Privacy meets Robustness: Unveiling the interplay between Differential Privacy and Robustness in Machine Learning

Xiyang Liu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Sewoong Oh, Chair

Simon Du

Pang Wei Koh

Program Authorized to Offer Degree:

Computer Science and Engineering

University of Washington

Abstract

Privacy meets Robustness: Unveiling the interplay between Differential Privacy and Robustness in Machine Learning

Xiyang Liu

Chair of the Supervisory Committee:
Professor Sewoong Oh
Paul G. Allen School of Computer Science and Engineering

The rapid advancement of machine learning over the past decade has been driven by the increasing availability of large-scale datasets. However, this growth has raised critical concerns regarding the privacy of individuals whose data is being used, as well as the robustness of algorithms against potentially malicious data corruption from unreliable sources. This thesis aims to explore the fundamental interplay between differential privacy (DP) and outlier robustness in machine learning.

This thesis investigates several canonical statistical problems to uncover the inherent connections between DP and robustness. The first problem addresses whether it is possible to develop algorithms that are both differentially private and robust to outliers without requiring additional data. We present the first efficient algorithm with sub-optimal sample complexity. Then, we introduce a unifying framework that achieves nearly optimal sample complexity, without considering computational efficiency, across various problems, including mean estimation, linear regression, covariance estimation, and principal component analysis (PCA). Finally, we propose two efficient algorithms that achieve near-optimal sample complexity for differentially private PCA and linear regression.

The findings of this research contribute to a deeper understanding of the interplay between privacy and robustness, providing new insights into the design of algorithms that are both

statistically optimal and computationally efficient for practical applications. The results presented in this thesis open avenues for further exploration into the protection of data privacy, particularly in high-dimensional and adversarial settings.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I have been extremely fortunate to have Sewoong Oh as my advisor and mentor throughout my academic journey. His guidance has been invaluable, not only in teaching me how to conduct research, develop research taste, and improve my communication skills, but also in nurturing my passion for meaningful work. I vividly remember many long research meetings where his enthusiasm for research sparked my curiosity and brought me happiness. From my time at UIUC to UW, he has always been kind and supportive, both in research and in life. Sewoong's patience and understanding have had a deep impact on me, helping me grow not only as a researcher but also as a person. I am truly grateful for the chance to learn from him—he is the role model I hope to follow in my future endeavors.

I want to express my gratitude to my long-term collaborator, Weihao Kong, for his mentorship throughout this thesis. I am also thankful to my brilliant co-authors: Mohammad Vahid Jamali, Ashok Vardhan Makkuva, Hessam Mahdavifar, Pramod Viswanath, Prateek Jain, Arun Sai Suggala, Gavin Brown, Jonathan Hayase, Samuel Hopkins, Juan C. Perdomo, and Adam Smith. This work would not have been possible without their help and effort. I extend my thanks to my committee members, Simon Du, Pang Wei Koh, Sheng Wang, and Maryam Fazel, for their suggestions and support during the completion of this thesis. I also wish to thank all the professors and teachers who have inspired my curiosity at various stages of my student life. Special thanks to Jerry Li for his course on differential privacy and robustness at UW, and to Gavin and Daogao for leading the differential privacy reading group at UW.

I also want to thank all my friends at UW and UIUC who made my time here so memorable.

Finally, I want to thank my family for their unconditional love and support.

Chapter 1

INTRODUCTION

The past decade has witnessed significant advancements in machine learning, largely driven by the proliferation of large-scale datasets. However, as these datasets have scaled, concerns regarding the privacy of the individuals represented in them have become increasingly prominent. At the same time, as data is sourced from a growing number of institutions, not all of which can be deemed trustworthy, there is a corresponding rise in concerns about the robustness of algorithms against malicious data corruption. Interestingly, privacy and robustness are inherently related, as both require algorithms to be insensitive to small perturbations in the dataset. The focus of this thesis is to investigate the fundamental connections and limitations between differential privacy and outlier robustness for machine learning algorithms. This chapter introduces the problem statement within the following framework of statistical estimation.

Estimating a parameter of a distribution is a canonical problem in statistics: given i.i.d. samples $S = \{x_1, x_2, \ldots, x_n\}$ from a distribution P_θ, which belongs to a known family $\mathcal{P}$ and is indexed by an unknown parameter θ, the goal is to find an estimator $\hat{\theta}$ that minimizes the distance $\ell(\theta, \hat{\theta})$. Two important desiderata for parameter estimation algorithms are differential privacy (DP) and robustness:

Differential Privacy: Introduced by Dwork et al. [78], DP has become the de facto standard for data privacy, widely used from U.S. Census data [2] to real-world commercial systems [189, 82, 84]. Informally, an estimator is considered DP if the likelihood of the (randomized) outcome does not change significantly when a single arbitrary entry is added or removed, as formally defined in Section 1.1. This strong privacy guarantee ensures that even if an adversary knows all other entries, they cannot confidently identify whether a specific individual's data

was included in the database, thereby providing *plausible deniability* for individual privacy.

Robustness: The *strong contamination model* considers scenarios where an α-fraction of the i.i.d. samples are adversarially corrupted, as formally defined in Section 1.1. Our objective is to develop a reliable estimator that still maintains high utility. Since the 1960s, Tukey and Huber have studied Gaussian mean estimation and linear regression under weaker corruption models-Huber's model [194, 14, 115]. However, these algorithms are computationally intractable for high-dimensional problems. Only recently, in 2016, the first polynomial-time algorithm that achieves optimal robustness was proposed [64, 150]. This breakthrough has sparked a flurry of research on robust estimation problems, including mean estimation [64, 73, 105, 106, 66], covariance estimation [50, 154], linear regression and sparse regression [31, 29, 23, 93, 174, 143, 63, 155, 138, 55, 170, 70, 137], principal component analysis [145, 121].

DP and robustness are intuitively related: both concepts require an algorithm to be stable or insensitive to small changes in the input. As early as 2009, Dwork and Lei observed that robust estimators could be adapted to provide privacy through the *propose-test-release (PTR)* framework [77]. This approach builds on the intuition that robust estimators, such as the median and truncated mean, are generally less sensitive than non-robust estimators like the mean, thus providing a good starting point for private estimators. However, satisfying DP guarantees is technically more challenging due to two main reasons: 1) DP requires guarantees in the worst-case scenario—meaning the algorithm must remain insensitive within a neighborhood around every input dataset, whereas a robust algorithm only needs to handle the average case around well-behaved distributions; 2) DP requires strict adherence to distributional distances, while robust algorithms focus primarily on final utility. These technical differences pose theoretical and practical challenges for applying robust estimators to private estimation problems, leading to the result that PTR only works in low-dimensional settings and does not provide optimal privacy guarantees for many high-dimensional canonical problems, such as subGaussian mean estimation and linear regression.

