\rho_e - r^*, \rho_s + r^*]$ with probability $1-\beta/T$, where $S = C_2(\log m + \log d)$.*

Proof. We want to apply the algorithm from Theorem 2.6.15. In order to do so, we must check that SDP is decreasing in r (Lemma 2.8.1) and has sensitivity 1 (Corollary 2.6.26). Furthermore, note that $\{r : f(r) \geq 0.91k, 0 \leq r \leq C_1\sqrt{d}\}$, and $\{r : f(r) \leq 0.94k, 0 \leq r \leq C_1\sqrt{d}\}$ are non-empty sets, which is true because of Corollary 2.6.19, and Corollary 2.6.32. Now we may apply Theorem 2.6.15 with

$$s \leftarrow 0.92k, \quad e \leftarrow 0.93k, \quad D \leftarrow C_1\sqrt{d}, \quad a \leftarrow r^* = 2\sqrt{k/n}\,.$$

Therefore, $S = \log(D/a) = \log\sqrt{\frac{C_1 dn}{4k}}$. Therefore, there exists a universal constant C_2 such that

$$S \leq C_2(\log d + \log m)\,.$$

It remains to bound Δ by $0.01k$, which is true by the assumption.

$$\begin{aligned}\Delta &= ST\log(ST/\beta)/\varepsilon \\ &\leq 0.01k\end{aligned}$$

Therefore, using the algorithm from Theorem 2.6.15 we constructed an algorithm that returns $\widehat{r} \in [\rho_e + r^*, \rho_s - r^*]$ with probability $1 - \beta/T$, using $\mathcal{O}(\log m + \log d)$ steps of the binary search, as desired. □

Lemma 2.6.17 (lower bound for distance estimate)**.** *Under the assumptions of Lemma 2.6.16, $\widehat{r} \geq \rho_e - r^*$ implies $\widehat{r} \geq (1 - M)\|\widetilde{\mu} - \mu\|_2$.*

Proof. Let Δ be the unit direction $\widetilde{\mu} - \mu$, then by Assumption 2.6.10 (which is implied by Assumption 2.6.9),

$$|\{i : \langle Z_i - \mu, \Delta\rangle \geq 5Mr^*\}| \leq k/M\,.$$

Therefore if we take the direction $-\Delta$, we have for $(1 - 1/M)k$ of the points that

$$(Z_i - \widetilde{\mu}, -\Delta) = \langle \widetilde{\mu} - \mu + \mu - Z_i, \Delta\rangle \geq \|\widetilde{\mu} - \mu\|_2 - 5Mr^*$$

Therefore $\mathrm{SDP}(\widetilde{\mu}, \|\widetilde{\mu} - \mu\|_2 - 5Mr^*, Z) \geq (1 - 1/M)k$, since SDP is the standard relaxation of QUAD. Now note that by Lemma 2.6.16, $\text{SDP-VAL}(\widetilde{\mu}, \widehat{r}, Z) \leq 0.94k$; therefore,

$$\mathrm{SDP}(\widetilde{\mu}, \|\widetilde{\mu} - \mu\|_2 - 5Mr^*, Z) \geq (1 - 1/M)k \geq 0.95k > 0.94k \geq \mathrm{SDP}(\widetilde{\mu}, \rho_e, Z)\,.$$

Now by monotonicity of SDP in the second parameter, we conclude that $\rho_e \geq \|\widetilde{\mu} - \mu\|_2 - 5Mr^*$. Therefore

$$\widehat{r} \geq \rho_e - r^* \geq \|\widetilde{\mu} - \mu\|_2 - (5M + 1)r^* \geq (1 - \frac{1}{M})\|\widetilde{\mu} - \mu\|_2\,,$$

as desired, where the final inequality comes from $\|\widetilde{\mu} - \mu\|_2 \geq M(5M + 1)r^*$. □

Lemma 2.6.18 (upper bound for distance estimation)**.** *Under the same assumptions as in Lemma 2.6.16, $\widehat{r} \leq \rho_s + r^*$ implies $\widehat{r} \leq 1.15\|\widetilde{\mu} - \mu\|_2$.*

Proof. The proof is similar to the proof in Lemma 5 from [24]. We show that $\mathrm{SDP}(\widetilde{\mu}, 1.14\|\widetilde{\mu} - \mu\|_2, Z) \leq 0.9k$. Suppose the contrary, then there exists some solution (b, v, B, W, V) that

achieves value larger than $0.9k$. Let $\mathcal{R}, \mathcal{T}$ be the sets from Lemma 2.6.11 where $\psi = 0.9$, $\psi' = (1 + 1/M)/1.14$. Now notice that

$$\frac{|\mathcal{T}| + |\mathcal{R}|}{k} = \frac{0.9 - \frac{1+1/M}{1.14}}{1 - \frac{1+1/M}{1.14}} + (0.9 - 1/M) > 1\,,$$

for $M \geq 86$, therefore there exists some $j \in \mathcal{R} \cap \mathcal{T}$. For this j,

$$\begin{aligned}(1 + 1/M)/1.14 \cdot (1.14\|\widetilde{\mu} - \mu\|_2) &\leq B_{j,j} 1.14\|\widetilde{\mu} - \mu\|_2 \\ &\leq \langle Z_j - x, W_j\rangle \\ &= \langle Z_i - \mu, W_j\rangle + \langle \mu - \widetilde{\mu}, W_j\rangle \\ &< B_{j,j} 5Mr^* + \|\widetilde{\mu} - \mu\|_2\,.\end{aligned}$$

Where the last step comes from $\|W_j\|_2 \leq 1$. This is easy to see in the pseudo-expectation language,

$$\begin{aligned}\|\tilde{\mathbf{E}}[b_j v]\|_2^2 &= \sum_\ell \tilde{\mathbf{E}}[b_j v_\ell]^2 \\ &\leq \sum_\ell \tilde{\mathbf{E}}[b_j^2]\,\tilde{\mathbf{E}}[v_\ell^2] && \text{(Cauchy-Schwarz)} \\ &\leq \tilde{\mathbf{E}}[\|v\|_2^2] && (\tilde{\mathbf{E}}[b_j^2] \leq 1) \\ &= 1\,.\end{aligned}$$

By rearranging the above inequality we obtain

$$B_{j,j} > 1/M \cdot \|\widetilde{\mu} - \mu\|_2 (5Mr^*)^{-1} > 1\,,$$

which is a contradiction. Therefore,

$$\mathrm{SDP}(\widetilde{\mu}, 1.14\|\widetilde{\mu} - \mu\|_2, Z) \leq 0.9k < 0.91k \leq \mathrm{SDP}(\widetilde{\mu}, \rho_s, Z)\,.$$

Now by monotonicity of SDP in the second parameter, we conclude that $\rho_s \leq 1.14\|\widetilde{\mu} - \mu\|_2$. Therefore

$$\widehat{r} \leq \rho_s + r^* \leq 1.14\|\widetilde{\mu} - \mu\|_2 + r^* \leq 1.15\|\widetilde{\mu} - \mu\|_2\,,$$

as desired, where the final inequality comes from $\|\widetilde{\mu} - \mu\|_2 \geq 100(5 \times 100 + 1)r^* \geq 100r^*$. □

Corollary 2.6.19 (existence of r with SDP less than $0.9k$)**.** *In the proof of Lemma 2.6.18, we have proven that under Assumption 2.6.9, and the assumption that $M \geq 86$,*

$$\mathrm{SDP}(\widetilde{\mu}, 1.14\|\mu - \widetilde{\mu}\|_2, Z) \leq 0.9k\,.$$

2.6.5 Gradient Estimation

Suppose we have a current estimate $\widetilde{\mu}$ and μ is a central point Moreover, $\widetilde{\mu}$ and μ are sufficiently far from each other. Our goal is to privately find a direction y such that $\langle y, (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2\rangle \geq \Omega(1)$. In order to do this we run the exponential mechanism over a ball of radius $1 - \Theta(1)$ in $\mathbb{R}^d$ with score function SDP-VAL. We prove that SDP-VAL has

low sensitivity (Section 2.6.5), is Lipschiz (Section 2.6.5), and is concave (Section 2.6.5). These properties ensure that we can sample from the corresponding distribution efficiently and that our algorithm is private. In other to prove the accuracy of the exponential mechanism (Section 2.6.5), we prove that there exists a large volume of points that have high scores (Section 2.6.5). If the exponential mechanism is accurate, it outputs a direction y from the ball that has high a high score. In Section 2.6.5, we show that if y has a high score it satisfies $|\langle y, (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2\rangle| \geq \Omega(1)$. We wrap up the guarantees of the exponential mechanism with SDP-VAL in Theorem 2.6.20.

Now in order to decide between y and $-y$, we run the exponential mechanism over these two candidates with score function being the number $Z_i - \widetilde{\mu}$'s that have positive inner product with each candidate. We show that at least one of the two candidates must have a high score and therefore with high probability it is outputted. Moreover the higher scoring candidate satisfies $\langle y, (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2\rangle \geq \Omega(1)$. See Theorem 2.6.21 for the theorem statement, and Section 2.6.5 for its proof.

Finally, we put together the guarantees of Theorem 2.6.20, and Theorem 2.6.21 to obtain Theorem 2.6.22.

Algorithm

Algorithm 6 GRADIENT-ESTIMATION

Input: Parameters: Data points $Z_1, \ldots, Z_k \in \mathbb{R}^d$, current estimate $\widetilde{\mu} \in \mathbb{R}^d$, estimate of the distance d

Operation:
1. Let $M = 1000$, $r = d/1.2$, $\zeta = 1/(2M)$.
2. Run the exponential mechanism over the ball of radius $1 - \zeta$ with score function $\textsc{Score}(y) = \text{SDP-VAL}(y; \widetilde{\mu}, r, Z)$ to obtain y_0.
3. Run the exponential mechanism with score function $\textsc{Score}(y) = \sum_{i=1}^{k} \mathbb{1}[\langle Z_i - \widetilde{\mu}, y\rangle > 0]$ over $\{y_0, -y_0\}$ to obtain $\widehat{y}$.

Output: $\widehat{y}$

Main Theorems

The following theorem guarantees that the exponential mechanism in the second line of Algorithm 6, is private and can be run efficiently and outputs a direction y_0 such that $|\langle y_0, \Delta\rangle| \geq \Omega 1$ with high probability.

Theorem 2.6.20 (SDP-VAL exponential mechanism)**.** *Let $Z_1 \ldots Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d$, $r^* = 2\sqrt{k/n}$, $M \geq 1000$, $d \geq 0$, $r = d/1.2$, $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$ Suppose Assumption 2.6.9 holds, $1.15\|\mu - \widetilde{\mu}\|_2 \geq d \geq 0.9\|\mu - \widetilde{\mu}\|_2$, $\|\mu - \widetilde{\mu}\|_2 \geq M(5Mr^*)$, $\zeta = 1/(2M)$. Then there exists a constant C'_M depending only on M such that the exponential mechanism in*

Algorithm 6 with score function $f(y) = \text{SDP-VAL}(y; \widetilde{\mu}, r, Z)$ *and privacy budget* ε/T*, over the ball of radius* $1-\zeta$ *in* $\mathbb{R}^d$*, returns* y_0 *such that if* $k \geq C'_M T(d + \log(T/\beta))/\varepsilon$*, then*

$$|\langle y_0, \Delta \rangle| \geq 0.1\,,$$

with probability $1 - T/\beta$*. Moreover there exists an algorithm taking time* $\text{poly}(n, d, 1/\varepsilon)$ *for sampling from the exponential mechanism.*

Proof. Combination of Lemma 2.6.25 (sensitivity, privacy), Theorem 2.6.35 (accuracy of the exponential mechanism), Theorem 2.6.23 (accuracy guarantee of SDP-VAL), and Lemma 2.6.36 (runtime of the exponential mechanism). See Section 2.6.5 to Section 2.6.5 for the full proof. □

From the previous theorem we know that with high probability $|\langle y_0, \widetilde{\mu} - \mu \rangle|\, \Omega(1)$ The following theorem guarantees that the third line of the algorithm is private and that it has accuracy. That is, conditioned on the success of the previous step, with high probability it outputs a vector $\widehat{y}$ from $\{y_0, -y_0\}$ such that $\langle \widehat{y}, (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2 \rangle \geq \Omega(1)$.

Theorem 2.6.21 (exponential mechanism for the two directions)**.** *Let* $Z_1, \ldots Z_k, \widetilde{\mu}, \mu, y_0 \in \mathbb{R}^d$*,* $r^* = 2\sqrt{k/n}$*,* $\|\mu - \widetilde{\mu}\|_2 \geq M(5Mr^*)$*,* $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$ *and* $M \geq 1000$*. Suppose* $|\langle y_0, \Delta \rangle| \geq 0.1$*, and that Assumption 2.6.9 holds. Then the exponential mechanism with score function* $\textsc{Score}(y) = \sum_{i=1}^k \mathbb{1}[\langle Z_i - \widetilde{\mu}, y \rangle > 0]$*, over* $\{y_0, -y_0\}$*, and privacy budget* ε/T*, as in Algorithm 6, returns* $\widehat{y}$ *such that if* $k \geq 2MT(\log 2 + \log T/\beta)/\varepsilon$*, then*

$$\langle \widehat{y}, \Delta \rangle \geq 0.1\,,$$

with probability $1 - \beta/T$*.*

Proof. See Section 2.6.5. □

Putting together Theorem 2.6.20, and Theorem 2.6.21 gives us the following algorithm, which guarantees that Algorithm 6 is private, has high utility with high probability, and that it can be run efficiently.

Theorem 2.6.22 (GRADIENT-ESTIMATION)**.** *Let* $Z_1, \ldots Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d$*,* $d \geq 0$*,* $r^* = 2\sqrt{k/n}$*,* $M \geq 1000$*,* $d \geq 0$*,* $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$*. Suppose Assumption 2.6.9 holds. Then there exists a constant* C''_M *depending only on* M *and* ε/T*-dp algorithm* GRADIENT-ESTIMATION *that if* $0.99\|\mu - \widetilde{\mu}\|_2 \leq d \leq 1.15\|\mu - \widetilde{\mu}\|_2$*,* $\|\mu - \widetilde{\mu}\|_2 \geq M(5Mr^*)$*, and* $k \geq C''_M T(d + \log T/\beta)/\varepsilon$*, returns* $\widehat{y}$ *such that*

$$\langle \widehat{y}, \Delta \rangle > 0.1\,,$$

with probability $1 - \beta/T$*. Moreover this algorithm has runtime* $\text{poly}(n, d, 1/\varepsilon)$*. We say that* GRADIENT-ESTIMATION *has been successful if the above guarantee holds.*

In the rest of this section we prove Theorem 2.6.20, and Theorem 2.6.21.