Since DP appears more stringent than robustness, one might be misled into thinking that privacy ensures robustness since DP guarantees that a single outlier cannot significantly alter the estimation. This intuition is true only in low-dimensional settings; in high dimensions, each corrupted data point can appear uncorrupted but still significantly shift the parameter when colluding [160].

The interplay between DP and robustness in parameter estimation offers both significant opportunities and challenges due to their similar intuitions but differing requirements and data models. These disparities pose intriguing questions for this thesis, particularly whether it is feasible to develop algorithms that simultaneously satisfy both robustness and privacy without additional samples. Furthermore, this research explores how robust estimators can enhance privacy protections in solving many open problems related to high-dimensional private estimation. Ultimately, our goal is to design algorithms that are not only statistically optimal but also computationally efficient for practical tasks such as PCA and linear regression.

1.1 Preliminaries

We first recall the definitions of differential privacy and the strong contamination model. We say two datasets S and S' of the same size are neighboring if the Hamming distance between them is at most one.

Definition 1.1.1 ([78]). *We say a randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if for all neighboring databases $S \sim S' \in \mathcal{X}^n$, and all $Y \subseteq \mathcal{Y}$, we have $\mathbb{P}(M(S) \in Y) \leq e^{\varepsilon}\mathbb{P}(M(S') \in Y) + \delta$.*

Definition 1.1.2 (Strong Contamination Model [64]). *Given a set $S_{\text{good}} = \{\tilde{x}_i \in \mathbb{R}^d\}_{i=1}^n$ of n data points, an adversary inspects all data points, selects $\alpha_{\text{corrupt}}n$ of the data points, and replaces them with arbitrary dataset S_{bad} of size $\alpha_{\text{corrupt}}n$. The resulting dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ is called α-corrupted dataset.*

Contributions: In this thesis, we aim to develop statistically optimal algorithms for canonical

problems in parameter estimation, such as mean estimation, linear regression, and PCA under differential privacy/robustness constraints. The contributions of this work are threefold:

1. In Chapter 2, we propose the first computationally efficient algorithm that achieves both robustness and privacy simultaneously.

2. In Chapter 3, focusing only on the statistical cost without concerning computational efficiency, we introduce a generic unifying framework, *High-dimensional Propose-Test-Release (HPTR)*, which leverages robust estimators to achieve not only optimal sample complexity for many previously open problems but also optimal robustness as a byproduct.

3. In Chapter 4 and Chapter 5, we provide both time-efficient and statistically optimal private estimators for PCA and linear regression.

Bibliographies: The result of Chapter 2 for private and robust mean estimation was published at [160]. The result of Chapter 3 for HPTR original was published at [161]. The result of Chapter 4 for DP-PCA was originally published at [159]. The result of Chapter 5 for label-robust differentially private linear regression was originally published at [158].

Chapter 2

DIFFERENTIALLY PRIVATE AND ROBUST MEAN ESTIMATION

2.1 Introduction

When releasing database statistics on a collection of entries from individuals, we would ideally like to make it impossible to reverse-engineer each individual's potentially sensitive information. Privacy-preserving techniques add just enough randomness tailored to the statistical task to guarantee protection. At the same time, it is becoming increasingly common to apply such techniques to databases collected from multiple sources, not all of which can be trusted. Emerging data access frameworks, such as federated analyses across users' devices or data silos [124], make it easier to temper with this collected dataset, leaving private statistical analyses vulnerable to a malicious corruption of a fraction of the data.

Differential privacy has emerged as a widely accepted de facto measure of privacy, which is now a standard in releasing the statistics of the U.S. Census data [2] statistics and also deployed in real-world commercial systems [189, 82, 84]. A statistical analysis is said to be *differentially private* if the likelihood of the (randomized) outcome does not change significantly when a single entry is replaced by another arbitrary entry (formally defined in §2.1.1). This provides a strong privacy guarantee: even a powerful adversary who knows all the other entries in the database cannot confidently identify whether a particular individual is participating in the database based on the outcome of the analysis, providing *plausible deniability*, central to protecting an individual's privacy. Despite more than a decade of literature focused in designing private mechanisms for various statistical and learning tasks, only recently have some of the most fundamental questions been resolved.

In this work, we focus on one of the most canonical problems in statistics: estimating the

mean of a distribution from i.i.d. samples. For distributions with unbounded support, such as sub-Gaussian and heavy-tailed distributions, fundamental trade-offs between accuracy, sample size, and privacy have only recently been identified [139, 129, 135, 4] and efficient private estimators proposed. However, these approaches are brittle when a fraction of the data is corrupted, posing a real threat – referred to as *data poisoning* attacks [48, 211] – and therefore emerging as a popular setting of recent algorithmic and mathematical breakthroughs [186, 64] in the defense of such attacks.

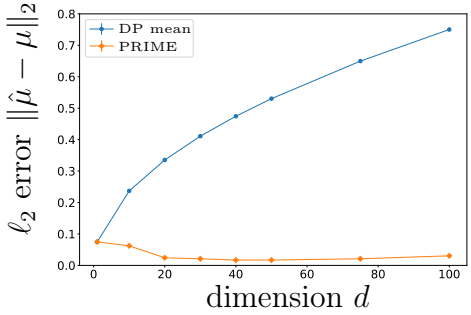

Figure 2.1: Private mean estimators (e.g., DP mean [129]) is vulnerable to adversarial corruption especially in high dimensions, while the proposed PRIME achieves robustness (and privacy) regardless of the dimension of the samples. Both are $(\varepsilon = 10, \delta = 0.01)$-DP and $\alpha = 0.05$ fraction of data is corrupted. Each data point is repeated 50 runs and standard error is shown as the error bar. Our implementation is available at `https://github.com/xiyangl3/robust_dp`.

One might be mislead to thinking that privacy ensures robustness since differential privacy guarantees that a single outlier cannot change the estimation too much. This intuition is true only in low dimensions where each sample has to be an obvious outlier to significantly change the mean. However, in high dimensions, each corrupted data point can look perfectly uncorrupted but still shift the mean significant when colluding together (Figure 2.1). Focusing on the canonical problem of mean estimation, we introduce novel algorithms that achieve

robustness and privacy simultaneously even when a fraction of data is corrupted arbitrarily by an adversary.

For such algorithms, there is a fundamental question of interest: do we need more samples to make private mean estimation also robust against adversarial corruption?

If we can afford exponential run-time in the dimension, we show that robustness can be achieved without extra cost in sample complexity. We introduce a novel estimator that nearly matches the known lower bound for a (non-robust) private mean estimation, as shown in Table 2.1. Its sole restriction is that the accuracy cannot surpass $\Omega(\alpha)$ when we have α fraction corrupted, which is necessary even for a (non-private) robust mean estimation with *infinite* samples. We nearly match this fundamental bound, achieving $O(\alpha\sqrt{\log(1/\alpha)})$ accuracy with an information theoretically optimal sample complexity.