SoS Analysis

Theorem 2.6.23 (accuracy guarantee of SDP-VAL)**.** *Let* $Z_1 \dots Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d$, $r^* = 2\sqrt{k/n}$, *and* $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$. *Suppose Assumption 2.6.9 holds and* $M \geq 1000$, $d \geq 0.99\|\mu - \widetilde{\mu}\|_2$, *and* $\|\mu - \widetilde{\mu}\|_2 \geq M(5Mr^*)$. *Moreover, suppose* SDP-VAL$(y; \widetilde{\mu}, d/1.2, Z) = \psi k$, *where* $\psi \geq 0.9$. *Then*

$$|\langle y, \Delta \rangle| \geq 0.1\,.$$

Proof. Let $r = d/1.2$, note that since $d \geq 0.99\|\mu - \widetilde{\mu}\|_2$, $r \geq 0.8\|\mu - \widetilde{\mu}\|_2$. Suppose $\tilde{\mathbf{E}}$ is the solution to SDP-VAL$(y; \widetilde{\mu}, r, Z)$ as in Definition 2.6.7. Therefore,

$$\begin{aligned} y &= \tilde{\mathbf{E}}\,[v] \\ \tilde{\mathbf{E}}\left[b_i^2\right] &= \tilde{\mathbf{E}}\,[b_i] && (2.5) \\ \sum_i \tilde{\mathbf{E}}\left[v_i^2\right] &\leq 1 && (2.6) \\ \tilde{\mathbf{E}}\left[b_i^2\right] \cdot r &\leq \langle Z_i - \widetilde{\mu}, \tilde{\mathbf{E}}\,[b_i v]\rangle\,. \end{aligned}$$

Moreover, y has score ψk, i.e.

$$\sum_i \tilde{\mathbf{E}}\left[b_i^2\right] = \psi k. \tag{2.7}$$

Let $\mathcal{T}$ be the set that its existence is guaranteed from Lemma 2.6.11. Then $|\mathcal{T}| \geq (\psi - 1/M)k$, and

$$\forall i \in \mathcal{T} : \tilde{\mathbf{E}}[\langle Z_i - \mu, b_i v\rangle] < 5Mr^*\,\tilde{\mathbf{E}}[b_i^2]\,. \tag{2.8}$$

Also note that

$$(2\psi - 1 - 1/M)k \leq \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i^2] \leq \psi k\,. \tag{2.9}$$

Therefore we may write

$$\begin{aligned} r\psi k &= r \cdot \sum_i \tilde{\mathbf{E}}[b_i^2] \\ &= r \cdot \left[\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i^2] + \sum_{i\notin\mathcal{T}} \tilde{\mathbf{E}}[b_i^2]\right] \\ &\leq \sum_{i\in\mathcal{T}} \langle Z_i - \widetilde{\mu}, \tilde{\mathbf{E}}[b_i v]\rangle + (1 - \psi + 1/M)kr \\ &= \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle Z_i - \widetilde{\mu}, v\rangle] + (1 - \psi + 1/M)kr \\ &= \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle Z_i - \mu, v\rangle] + \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle \mu - \widetilde{\mu}, v - \tilde{\mathbf{E}}[v]\rangle] + \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i]\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle + (1 - \psi + 1/M)kr\,. \end{aligned}$$

Now we need to bound each term separately.

1. $\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle Z_i - \mu, v\rangle]$.

$$\begin{aligned} \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle Z_i - \mu, v\rangle] &\leq 5Mr^* \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}\left[b_i^2\right] && \text{(Equation (2.8))} \\ &\leq 5Mkr^* \end{aligned}$$

2. $\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle \mu - \widetilde{\mu}, v - \tilde{\mathbf{E}}[v]\rangle]$.

$$\begin{aligned}
\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i\langle u - \widetilde{\mu}, v - \tilde{\mathbf{E}}[v]\rangle] &\le \langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[(\sum_{i\in\mathcal{T}} b_i)(v - \tilde{\mathbf{E}}[v])]\rangle \\
&\le \langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[\sum_{i\in\mathcal{T}} (b_i - \tilde{\mathbf{E}}[b_i])(v - \tilde{\mathbf{E}}[v])]\rangle \\
&\le \|\mu - \widetilde{\mu}\|_2 \cdot \|\tilde{\mathbf{E}}[\sum_{i\in\mathcal{T}} (b_i - \tilde{\mathbf{E}}[b_i])(v - \tilde{\mathbf{E}}[v])]\|_2 \\
&\le \|\mu - \widetilde{\mu}\|_2 \cdot \sqrt{\tilde{\mathbf{E}}[(\sum_{i\in\mathcal{T}} (b_i - \tilde{\mathbf{E}}[b_i]))^2] \cdot \tilde{\mathbf{E}}[\|v - \tilde{\mathbf{E}}[v]\|_2^2]} \\
&\qquad\qquad \text{(Lemma 2.9.4)}
\end{aligned}$$

Now we need to bound the two last terms.

$$\begin{aligned}
\tilde{\mathbf{E}}\left[\left(\sum_{i\in\mathcal{T}} (b_i - \tilde{\mathbf{E}}[b_i])\right)^2\right] &= \tilde{\mathbf{E}}\left[\left(\sum_{i\in\mathcal{T}} b_i - \tilde{\mathbf{E}}[\sum_{i\in\mathcal{T}} b_i]\right)^2\right] \\
&= \tilde{\mathbf{E}}[(\sum_{i\in\mathcal{T}} b_i)^2] - \tilde{\mathbf{E}}[\sum_{i\in\mathcal{T}} b_i]^2 && \text{(Lemma 2.9.5)} \\
&\le \tilde{\mathbf{E}}[k \cdot \sum_{i\in\mathcal{T}} b_i^2] - \left(\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i]\right)^2 && \text{(Cauchy-Schwarz)} \\
&= k \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i^2] - \left(\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i^2]\right)^2 && \text{(Equation (2.5))} \\
&\le \psi k^2 - (2\psi - 1 - /1M)^2 k^2 && \text{(Equation (2.9))} \\
&= k^2(\psi - (2\psi - 1 - 1/M)^2)
\end{aligned}$$

$$\begin{aligned}
\tilde{\mathbf{E}}[\|v - \tilde{\mathbf{E}}[v]\|_2^2] &= \tilde{\mathbf{E}}[\|v\|_2^2] - \|\tilde{\mathbf{E}}[v]\|_2^2 && \text{(Lemma 2.9.5)} \\
&\le \tilde{\mathbf{E}}[\|v\|_2^2] \\
&\le 1 && \text{(Equation (2.6))}
\end{aligned}$$

Therefore,

$$\sum_i \tilde{\mathbf{E}}[b_i\langle u - \widetilde{\mu}, v - \tilde{\mathbf{E}}[v]\rangle] \le k\sqrt{\psi - (2\psi - 1 - 1/M)^2} \cdot \|\mu - \widetilde{\mu}\|_2 .$$

3. $\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i]\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle$.

$$\begin{aligned}
\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i] &= \sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i^2] && \text{(Equation (2.5))} \\
&\le \psi k && \text{(Equation (2.9))}
\end{aligned}$$

Therefore,

$$\sum_{i\in\mathcal{T}} \tilde{\mathbf{E}}[b_i]\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle \le \psi k |\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle| .$$

Putting all parts together we obtain

$$(2\psi - 1 - 1/M)rk \leq \sum_i \tilde{\mathbf{E}}[b_i\langle Z_i - \mu, v\rangle] + \sum_i \tilde{\mathbf{E}}[b_i\langle \mu - \widetilde{\mu}, v - \tilde{\mathbf{E}}[v]\rangle] + \sum_i \tilde{\mathbf{E}}[b_i]\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle$$
$$\leq 5Mkr^* + k\sqrt{\psi - (2\psi - 1 - 1/M)^2} \cdot \|\mu - \widetilde{\mu}\|_2 + \psi k|\langle \mu - \widetilde{\mu}, \tilde{\mathbf{E}}[v]\rangle| \,.$$

Therefore if $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$, $\tilde{d} = \|\mu - \widetilde{\mu}\|_2$.

$$\frac{1}{\psi} \cdot [(2\psi - 1 - 1/M) \cdot \frac{r}{\tilde{d}} - \frac{5Mr^*}{\tilde{d}} - \sqrt{\psi - (2\psi - 1 - 1/M)^2}] \leq |\langle \Delta, y\rangle| \,.$$

Now since $0.9 \leq \psi \leq 1, M \geq 1000, r \geq 0.8\tilde{d}, \tilde{d} \geq M(5Mr^*)$, we get that

$$|\langle \Delta, y\rangle| \geq 0.65(2\psi - 1 - 1/M) - \sqrt{\psi - (2\psi - 1 - 1/M)^2} + 0.15(2\psi - 1 - 1/M) - \frac{5Mr^*}{\tilde{d}}$$
$$\geq 0.15(2 \times 0.9 - 1 - 1/1000) - 1/1000$$
$$> 0.1 \,.$$

□

The following corollary is useful in Section 2.6.3.

Corollary 2.6.24 (SDP far)**.** *Let* $Z_1, \ldots, Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d, r^* = 2\sqrt{k/n}$*. Suppose Assumption 2.6.9 holds and* $M \geq 1000$*. Suppose* $\mathrm{SDP}(\widetilde{\mu}, r, Z) \geq 0.9k$*. Then*

$$\|\mu - \widetilde{\mu}\|_2 \geq \frac{0.7r - 5Mr^*}{2} \,.$$

Proof. From the proof of Theorem 2.6.23, we have that

$$2\|\mu - \widetilde{\mu}\|_2 + 5kr^* \geq rk(2 \times 0.9 - 1 - 1/M) \,.$$

Rearranging the terms and noting that $M \leq 1000$, completes the proof. □

Analysis of the SDP-VAL Exponential Mechanism

In order to analyze the exponential mechanism we need to analyze the following properties of SDP-VAL.

1. Sensitivity: for privacy analysis of the exponential mechanism.
2. Lipschitzness: for runtime analysis of the exponential mechanism.
3. Concavity: for runtime analysis of the exponential mechanism.
4. Volume of the set of *optimal* points: for accuracy analysis of the exponential mechanism.

In the following subsections we analyze these properties.

Sensitivity

Lemma 2.6.25 (SDP-VAL sensitivity)**.** $\Delta_{\text{SDP-VAL}} \leq 1$.

Proof. We want to prove that for every two neighboring datasets Z and Z'

$$\max_{y,\widetilde{\mu},r} \max_{Z,Z'} |\text{SDP-VAL}(y;\widetilde{\mu},rZ) - \text{SDP-VAL}(y;\widetilde{\mu},r,Z')| \leq 1\,.$$

If we prove that for every $y, \widetilde{\mu}, r$ and for every neighboring datasets Z, Z'

$$\text{SDP-VAL}(y;\widetilde{\mu},r,Z) - 1 \leq \text{SDP-VAL}(y;\widetilde{\mu},r,Z'),$$

by symmetry, our desired statement will be implied. Suppose that the optimal solution for $(y;\widetilde{\mu},r,Z)$ is $(1,b,B,v,V,W)$. Without loss of generality, suppose that Z and Z' differ in the first index. Now consider $(1,b',B',v,V,W')$, where it is equal to $(1,b,B,v,V,W)$ everywhere, except at $b_1, B_{1,\cdot}, B_{\cdot,1}, W_{1,\cdot}$, where it is equal to zero, or in other words, in the second column and row of the PSD matrix. Then $(1,b',B',v,V,W')$ will still satisfy the linear inequalities. Moreover the corresponding matrix for $(1,b',B',v,V,W')$, will be positive semidefinite, by Lemma 2.9.1. Now note that $\operatorname{Tr} B' \geq \operatorname{Tr} B - 1$, and therefore,

$$\text{SDP-VAL}(y;\widetilde{\mu},r,Z) - 1 \leq \text{SDP-VAL}(y;\widetilde{\mu},r,Z')\,,$$

which completes the proof. □

The following corollary is useful in the analysis in Section 2.6.4.

Corollary 2.6.26 (SDP sensitivity)**.** $\Delta_{\text{SDP}} \leq 1$.

Proof. Same as Lemma 2.6.25. □

Lipschitzness

Lemma 2.6.27. *Let $Z_1,\ldots,Z_k,\widetilde{\mu} \in \mathbb{R}^d, r \in \mathbb{R}, w > 0$. Then there exists $\alpha \in \mathbb{R}^d; \beta \in \mathbb{R}^k; \gamma \in \mathbb{R}; \lambda \in \mathbb{R}^k_{\geq 0}$, and a degree-2 SoS polynomial $\sigma \in \mathbb{R}[b_1\ldots,b_k,v_1,\ldots,v_k]_{\leq 2}$ such that the following polynomial identity holds in variables $b_1,\ldots,b_k,v_1,\ldots,v_d$.*

$$\begin{aligned}\text{SDP-VAL}(y;\widetilde{\mu},r,Z) + \omega - \sum_{i=1}^k b_i^2 = \sum_{i=1}^d \alpha_i(y_i - v_i) + \sum_{i=1}^k \beta_i(b_i - b_i^2) + \gamma(1 - \|v\|_2^2) \\ + \sum_{i=1}^k \lambda_i(\langle Z_i - \widetilde{\mu}, bv\rangle - b_i^2 r) + \sigma(b,v)\end{aligned}$$

This lemma is a direct result of duality of pseudo-distributions and sos proofs (e.g. see [14]). Note that the constraints on b and v in the optimization problem imply boundedness (Lemma 2.8.4) and therefore our problem satisfies the duality conditions.

Now we use this lemma to prove the following theorem.

Lemma 2.6.28 (SDP-VAL Lipschitzness)**.** *$f(y) := \text{SDP-VAL}(y; \widetilde{\mu}, r, Z)$ is $L = \sqrt{d}k/\zeta$-Lipschitz over the ball of radius $(1-\zeta)$ in $\mathbb{R}^d$, where $0 < \zeta \leq 1$.*

Proof. Let $\omega > 0$, then by Lemma 2.6.27, there exists $\alpha \in \mathbb{R}^d; \beta \in \mathbb{R}^k; \gamma \in \mathbb{R}; \lambda \in \mathbb{R}^k_{\geq 0}$ and a degree-2 SoS polynomial σ such that

$$\begin{aligned}\text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + \omega - \sum_{i=1}^{k} b_i^2 = \sum_{i=1}^{d} \alpha_i(y_i - v_i) + \sum_{i=1}^{k} \beta_i(b_i - b_i^2) + \gamma(1 - \|v\|_2^2) \\ + \sum_{i=1}^{k} \lambda_i(\langle Z_i - \widetilde{\mu}, bv\rangle - b_i^2 r) + \sigma(b, v)\,.\end{aligned}$$

First we prove that α_is are bounded. Note that the above equality, is a polynomial identity, therefore we may pick different values for b and v, and the equality should still hold. For $1 \leq j \leq k$ let $b = 0, v = y + te_j$ where t is chosen such that $\|y + te_j\|_2 = 1$. Such t exists because $\|y\|_2 < 1 - \zeta$. Moreover $t > \zeta$ due to the triangle inequality. Now if we plug b and v into the above equality we obtain

$$\begin{aligned}\text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + \omega &= -\alpha_j t + \sigma(0, y + te_j) && (\sigma \geq 0, \gamma_i \geq 0) \\ &\geq -\alpha_j t\,.\end{aligned}$$

Therefore, for every $1 \leq j \leq k$,

$$\alpha_j \geq -\frac{k + \omega}{\zeta}\,.$$

Similarly by letting $b = 0, v = y - e_j t$, we may obtain $\alpha_j \leq (k + \omega)/\zeta$, therefore $|\alpha_j| \leq (k + \omega)/\zeta$, and

$$\|\alpha\|_2 \leq \sqrt{d}(k + \omega)/\zeta\,.$$

Now that we have bounded α, we can move on to proving Lipschitzness. Let

$$\begin{aligned}g(\alpha, \beta, \gamma, \lambda, \sigma; y; \widetilde{\mu}, r, Z) = \sum_{i=1}^{k} b_i^2 + \sum_{i=1}^{d} \alpha_i(y_i - v_i) + \sum_{i=1}^{k} \beta_i(b_i - b_i^2) + \gamma(1 - \|v\|_2^2) \\ + \sum_{i=1}^{k} \lambda_i(\langle Z_i - \widetilde{\mu}, bv\rangle - b_i^2 r) + \sigma(b, v)\,.\end{aligned}$$

Note that the co-domain of g is the set of degree-2 polynomials in (b, v). Suppose $y' = y + \delta u$, where $\|u\|_2 = 1$. In order to prove Lipschitzness, we prove that there exists $\alpha', \beta', \gamma', \lambda', \sigma'$ such that $g(\alpha', \beta', \gamma', \lambda', \sigma'; y'; \widetilde{\mu}, r, Z)$ is a degree-0 polynomial (i.e. it is a number in $\mathbb{R}$), and is at most $\text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + L\delta + h(\omega)$, where $\lim_{\omega \to 0} h(\omega) = 0$. If we do so, by pseudo-expectation and SoS duality, we can conclude that for every $\omega > 0$, $\text{SDP-VAL}(y'; \widetilde{\mu}, r, Z) < \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + L\delta + h(\omega)$. Taking $\lim_{\omega \to 0}$ from both sides gives us $\text{SDP-VAL}(y'; \widetilde{\mu}, r, Z) \leq \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + L\delta$. By symmetry, this implies that $|\text{SDP-VAL}(y'; \widetilde{\mu}, r, Z) - \text{SDP-VAL}(y; \widetilde{\mu}, r, Z)| \leq L\delta$ and that f is L-Lipschitz in y. It remains to construct a dual solution that satisfies the described properties. To do so,

take $(\alpha', \beta', \gamma', \lambda', \sigma') = (\alpha, \beta, \gamma, \lambda, \sigma)$. Then

$$\begin{aligned} g(\alpha', \beta', \gamma', \lambda', \sigma'; y'; \widetilde{\mu}, r, Z) &= g(\alpha, \beta, \gamma, \lambda, \sigma; y; \widetilde{\mu}, r, Z) + \delta \sum_{i=1}^{d} \alpha_i u_i \\ &\leq \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + \omega + \delta \|\alpha\|_2 \\ &\leq \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + \omega + \delta \left(\sqrt{d} k/\zeta + \sqrt{d}\omega/\zeta \right) \\ &= \text{SDP-VAL}(y; \widetilde{\mu}, r, Z) + \delta\sqrt{d}k/\zeta + \omega(\delta\sqrt{d}/\zeta + 1)\,. \end{aligned}$$

as desired. Therefore f is $L = \sqrt{d}k/\zeta$-Lipschitz. □

Concavity

Corollary 2.6.29 (SDP-VAL concavity)**.** *Let $Z_1, \ldots, Z_k, \widetilde{\mu} \in \mathbb{R}^d$, $r \geq 0$. Let $f(y) :=$ SDP-VAL$(y; \widetilde{\mu}, r, Z)$. Then f is concave.*

Proof. Direct result of Lemma 2.9.2. □

Volume

In this section we prove the following lemma.

Lemma 2.6.30 (SDP-VAL volume ratio)**.** *Let $Z_1, \ldots, Z_k, \widetilde{\mu}, \mu \in \mathbb{R}^d, r^* = 2\sqrt{k/n}, 1/(2M) \geq \zeta \geq 0$. Suppose $\|\widetilde{\mu} - \mu\|_2 \geq M(5M)r^*$, $M \geq 1000$, $d \leq 1.15\|\mu - \widetilde{\mu}\|_2$,and that Assumption 2.6.9 holds. Then there exists a set $\mathcal{H}^*$ inside $\mathbb{B}^d_{1-\zeta}$, the ball of radius $1-\zeta$ in $\mathbb{R}^d$, and a constant C_M depending only on M such that*

$$\log \frac{\text{vol}(\mathbb{B}^d_{1-\zeta})}{\text{vol}(\mathcal{H}^*)} \leq C_M d\,,$$

and

$$\inf_{y \in \mathcal{H}^*} \text{SDP-VAL}(y; \widetilde{\mu}, d/1.2, Z) \geq (1 - \frac{1}{M})k\,.$$

Proof. Putting together Lemma 2.6.31, Lemma 2.6.33, and Corollary 2.6.34, and setting $\gamma = \theta = 1/M$, and noting that $d \leq 1.15\|\mu - \widetilde{\mu}\|_2$ implies

$$\frac{d}{1.2} \leq 0.96\|\mu - \widetilde{\mu}\|_2 \leq (1 - 1/M)(1 - 2/M)\|\mu - \widetilde{\mu}\|_2\,.$$

Also note that from Corollary 2.6.34, $C_M \geq \mathcal{O}(\log M)$ should suffice. □

The following lemma tells us that there's a set of good points. Later in this section we show that this set has a large volume inside the ball of radius $1 - \zeta$.

Lemma 2.6.31 (goodness of the set)**.** *Let $Z_1 \ldots, Z_k, \widetilde{\mu} \in \mathbb{R}^d$, $r^* = 2\sqrt{k/n}$. Assume Assumption 2.6.9 holds. Let $\tilde{d} = \|\mu - \widetilde{\mu}\|_2$, $\Delta = (\mu - \widetilde{\mu})/\|\mu - \widetilde{\mu}\|_2$. Then for any $0 \leq \gamma, \theta \leq 1$, if $\langle u, \Delta \rangle \geq 1 - \gamma$, $\|u\|_2 = 1$, $\tilde{d} \geq M(5M)r^*$, $(1-\theta)(1-\gamma-1/M)\tilde{d} \geq r$,*

$$\forall \alpha \in [1-\theta, 1] : \text{SDP-VAL}(\alpha u; \widetilde{\mu}, r, Z) \geq (1 - \frac{1}{M})k\,.$$

Proof. By Assumption 2.6.10 (which is a result of Assumption 2.6.9) we know that for the direction u there exists a set S_{-u} such that $|S_{-u}| \geq k(1-1/M)$ and

$$\forall i \in S_{-u} : \langle Z_i - \mu, -u \rangle \leq 5Mr^* \,.$$

Now for every $i \in S_{-u}$, we have that

$$\begin{aligned}
\langle Z_i - \widetilde{\mu}, u \rangle &= \langle Z_i - \mu, u \rangle + \langle \mu - \widetilde{\mu}, u \rangle \\
&\geq -5Mr^* + \|\mu - \widetilde{\mu}\|_2 \langle u, \Delta \rangle \\
&\geq (1-\gamma)\tilde{d} - 5Mr^* \\
&\geq (1-\gamma-1/M)\tilde{d} \,.
\end{aligned}$$

Therefore for at least $k(1-1/M)$ indices

$$\langle Z_i - \widetilde{\mu}, \alpha u \rangle \geq (1-\theta)(1-\gamma-1/M)\tilde{d} \geq r \,,$$

which completes the proof by Lemma 2.8.5. □

The following corollary is useful in the analysis in Section 2.6.4.

Corollary 2.6.32. *Let $Z_1, \ldots, Z_k, \widetilde{\mu} \in \mathbb{R}^d$, and $r^* = 2\sqrt{k/n}$. Assume Assumption 2.6.9, holds, $\|\mu - \widetilde{\mu}\|_2 \geq M(5M)r^*$. Then*

$$\mathrm{SDP}(\widetilde{\mu}, (1 - \frac{1}{M})\|\mu - \widetilde{\mu}\|_2, Z) \geq (1 - \frac{1}{M})k \,.$$

The following lemma sates that the set defined in Lemma 2.6.31, has a high volume.

Lemma 2.6.33 (size of the good set)**.** *Suppose $\Delta \in \mathbb{R}^d$ such that $\|\Delta\|_2 = 1$, $0 \leq \gamma \leq 1$ and $0 \leq \zeta \leq \theta \leq 1$. Let*

$$S_{\theta,\zeta,\gamma,\Delta} = \left\{ \alpha u \;\middle|\; \langle u, \Delta \rangle \geq 1-\gamma, 1-\theta \leq \alpha \leq 1-\zeta, \|u\|_2 = 1, u \in \mathbb{R}^d \right\} .$$

Let $\mathbb{B}_r^d$ denote the unit ball of radius r in $\mathbb{R}^d$. Then

$$\frac{\mathrm{vol}(S_{\theta,\zeta,\gamma,\Delta})}{\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)} \geq \left(1 - \left(\frac{1-\theta}{1-\zeta}\right)^d\right) \cdot \frac{1}{2} \left(\frac{\sqrt{1-(1-\gamma)^2}}{2}\right)^{d-1} .$$

Proof. In order to prove this lemma we lower bound $\mathrm{vol}(S_{\theta,\zeta,\gamma,\Delta})$, / $\mathrm{vol}(S_{1,\zeta,\gamma,\Delta})$, $\mathrm{vol}(S_{1,\zeta,\gamma,\Delta})$, / $\mathrm{vol}(S_{1,0,\gamma,\Delta})$, $\mathrm{vol}(S_{1,0,\gamma,\Delta})$/ $\mathrm{vol}(\mathbb{B}_1^d)$, $\mathrm{vol}(\mathbb{B}_1^d)$/ $\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)$

Note that $S_{1,\theta,\gamma,\Delta}$ is just a scaling of the set $S_{1,\zeta,\gamma,\Delta}$ with respect to the origin with ratio $(1-\theta)/(1-\zeta)$. Therefore

$$\frac{\mathrm{vol}(S_{1,\theta,\gamma,\Delta})}{\mathrm{vol}(S_{1,\zeta,\gamma,\Delta})} = \left(\frac{1-\theta}{1-\zeta}\right)^d ,$$

and

$$\frac{\mathrm{vol}(S_{\theta,\zeta,\gamma,\Delta})}{\mathrm{vol}(S_{1,\zeta,\gamma,\Delta})} = 1 - \left(\frac{1-\theta}{1-\zeta}\right)^d .$$

Similarly,

$$\frac{\mathrm{vol}(S_{1,\zeta,\gamma,\Delta})}{\mathrm{vol}(S_{1,0,\gamma,\Delta})} = (1-\zeta)^d \,.$$

To bound $\mathrm{vol}(S_{1,0,\gamma,\Delta})/\mathrm{vol}(\mathbb{B}_1^d)$, note that this ratio is equal to the ratio of the surface area of the corresponding cap of $S_{1,0,\gamma,\Delta}$ to the surface area of the unit sphere, therefore by Fact 2.9.8,

$$\frac{\mathrm{vol}(S_{1,0,\gamma,\Delta})}{\mathrm{vol}(\mathbb{B}_1^d)} \geq \frac{1}{2}\left(\frac{\sqrt{1-(1-\gamma)^2}}{2}\right)^{d-1} \,.$$

Now it remains to lower bound $\mathrm{vol}(\mathbb{B}_1^d)/\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)$. Note that $\mathbb{B}_1^d$ is a scaling of the set $\mathbb{B}_{1-\zeta}^d$ with respect to the origin with ratio $1/(1-\zeta)$, therefore,

$$\frac{\mathrm{vol}(\mathbb{B}_1^d)}{\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)} = (1-\zeta)^{-d} \,.$$

Multiplying the bounds gives us the desired inequality. □

Corollary 2.6.34 (size of the good set)**.** *Under the same assumptions as Lemma 2.6.33,*

$$\frac{\mathrm{vol}(S_{\theta,\zeta,\gamma,\Delta})}{\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)} \geq \frac{\theta-\zeta}{2}\cdot(\gamma^{1/2}/2)^{d-1} \,,$$

More specifically, for every θ,ζ,γ, there exists some $C_{\theta,\zeta,\gamma}\geq 0$, such that

$$\log\frac{\mathrm{vol}(\mathbb{B}_{1-\zeta}^d)}{\mathrm{vol}(S_{\theta,\zeta,\gamma,\Delta})} \leq C_{\theta,\zeta,\gamma}d \,.$$

Accuracy Analysis of the Exponential Mechanism for SDP-VAL

Theorem 2.6.35 (accuracy guarantee of the exponential mechanism for SDP-VAL)**.** *Let $Z_1\ldots Z_k,\mu,\widetilde{\mu}\in\mathbb{R}^d$, $r^*=2\sqrt{k/n}$, $1/(2M)\geq\zeta\geq 0$. Suppose $M\geq 1000$, $d\leq 1.15\|\mu-\widetilde{\mu}\|_2$, $r=d/1.2$, and that Assumption 2.6.9 holds. Let $\mathcal{M}_E$ be the exponential mechanism when instantiated with the score function $f(y)=\text{SDP-VAL}(y;\widetilde{\mu},r,Z)$ and privacy budget ε/T, over the ball of radius $1-\zeta$ in $\mathbb{R}^d$, as in Algorithm 6. Let y_0 denote the vector returned by the exponential mechanism, then there exists a constant C_M depending only on M such that*

$$\text{SDP-VAL}(y)\geq(1-\frac{1}{M})k-\frac{2T}{\varepsilon}\left(C_M d+\log(T/\beta)\right) \,,$$

with probability $1-\beta/T$. More specifically, if $k\geq 2TM(C_M d+\log(T/\beta))/\varepsilon$, then

$$\text{SDP-VAL}(y)\geq(1-\frac{2}{M})k \,,$$

with probability $1-\beta/T$.

Proof. This is a direct result of the volume based exponential mechanism, and our arguments about the volume of the good set and sensitivity i.e. Theorem 2.3.1, Lemma 2.6.30 and Lemma 2.6.25. □

Runtime Analysis of the Exponential Mechanism for SDP-VAL

Lemma 2.6.36 (runtime of the exponential mechanism)**.** *There exists a sampling scheme for the exponential mechanism for* SDP-VAL *in Algorithm 6 that takes time* $\mathrm{poly}(n, d, 1/\varepsilon)$.

This lemma is a direct application of Theorem 2.3.2's sampling algorithm. Note that we have shown that SDP-VAL has Lipschitzness (Lemma 2.6.28), concavity (Corollary 2.6.29), and low sensitivity (Lemma 2.6.25), therefore we are allowed to use the mentioned result.

Proof of Theorem 2.6.21

First let's prove the following lemma.

Lemma 2.6.37. *Let* $Z_1 \ldots, Z_k, \mu, \widetilde{\mu}, y_0 \in \mathbb{R}^d$, $\tilde{d} = \|\mu - \widetilde{\mu}\|_2, \Delta = |\mu - \widetilde{\mu}| / \tilde{d}$. *Suppose* $|\langle y, \Delta\rangle| > 0.1$, $\langle y, \Delta\rangle > 0$, $\|\mu - \widetilde{\mu}\|_2 \geq M(5Mr^*)$, $M \geq 1000$, *and that Assumption 2.6.10 holds. Then*

$$|\{i \mid \langle Z_i - \widetilde{\mu}, y\rangle > 0\}| \geq (1 - 1/M)k\,.$$

Proof. The proof is similar to [24]. By Assumption 2.6.10, we know that there exists a set G, such that $|G| \geq (1 - 1/M)k$ and

$$\forall \in G \;:\; \langle Z_i - \mu, -y\rangle \leq 5Mr^*\,.$$

Therefore, for every $i \in G$,

$$\begin{aligned}\langle Z_i - \widetilde{\mu}, y\rangle &= \langle Z_i - \mu, y\rangle + \langle \mu - \widetilde{\mu}, y\rangle \\ &\geq -5Mr^* + \langle \Delta, y\rangle \|\mu - \widetilde{\mu}\|_2 \\ &\geq -5Mr^* + 0.1M(5Mr^*) \\ &> 0\,,\end{aligned}$$

which completes the proof. □

Now we use this lemma to prove Theorem 2.6.21.

Proof of Theorem 2.6.21. Since either $\langle y_0, \Delta\rangle > 0$, or $\langle -y_0, \Delta\rangle > 0$, by Lemma 2.6.37, and the guarantee of the exponential mechanism, the exponential mechanism returns $\widehat{y}$ such that

$$\textsc{Score}(\widehat{y}) \geq (1 - 1/M)k - 2T/\varepsilon(\log 2 + \log T/\beta)\,,$$

with probability $1 - \beta/T$. Since $k \geq 2MT(\log 2 + \log T/\beta)/\varepsilon$, the exponential mechanism returns $\widehat{y}$ such that

$$\textsc{Score}(\widehat{y}) \geq (1 - 2/M)k\,,$$

with probability $1 - \beta/T$. If $\textsc{Score}(\widehat{y}) \geq (1 - 2/M)k$, then by Lemma 2.6.37, and the assumption that $M \geq 1000$, $\textsc{Score}(\widehat{y}) \geq (1 - 2/M)k$ implies that $\langle \widehat{y}, \Delta\rangle > 0$, as desired. □

2.6.6 Gradient Descent

We show that if the algorithm has not stopped in a round of the algorithm and that every step of the algorithm i.e. HALT-ESTIMATION, DISTANCE-ESTIMATION, and GRADIENT-ESTIMATION, have been successful, then we move closer to μ (the central point) linearly (Lemma 2.6.38).