Theorem 1 (Informal version of Theorem 8, exponential time algorithm for sub-Gaussian distributions)**.** *When α fraction of the data is arbitrarily corrupted from n samples drawn from a d-dimensional sub-Gaussian distribution with mean μ and an identity sub-Gaussian parameter, if $n = \widetilde{\Omega}(d/\alpha^2 + (d + d^{1/2}\log(1/\delta))/(\alpha\varepsilon))$ then Algorithm 8 is (ε, δ)-differentially private and achieves $\|\hat{\mu} - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ with high probability.*

We introduce PRIME (PRIvate and robust Mean Estimation) in Algorithm 5 to make robust and private mean estimation computationally efficient. It requires a run-time of only $\widetilde{O}(d^3 + nd^2)$, but at the cost of requiring extra $d^{1/2}$ factor larger number of samples. This cannot be improved upon with current techniques since efficient robust estimators rely on the top PCA directions of the covariance matrix to detect outliers. [209] showed that $\widetilde{\Omega}(d^{3/2})$ samples are necessary to compute PCA directions while preserving (ε, δ)-differential privacy. It remains an open question if this $\widetilde{\Omega}(d^{3/2}/(\alpha\varepsilon))$ bottleneck is fundamental because no matching lower bound is currently known for the differentially private robust mean estimation problem.

Theorem 2 (Informal version of Theorem 7, efficient algorithm for sub-Gaussian distributions)**.** *Under the assumption of Theorem 1, if $n = \widetilde{\Omega}(d/\alpha^2 + (d^{3/2}\log(1/\delta))/(\alpha\varepsilon))$ then PRIME is*

(ε, δ)-differentially private and achieves $\|\hat{\mu} - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ with high probability.

	Upper bound (poly-time)	Upper bound (exp-time)	Lower bound
(ε, δ)-DP [36, 129]	$\widetilde{O}(\frac{d}{\alpha^2} + \frac{d\log^{1/2}(1/\delta)}{\alpha\varepsilon})$	$\widetilde{O}(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon})^{\clubsuit}$	$\widetilde{\Omega}(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon})^{\spadesuit}$
α-corruption [73]	$\widetilde{O}(\frac{d}{\alpha^2})$	$\widetilde{O}(\frac{d}{\alpha^2})$	$\Omega(\frac{d}{\alpha^2})$
α-corruption and (ε, δ)-DP (this work)	$\widetilde{O}\big(\frac{d}{\alpha^2} + \frac{d^{3/2}\log(1/\delta)}{\alpha\varepsilon}\big)$ [Theorem 7]	$\widetilde{O}(\frac{d}{\alpha^2} + \frac{d+d^{1/2}\log(1/\delta)}{\alpha\varepsilon})$ [Theorem 8]	$\widetilde{\Omega}(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon})^{\spadesuit}$ [129]

Table 2.1: For the fundamental task of learning the mean $\mu \in [-R, R]^d$ of a *sub-Gaussian* distribution with a known covariance, we list the sufficient or necessary conditions on the sample sizes to achieve an error $\|\hat{\mu} - \mu\|_2 = \widetilde{O}(\alpha)$ under (ε, δ)-differential privacy (DP), corruption of an α-fraction of samples, and both. $^{\clubsuit}$ requires the distribution to be a Gaussian and $^{\spadesuit}$ requires $\delta \leq \sqrt{d}/n$.

When the samples are drawn from a distribution with a bounded covariance, parameters of Algorithm 8 can be modified to nearly match the optimal sample complexity of (non-robust) private mean estimation in Table 2.2. This algorithm also matches the fundamental limit on the accuracy of (non-private) robust estimation, which in this case is $\Omega(\alpha^{1/2})$.

Theorem 3 (Informal version of Theorem 9, exponential time algorithm for covariance bounded distributions)**.** *When α fraction of the data is arbitrarily corrupted from n samples drawn from a d-dimensional distribution with mean μ and covariance $\Sigma \preceq \mathbf{I}$, if $n = \widetilde{\Omega}((d + d^{1/2}\log(1/\delta))/(\alpha\varepsilon) + d^{1/2}\log^{3/2}(1/\delta)/\varepsilon)$ then Algorithm 8 is (ε, δ)-differentially private and achieves $\|\hat{\mu} - \mu\|_2 = O(\alpha^{1/2})$ with high probability.*

The proposed PRIME-HT for covariance bounded distributions achieve computational efficiency at the cost of an extra factor of $d^{1/2}$ in sample size. This bottleneck is also due to DP-PCA, and it remains open whether this gap can bee closed by an efficient estimator.

Theorem 4 (Informal version of Theorem 10, efficient algorithm for covariance bounded distributions)**.** *Under the assumptions of Theorem 3, if* $n = \widetilde{\Omega}((d^{3/2}\log(1/\delta))/(\alpha\varepsilon))$ *then* PRIME-HT *is* (ε, δ)*-differentially private and achieves* $\|\hat{\mu} - \mu\|_2 = O(\alpha^{1/2})$ *with high probability.*

	Upper bound (poly-time)	Upper bound (exp-time)	Lower bound
(ε, δ)-DP [25, 135]	$\widetilde{O}(\frac{d\log^{1/2}(1/\delta)}{\alpha\varepsilon})$	$\widetilde{O}(\frac{d}{\alpha\varepsilon})$	$\Omega(\frac{d}{\alpha\varepsilon})$
α-corruption [73]	$\widetilde{O}(\frac{d}{\alpha})$	$\widetilde{O}(\frac{d}{\alpha})$	$\Omega(\frac{d}{\alpha})$
α-corruption and (ε, δ)-DP (this chapter)	$\widetilde{O}(\frac{d^{3/2}\log(1/\delta)}{\alpha\varepsilon})$ [Theorem 10]	$\widetilde{O}(\frac{d+(d^{1/2}+\alpha\, d^{1/2}\log^{1/2}(1/\delta))\log(1/\delta)}{\alpha\varepsilon})$ [Theorem 9]	$\Omega(\frac{d}{\alpha\varepsilon})$ [25, 135]

Table 2.2: For the fundamental task of learning the mean $\mu \in [-R, R]^d$ of a covariance bounded distribution, we list the sufficient or necessary conditions on the sample size to achieve an error $\|\hat{\mu} - \mu\|_2 = O(\alpha^{1/2})$ under (ε, δ)-differential privacy (DP), corruption of an α-fraction of samples, and both.

Contributions. We introduce a novel *private and robust mean estimator* that achieves optimal guarantees but takes an *exponential* run-time (Algorithm 8). Its innovation is leveraging on the *resilience* property of well-behaved distributions not only to estimate the mean robustly (which is the standard use of the property) but also to adaptively bound the sensitivity of the estimator, thus achieving optimal privacy.

We introduce new *efficient private and robust mean estimators* (Algorithms 5 and 9). These algorithms critically rely on the private version of matrix multiplicative weight filtering to limit the number of dataset accesses to its minimum. Finally, we present a novel *private 1-dimensional filter* that uses a private histogram on a set of carefully chosen intervals to guarantee improvement in each step.

2.1.1 Preliminaries

Differential privacy provides a formal mathematical metric for measuring privacy leakage when a dataset is accessed with a query.

Definition 2.1.1 (Differential privacy [78]). *Given two datasets $S = \{x_i\}_{i=1}^n$ and $S' = \{x'_i\}_{i=1}^n$, we say S and S' are* neighboring *if they differ in at most one entry, which is denoted by $S \sim S'$. For an output of a stochastic query q on a database, we say q satisfies* (ε, δ)-differential privacy *for some $\varepsilon > 0$ and $\delta \in (0, 1)$ if $\mathbb{P}(q(S) \in A) \leq e^{\varepsilon}\mathbb{P}(q(S') \in A) + \delta$ for all neighboring databases $S \sim S'$ and all subset A in the range of the query.*

By introducing enough randomness when answering a query, we can achieve small values of ε and δ (and hence strong privacy). This ensures that the query output does not reveal whether a single person participated in the dataset or not with high confidence, to a powerful adversary who knows all the other entries of the dataset. The main building block of our proposed algorithms is output perturbation. Let $z \sim \mathrm{Lap}(b)$ denote a random vector whose entries are i.i.d. sampled from Laplace distribution with pdf $(1/2b)e^{-|z|/b}$. Let $z \sim \mathcal{N}(\mu, \Sigma)$ denote a Gaussian random vector with mean μ and covariance Σ.

Definition 2.1.2 (Output perturbation). *The* sensitivity *of a non-private query $f(S) \in \mathbb{R}^k$ is defined as $\Delta_p = \sup_{S \sim S'} \|f(S) - f(S')\|_p$ for a norm $\|x\|_p = (\sum_{i \in [k]} |x_i|^p)^{1/p}$. For $p = 1$, the Laplace mechanism outputs $f(S) + \mathrm{Lap}(\Delta_1/\varepsilon)$ and achieves $(\varepsilon, 0)$-differential privacy [78]. For $p = 2$, the Gaussian mechanism outputs $f(S) + \mathcal{N}(0, (\Delta_2(\sqrt{2\log(1.25/\delta)})/\varepsilon)^2 \mathbf{I})$ and achieves (ε, δ)-differential privacy [79].*

Other output perturbation mechanisms include the exponential mechanism [164] (which we explain in detail in §2.2) and staircase mechanisms [94, 125] (which achieves the minimum variance).

Private statistical analysis. Traditional private data analyses require bounded support of the samples to leverage the resulting bounded sensitivity. For example, each entry is constrained to have finite ℓ_2 norm in standard private principal component analysis [46],

which does not apply to Gaussian samples. Fundamentally departing from these approaches, [139] first established an optimal mean estimation of Gaussian samples with *unbounded* support. The breakthrough is in first adaptively estimating the range of the data using a private histogram, thus bounding the support and the resulting sensitivity. This spurred the design of private algorithms for high-dimensional mean and covariance estimation [129, 32], heavy-tailed mean estimation [135], learning mixture of Gaussian [134], learning Markov random fields [214], and statistical testing [42]. Under the Gaussian distribution with no adversary, [4] achieves an accuracy of $\|\hat{\mu} - \mu\|_2 \leq \tilde{\alpha}$ with the best known sample complexity of $n = \widetilde{O}((d/\tilde{\alpha}^2) + (d/\tilde{\alpha}\varepsilon) + (1/\varepsilon)\log(1/\delta))$ while guaranteeing (ε, δ)-differential privacy. This nearly matches the known lower bounds of $\Omega(d/\tilde{\alpha}^2)$ for non-private finite sample complexity, $\widetilde{\Omega}((1/\varepsilon)\min\{\log(1/\delta), \log(R)\})$ for privately learning one-dimensional unit variance Gaussian [139], and $\widetilde{\Omega}(d/\tilde{\alpha}\varepsilon)$ for multi-dimensional Gaussian estimation [129]. However, this does not generalize to sub-Gaussian distributions and [4] does not provide a tractable algorithm. A polynomial time algorithm is proposed in [129] that achieves a slightly worse sample complexity of $\widetilde{O}((d/\tilde{\alpha}^2) + (d\log^{1/2}(1/\delta)/\tilde{\alpha}\varepsilon))$, which can also seamlessly generalized to sub-Gaussian distributions. For estimating the mean of a *covariance bounded* distributions up to an accuracy of $\|\hat{\mu} - \mu\|_2 = O(\tilde{\alpha}^{1/2})$, [135] shows that $\Omega(d/(\tilde{\alpha}\varepsilon))$ samples are necessary and provides an efficient algorithm matching this up to a factor of $\log^{1/2}(1/\delta)$. In the same paper, an inefficient algorithm based on the exponential mechanism with a tournament-based scoring is proposed, that achieves the optimal sample complexity with pure $(\varepsilon, \delta = 0)$-DP. It might be possible to extend this approach to design an robust and DP mean estimator (with exponential run-time).

With a similar motivation as this work but for a different problem of learning half-spaces robustly and privately, [96] provides the fundamental limits on the sample complexity and proposes efficient algorithms matching those information theoretic lower bounds. The approach is a variation of the margin perceptron algorithm, and uses batch sampling together with Laplace and Gaussian mechanisms.

Robust estimation. Designing robust estimators under the presence of outliers has been considered by statistics community since 1960s [194, 14, 115]. Recently, [62, 150] give the first polynomial time algorithm for mean and covariance estimation with no (or very weak) dependency on the dimensionality in the estimation error. Since then, there has been a flurry of research on robust estimation problems, including mean estimation [64, 73, 105, 106, 66], covariance estimation [50, 154], linear regression and sparse regression [31, 29, 23, 93, 174, 143, 63, 155, 138, 55, 170, 70, 137], principal component analysis [145, 121], mixture models [61, 122, 147, 110] and list-decodable learning [69, 175, 43, 21, 52]. See [67] for a survey of recent work.

One line of work that is particularly related to our algorithm PRIME is [49, 73, 57, 50, 52], which leverages the ideas from matrix multiplicative weight and fast SDP solver to achieve faster, sometimes nearly linear time, algorithms for mean and covariance estimation. In PRIME, we use a matrix multiplicative weight approach similar to [73] to reduce the iteration complexity to logarithmic, which enables us to achieve the $d^{3/2}$ dependency in the sample complexity.

The concept of *resilience* is introduced in [186] as a sufficient condition such that learning in the presence of adversarial corruption is information-theoretically possible. The idea of resilience is later generalized in [217] for a wider range of adversarial corruption models. While there exists a simple exponential time robust estimation algorithm under resilience conditions, it is challenging to achieve differential privacy due to high sensitivity. We propose a novel approach to leverage the resilience property in our exponential time algorithm for sub-gaussian and heavy-tailed distributions.