Finally we conclude that if all of the steps of the algorithm are successful then either we halt and return a point with optimal distance or we get closer to the central point μ, linearly (Theorem 2.6.39)

Lemma 2.6.38 (Gradient Descent Step)**.** *Let $Z_1, \ldots, Z_k, \widetilde{\mu}_{t-1}, \mu \in \mathbb{R}^d$, $\Delta_{t-1} = (\mu - \widetilde{\mu}_{t-1})/\|\mu - \widetilde{\mu}_{t-1}\|_2$, the step-size $\eta = 0.075$. Suppose in the t-th round inside the for loop in Algorithm 3, we have not halted, and all of* HALT-ESTIMATION, DISTANCE-ESTIMATION, *and* GRADIENT-ESTIMATION, *have been successful, i.e. the guarantees in Theorem 2.6.13, Theorem 2.6.14, and Theorem 2.6.22 are satisfied and*

$$0.99\|\widetilde{\mu}_{t-1} - \mu\|_2 \leq d_t \leq 1.15\|\widetilde{\mu}_{t-1} - \mu\|_2, \quad \langle g_t, \Delta_{t-1}\rangle > 0.1, \quad \|g_t\|_2 \leq 1\,.$$

Then

$$\|\widetilde{\mu}_t - \mu\|_2 \leq 0.999\|\widetilde{\mu}_{t-1} - \mu\|_2\,.$$

Proof. The proof is similar to [24]. Let $\tilde{d}_{t-1} = \|\widetilde{\mu}_{t-1} - \mu\|_2$. Then

$$\begin{aligned}
\|\widetilde{\mu}_t - \mu\|_2^2 &= \|\widetilde{\mu}_t - \widetilde{\mu}_{t-1} + \widetilde{\mu}_{t-1} - \mu\|_2^2 \\
&= \|g_t d_t \eta\|_2^2 + \tilde{d}_{t-1}^2 + 2\langle g_2 d_t \eta, \widetilde{\mu}_{t-1} - \mu\rangle \\
&\leq \tilde{d}_{t-1}^2(1 + 1.15^2\eta^2) + 2\langle g_2 d_t \eta, -\Delta_{t-1}\rangle \\
&\leq \tilde{d}_{t-1}^2(1 + 1.15^2\eta^2 - 2 \times 0.99 \times 0.1 \times \eta) \\
&= 0.993\|\widetilde{\mu}_{t-1} - \mu\|_2^2\,.
\end{aligned}$$

Taking square root gives us the final result. □

Theorem 2.6.39 (Gradient Descent)**.** *Let $Z_1, \ldots Z_k, \mu_0 \in \mathbb{R}^d$ $M \geq 1000$, $r^* = 2\sqrt{k/n}$. Suppose the guarantees in Theorem 2.6.13, Theorem 2.6.14, and Theorem 2.6.22, have been successful in every round t. Then two outcomes are possible, either Algorithm 3 doesn't halt and returns $\widetilde{\mu}^*$ such that*

$$\|\widetilde{\mu}^* - \mu\|_2 \leq (0.999)^T\|\mu_0 - \mu\|_2\,,$$

or Algorithm 3, halts and returns $\widetilde{\mu}^*$ such that

$$\|\widetilde{\mu}^* - \mu\|_2 \leq (4M(5M+1) + 15M)r^*\,.$$

Proof. Putting together Lemma 2.6.38, and Theorem 2.6.13. □

2.6.7 Robustness

Remark 2.6.40 (Robustness of Fine Estimation)**.** *Note that* SDP*, and* SDP-VAL*, and our quadratic programs all have sensitivity* 1 *by Lemma 2.6.25, and Corollary 2.6.26. There exists a universal constant C_0 (say $C_0 \le 1/1000$), such that our our arguments would still hold under $C_0 k$-fraction corruption of the bucketed means $(Z_1, \dots, Z_k)$. Now suppose that $m = C_0/\eta_0$, then under $\eta \le \eta_0$ corruption of the samples $(X_1, \dots X_n)$,* SDP*,* SDP-VAL*, and our quadratic programs are perturbed by at most*

$$\eta n \le \eta_0 m k \le C_0 k \,.$$

Therefore, the guarantees of the algorithm will still hold. Moreover, in this setting, by Theorem 2.6.12, we obtain an estimate $\widetilde{\mu}^*$, where $\|\mu - \widetilde{\mu}^*\|_2 \le \mathcal{O}(\frac{1}{\sqrt{m}}) = \mathcal{O}(\sqrt{\eta_0})$, as desired.

2.7 Lower Bounds

In this section, we prove lower bounds for mean estimation. While lower bounds for these settings were known previously [48, 15, 20, 64], in contrast to prior work, we focus explicitly on the failure probability β. We first establish a general purpose lower bound framework, and then apply it to heavy-tailed and Gaussian distributions. Technically, we follow the standard packing lower bound technique.

Theorem 2.7.1. *Let $\mathcal{P} = \{P_1, \dots, P_m\}$ be a set of distributions, and P_O be a distribution such that for every $P_i \in \mathcal{P}$, $\|P_i - P_O\|_{\mathrm{TV}} \le \gamma$. Let $\mathcal{G} = \{G_1, \dots, G_m\}$ be a collection of disjoint subsets of some set $\mathcal{Y}$. If there is an ε-DP algorithm M such that $\mathbb{P}_{X \sim P_i^n}[M(X) \in G_i] \ge \beta$ for all $i \in [m]$, then*

$$n \ge \Omega\left(\frac{\log m + \log(1/\beta)}{\gamma\,(e^{2\varepsilon} - 1)}\right) .$$

Note that for the usual regime $\varepsilon \le 1$, we can replace the $e^{2\varepsilon} - 1$ in the denominator with ε.

Proof. Assume there exists some mechanism M that satisfies the theorem's assumptions. By the definition of ε-DP and group privacy, we know that for any two datasets D, D', and any $i \in [m]$ we have

$$\mathbb{P}\left[M(D) \in G_i\right] \le e^{\varepsilon \#[D \neq D']}\, \mathbb{P}\left[M(D') \in G_i\right] \,, \tag{2.10}$$

where $\#\left[D \neq D'\right]$ denotes the Hamming distance between D and D'. Note that for the time being, D and D' are fixed datasets (as opposed to random variables), and $0 \le \#\left[D \neq D'\right] \le n$.

Basic facts about coupling imply that, for any two distributions $P_i, P_j \in \mathcal{P}$, there exists a coupling $\omega_{i,j}^*$ of P_i^n and P_j^n such that if $(X_i, X_j) \sim \omega_{i,j}^*$, we have

$$\#\left[X_i \neq X_j\right] \sim \mathrm{Binomial}\left(n, \|P_i - P_j\|_{\mathrm{TV}}\right) . \tag{2.11}$$

Note that by the triangle inequality (through P_O), we have that $\|P_i - P_j\|_{\mathrm{TV}} \le 2\gamma$.

By taking expected value of both sides of Equation (2.10), we get

$$\mathbf{E}_{(X_i,X_j)\sim\omega^*}\left[\mathbb{P}\left[M(X_j)\in G_j\right]\right] \le \mathbf{E}_{(X_i,X_j)\sim\omega^*}\left[e^{\varepsilon\#[X_i\neq X_j]}\cdot\mathbb{P}\left[M(X_i)\in G_j\right]\right], \tag{2.12}$$

for any two P_i, P_j.

For every $t\in[m]$, define $Z_t^{(i)} = \mathbb{P}_M\left[M(X_i)\in G_t\right]$. Note that the probability here is only over M, $Z_t^{(i)}$ remains a random variable which depends on the sampling $X_i\sim P_i^n$. Stated differently, to evaluate the random variable $Z_t^{(i)}$, one first samples $X_i\sim P_i^n$, and then computes $\mathbb{P}_M\left[M(X_i)\in G_t\right]$ for the realized dataset X_i. Since the G_t's are disjoint, the law of total probability implies $\sum_t Z_t^{(i)}\le 1$ for any $i\in[m]$. By the accuracy guarantee of the mechanism M, we know that

$$\mathbf{E}_{X_i\sim P_i^n}\left[Z_i^{(i)}\right] = \mathop{\mathbb{P}}_{X_i\sim P_i^n, M}\left[M\left(X_i\right)\in G_t\right] \ge 1-\beta,$$

therefore $\mathbf{E}_{X_i\sim P_i^n}\left[\sum_{t\neq i} Z_t^{(i)}\right]\le\beta$. Since $Z_t^{(i)}\le 1$,

$$\sum_{t\neq i}\mathbf{E}\left[\left(Z_t^{(i)}\right)^2\right] \le \sum_{t\neq i}\mathbf{E}\left[Z_t^{(i)}\right]\le\beta.$$

Therefore by the pigeonhole principle, for every i, there exists some ℓ such that

$$\mathbf{E}_{X_i\sim P_i^n}\left[\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]^2\right] = \mathbf{E}_{X_i\sim P_i^n}\left[\left(Z_\ell^{(i)}\right)^2\right]\le\frac{\beta}{m-1}. \tag{2.13}$$

Now we take $j=\ell$ in Equation (2.12). We get

$$\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[\mathbb{P}\left[M\left(X_\ell\right)\in G_\ell\right]\right] \le \mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[e^{\varepsilon\#[X_i\neq X_\ell}\cdot\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]\right].$$

By the accuracy guarantee of M we know that we can lower bound the left-hand side by $1-\beta$. We upper bound the right-hand side by Hölder's inequality.

$$\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[e^{\varepsilon\#[X_i\neq X_\ell]}\cdot\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]\right] \le \mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[e^{2\varepsilon\#[X_i\neq X_\ell]}\right]^{1/2}\cdot\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]^2\right]^{1/2}.$$

For the second term, Equation (2.13) implies that

$$\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]^2\right] = \mathbf{E}_{X_i\sim P_i^n}\left[\mathbb{P}\left[M\left(X_i\right)\in G_\ell\right]^2\right]\le\frac{\beta}{m-1}.$$

It remains to bound the first term. By Equation (2.11), we know that $\#\left[X_i\neq X_\ell\right]\sim$ Binomial $\left(n, \|P_i-P_\ell\|_{\mathrm{TV}}\right)$, and thus, this term is just the moment-generating function of the Binomial distribution, evaluated at the point 2ε. Since the MGF of Binomial(n,p) is $(1-p+pe^t)^n$,

$$\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[e^{2\varepsilon\#[X_i\neq X_\ell]}\right]^{1/2} = \left(1-p+pe^{2\varepsilon}\right)^{n/2}\le\exp\left(\frac{n}{2}\cdot\log\left(1+2\gamma\left(e^{2\varepsilon}-1\right)\right)\right).$$

Note that $x\ge\log(1+x)$, therefore

$$\mathbf{E}_{(X_i,X_\ell)\sim\omega^*}\left[e^{2\varepsilon\#[X_i\neq X_\ell]}\right]^{1/2}\le\exp\left(n\gamma\cdot\left(e^{2\varepsilon}-1\right)\right).$$

Combining these derivations,

$$1-\beta \le \exp\left(n\gamma\left(e^{2\varepsilon}-1\right)\right)\cdot\sqrt{\frac{\beta}{m-1}}\,.$$

Therefore,

$$n \ge \Omega\left(\frac{\log\left(\frac{(1-\beta)^2}{\beta}\right)+\log m}{\gamma\left(e^{2\varepsilon}-1\right)}\right).$$

Since $\beta < 1$, we can conclude

$$n \ge \Omega\left(\frac{\log m+\log\frac{1}{\beta}}{\gamma\left(e^{2\varepsilon}-1\right)}\right).$$

□

Now, we apply this theorem to derive lower bounds for mean estimation for distributions with bounded moments.

Theorem 2.7.2. *Suppose there exists an ε-DP algorithm M such that for every distribution D on $\mathbb{R}^d$ such that $\|\mathbf{E}_{X\sim D} X\| \le R$ and $\mathbf{E}_{X\sim D}[|\langle X-\mathbf{E}_{X\sim D}[X], v\rangle|^k] \le 1$ for some $k \ge 2$ and every unit vector $v \in \mathbb{R}^d$, given $X_1,\dots,X_n \sim D$, with probability $1-\beta$ the algorithm M outputs $\hat{\mu}$ such that $\|\hat{\mu}-\mathbf{E}_{X\sim D} X\| \le \alpha$ where α is smaller than some absolute constant. Then we have that*

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

Proof. The first term is folklore from the non-private setting and obtains even for estimating the mean of a Gaussian.

To obtain the last term, consider a maximum 1-packing of the ball with ℓ_2-radius R. Standard results imply that this will be of size $R^{\Omega(d)}$. Consider a point mass distribution at each of these points. Each distribution clearly satisfies the mean and bounded moment condition in the theorem statement. We apply Theorem 2.7.1 with $\mathcal{P}$ being this set of distributions, noting that $\gamma \le 1$ trivially. Combining the 1-packing property with our accuracy requirement that $\alpha \le 1/2$, we get that the set of sets $\mathcal{G}$ are indeed disjoint. This gives the third term.

To obtain the second term, we employ a construction of Barber and Duchi [15]. We consider a set of distributions parameterized by unit vectors $v \in \mathbb{R}^d$. The distribution parameterized by v has mass $1-25\alpha^{\frac{k}{k-1}}$ at the origin and mass $25\alpha^{\frac{k}{k-1}}$ on the point $\frac{1}{6}\alpha^{-\frac{1}{k-1}}v$.

We show that each of these distributions satisfies the conditions of the theorem statement. The mean of this distribution is $25\alpha^{\frac{k}{k-1}}\frac{1}{6}\alpha^{-\frac{1}{k-1}}v = \frac{25}{6}\alpha v$, which has ℓ_2 norm $\le R$ for R larger than an absolute constant. We check the bounded moment condition for an

arbitrary unit vector u:

$$\begin{aligned}\mathbf{E}[|\langle X - \mathbf{E}[X], u\rangle|^k] &\leq \mathbf{E}[|\langle X - \mathbf{E}[X], v\rangle|^k] \\ &\leq (1 - 25\alpha^{\frac{k}{k-1}}) \left(\frac{25}{6}\alpha\right)^k + 25\alpha^{\frac{k}{k-1}} \left(\frac{1}{6}\alpha^{-\frac{1}{k-1}}\right)^k \\ &\leq 1\,.\end{aligned}$$

The last inequality uses the bounds on α and k. Therefore, each such distribution satisfies the theorem conditions.

Now, we look to apply Theorem 2.7.1. Let $\mathcal{P}$ be the set of distributions parameterized by a maximum set of points V on the unit sphere such that for each $v, v' \in V$, we have that $\|v - v'\| \geq 1/2$. Standard results imply that this set will be of size at least $2^{\Omega(d)}$. All distributions are within distance $\gamma = 25\alpha^{\frac{k}{k-1}}$ from the point mass distribution on the origin. Finally, since for each pair v, v' we have $\|v - v'\| \geq 1/2$, this implies that the means of any two distributions are at distance $\geq \frac{25}{12}\alpha$ from each other, and the balls of radius α around each mean are disjoint. Letting these be the G_i's, we conclude the theorem statement using Theorem 2.7.1. □

We can similarly derive a lower bound for mean estimation of Gaussian distributions.

Theorem 2.7.3. *Suppose there exists an ε-DP algorithm M such that for every Gaussian distribution $\mathcal{N}(\mu, I)$ on $\mathbb{R}^d$ such that $\|\mu\| \leq R$, given $X_1, \ldots, X_n \sim D$, with probability $1 - \beta$ the algorithm M outputs $\hat{\mu}$ such that $\|\hat{\mu} - \mathbf{E}_{X\sim D} X\| \leq \alpha$ where α is smaller than some absolute constant. Then we have that*

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

Proof. The first term is again standard from the non-private literature, and the third term can be obtained akin to in the proof of Theorem 2.7.2 (though we consider a Gaussian with each center, rather than a point mass).

The second term is also similar to that of Theorem 2.7.2. The construction will involve the set of Gaussians with mean $4\alpha v$ where v is a unit vector. These are Gaussian with bounded mean, thus satisfying the conditions of the theorem statement.

As before, we try to apply Theorem 2.7.1. We let $\mathcal{P}$ be the set of $2^{\Omega(d)}$ distributions associated with a packing of unit vectors V such that for each $v, v' \in V$, we have that $\|v - v'\| \geq 1/2$. Each pair of means is at distance $\geq 2\alpha$, and thus the balls of radius α around each mean are disjoint. The proof is concluded by noting that all these distributions are at total variation distance $\gamma = O(\alpha)$ from $N(0, I)$, due to the standard fact that ℓ_2 distance between means and total variation distance are equivalent (up to constant factors) for identity-covariance Gaussians when the means are at distance smaller than an absolute constant. □

2.8 Deferred Proofs

Lemma 2.8.1 (SDP decreasing)**.** SDP *(Definition 2.6.4) is decreasing in r.*

Proof. Suppose $r > r'$. Our goal is to prove that $\text{SDP}(\widetilde{\mu}, r, Z) \leq \text{SDP}(\widetilde{\mu}, r', Z)$. Suppose X is the solution that maximizes $\text{SDP}(\widetilde{\mu}, r, Z)$. Then X is a feasible solution to $\text{SDP}(\widetilde{\mu}, r', Z)$. Therefore $\text{SDP}(\widetilde{\mu}, r, Z) \leq \text{SDP}(\widetilde{\mu}, r', Z)$. □

Lemma 2.8.2. $\{b^2 = b\} \vdash_2 b^2 \leq 1$.