2.1.2 Problem formulation

We are given n samples from a sub-Gaussian distribution with a known covariance but unknown mean, and α fraction of the samples are corrupted by an adversary. Our goal is to estimate the unknown mean. We follow the standard definition of adversary in [64], which can adaptively choose which samples to corrupt and arbitrarily replace them with any data points.

The main challenge is in achieving the near optimal accuracy of $\|\hat{\mu}(S)-\mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ while preserving (ε, δ)-differential privacy.

Assumption 1 (α-corrupted sub-Gaussian model)**.** *An uncorrupted dataset S_{good} consists of n i.i.d. samples from a d-dimensional sub-Gaussian distribution with mean $\mu \in [-R, R]^d$ and covariance $\mathbb{E}[xx^\top] = \mathbf{I}_d$, which is 1-sub-Gaussian, i.e., $\mathbb{E}[\exp(v^\top x)] \le \exp(\|v\|_2^2/2)$. For some $\alpha \in (0, 1/2)$, we are given a corrupted dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ where an adversary adaptively inspects all the samples in S_{good}, removes αn of them, and replaces them with S_{bad} which are αn arbitrary points in $\mathbb{R}^d$.*

Similarly, we consider the same problem for heavy-tailed distributions under the bounded covariance assumption.

Assumption 2 (α-corrupted bounded covariance model)**.** *An uncorrupted dataset S_{good} consists of n i.i.d. samples from a distribution D with mean $\mu \in [-R, R]^d$ and covariance $\Sigma \preceq \mathbf{I}$. For some $\alpha \in (0, 1/2)$, we are given a corrupted dataset $S = \{x_i\}_{i=1}^n$ where an adversary adaptively inspects all the samples in S_{good}, removes αn of them and replaces them with S_{bad} which are αn arbitrary points in $\mathbb{R}^d$.*

Notations. Let $[n] = \{1, 2, \ldots, n\}$. For $x \in \mathbb{R}^d$, we use $\|x\|_2 = (\sum_{i\in[d]}(x_i)^2)^{1/2}$ to denote the Euclidean norm. For $X \in \mathbb{R}^{d\times d}$, we use $\|X\|_2 = \max_{\|v\|_2=1} \|Xv\|_2$ to denote the spectral norm. The $d \times d$ identity matrix is $\mathbf{I}_d$. Whenever it is clear from context, we use S to denote both a set of data points and also the set of indices of those data points. $\widetilde{O}$ and $\widetilde{\Omega}$ hide poly-logarithmic factors in $d, n, 1/\alpha$, R, and the failure probability.

Outline. We present our results for sub-Gaussian distribution first. We provide a background on existing approaches in §2.2. We introduce an efficient algorithm for mean estimation in §2.3. We then introduce an exponential time algorithm with near optimal guarantee in §2.4. Analogous results for heavy-tailed distributions are presented in in §2.5.

2.2 Background on exponential time approaches for Gaussian distributions

In this section, we provide a background on exponential time algorithms that achieve optimal guarantees but only applies to and heavily relies on the assumption that samples are drawn from a *Gaussian* distribution. In §2.4, we introduce a novel exponential time approach that seamlessly generalizes to both sub-Gaussian and covariance-bounded distributions.

We introduce Algorithm 1, achieving the optimal sample complexity of $\widetilde{O}(d/\min\{\alpha\varepsilon, \alpha^2\})$ (Theorem 5). The main idea is to find an approximate Tukey median (which is known to be a robust estimate of the mean [218]), using the exponential mechanism of [164] to preserve privacy.

Tukey median set. For any set of points $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ and $\hat{\mu} \in \mathbb{R}^d$, the *Tukey depth* is defined as the minimal empirical probability density on one side of a hyperplane that includes $\hat{\mu}$:

$$D_{\text{Tukey}}(S, \hat{\mu}) \;=\; \inf_{v \in \mathbb{R}^d} \mathbb{P}_{x \sim \hat{p}_n}(v^\top (x - \hat{\mu}) \geq 0) \;,$$

where $\hat{p}_n$ is the empirical distribution of S. The *Tukey median set* is defined as the set of points achieving the maximum Tukey depth, which might not be unique. Tukey median reduces to median for $d = 1$, and is a natural generalization of the median for $d > 1$. Inheriting robustness of one-dimensional median, Tukey median is known to be a robust estimator of the multi-dimensional mean under an adversarial perturbation. In particular, under our model, it achieves the optimal sample complexity and accuracy. This optimality follows from the well-known fact that the sample complexity of $O((1/\alpha^2)(d + \log(1/\zeta)))$ cannot be improved upon even if we have no corruption, and the fact that the accuracy of $O(\alpha)$ cannot be improved upon even if we have infinite samples [218]. However, finding a Tukey median takes exponential time scaling as $\tilde{O}(n^d)$ [157].

Corollary 2.2.1 (Corollary of [218, Theorem 3])**.** *For a dataset of n i.i.d. samples from a d-dimensional Gaussian distribution $\mathcal{N}(\mu, \mathbf{I}_d)$, an adversary corrupts an $\alpha \in (0, 1/4)$ fraction*

of the samples as defined in Assumption 1. Then, any $\hat{\mu}$ in the Tukey median set of a corrupted dataset S satisfies $\|\hat{\mu}-\mu\|_2 = O(\alpha)$ with probability at least $1-\zeta$ if $n = \Omega((1/\alpha^2)(d+\log(1/\zeta)))$.

Exponential mechanism. The exponential mechanism was introduced in [164] to elicit approximate truthfulness and remains one of the most popular private mechanisms due to its broad applicability. It can seamlessly handle queries with non-numeric outputs, such as routing a flow or finding a graph. Consider a utility function $u(S, \hat{\mu}) \in \mathbb{R}$ on a dataset S and a variable $\hat{\mu}$, where higher utility is preferred. Instead of truthfully outputting $\arg\max_{\hat{\mu}} u(S, \hat{\mu})$, the exponential mechanism outputs a randomized approximate maximizer sampled from the following distribution:

$$r_S(\hat{\mu}) \;=\; \frac{1}{Z_S} e^{\frac{\varepsilon}{2\,\Delta_u} u(S,\hat{\mu})}\,, \tag{2.1}$$

where $\Delta_u = \max_{\hat{\mu}, S\sim S'} |u(S,\hat{\mu}) - u(S',\hat{\mu})|$ is the sensitivity of u (from Definition 2.1.2) and Z_S ensures normalization to one. This mechanism is $(\varepsilon, 0)$-differentially private, since $e^{\frac{\varepsilon}{2\Delta_u}|u(S,\hat{\mu})-u(S',\hat{\mu})|} \le e^{\varepsilon/2}$ and $e^{-\varepsilon/2} \le Z_S/Z_{S'} \le e^{\varepsilon/2}$.