Proof.

$$(1-b)^2 + 2(b-b^2) = 1 - b^2 \implies b^2 = b \vdash_2 b^2 \leq 1\,.$$

□

Lemma 2.8.3 (boundedness of the solutions of Definition 2.5.13)**.** *The constraints of* COARSE-SDP*, as defined in Definition 2.5.13, imply that* $\|(b,v)\|_2^2 \leq (2R + R/1000)^2 + k$.

Proof. From the constraints we have that $\|v\|_2^2 \leq (2R + R/1000)^2$. Furthermore by Lemma 2.8.2, we know that $\|b\|_2^2 \leq k$. Therefore $\|(b,v)\|_2^2 \leq (2R + R/1000)^2 + k$. □

Lemma 2.8.4 (boundedness of the solutions of Definition 2.6.7)**.** *The constraints of* SDP-VAL*, as defined in (Definition 2.6.3) imply that* $\|(b,v)\|_2^2 \leq k + 1$.

Proof. From the constraints we have that $\|v\|_2^2 = 1$. Furthermore by Lemma 2.8.2, we know that $\|b\|_2^2 \leq k$. Therefore $\|(b,v)\|_2^2 \leq k + 1$ □

Lemma 2.8.5 (QUAD-VAL$(y) \leq$ SDP-VAL(y))**.** *For every* $Z_1, \ldots Z_k, \widetilde{\mu} \in \mathbb{R}^d$*, and* $r \geq 0$*,*

$$\text{QUAD-VAL}(y; \widetilde{\mu}, r, Z) \leq \text{SDP-VAL}(y; \widetilde{\mu}, r, Z)\,.$$

Proof. Suppose (b, v) is a solution that maximizes the quadratic optimization problem of QUAD-VAL(y). Therefore $\sum_{i=1}^k b_i^2 = \text{QUAD-VAL}(y)$. Let $X = (1, b, v) \otimes (1, b, v)$, then X will be PSD. Suppose

$$X = \begin{bmatrix} 1 & b^\mathsf{T} & v^\mathsf{T} \\ b & B & W \\ v & W^T & V \end{bmatrix},$$

then X has $\text{Tr}(B) = \sum_{i=1}^k b_i^2 = \text{QUAD-VAL}(y)$, and satisfies every constraint in the SDP-VAL optimization problem (Definition 2.6.7), except for $\text{Tr}(V) = 1$. This is because $\|y\|_2$ may be smaller than 1. Let

$$X' = \begin{bmatrix} 1 & b^\mathsf{T} & v^\mathsf{T} \\ b & B & W \\ v & W^T & V' = V/\text{Tr}(V) \end{bmatrix}.$$

Now the objective function's value and all of the constraints will be the same as X, except for the $\text{Tr}(V) = 1$ constraint, which is now satisfied. It remains to argue that X' is PSD, which is a direct result of Lemma 2.8.6. □

Lemma 2.8.6. *If* $M = \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}$ *is PSD, then for any* $k \geq 1$*,* $M' = \begin{bmatrix} A & B \\ B^\mathsf{T} & kC \end{bmatrix}$ *is also PSD.*

Proof. We prove the third condition in Lemma 2.9.6 for M'. Note that $kC^\dagger = C^\dagger/k$. From $M \succcurlyeq 0$,

- $A \succcurlyeq 0$,
- $C \succcurlyeq 0$,
- $(I - CC^\dagger)B^\mathsf{T} = 0$,
- $A - BC^\dagger B^\mathsf{T} \succcurlyeq 0$,

and we want to prove that

- $kC \succcurlyeq 0$,
- $(I - kC(kC)^\dagger)B^\mathsf{T} = 0$,
- $A - B(kC)^\dagger B^\mathsf{T} \succcurlyeq 0$.

The first two are obvious. For the last one, note that

$$A \succcurlyeq 0 \implies (k-1)A \succcurlyeq 0 \implies kA \succcurlyeq A\,.$$

Therefore

$$kA \succcurlyeq A \succcurlyeq BC^\dagger B^\mathsf{T}\,,$$

and

$$A - BC^\dagger B^\mathsf{T}/k \succcurlyeq 0\,,$$

as desired. □

2.9 Lemmata

Lemma 2.9.1. *Suppose $M \succcurlyeq 0$ is a matrix in $\mathbb{R}^{d\times d}$, $1 \le k \le d$. Let N be a $d \times d$ matrix such that for every $1 \le i, j \le d$, $N_{i,j} = M_{i,j}$, if $i \ne k$ and $j \ne k$, and $N_{i,j} = 0$, if $i = k$, or $j = k$. Then $N \succcurlyeq 0$.*

Proof.

$$\begin{aligned}\min_{\|z\|_2=1} z^\mathsf{T} N z &= \min_{\|z\|_2=1, z_k=0} zN^\mathsf{T} z \\ &= \min_{\|z\|_2=1, z_k=0} zM^\mathsf{T} z \\ &\ge \min_{\|z\|_2=1} zM^\mathsf{T} z \\ &\ge 0\end{aligned}$$

□

Lemma 2.9.2. *Suppose $\mathcal{K}$ is a closed convex set, then $f(y) = \max_{\substack{x\in\mathcal{K}\\ Ax=y}} \langle x, c\rangle$ is concave.*

Proof. We need to prove that for $\alpha \in [0,1]$

$$\alpha f(y_1) + (1-\alpha) f(y_2) \leq f(\alpha y_1 + (1-\alpha) y_2)\,.$$

Suppose $f(y_1)$ is maximized at x_1 and $f(y_2)$ is maximized at x_2. Let $z = \alpha x_1 + (1-\alpha) x_2 \in \mathcal{K}$, then $Az = \alpha y_1 + (1-\alpha) y_2$ and

$$f(\alpha y_1 + (1-\alpha) y_2) \geq \langle z, c \rangle\,,$$

which implies the desired inequality. □

The following lemma is the ordinary binary search.

Lemma 2.9.3 (Non Private Binary Interval Search)**.** *Suppose an interval $[0, D]$ is given. For an interval $[s, e] \subseteq [0, D]$, and for an accuracy parameter a, the binary interval search algorithm returns a point x_T after $T = \log(D/a)$ steps of the following binary search, where x_i is the parameter of the binary search at step i.*

1. *If $x_i < s$, move right.*
2. *If $s \leq x_i \leq e$, move in an arbitrary direction. Note that naturally in a binary search we would terminate in this step, but here we assume that we might go to any direction.*
3. *if $x_i > e$, move left.*

Then $x_T \in [s - a, e + a]$.

Lemma 2.9.4 (Cauchy-Schwarz inequality)**.** *If (f, g) is a level-2 fictitious random variable over $\mathbb{R}^{\mathcal{U}} \times \mathbb{R}^{\mathcal{U}}$, then*

$$\tilde{\mathbf{E}}_{f,g} \langle f, g \rangle \leq \sqrt{\tilde{\mathbf{E}}_f \|f\|_2^2} \cdot \sqrt{\tilde{\mathbf{E}}_g \|g\|_2^2}\,.$$

Lemma 2.9.5 (pseudo-covariance)**.** *If $\mathcal{X}$ is a degree-2 pseudo-distribution over $\mathbb{R}^d$, then*

$$\tilde{\mathbf{E}}_{\mathcal{X}} \left[\left\| x - \tilde{\mathbf{E}}_{\mathcal{X}} [x] \right\|_2^2 \right] = \tilde{\mathbf{E}}_{\mathcal{X}} \left[\|x\|_2^2 \right] - \|\tilde{\mathbf{E}}_{\mathcal{X}} [x]\|_2^2\,.$$

Proof.

$$\begin{aligned}
\tilde{\mathbf{E}}_{\mathcal{X}} \left[\left\| x - \tilde{\mathbf{E}}_{\mathcal{X}} [x] \right\|_2^2 \right] &= \sum_i \tilde{\mathbf{E}}_{\mathcal{X}} \left[\left(x_i - \tilde{\mathbf{E}}_{\mathcal{X}} [x_i] \right)^2 \right] \\
&= \sum_i \tilde{\mathbf{E}}_{\mathcal{X}} \left[x_i^2 \right] - \left(\tilde{\mathbf{E}}_{\mathcal{X}} [x_i] \right)^2 \\
&= \tilde{\mathbf{E}}_{\mathcal{X}} \left[\|x\|_2^2 \right] - \|\tilde{\mathbf{E}}_{\mathcal{X}} [x]\|_2^2
\end{aligned}$$

□

Lemma 2.9.6 (PSD 2×2 block matrix)**.** *Given any matrix, $M = \begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix}$, the following conditions are equivalent:*

1. $M \succcurlyeq 0$ *(M is positive semidefinite).*

2. $A \succcurlyeq 0, \quad (I - AA^\dagger)B = 0, \quad C - B^\mathsf{T} A^\dagger B \succcurlyeq 0.$

3. $C \succcurlyeq 0, \quad (I - CC^\dagger)B^\mathsf{T} = 0, \quad A - BC^\dagger B^\mathsf{T} \succcurlyeq 0.$

Fact 2.9.7 (SoS Triangle Inequality)**.** *Let x, y be indeterminates. Let t be a power of 2. Then*

$$\vdash_t (x+y)^2 \leq 2^{t-1}(x^t + y^t)\,.$$

Fact 2.9.8 (lower bound for spherical caps [13])**.** *For $0 \leq r \leq 2$, a cap of radius r on $\mathbb{S}^{n-1}$ has measure at least $\frac{1}{2}(r/2)^{n-1}$.*

Chapter 3

Robustness Implies Privacy: Learning Gaussian Distributions

3.1 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)* [39] 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" [38]. More recent works continue to leverage ideas from robust estimation to design private estimation procedures [20, 64, 19, 89, 69, 79, 51, 45, 90] – 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 suboptimal 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 [51] 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 [**?**] and mean estimation under other bounded-moment assumptions [**?** **?**].

Conclusions on Robust versus Private Estimation: Recent work [45] 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 3.2 we discuss situations where optimal robust estimators don't necessarily yield optimal private ones, at least in a black-box way.

3.1.1 Results

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

Definition 3.1.1 (Differential Privacy (DP) [39, 37])**.** *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 3.1.2 (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.*

Learning High-Dimensional Gaussian Distributions in TV Distance

We begin with learning Gaussians in total variation distance.

Theorem 3.1.3 (Learning Arbitrary Gaussians, Pure DP, Subsection 3.7.4)**.** *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 [58],[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 3.1.3. Under (ε, δ)-DP these bounds are avoidable. But, obtaining a polynomial-time (ε, δ)-DP

[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 [33]; the same goes for our other results on learning Gaussians.

[2] replacing the Gaussian mechanism with the Laplace mechanism

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 [6, 62, 69, 94] (see Table 3.1). 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 [6, 94].

Theorem 3.1.4 (Learning Arbitrary Gaussians, (ε, δ)-DP, Subsection 3.7.4)**.** *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 3.1.3 and 3.1.4 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.

Estimating the Mean of a Subgaussian Distribution

Mean estimation in high dimensions subject to differential privacy has also received substantial recent attention [67, 58, 21, 20, 64, 77, 19, 79, 51]. 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 [20, 64]. 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 [58, 51] (see Table 3.2). We give the first nearly-sample-optimal pure-DP algorithm:

Theorem 3.1.5 (Estimating the Mean of a Spherical Subgaussian Distribution, Theorem 3.5.1)**.** *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*

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

[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 $\log K$.

It is natural to ask whether the identity-covariance assumption can be removed from Theorem 3.1.5, 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 [**?**].

There is also an analogue (Theorem 3.5.2) for polynomial-time mean estimation subject to (ε, δ)-DP without the $\|\mu\| \le R$ assumption, using $\tilde{O}(\frac{d}{\alpha\varepsilon} + \frac{d}{\alpha^2} + \frac{\log 1/\delta}{\varepsilon})$ samples. We obtain this result from our approx-DP framework similar to proving Theorem 3.1.4: one could alternatively combine Theorem 3.1.5 with an (ε, δ)-DP procedure that obtains an $O(d)$-accurate estimate, such as [43].

Finally, we note that Theorems 3.1.3 and 3.1.5 are known to be near-optimal from standard packing lower bounds [20], and Theorem 3.1.4 and Theorem 3.5.2 are also known to be near-optimal, via the technique of fingerprinting [58, 60], except, as in Theorems 3.1.3 and 3.1.4, 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 3.4.1, 3.4.2, which give general tools for turning SoS-based robust estimators into private ones.

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[67]	$\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\min(\log K, \log \delta^{-1})}{\varepsilon}$, $d = 1$	No	Yes	Pure/Approximate
[58]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d^{3/2}\sqrt{\log K \log \delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[20]	$\frac{d^2}{\alpha^2} + \frac{d^2 \log K}{\alpha\varepsilon}$	Optimal	No	Pure
[1]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[79]	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[62]	$\frac{d^2}{\alpha^2} + \left(\frac{d^2}{\alpha\varepsilon} + \frac{d^{5/2}}{\varepsilon}\right) \cdot (\log \delta^{-1})^{O(1)}$	No	Yes	Approximate
[69]	$\frac{d^8}{\alpha^4} \cdot \left(\frac{\log \delta^{-1}}{\varepsilon}\right)^6$	Suboptimal	Yes	Approximate
[6, 94]	$\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log \delta^{-1}}}{\alpha\varepsilon} + \frac{d \log \delta^{-1}}{\varepsilon}$	No	Yes	Approximate
[6, 94]	$\frac{d^{3.5} \log \delta^{-1}}{\alpha^3\varepsilon}$	Optimal	Yes	Approximate
Thm 3.1.3	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{d^2 \log K}{\varepsilon}$	Optimal	Yes	Pure
Thm 3.1.4	$\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log \delta^{-1}}{\varepsilon}$	Optimal	Yes	Approximate

Table 3.1: 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.

3.1.2 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 [**?**]), and on high-dimensional algorithmic robust statistics (see [**?**]). We discuss other works at the intersections of these areas.

Paper	Sample Complexity	Robust?	Poly-time?	Privacy
[67]	$\frac{1}{\alpha^2}+\frac{1}{\alpha\varepsilon}+\frac{\min(\log R,\log\delta^{-1})}{\varepsilon}$, $d=1$	No	Yes	Pure/Approximate
[58]	$\frac{d}{\alpha^2}+\frac{d\sqrt{\log\delta^{-1}}}{\alpha\varepsilon}+\frac{\sqrt{d\log R\log\delta^{-1}}}{\varepsilon}$	No	Yes	Concentrated
[20]	$\frac{d}{\alpha^2}+\frac{d\log R}{\alpha\varepsilon}$	Optimal	No	Pure
[64]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{d\log R}{\varepsilon}$	Optimal	No	Pure
[1]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\varepsilon}$	Optimal	No	Approximate
[77]	$\frac{d}{\alpha^2}+\frac{d^{3/2}\log\delta^{-1}}{\alpha\varepsilon}$	Optimal	Yes	Approximate
[20, 79]	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\alpha\varepsilon}$	Optimal	No	Approximate
[51]	$\frac{d}{\alpha^2\varepsilon}+\frac{d\log R}{\varepsilon}$	Suboptimal	Yes	Pure
Theorem 3.1.5	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{d\log R}{\varepsilon}$	Optimal	Yes	Pure
Theorem 3.5.2	$\frac{d}{\alpha^2}+\frac{d}{\alpha\varepsilon}+\frac{\log\delta^{-1}}{\varepsilon}$	Optimal	Yes	Approximate

Table 3.2: 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.

Private and Robust Estimators: [38] first used robust statistics primitives to design private algorithms, a tradition continued by [20, 64, 79, 19, 89, 69, 51]. Other works from the Statistics community also investigate connections between robustness and privacy [9, 10, 90, 91], including local differential privacy [76]. Our black-box reduction from privacy to robustness can be seen as a generalization of methods of [20, 64], 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 [20, 47, 77, 43, 6, 69, 94, 79]; see Tables 3.1, 3.2.

Recall that [45] 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: [51] and [69] pioneer the use of SoS for private algorithm design. [51] gives a polynomial-time algorithm for pure-DP mean estimation under a bounded covariance assumption, using $\frac{d}{\alpha^2\varepsilon}$ samples, and [69] gives a $\approx d^8$-sample (ε,δ)-DP algorithm for learning d-dimensional Gaussians. [45] uses SoS for private *sparse* mean estimation.

On a technical level, our work most resembles [51]; we also employ SoS SDPs as score functions and leverage tools from log-concave sampling. However, there are fundamental roadblocks to using [51]'s strategy for converting SoS proofs into private algorithms in settings beyond mean estimation under bounded covariance, as we discuss in Section 3.2. We provide a blueprint for converting a much wider range of SoS-based robust algorithms to private ones.

Inverse Sensitivity Mechanism: In [8, 7], 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 (3.1). However, the focus of their work is rather different, as they investigate applications to instance-optimal private estimation, whereas our goal is to understand private estimation through the lens of robustness. Furthermore, their study is centered on one-dimensional statistics, and their analysis is not black-box.

Contemporaneous work: In independent and simultaneous work, Alabi, Kothari, Tankala, Venkat, and Zhang also design efficient robust and private algorithms for learning high-dimensional Gaussians with nearly-optimal sample complexity with respect to dimension; however, their algorithms require $\mathrm{poly}(1/\varepsilon, \log 1/\delta, 1/\alpha)$-factors more samples than those we present [3].

3.2 Techniques

3.2.1 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 [84]: 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:

$$\mathbb{P}(\theta) \propto \exp(-\varepsilon \cdot \mathrm{score}_X(\theta)) \text{ where } \mathrm{score}_X(\theta) = \min\{d(X, X') \,:\, \|\hat{\theta}(X') - \theta\| \le \alpha(\eta_0)\}, \tag{3.1}$$

where $B_{\|\cdot\|}$ is the unit ball of $\|\cdot\|$. In words: the mechanism assigns each θ within distance $\alpha(\eta_0)$ of Θ a score given by the number of input samples which would have to be changed to obtain a dataset X' for which the robust estimator $\hat{\theta}(X')$ is close to θ, and samples θ with probability $\propto \exp(-\varepsilon \cdot \mathrm{score}_X(\theta))$. 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 [48]. Later, we use a *truncated* version of (3.1) 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

[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.

(approximately) θ appears to be folklore; see for instance the *inverse sensitivity mechanism* of [8]. Our contribution is (a) to show that for (3.1) 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 (3.1) 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 $\eta_0 = \Theta(\alpha^2)$ [28, 92].

The exponential mechanism enjoys $(2\varepsilon, 0)$-DP, but the question of utility remains. Suppose that $X_1, \ldots, X_n \sim p_{\theta^*}$. 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 3.2.1. *Suppose a dataset $X_1, \ldots, X_n \sim p_{\theta^*}$, 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 (3.1) has $\|\theta - \theta^*\| \leq 2\alpha(\eta_0)$ with probability at least $1 - 2\beta$, if*

$$n \geq \max_{\eta_0 \leq \eta \leq 1} \frac{D \cdot \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log(1/\beta) + O(\log \eta n)}{\eta \varepsilon}. \tag{3.2}$$

Observe that the $O(\log \eta n)$ term in (3.2) is negligible compared to $D \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} \geq D \log 2$ if $n \ll 2^D$.

The sample complexity in (3.2) 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 3.2.1, in which the dataset $X_1, \ldots, X_n$ is a *contamination* of i.i.d. samples from p_{θ^*}, follows by a similar proof. See the full version of the paper [52] for the proof.

Consequences of Lemma 3.2.1: Applied to robust mean estimators with optimal error rates under bounded k-th moment assumptions, for any $k \geq 2$, Lemma 3.2.1 gives optimal pure-DP estimators under those same assumptions, recovering the main results of [64], applied to robust linear regression (with known covariance) [34], it yields a pure-DP analogue of the nearly-optimal regression result of [77], 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 [77, 19], using a robust estimator of mean and covariance.

To illustrate, we apply Lemma 3.2.1 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

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

η-contamination with error $\|\hat{\mu} - \mu\| \le O(\alpha + \eta)$, if $\eta < 1/2$. For ε-DP guarantees, we need to restrict to the case of $\|\mu\| \le R$ for some (large) $R > 0$; we will assume that even for $\eta \ge 1/2$, $\|\hat{\mu}\| \le R$.

Plugging such a robust $\hat{\mu}$ into Lemma 3.2.1, and choosing $\eta_0 = \alpha$, there are two interesting cases: $\eta = O(\eta_0)$ and $\eta = 1$. In the former, $\alpha(2\eta_0)/\alpha(\eta_0) = O(1)$, so we get the requirement $n \ge O(\frac{d+\log(1/\beta)}{\alpha\varepsilon})$, and in the latter $\alpha(1) = R$, so we get the additional requirement $n \ge \frac{d \log R}{\varepsilon}$, meaning that we obtained an ε-DP estimator with accuracy $O(\alpha)$ using n samples,

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

This is tight up to constants [48, 20]. 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 3.2.1 Loose? More refined analyses of the construction (3.1) 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 3.2.1 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 3.2.1 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 **??** that $\tilde{O}(\frac{\log d}{\alpha^2} + \frac{d^{2/3}}{\alpha\varepsilon^{2/3}} + \frac{\sqrt{d}}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon})$ samples are enough, in the pure-DP setting.

From Robustness to (ε, δ)-DP: If $\hat{\theta}$ has a nontrivial *breakdown point* – i.e., a fraction of corruptions η beyond which it admits no error guarantees, then Lemma 3.2.1 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 3.2.1.

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 [79], and the approach of [19], for building approx-DP estimators from robust estimation primitives; we can recover their results on covariance-aware mean estimation and linear regression with (ε, δ)-DP guarantees. This approach carries the advantages of black-box-ness and potential polynomial-time implementability, since SoS-based robust estimators for mean and covariance have the required worst-case behavior.

Consider again a deterministic robust estimator $\hat{\theta}$: datasets $\to \Theta \cup \{\text{REJECT}\}$ for a parameter $\theta \in \mathbb{R}^d$, which takes n inputs and returns either some element of Θ or REJECT. Let $\mathcal{P}$ be a distribution family, $\|\cdot\|$ be a norm, $\alpha : [0,1] \to \mathbb{R}$ be a nondecreasing function, $n \in \mathbb{N}$, and $\eta_0, \eta^* \in [0,1]$. We continue to employ $\text{SCORE}_X(\theta)$ as defined in (3.1). Suppose as before that with probability $1-\beta$ over samples $X_1, \ldots, X_n \sim p_\theta \in \mathcal{P}$, for every $\eta < \eta^*$, given any η-corruption of $X_1, \ldots, X_n$, $\|\hat{\theta} - \theta\| \le \alpha(\eta)$. And, suppose that $\hat{\theta}$ has the following worst-case robustness property: for *any* input $X = X_1, \ldots, X_n$, if $\hat{\theta}(X) \neq$ REJECT, then for every $\eta < \eta^*$, given any η-corruption X' of X, either $\hat{\theta}(X') =$ REJECT, or $\|\hat{\theta}(X') - \hat{\theta}(X)\| \le \alpha(\eta^*)$.

Lemma 3.2.2. *Let $\eta_0 < \eta^* \in [0,1]$ be such that $\eta^* n$ is a sufficiently large constant. For every $\varepsilon, \delta > 0$, there is an $(O(\varepsilon), O(e^{2\varepsilon}\delta))$-DP mechanism which, for any θ^*, takes $X_1, \ldots, X_n \sim p_{\theta^*}$ and with probability $1-\beta$ outputs θ such that $\|\theta - \theta^*\| \le 2\alpha(\eta_0)$, if*

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

For a proof of Lemma 3.2.2 we refer the reader to the full version of the paper [52].

3.2.2 Algorithms

Even if the robust estimator $\hat{\theta}$ can be computed in polynomial time, the sampling problem in (3.1) 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 3.2.3 (SoS Proof)**.** *Let $p_1(x) \ge 0, \ldots, p_m(x) \ge 0$ be a system of polynomial inequlities in variables $x_1, \ldots, x_n$. An inequality $q(x) \ge 0$ has a* degree d SoS proof *from $p_1 \ge 0, \ldots, p_m \ge 0$, written $\{p_1 \ge 0, \ldots, p_m \ge 0\} \vdash_d^x q \ge 0$, if for each multiset $S \subseteq [m]$ there exists a sum of squares polynomial $q_S(x)$, such that $\deg(q_S(x) \cdot \prod_{i \in S} p_i(x) \le d)$ 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 3.2.4 (Pseudoexpectation)**.** *Let $\mathbb{R}[x]_{\le d}$ be the set of degree at most d polynomials in variables $x_1, \ldots, x_n$. A linear operator $\tilde{\mathbf{E}} : \mathbb{R}[x]_{\le d} \to \mathbb{R}$ is a degree d* pseudoexpectation *if $\tilde{\mathbf{E}} 1 = 1$ and $\tilde{\mathbf{E}} p^2 \ge 0$ for any p of degree at most $d/2$. A pseudoexpectation $\tilde{\mathbf{E}}$* satisfies

a system of polynomial inequalities $p_1 \geq 0, \dots, p_m \geq 0$, written $\tilde{\mathbf{E}} \models p_1 \geq 0, \dots, p_m \geq 0$, 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 $\mathcal{A}(X, \theta, z) \vdash_{O(1)} \langle \theta - \theta^*, v \rangle \leq \alpha$ for every v in the dual ball of $\|\cdot\|$.

To give a robust estimation algorithm, on input η-corrupted X, we can obtain $\tilde{\mathbf{E}}$ 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 $\mathcal{A} \vdash^{\theta,z}_{O(1)} \langle \theta - \theta^*, v \rangle \leq \alpha$, we get $\|\tilde{\mathbf{E}}\, \theta - \theta^*\| \leq \alpha$.

Lemma 3.2.5 (Informal, implicit in [68])**.** *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 $\alpha = \tilde{O}(\eta)$.*

Robustness to Privacy, Algorithmically

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

$$\text{score}_X(\theta) = \min\{d(X, X') \,:\, \|\text{KMZ}(Y) - \theta\| \leq \alpha\}\,, \tag{3.3}$$

much less sample from the distribution (3.1). 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 3.2.1 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 $X'_1, \dots, X'_n$, $w_1, \dots, w_n$, θ', and consider

$$\mathcal{B}_t = \left\{ w_i^2 = w_i,\ \sum_{i=1}^n w_i = n - t,\ w_i X_i = w_i X'_i, \right\} \cup \mathcal{A}(X', \theta', z)\,, \tag{3.4}$$

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

$$\text{SoS-score}_X(\theta) = \min t \text{ s.t. } \exists \text{ degree } O(1)\ \tilde{\mathbf{E}} \text{ in variables } X', w, \theta', z,\ \tilde{\mathbf{E}} \models \mathcal{B}_t,\ \|\tilde{\mathbf{E}}\, \theta' - \theta\| \leq \alpha\,. \tag{3.5}$$

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

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 3.2.1 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 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 $\mathcal{B}_t \vdash_{O(1)}^{X',\theta',w,z} \langle \theta' - \theta^*, v\rangle \le \tilde{O}(\eta)$. Luckily, the SoS proofs of [68] readily generalize to show this.

In fact, [68]'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 (3.4), (3.5), while closely following our black-box reduction strategy, contain an unnecessary layer of indirection. When we implement this strategy in detail in Sections 3.5, 3.6, and 3.7, we remove this indirection for simplicity.

On "Satisfies": An important technical difference between our score function and that of [51] 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 $\|\tilde{\mathbf{E}}\,\theta' - \theta^*\| \ge \Omega(R)$, 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 [88]. 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 3.8. Ultimately, we find a further-relaxed score function $\text{SoS-score}'_X$, 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 $\text{SoS-score}_X(\theta)$ 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 [51]. 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 $\sum_{i=1}^n w_i \ge n - \frac{1}{2}(t_0 + t_1)$, even though it does have $\frac{1}{2}(\tilde{\mathbf{E}}_0 + \tilde{\mathbf{E}}_1)[\sum_{i=1}^n w_i] \ge n - \frac{1}{2}(t_0 + t_1)$.

$\text{SoS-score}_X(\theta)$ is *quasi-convex* in θ, meaning that its sub-level sets $S_t = \{\theta : \text{SoS-score}_X(\theta) \le t\}$ are convex for all t. This is good news: if we discretize the range of possible scores $[0, n]$ into $t_1, \dots, t_{n^{O(1)}}$ (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$_X(\theta) \le t_i$, 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 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.

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 (3.1). Theorems 3.4.1 and 3.4.2 capture this strategy formally.

3.3 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 $r \ge 0$, 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 3.3.1 (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), $|f(\theta, \mathcal{X}) - f(\theta, \mathcal{X}')| \le \Delta$. We will implicitly assume that sensitivity is with respect to the dataset.*

Definition 3.3.2 (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*

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

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.

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

Definition 3.3.3 (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 $\|\mu - \mu'\|_\Sigma = \alpha \le 1$, then $d_{\mathrm{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma)) = \Theta(\alpha)$, and if Σ, Σ' have Mahalanobis distance $\alpha \le 1$, 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 3.3.4 (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 3.3.5 (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 3.3.6 (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 3.3.7 (monomial vector)**.** *A monomial vector of degree d is a $n^{O(d)}$-dimensional vector $v_d(x)$ indexed by multisets $S \subseteq [n]$, $|S| \le d$, where the entry $v_d(x)_S$ is the monomial*

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

Remark 3.3.8. *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 Section 3.8.*

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 3.3.9 (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

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

3.4 A General Private Sampling Algorithm

In this section, we state 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 3.4.1), or an approx-DP algorithm (Theorem 3.4.2), 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 3.4.1 and 3.4.2 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 3.4.1. *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 $\theta \in \Theta$, $|\mathcal{S}(\theta, \mathcal{Y}) - \mathcal{S}(\theta, \mathcal{Y}')| \leq 1$.*
- *(Quasi-Convexity) For any fixed dataset $\mathcal{Y}$, any $\theta, \theta' \in \Theta$, and any $0 \leq \lambda \leq 1$, we have that $\mathcal{S}(\lambda\theta + (1-\lambda)\theta', \mathcal{Y}) \leq \max(\mathcal{S}(\theta, \mathcal{Y}), \mathcal{S}(\theta', \mathcal{Y}))$.*
- *(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, $\mathcal{S}(\theta', \mathcal{Y}) \leq T + 1$, in time $\mathrm{poly}(n, d, \log \frac{R}{r})$.*

- *(Volume) For any given dataset $\mathcal{Y}$ and $\eta' \geq \eta$, 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 $\mathcal{S}(\theta, \mathcal{Y}) \leq \eta n$ and if $n \geq \Omega\left(\max_{\eta' : \eta \leq \eta' \leq 1} \frac{\log(V_{\eta'}(\mathcal{Y})/V_{\eta}(\mathcal{Y})) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right)$, 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 state a similar result for approx-DP algorithms.

Theorem 3.4.2. *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 3.4.1.*

In addition, fix some parameter $\eta^* \in [10\eta, 1]$. Suppose that $n \geq \Omega\left(\frac{\log(1/\delta) + \log(V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}))}{\varepsilon \cdot \eta^*}\right)$ for all $\mathcal{Y}$ such that there exists θ with $\mathcal{S}(\theta, \mathcal{Y}) \leq 0.7\eta^* n$. 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 $\mathcal{S}(\theta, \mathcal{Y}) \leq \eta n$ and if $n \geq \Omega\left(\max_{\eta' : \eta \leq \eta' \leq \eta^*} \frac{\log(V_{\eta'}(\mathcal{Y})/V_{\eta}(\mathcal{Y})) + \log(1/(\beta \cdot \eta))}{\varepsilon \cdot \eta'}\right)$, 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.