Proposition 2.2.2 ([164, Theorem 6])**.** *The sampled $\hat{\mu}$ from the distribution (2.1) is $(\varepsilon, 0)$-differentially private.*

This naturally leads to the following algorithm. The privacy guarantee follows immediately since the Tukey depth has sensitivity $1/n$, i.e., $|\, D_{\text{Tukey}}(S_n, \hat{\mu}) - D_{\text{Tukey}}(S'_n, \hat{\mu})\,| \le 1/n$ for all $\hat{\mu} \in \mathbb{R}^d$ and two neighboring databases $S_n \sim S'_n$ of size n.

Algorithm 1: Private Tukey median

1 Output a random data point $\hat{\mu} \in [-2R, 2R]^d$ sampled from a density $r(\hat{\mu}) \;\propto\; e^{(1/2)\varepsilon n D_{\text{Tukey}}(S,\hat{\mu})}$.

The private Tukey median achieves the following near optimal guarantee, whose proof is provided in §A.1. The accuracy of $O(\alpha)$ and sample complexity of $n = \Omega((1/\alpha^2)(d+\log(1/\zeta)))$ cannot be improved even without privacy (cf. Corollary 2.2.1), and $n = \tilde{\Omega}(d/(\alpha\varepsilon))$ is necessary even without any corruption [129, Theorem 6.5].

Theorem 5. *Under the hypotheses of Corollary 2.2.1, there exists a universal constant* $c > 0$ *such that if* $\mu \in [-R, R]^d$, $\alpha \leq \min\{c, R\}$ *and* $n = \Omega((1/\alpha^2)(d + \log(1/\zeta)) + (1/\alpha\varepsilon)d\log(dR/\zeta\alpha))$, *then Algorithm 1 is* $(\varepsilon, 0)$*-differentially private and achieves* $\|\hat{\mu} - \mu\|_2 = O(\alpha)$ *with probability* $1 - \zeta$.

The private Tukey median, however, is a conceptual algorithm since we cannot sample from $r(\hat{\mu})$. The $\mathcal{A}_{FindTukey}$ algorithm from [28] approximately finds the Tukey median privately. This achieves $O(\alpha)$ accuracy with $n = \tilde{\Omega}(d^{3/2}\log(1/\delta)/(\alpha\varepsilon) + (1/\alpha^2)(d + \log(1/\zeta)))$, but it still requires a runtime of $O(n^{\text{poly}(d)})$. Alternatively, we can sample from an α-cover of $[-2R, 2R]^d$, which has $O((dR/\alpha)^d)$ points. However, evaluating the Tukey depth of a point is an NP-hard problem [9], requiring a runtime of $\tilde{O}(n^{d-1})$ [156]. The runtime of the discretized private Tukey median is $\tilde{O}(n^{-1}(dnR/\alpha)^d)$. Similarly, [36] introduced an exponential mechanism over the α-cover with a novel utility function achieving the same guarantee as Theorem 5, but this requires a runtime of $O(n(dR/\alpha)^{2d})$.

2.3 Efficient algorithms for private and robust mean estimation

A major challenge in making a robust estimation algorithm private is the high sensitivity of the iterates as we show in §2.3.1. Instead, we propose making only the first and second order statistics private, hence significantly reducing the sensitivity in §2.3.2. However, the $O(d)$ number of iterations is prohibitive because the privacy leakage compounds over those iterations. We therefore propose PRIME (PRIvate and robust Mean Estimation), which uses a matrix multiplicative weights approach to reduce the number of iterations down to $O((\log d)^2)$; see §2.3.3.

Algorithm 2: Non-private robust mean estimation [153]

Input: $S = \{x_i\}_{i=1}^n$, $\alpha \in (0,1)$, $S_0 = [n]$

for $t = 1, \ldots$ **do**

  if $\|\sum_{i \in S_{t-1}} (x_i - \mu_{t-1})(x_i - \mu_{t-1})^\top - \mathbf{I}\|_2 < C\alpha \log(1/\alpha)$ **then**

    Output: $\hat{\mu} = \sum_{i \in S_{t-1}} x_i$

  else

    $\mu_t \leftarrow (1/|S_{t-1}|) \sum_{i \in S_{t-1}} x_i$

    $v_t \leftarrow$ 1st principal direction of $(\{(x_i - \mu_t)\}_{i \in S_{t-1}})$

    $Z_t \leftarrow \mathrm{Unif}([0,1])$

    $S_t \leftarrow S_{t-1} \setminus \{i \,|\, i \in \mathcal{T}_{2\alpha}$ for $\{\tau_j = (v_t^\top (x_j - \mu_t))^2\}_{j \in S_{t-1}}$ and $\tau_i \geq Z_t \max_{j \in S_{t-1}} (v_t^\top (x_j - \mu_t))^2\}$, where $\mathcal{T}_{2\alpha}$ is defined in Definition 2.3.1.

2.3.1 Background on robust mean estimation

Non-private robust mean estimation approaches, such as Algorithm 2, recursively apply a *filter*: $S_t = F(S_{t-1})$. Given a dataset $S = \{x_i\}_{i=1}^n$, the set $S_t \subseteq [n]$ is updated starting with $S_0 = [n]$. At each step, the filter attempts to detect the corrupted data points and remove them. The filter focuses on the direction of the current principal component and removes data points with probability proportional to their variance, but only does so for those in the largest $n\alpha$ subset of remaining points, defined as follows. The tie-breaking rule is not essential for robust estimation, but is critical for proving differential privacy, as shown in §A.5.1.

Definition 2.3.1 (Subset of the largest α fraction)**.** *Given a set of scalar values* $\{\tau_i = \langle V, (x_i - \mu)(x_i - \mu)^\top \rangle\}_{i \in S'}$ *for a subset* $S' \subseteq [n]$*, define the sorted list* π *of* S' *such that* $\tau_{\pi(i)} \geq \tau_{\pi(i+1)}$ *for all* $i \in [|S'| - 1]$*. When there is a tie such that* $\tau_i = \tau_j$*, it is broken by* $\pi^{-1}(i) \leq \pi^{-1}(j) \Leftrightarrow x_{i,1} \geq x_{j,1}$*. Further ties are broken by comparing the remaining entries of* x_i *and* x_j*, in an increasing order of the coordinate. If* $x_i = x_j$ *,then the tie is broken arbitrarily. We define* $\mathcal{T}_\alpha = \{\pi(1), \ldots, \pi(\lceil n\alpha \rceil)\}$ *to be the set of largest* $\lceil n\alpha \rceil$ *valued samples.*

Removing data points with probability proportional to their variance ensures that we

remove more corrupted samples than the clean samples, while reducing the covariance. Hence, we do not remove more than αn clean samples (on average) before removing all the corrupted ones. When the covariance is sufficiently reduced (line 2 in Algorithm 2), the following key technical lemma ensures that our estimate is accurate.