For proofs of Theorems 3.4.1 and 3.4.2 we refer the reader to full version of the paper [52]. In the following sections we apply this generic framework repeatedly to solve Gaussian estimation.

3.5 Estimating the Mean of a Gaussian

3.5.1 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 3.5.1 (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 3.5.2 (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 3.5.2 by combining Theorem 3.5.1 with [43, Corollary 5] (or alternatively [46, 94]), 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 3.5.1 and Theorem 3.5.2.

3.5.2 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 3.5.3 (Resilience of First and Second Moments, Proposition 3.3 in [31])**.** *Let $n \geq O((d+\log(1/\beta))/\alpha^2)$, 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 $\mathbf{E}_i\, b_i \geq 1-\eta$ and all unit vectors $v \in \mathbb{R}^d$, we have*

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

In addition,

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

Corollary 3.5.4. *Let $\mu, \mathcal{D}, \{x_i\}, \alpha, \eta$ be as in Lemma 3.5.3. Then, with probability $1-\beta$, the following all hold for all unit vectors v simultaneously.*

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

Proof. Fix a vector v and let $z_i := \langle x_i - \mu, v\rangle$. Suppose the events of Lemma 3.5.3 hold.

Parts 1 and 2 are immediate from Lemma 3.5.3, by setting $b_i = 1$ for all i. Part 3 follows by setting $a_i = 1 - b_i$, and then noticing that $\left|\frac{1}{n}\sum_{i=1}^n a_i z_i\right| \leq \left|\frac{1}{n}\sum_{i=1}^n z_i\right| + \left|\frac{1}{n}\sum_{i=1}^n b_i z_i\right| \leq \widetilde{O}(\eta)$.

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

Remark 3.5.5. *The conditions in Corollary 3.5.4 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 3.5.4.*

We also note that if $\mu, \{x_i\}$ satisfy Corollary 3.5.4, then for all symmetric H with $\|H - I\|_{op} \leq \alpha$, $H\mu, \{Hx_i\}$ also satisfies Corollary 3.5.4 (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, $|\mathbf{E}_i \langle Hx_i, v\rangle| = |\mathbf{E}_i \langle x_i, Hv\rangle| \leq \alpha \cdot \|Hv\|_2 \leq \alpha \cdot (1+\alpha) \leq 2\alpha$. We can repeat the same argument for the 2nd, 3rd, and 4th conditions.

3.5.3 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 [68].

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.)

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

It can be proven that if n is as in Lemma 3.5.3, 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 $\tilde{\mathbf{E}}$.

3.5.4 Score Function and its Properties

Our goal is to use Theorem 3.4.1, 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 3.5.3 to deal with approximate pseudoexpectations.

Definition 3.5.6 (Certifiable Mean)**.** *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 $\widetilde{\mu} \in \mathbb{R}^d$. 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 $[\ldots]_{j,k}$ 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 3.5.7 (Score Function)**.** *Let* $\mathbb{B}_R^d$ *denote the ball of radius* R *in* $\mathbb{R}^d$ *centered at the origin. Let* $\alpha, \tau, \phi, T \in \mathbb{R}^{\geq 0}$*,* $y_1, \ldots y_n \in \mathbb{R}^d$ *(with* $\mathcal{Y} = \{y_1, \ldots, y_n\}$*) and* $\widetilde{\mu} \in \mathbb{R}^d$*. We define the score function* $\mathcal{S} : \mathbb{B}_R^d \to \mathbb{R}$ *as*

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

In the rest of this section we will prove the following properties for this score function. This will allow us to use Theorem 3.4.1.

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

Sensitivity

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

Lemma 3.5.8 (score function upper bound)**.** *The value of the score function* $\mathcal{S}$ *defined in Definition 3.5.7 is less than or equal to* n.

[8] See Lemma 3.5.12 for more details on how large we require R' to be.

Proof. It suffices to show that in Definition 3.5.6 for $T = n$, there exists a linear functional $\mathcal{L}$ such that the constraints of Definition 3.5.6 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 3.5.7 is at most n. □

Lemma 3.5.9 (sensitivity)**.** *The score function $\mathcal{S}$ as defined in Definition 3.5.7 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 3.5.6 are satisfied for $\mathcal{L}$, $\mathcal{Y}$, and T. If we construct a new functional $\mathcal{L}'$ such that the constraints of Definition 3.5.6 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 $\|p\|_2 \leq 1$ and $p = q + w_j r$, where q does not contain a monomial containing w_j.

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

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

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

- $\forall i : \mathcal{L}'(w_i^2 - w_i)p^2 \in [-\tau \cdot (T+1), \tau \cdot (T+1)]$. 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.

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

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

  To bound the first term, we have that $\mathcal{L}(\Sigma w_i - n + T)q^2 \geq -5\tau \cdot T \cdot n$. 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 $\|(1-w_j)q\|_2 \leq 2\|q\|_2 \leq 2$, so $\mathcal{L}[(1-w_j)^2 q^2] \geq -4\tau \cdot T$, and $\mathcal{L}[(w_j - w_j^2)q^2] \geq -\tau \cdot T$. 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 $-5\tau \cdot T \cdot n - 5\tau \cdot T \geq -5\tau \cdot (T+1) \cdot n$, which is trivial by Lemma 3.5.8.

- $\forall j, k : \mathcal{L}'\left(\left[\frac{1}{n}\sum_i (x_i' - \mu')(x_i' - \mu')^\mathsf{T} + MM^\mathsf{T} - (1+\alpha)I\right]_{j,k} p^2\right) \in [-\tau \cdot T, \tau \cdot T]$, 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 3.5.6 for y', and $T+1$. Finally, note that $\|\mathcal{R}(\mathcal{L}')\|_2 \leq \|\mathcal{R}(\mathcal{L})\|_2$ clearly holds. Therefore the score function $\mathcal{S}$ has sensitivity 1 with respect to its first input. □

Quasi-convexity

Lemma 3.5.10 (quasi-convexity)**.** *The score function $\mathcal{S}$ as defined in Definition 3.5.7 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 3.5.6 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 3.5.6, 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 3.5.6 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

$$\begin{aligned}\mathcal{L}_3\left(\sum w_i - n + T_3\right)p^2 &\geq (\lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2)\left(\sum w_i - n + T_2\right)p^2 \\ &= \lambda\mathcal{L}_1\left(\sum w_i - n + T_1\right)p^2 + (1-\lambda)\mathcal{L}_2\left(\sum w_i - n + T_2\right)p^2 + \lambda(T_2 - T_1)\mathcal{L}_1 p^2 \\ &\geq -5\tau \cdot n(\lambda T_1 + (1-\lambda)T_2) - \lambda(T_2 - T_1) \cdot \tau \cdot T_1 \\ &\geq -5\tau \cdot n(\lambda T_1 + (1-\lambda)T_2 + \lambda(T_2 - T_1)) \\ &\qquad\qquad (n \geq T_1, \text{ Lemma 3.5.8}) \\ &= -5\tau \cdot T_3 \cdot n,\end{aligned}$$

as desired. □

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 3.5.11. *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 3.5.11 essentially follows from the same argument as in [68], with slight modifications to deal with our modified score function. For a detailed proof we refer the reader to [52].

Volume of Good Points

Lemma 3.5.12. *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 3.5.3, 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, $\frac{1}{n}\sum_{i=1}^n \langle x_i - \mu', v\rangle^2 \leq 1 + \widetilde{O}(\eta)$ 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 $\|\mathcal{R}(\mathcal{L})\|_2 \leq (Rnd)^{O(1)}$. □

Lemma 3.5.13. *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 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 3.5.9), 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 $\alpha = \widetilde{O}(\eta) \leq \alpha'$, we have that by Lemma 3.5.11 that any $(\alpha, \tau, \phi, T + \eta n)$-certifiable mean $\widetilde{\mu}$ must satisfy $\|\widetilde{\mu} - \mu\|_2 \leq O(\alpha') \leq \widetilde{O}(T/n)$. □

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 $\eta \leq \eta' \leq \eta^*$, 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 $\|\widetilde{\mu}\|_2 \leq R$, is at most $O(R/\sqrt{d})^d$.

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 3.8.7, deferred to Section 3.8, that for the score in Definition 3.5.7, defined by the constraints in Definition 3.5.6, we can compute it up to error γ in time $\text{poly}(n, d, \log R, \log \gamma^{-1})$. Hence, this verifies the "efficiently computable" criterion for Theorem 3.4.1.

Efficient Finding of Low-Scoring Point

Verifying the "robust algorithm finds low-scoring point" criterion is also direct from Theorem 3.8.7. 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 3.8.7 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 $\min_{\widetilde{\mu}} \mathcal{S}(\widetilde{\mu}, \mathcal{Y}) + 1$.

3.5.5 Proof of Theorem 3.5.1

We apply Theorem 3.4.1, using the score function defined in Definition 3.5.7. Indeed, for $r = \phi = \alpha/\sqrt{d}$, we have verified all conditions, as long as $n \geq O((d + \log(1/\beta))/\alpha^2)$. Therefore, we have an ε-DP algorithm running in time $\text{poly}(n, d, \log \frac{R}{\alpha}) = \text{poly}(n, d)$ that finds a candidate mean $\widetilde{\mu}$ of score at most $2\eta n$, as long as

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

Using Lemmas 3.5.12 and 3.5.13, we have that for $\eta' \leq \eta^*$ for some $\eta^* = \Omega(1)$, $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) = (\widetilde{O}(\eta')/\eta)^d \leq (O(1/\eta))^d$. For $\eta' > \eta^*$, we have that $V_{\eta'}(\mathcal{Y})/V_\eta(\mathcal{Y}) \leq$

$(O(R/\eta))^d$. So overall, it suffices for

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

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 3.5.11, 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 3.5.11 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.

3.5.6 The approx-DP setting

In this subsection, we prove Theorem 3.5.2. In this setting, the score function is identical, but we can afford fewer samples as we apply the algorithm of Theorem 3.4.2 instead of Theorem 3.4.1. The main additional thing we must check is that for *any* dataset $\mathcal{Y}$, if $\mathcal{S}(\Sigma, \mathcal{Y}) \leq 0.7\eta^* n$ 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. For the proof see [52].

Lemma 3.5.14. *Fix η^* to be a sufficiently small constant, and $T = \eta^* n$. Also, suppose $\phi \leq \alpha/\sqrt{d}$. 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 $\|\widetilde{\mu}_1 - \widetilde{\mu}_2\|_2 \leq O(1)$.*

As a corollary of Lemma 3.5.14, we have the following result.

Corollary 3.5.15. *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 $\phi \leq \alpha/\sqrt{d}$. Then, the set of $(\alpha, \tau, \phi, 0.8\eta^* n)$-certifiable means contains all $\widetilde{\mu}$ such that $\|\widetilde{\mu} - \hat{\mu}\|_\infty \leq \phi$, and any $(\alpha, \tau, \phi, \eta^* n)$-certifiable mean $\widetilde{\mu}$ must satisfy $\|\widetilde{\mu} - \hat{\mu}\| \leq O(1)$.*

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 $\|\hat{\mu} - \widetilde{\mu}\|_\infty \leq \phi$ 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 3.5.14. □

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 $\phi^d \geq \Omega(1/\sqrt{d})^d$. So, the ratio $V_{\eta^*}(\mathcal{Y})/V_{0.8\eta^*}(\mathcal{Y}) \leq O(1/\alpha)^d$.

We now prove Theorem 3.5.2, by applying Theorem 3.4.2. 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 $\|\mu\| \leq R$, 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 3.5.15, 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 $n \geq O\left(\frac{\log(1/\delta) + d\log(1/\alpha)}{\varepsilon}\right)$, the algorithm of Theorem 3.4.2 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 $n \geq \widetilde{O}((d + \log(1/\beta))/\eta^2)$ 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

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

This concludes the proof of Theorem 3.5.2.

3.6 Preconditioning the Gaussian

3.6.1 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 3.6.1 (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 3.6.2 (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.

3.6.2 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 3.6.3. *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 [27], which means the number of samples required is $d^{2+\Omega(1)}$ if we want exponentially small failure probability. We prove Lemma 3.6.3 in **??**.

Remark 3.6.4. *As in the mean estimation case, Lemma 3.6.3 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.)*

3.6.3 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 3.6.3, with probability $1 - \beta$ over the choice of x_i's, if we output $\tilde{\mathbf{E}}\,\Sigma'$, then $\|\Sigma^{-1/2}(\tilde{\mathbf{E}}\,\Sigma')\Sigma^{-1/2} - I\|_{op} = \widetilde{O}(\eta)$.

3.6.4 Score Function and its Properties

Our goal is to use Theorem 3.4.1, so we relax the pseudo-expectation from the robust algorithm to a linear operator that behaves as an approximate pseudoexpectation.

Definition 3.6.5 (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 $\|\mathcal{R}(\mathcal{L})\|_2 \leq R' + T \cdot \tau$ *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 3.6.6 (Score Function)**.** *Let* $\alpha, \tau, T \in \mathbb{R}^{\geq 0}$, $y_1, \ldots, y_n \in \mathbb{R}^d$ *(with* $\mathcal{Y} := \{y_1, \ldots, y_n\}$*), and* $\widetilde{\Sigma} \in \mathbb{R}^{d \times d}$ *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 3.4.1.

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.

The proofs in rest of this subsection are often very similar to the proofs in the mean estimation section. In this section we only include the statements of lemmas and the main steps and refer the reader to the full version of the paper for the full proofs [52].

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 [72] for *certifiable* fourth moment bounds, which we will need to verify Condition 3 of Definition 3.6.5, 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.

Lemma 3.6.7. *Suppose $n \geq \widetilde{O}\left(\frac{d^2+\log^2(1/\beta)}{\eta^2}\right)$, and let $\alpha = \widetilde{O}(\eta)$. Let $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, and let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ represent an η-corruption of $\mathcal{X}$. Then, for any $\tau \geq 0$ 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., $\|(\Sigma')^{-1/2}\widetilde{\Sigma}(\Sigma')^{-1/2} - I\|_{op} \leq \alpha$) is an (α, τ, T)-certifiable covariance for $\mathcal{Y}$.*

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 3.6.7.

Lemma 3.6.8 (score function upper bound)**.** *The value of the score function $\mathcal{S}$ defined in Definition 3.6.6 is less than or equal to n.*

Lemma 3.6.9 (sensitivity)**.** *The score function $\mathcal{S}$ as defined in Definition 3.6.6 has sensitivity 1 with respect to its first input.*

Quasi-convexity

Lemma 3.6.10 (quasi-convexity)**.** *The score function $\mathcal{S}$ as defined in Definition 3.6.6 is quasi-convex in its second input, $\widetilde{\Sigma}$.*

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., $\|\Sigma^{-1/2}\widetilde{\Sigma}\Sigma^{-1/2} - I\|_{op} \leq O(\alpha)$.

Lemma 3.6.11. *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 $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, 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 [68].

Volume of Good Points

Lemma 3.6.12. *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 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.*

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 $V_{\eta'} \le O(1)^{d^2} \cdot V_\Sigma$ for $\eta' \in [\eta, \eta^*]$. In addition, by Constraint 5 of Definition 3.6.5, 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 $\eta' \in [\eta^*, 1]$, $V_{\eta'} \le O(K^2)^{d^2} \cdot V_\Sigma = e^{O(d^2 \log K)} \cdot V_\Sigma$.

Finally, by Lemma 3.6.7, 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 $n \ge \widetilde{O}((d^2 + \log^2(1/\beta))/\eta^2)$, $\frac{1}{n}\sum x_i x_i^\top$ 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 $\eta \cdot n$. Therefore, $V_\eta \ge V_\Sigma \cdot (\alpha/10)^{d^2}$.

Efficient Computability

As in the mean estimation case, we apply Theorem 3.8.7 in Section 3.8. 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 3.8.7. So, we have efficient computability.

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 3.8.7 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 $\min_{\widetilde{\Sigma}} \mathcal{S}(\widetilde{\Sigma}, \mathcal{Y}) + 1$.