Lemma 2.3.2 (Corollary of [73, Lemma 4.6]). *Under Assumption 1, if* $n = \Omega((d + \log(1/\zeta))/(\alpha^2 \log(1/\alpha)))$, *then with probability* $1 - \zeta$ *we have*

$$\|\mu(T) - \mu\|_2 \;=\; O\left(\sqrt{\alpha\left(\|M(T) - \mathbf{I}\|_2 + \alpha \log(1/\alpha)\right)} + \alpha\sqrt{\log(1/\alpha)}\right) ,$$

for any $T \subseteq S$ *such that* $\left(n - |T \cap S_{\text{good}}|\right) = O(\alpha n)$, *where* $M(T) \triangleq (1/n)\sum_{i\in T}(x_i - \mu(T))(x_i - \mu(T))^\top$, $\mu(T) \triangleq (1/|T|)\sum_{i\in T} x_i$, S *denotes the entire (corrupted) dataset, and* S_{good} *is the original set of clean data, as defined in Assumption 1.*

Using this lemma, we can show that this algorithm achieves the near-optimal sample complexity that nearly matches that of Corollary 2.2.1 up to a $\log(1/\alpha)$ factor.

Proposition 2.3.3 (Corollary of [153, Theorem 2.1]). *Under assumption 1, Algorithm 2 achieves accuracy* $\|\hat{\mu} - \mu\|_2 \le O(\alpha\sqrt{\log(1/\alpha)})$ *with probability* 0.9 *if* $n \ge \widetilde{\Omega}(d/\alpha^2)$.

To get a differentially private robust mean, a naive attempt is to apply a standard output perturbation mechanism to $\hat{\mu}$. However, this is challenging since the end-to-end sensitivity is intractable. The standard recipe to circumvent this is to make the current "state" S_t private at every iteration. Once S_{t-1} is private (hence, public knowledge), making the next "state" S_t private is simpler. We only need to analyze the sensitivity of a single step and apply some output perturbation mechanism with $(\varepsilon_t, \delta_t)$. End-to-end privacy is guaranteed by accounting for all these $(\varepsilon_t, \delta_t)$'s using advanced composition [127]. This recipe has been quite successful, for example, in training neural networks with (stochastic) gradient descent [182, 1, 124], where the current state can be the optimization variable $\mathbf{x}_t$. However, for Algorithm 2, this standard recipe fails, since in our case, state S_t is a set and has large sensitivity. Changing a single data point in S_{t-1} can significantly alter which (and how many) samples are filtered out at that step because the principal direction v_t can change dramatically.

2.3.2 Making the mean and the principal component private

To reduce the sensitivity of intermediate iterates in Algorithm 2, we propose making private only the mean μ_t and the top principal direction v_t. To this end, we introduce DPRANGE and DPFILTER in Algorithm 3, which achieves the following guarantee. This follows from Lemmas 2.3.5 and 2.3.7, and for completeness we provide a proof in §A.3.4.

Theorem 6. *Algorithm 3 is (ε, δ)-differentially private if $n = \tilde{\Omega}((T^{1/2}/\varepsilon)(\log(1/\delta))^{3/2})$. Under Assumption 1, there exists a universal constant $c \in (0, 0.1)$ such that if $\alpha \le c$, $n = \widetilde{\Omega}\left((d/\alpha^2) + d^{2.5}\log(1/\delta)/(\varepsilon\alpha^{1.5})\right)$ and $T = \widetilde{\Theta}(d^2/\alpha)$, then Algorithm 3 achieves $\|\hat{\mu} - \mu\|_2 \le O(\alpha\sqrt{\log(1/\alpha)})$ with probability 0.9.*

The first term $O(d/\alpha^2)$ in the sample complexity is optimal (cf. Theorem 5), but there is a factor of d gap in the second term because DPFILTER runs for $O(d)$ iterations in the worst-case. According to the advanced composition (Lemma 2.3.4), each iteration is allowed only a privacy budget of only $(O(\varepsilon/\sqrt{d\log(1/\delta)}), O(\delta/d))$ to ensure the end-to-end guarantee of $(0.99\varepsilon, 0.99\delta)$-DP (line 3, Algorithm 3). Therefore, we introduce DPMMWFILTER in §2.3.3 to reduce the number of iterations to $O((\log d)^2)$ and significantly decrease sample complexity.

Lemma 2.3.4 (Composition theorem of [127, Theorem 3.4])**.** *For $\varepsilon \le 0.9$, an end-to-end guarantee of (ε, δ)-differential privacy is satisfied if a dataset is accessed k times, each with a $(\varepsilon/2\sqrt{2k\log(2/\delta)}, \delta/2k)$-differential private mechanism.*

Algorithm 3: Private iterative filtering

Input: $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$, range $[-R, R]^d$, adversarial fraction $\alpha \in (0, 1/2)$, target probability $\eta \in (0, 1)$, number of iterations $T \in \mathbb{Z}_+$, target privacy (ε, δ)

1 $(\bar{x}, B) \leftarrow$ DPRANGE$(\{x_i\}_{i=1}^n, R, 0.01\varepsilon, 0.01\delta)$ [Algorithm 14]

2 Clip the data points: $\tilde{x}_i \leftarrow \mathcal{P}_{\bar{x}+[-B/2,B/2]^d}(x_i)$, for all $i \in [n]$

3 $\hat{\mu} \leftarrow$ DPFILTER$(\{\tilde{x}_i\}_{i=1}^n, \alpha, T, 0.99\varepsilon, 0.99\delta)$ [Algorithm 4]

Output: $\hat{\mu}$

2.3.2.1 Proof sketch and algorithm detail

DPRANGE, introduced in [139], returns a hypercube $\bar{x} + [-B, B]^d$ that is guaranteed to include all uncorrupted samples, while preserving privacy. In the following lemma, we show that DPRANGE is also *robust* to adversarial corruption. Such adaptive bounding of the support is critical in privacy analysis of the subsequent steps. We clip all data points by projecting all the points with $\mathcal{P}_{\bar{x}+[-B/2,B/2]^d}(x) = \arg\min_{y \in \bar{x}+[-B/2,B/2]^d} \|y - x\|_2$ to lie inside the hypercube and pass them to DPFILTER for filtering. The algorithm and a proof are provided in §A.2. Perhaps surprisingly, there is no dependence in R for $R > 1/\delta$, which is achieved by utilizing the private histogram mechanism from [198, 37].