3.6.5 Proof of Theorem 3.6.1 and 3.6.2

We apply Theorem 3.4.1, using the score function defined in Definition 3.6.6 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

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

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 3.6.11, 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.

The approx-DP setting. The proof is similar to mean estimation. Instead we would need to prove the following lemma and the rest would be identical. See [52] for a full proof.

Lemma 3.6.13. *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 $\widetilde{\Sigma}_2 \preccurlyeq O(1)\cdot\widetilde{\Sigma}_1$.*

3.7 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 3.7.1 (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 $\mathcal{N}(\boldsymbol{0}, \Sigma)$ 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 3.7.1 with Theorem 3.5.1 and Theorem 3.6.1 (or Theorem 3.5.2 and Theorem 3.6.2), we will be able to prove our main results on privately learning Gaussians up to low total variation distance, namely Theorems 3.1.3 and 3.1.4. We prove these theorems in Section 3.7.4.

3.7.1 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 $\frac{1}{n}\sum (x_i')(x_i')^\top$: note that S' is the flattening of Σ'. It can be proven that if n is as in Lemma 3.6.3, with probability $1-\beta$ over the choice of x_i's, if we output $\tilde{\mathbf{E}}\,\Sigma'$, then $\|\tilde{\mathbf{E}}\,\Sigma' - \Sigma\|_F = \widetilde{O}(\eta)$.

3.7.2 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 3.7.2 (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 3.7.3 (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*

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

Similar to the cases for mean estimation and covariance estimation in spectral distance, we need to check the constraints of Theorem 3.4.1, we omit the details and refer the reader to the full version of the paper for checking each of the constraints and their proofs [52].

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 3.7.4. *Suppose that $n \geq \widetilde{O}\left(\frac{d^2 + \log^2(1/\beta)}{\eta^2}\right)$ and $\alpha = \widetilde{O}(\eta)$. Let $\mathcal{X} = \{x_1, \ldots, x_n\} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma)$, 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.*

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.

Lemma 3.7.5. *Let $\alpha = \widetilde{O}(\eta)$ and suppose α, η are bounded by a sufficiently small constant. Let $n \geq \frac{d^2 + \log^2(1/\beta)}{\alpha^2}$, and $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, for $\Sigma \in \mathbb{R}^{d \times d}$ with $(1 - \alpha)I \preccurlyeq \Sigma \preccurlyeq (1 + \alpha)I$.*

Then, for any $\alpha^* \leq \alpha$, 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 $\phi \leq \alpha/d$ must satisfy $\|\widetilde{\Sigma} - \Sigma\|_F \leq O(\alpha)$.

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 3.7.4.

Lemma 3.7.6. *Let $\mathcal{X} = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ (where $(1 - \alpha)I \preccurlyeq \Sigma \preccurlyeq (1 + \alpha)I$), 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, ϕ, T)-certifiable covariance $\widetilde{\Sigma}$ with respect to $\mathcal{Y}$, for $\phi = \alpha/d$, satisfies $\|\Sigma - \widetilde{\Sigma}\|_F \leq \widetilde{O}(T/n)$.*

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 $d^{O(d^2)}$.

3.7.3 Proof of Theorem 3.7.1

We apply Theorem 3.4.1, using the score function defined in Definition 3.7.3. 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 3.7.4 and 3.7.6, and by the commentary after Lemma 3.7.6, 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 3.7.5, 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.

3.7.4 Proof of Theorems 3.1.3 and 3.1.4

By combining Theorems 3.6.1, 3.7.1, and 3.5.1, we are able to prove Theorem 3.1.3. The proof of Theorem 3.1.4 similar: this time, we combine Theorems 3.6.2, 3.7.1, and 3.5.2.

Proof sketch. First learn $\widetilde{\Sigma}_1$ such that $\|\Sigma^{-1/2} \widetilde{\Sigma}_1 \Sigma^{-1/2} - I\|_{op} \le \alpha$ using the preconditioning algorithms. Then replace each datapoint y_i with $\hat{y}_i = \widetilde{\Sigma}_1^{-1/2} y_i$. Then learn some $\hat{\Sigma}$ such that $\Sigma, \hat{\Sigma}$ are close in Mahalanobis distance. Finally, redefine $\hat{y}_i = \hat{\Sigma}^{-1/2} y_i$ and learn $\hat{\mu}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}) \le \mathcal{O}(\alpha)$. □

For a full proof see [52].

3.8 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 3.8.2.

Assumption 3.8.1. *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 3.8.2 (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 3.8.1. 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 $\mathcal{L} \models_\tau Q_T$ 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 3.8.3. *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* $S = \varnothing$ *index, which corresponds to the monomial* 1.

Lemma 3.8.4 (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 3.8.1, with fixed parameter* T. *Let* $\mathcal{R}(\mathcal{L})_{\overline{\varnothing}}$ *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* $\|\mathcal{R}(\mathcal{L})_{\overline{\varnothing}}\|_2 \leq R'$, *and* $\mathcal{L} \vDash_{\tau,O(\gamma)} Q_T$; *or,*

2. *shows that the volume of representations of functionals* $\mathcal{L}$ *such that* $\|\mathcal{R}(\mathcal{L})_{\overline{\varnothing}}\|_2 \leq R'$ *and* $\mathcal{L} \vDash_\tau Q_T$, *when projected to the entries* $S \neq \varnothing$, *is less than the volume of a ball of radius* r,

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

In essence, we use reductions to semi-definite programs. For a textbook treatment of this approach see Chapter 3 of [44].

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

$$K = \{\mathcal{R}(\mathcal{L}) \mid \mathcal{L} \vDash_\tau Q_T\}, K_{\overline{\varnothing}} = \{\mathcal{R}(\mathcal{L})_{\overline{\varnothing}} \mid \mathcal{L} \vDash_\tau Q_T\}.$$

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

1. $K_{\overline{\varnothing}}$ is convex; and,

2. $K_{\overline{\varnothing}}$ 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 $K_{\overline{\varnothing}}$ 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 $\|M_3\|_2 \leq \alpha\|M_2\|_2 + (1-\alpha)\|M_2\|_2 \leq R$. 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. $\mathcal{L}_3 q = \alpha\mathcal{L}_1 q + (1-\alpha)\mathcal{L}_2 q \geq -\tau \cdot T$, 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 $M \in K_{\overline{\varnothing}}$, or not. Let M' be equal to M with the additional entry $M'_{\emptyset} = 1$. Then it is easy to see that $M \in K_{\overline{\varnothing}}$, 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 3.8.2.

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

PSD Constraints. These constraints are in the following form.

$\mathcal{L}qh^2 \geq -\tau \cdot T$, for every polynomial h where $\|\mathcal{R}(h)\|_2 \leq 1$, and $\deg q + 2\deg h \leq d$, and for every polynomial q that is either 1 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 $\text{poly}(n^d, \log\frac{1}{\tau\cdot\gamma})$. 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 $\mathcal{L} \vDash_{\tau,\mathcal{O}(\gamma)}$, 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 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 $\text{poly}(n^d, \log \frac{1}{\tau \cdot \gamma})$. 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 $\mathcal{L} \vDash_{\tau, \mathcal{O}(\gamma) Q_T}$, 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 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 $c^\mathsf{T} X c < -\tau \cdot (T + \gamma)$. 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 $\langle M, H_{\overline{\varnothing}} \rangle = \langle M', H \rangle - H_\varnothing < -\tau \cdot (T + \gamma) - H_\varnothing$. Similarly, if $N \in K_{\overline{\varnothing}}$, 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 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 $\langle M, H_{\overline{\varnothing}} \rangle = \langle M', H \rangle - H_\varnothing = \|H\|_2^2 - 1$. Moreover, for every $N \in K_{\overline{\varnothing}}$, we have that $\langle N, H_{\overline{\varnothing}} \rangle \leq R\|H\|_2 - 1$. Therefore $H_{\overline{\varnothing}}$ is a separating hyperplane. □

Lemma 3.8.5 (robust satisfiability)**.** *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption 3.8.1. 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*

$$\{\mathcal{R}(\mathcal{L})_{\overline{\varnothing}} \mid \mathcal{L} \in \mathcal{F}\}$$

contains a full-dimensional ball of radius $r = \operatorname{poly}(1/\operatorname{poly}(n^d), \tau, \gamma, 1/k, 1/\|R(Q_{T_0})\|_\infty)$, 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 $\|E_{\overline{\varnothing}}\|_2 \leq r$ and $E_\varnothing = 0$. 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 $E_{\overline{\varnothing}}$, where $\|E_{\overline{\varnothing}}\|_2 \leq r$, 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

$$\mathcal{L}h^2 = \mathcal{L}_0 h^2 + \langle E, \mathcal{R}(h^2)\rangle \geq -\tau \cdot T_0 - r\operatorname{poly}(n^d).$$

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

$$\begin{aligned}\mathcal{L}qh^2 &= \mathcal{L}_0 qh^2 + \langle E, \mathcal{R}(qh^2)\rangle \\ &\geq -\tau \cdot T_0 - r\|\mathcal{R}(q)\|_\infty \cdot \operatorname{poly}(n^d).\end{aligned}$$

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

$$\begin{aligned}\mathcal{L}q_{T_0+\gamma}h^2 &= \mathcal{L}(q_{T_0} + c\gamma)h^2 \\ &= \mathcal{L}_0 q_{T_0}h^2 + c\gamma\mathcal{L}h^2 + \langle E, \mathcal{R}(q_{T_0}h^2)\rangle \\ &\geq -\tau \cdot T_0 - c\gamma\left(\tau \cdot T_0 + r\operatorname{poly}(n^d)\right) - r \cdot \|\mathcal{R}(q_{T_0})\|_\infty \cdot \operatorname{poly}(n^d)\end{aligned}$$

6. Let $\mathcal{E}$ be the corresponding linear functional for E. For every $k \times k$ matrix PSD constraint q, we have that

$$\begin{aligned}\|\mathcal{E}q\|_2 &\leq \sqrt{k}\|\mathcal{E}q\|_\infty \\ &= \sqrt{k}\max_{i,j}|\langle E, \mathcal{R}(q_{i,j})\rangle| \\ &\leq r \cdot \sqrt{k} \cdot \operatorname{poly}(n^d) \cdot \max_{i,j}\|\mathcal{R}(q_{i,j})\|_\infty.\end{aligned}$$

   Therefore

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

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

Therefore it suffices to take r such that

1. $r \cdot \mathrm{poly}(n^d) \le \tau\gamma$. In order to do this take $r \le \tau\gamma/\mathrm{poly}(n^d)$.
2. $c\gamma r\,\mathrm{poly}(n^d) + r \cdot \|\mathcal{R}(q_{T_0})\|_\infty \cdot \mathrm{poly}(n^d) \le \tau\gamma/2$. In order to do this take r to be
$$r \le \frac{\tau}{4\,\mathrm{poly}(n^d)} \cdot \min\left(\frac{\gamma}{\|\mathcal{R}(q_{T_0})\|_\infty}, \frac{1}{T_{\max}}\right),$$
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 3.8.6. *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption 3.8.1. 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 3.8.7 (computability of score functions)**.** *Consider the family of polynomial constraints $\{Q_T\}$ of up to degree d over $\mathbb{R}^n$ as in Assumption 3.8.1. Let*

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

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 $\mathcal{O}(\gamma)$. Note that R is as in Definition 3.8.2.

Proof. We apply binary search in order to estimate T_0. Suppose T is given, run the ellipsoid algorithm from Lemma 3.8.4, 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 3.8.5. If we are in the first case, by Lemma 3.8.6 we know that $\mathcal{L} \vDash_\tau Q_{T+4\gamma}$. Therefore, $T + 4\gamma \geq T_0$, 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 3.8.5 there should exists a ball of radius $r(\gamma)$. This gives us an efficient algorithm for approximating the score function. □

References

[1] 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.

[2] Ishaq Aden-Ali, Hassan Ashtiani, and Christopher Liaw. Privately learning mixtures of axis-aligned gaussians. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

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

[4] Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially private covariance estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 14190–14199. Curran Associates, Inc., 2019.

[5] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. *arXiv preprint arXiv:2111.11320*, 2021.

[6] 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.

[7] 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.

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

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

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

[11] Marco Avella-Medina and Victor-Emmanuel Brunel. Differentially private sub-Gaussian location estimators. *arXiv preprint arXiv:1906.11923*, 2019.

[12] Brendan Avent, Yatharth Dubey, and Aleksandra Korolova. The power of the hybrid model for mean estimation. *Proceedings on Privacy Enhancing Technologies*, 2020(4):48–68, 2019.

[13] Keith Ball. An elementary introduction to modern convex geometry. 1997.

[14] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*, 2016.

[15] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[16] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society.

[17] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 14475–14485. Curran Associates, Inc., 2020.

[18] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '12, pages 410–419, Washington, DC, USA, 2012. IEEE Computer Society.

[19] 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.

[20] 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.

[21] 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.

[22] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM.

[23] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

[24] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-Gaussian rates. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 786–806, 2019.

[25] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[26] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

[27] 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.

[28] 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.

[29] 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.

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

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

[32] Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 1830–1840. Curran Associates, Inc., 2020.

[33] 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.

[34] 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.

[35] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

[36] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *arXiv preprint arXiv:2001.02285*, 2020.

[37] 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.

[38] 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.

[39] 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.

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

[41] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 650–669, Washington, DC, USA, 2015. IEEE Computer Society.

[42] Marek Eliáš, Michael Kapralov, Janardhan Kulkarni, and Yin Tat Lee. Differentially private release of synthetic graphs. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, pages 560–578, Philadelphia, PA, USA, 2020. SIAM.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

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

[51] 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.

[52] Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation, 2022.

[53] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

[54] 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.

[55] Samuel B Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 11902–11912. Curran Associates, Inc., 2020.

[56] Justin Hsu, Aaron Roth, Tim Roughgarden, and Jonathan Ullman. Privately solving linear programs. In *International Colloquium on Automata, Languages, and Programming*, pages 612–624. Springer, 2014.

[57] Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[58] 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.

[59] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

[60] 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.

[61] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Annual Conference Computational Learning Theory*, 2021.

[62] 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.

[63] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated Gaussians. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 168–180. Curran Associates, Inc., 2019.

[64] 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.

[65] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[66] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1395–1414, Philadelphia, PA, USA, 2013. SIAM.

[67] 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.

[68] 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.

[69] 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.

[70] Pravesh K Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*, 2017.

[71] Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

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

[73] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 665–674, Washington, DC, USA, 2016. IEEE Computer Society.

[74] Jonathan Leake, Colin McSwiggen, and Nisheeth K Vishnoi. Sampling matrices from harish-chandra–itzykson–zuber densities with applications to quantum inference and differential privacy. In *Proceedings of the 53nd Annual ACM Symposium on the Theory of Computing*, STOC '21, pages 1384–1397, New York, NY, USA, 2021. ACM.

[75] Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

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

[77] 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.

[78] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. *arXiv preprint arXiv:2111.06578*, 2021.

[79] 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.

[80] Yuhan Liu, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: User vs item-level privacy. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20. Curran Associates, Inc., 2020.

[81] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[82] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[83] Oren Mangoubi and Nisheeth K. Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees and applications to differentially private optimization. *arXiv preprint arXiv:2111.04089*, 2021.

[84] 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.

[85] Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–1673. PMLR, 2017.

[86] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

[87] Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3389–3423. World Scientific, 2018.

[88] Prasad Raghavendra and Benjamin Weitz. On the bit complexity of sum-of-squares proofs. In *International Colloquium on Automata, Languages and Programming*, 2017.

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

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

[91] 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.

[92] 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.

[93] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 1588–1628, 2015.

[94] 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.

[95] Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ilias Zadik. Optimal private median estimation under minimal distributional assumptions. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 3301–3311. Curran Associates, Inc., 2020.

[96] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 10081–10091. JMLR, Inc., 2020.

[97] Benjamin Weitz. *Polynomial proof systems, effective derivations, and their applications in the sum-of-squares hierarchy*. University of California, Berkeley, 2017.

[98] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning Markov random fields. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 11129–11140. JMLR, Inc., 2020.