Lemma 2.3.5. DPRANGE$(S, R, \varepsilon, \delta)$ *(Algorithm 14 in §A.2) is* (ε, δ)*-differentially private. Under Assumption 1,* DPRANGE$(S, R, \varepsilon, \delta)$ *returns* $(\bar{x}, B)$ *such that if* $n = \Omega\left((\sqrt{d\log(1/\delta)}/\varepsilon)\min\left(\log(dR/\zeta), \log(d/\zeta\delta)\right)\right)$ *and* $\alpha < 0.1$*, then all uncorrupted samples in* S *are in* $\bar{x} + [-B, B]^d$ *with probability* $1 - \zeta$*.*

Algorithm 4: Differentially private filtering (DPFILTER)

Input: $S = \{x_i \in \bar{x} + [-B/2, B/2]^d\}_{i=1}^n$, $\alpha \in (0, 1/2)$, $T = \widetilde{\Theta}(dB^2)$, (ε, δ)

$S_0 \leftarrow [n]$, $\varepsilon_1 \leftarrow \min\{\varepsilon, 0.9\}/(4\sqrt{2T\log(2/\delta)})$, $\delta_1 \leftarrow \delta/(8T)$

for $t = 1, \ldots, T$ **do**

  $n_t \leftarrow |S_{t-1}| + \mathrm{Lap}(1/\varepsilon_1)$

  if $n_t < 3n/4$ **then**

    terminate

  $\mu_t \leftarrow (1/|S_{t-1}|)\sum_{i\in S_{t-1}} x_i + \mathrm{Lap}(2B/(n\,\varepsilon_1))$

  $\lambda_t \leftarrow \|(1/n)\sum_{i\in S_{t-1}}(x_i - \mu_t)(x_i - \mu_t)^\top - \mathbf{I}\|_2 + \mathrm{Lap}(2B^2 d/(n\varepsilon_1))$

  if $\lambda_t \le (C - 0.01)\alpha\log(1/\alpha)$ **then**

    Output: μ_t

  $v_t \leftarrow$ top singular vector of $\Sigma_{t-1} \triangleq$

  $\frac{1}{n}\sum_{i\in S_{t-1}}(x_i - \mu_t)(x_i - \mu_t)^\top + \mathcal{N}(0, (B^2 d\sqrt{2\log(1.25/\delta)}/(n\varepsilon_1))^2 \mathbf{I}_{d^2\times d^2})$

  $Z_t \leftarrow \mathrm{Unif}([0,1])$

  $S_t \leftarrow S_{t-1} \setminus \{i \,|\, i \in \mathcal{T}_{2\alpha}$ for $\{\tau_j = (v_t^\top(x_j - \mu_t))^2\}_{j\in S_{t-1}}$ and $\tau_i \ge d\,B^2\,Z_t\}$, where $\mathcal{T}_{2\alpha}$ is defined in Definition 2.3.1.

In DPFILTER, we make only the mean μ_t and the top principal direction v_t private to decrease sensitivity. The analysis is now more challenging since (μ_t, v_t) depends on all past iterates $\{(\mu_j, v_j)\}_{j=1}^{t-1}$ and internal randomness $\{Z_j\}_{j=1}^{t-1}$. To decrease the sensitivity, we modify the filter in line 11 to use the maximum support dB^2 (which is data independent) instead of the maximum contribution $\max_i (v_t^\top(x_i - \mu_t))^2$ (which is data dependent and sensitive). While one data point can significantly change $\max_i (v_t^\top(x_i - \mu_t))^2$ and the output of one step of the filter in Algorithm 2, the sensitivity of the proposed filter is bounded conditioned on all past $\{(\mu_j, v_j)\}_{j=1}^{t-1}$, as we show in the following lemma. This follows from the fact that conditioned on (μ_j, v_j), the proposed filter is a contraction. We provide a proof in §A.3.

Lemma 2.3.6. *Let $S_t(\mathcal{S})$ denote the resulting subset of samples after t iterations of the filtering in* DPFILTER *are applied to a dataset $\mathcal{S}$ using* fixed *parameters $\{(\mu_j, v_j, Z_j)\}_{j=1}^t$. Then, we have $d_\triangle(S_t(\mathcal{S}), S_t(\mathcal{S}')) \le d_\triangle(\mathcal{S}, \mathcal{S}')$, where $d_\triangle(\mathcal{S}, \mathcal{S}') \triangleq \max\{|\mathcal{S} \setminus \mathcal{S}'|, |\mathcal{S}' \setminus \mathcal{S}|\}$.*

Recall that two datasets are neighboring, i.e., $\mathcal{S} \sim \mathcal{S}'$, iff $d_{\triangle}(\mathcal{S}, \mathcal{S}') \leq 1$. This lemma implies that if two datasets are neighboring, then they are still neighboring after filtering with the same parameters, no matter how many times we filter them. Hence, we can use standard mechanisms in the Laplace mechanism for private μ_t (line 6) and in the private PCA for v_t (line 9). Analyzing the utility of this algorithm, we get the following guarantee, which follows from Theorem 25 and Lemma A.6.3 in the appendix. Putting together Lemmas 2.3.5 and 2.3.7, we get the desired result in Theorem 6.

Lemma 2.3.7. DPFILTER$(S, \alpha, T, \varepsilon, \delta)$ *is* (ε, δ)*-differentially private if* $n = \widetilde{\Omega}((T^{1/2}/\varepsilon)(\log(1/\delta))^{3/2})$. *Under the hypotheses of Theorem 6,* DPFILTER$(S, \alpha, T = \widetilde{\Theta}(B^2 d), \varepsilon, \delta)$ *achieves* $\|\hat{\mu} - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ *with probability* 0.9, *if* $n = \widetilde{\Omega}(d/\alpha^2 + B^2 d^2 \log(1/\delta)/(\varepsilon\alpha))$ *and* B *is large enough such that the original uncorrupted samples are inside the hypercube* $\bar{x} + [-B/2, B/2]^d$.

2.3.3 PRIME: novel private and robust mean estimation algorithm

PRIvate and robust Mean Estimation (PRIME) replaces the DPFILTER with the DPMMWFILTER of Algorithm 6, which uses a matrix multiplicative weights approach from [73] to dramatically reduce the number of iterations (from $\widetilde{O}(d^2)$ to $O((\log d)^2)$) and improves sample complexity, as follows. We provide a proof in §A.5.

Algorithm 5: PRIvate and robust Mean Estimation (PRIME)

Input: $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$, range $[-R, R]^d$, adversarial fraction $\alpha \in (0, 1/2)$, number of iterations $T_1 = O(\log d), T_2 = O(\log d)$, target privacy (ε, δ)

1 $(\bar{x}, B) \leftarrow$ DPRANGE$(\{x_i\}_{i=1}^n, R, 0.01\varepsilon, 0.01\delta)$ [Algorithm 14 in §A.2]

2 Clip the data points: $\tilde{x}_i \leftarrow \mathcal{P}_{\bar{x}+[-B/2,B/2]^d}(x_i)$, for all $i \in [n]$

3 $\hat{\mu} \leftarrow$ DPMMWFILTER$(\{\tilde{x}_i\}_{i=1}^n, \alpha, T_1, T_2, 0.99\varepsilon, 0.99\delta)$ [Algorithm 6]

Output: $\hat{\mu}$

Theorem 7. *PRIME is* (ε, δ)*-differentially private if* $n = \tilde{\Omega}((1/\varepsilon)\log(1/\delta))$. *Under Assumption 1 there exists a universal constant* $c \in (0, 0.1)$ *such that if* $\alpha \leq c$, $n = \widetilde{\Omega}((d/\alpha^2) + (d^{3/2}/(\varepsilon\alpha))\log(1/\delta))$, $T_1 = \Omega(\log d)$, *and* $T_2 = \Omega(\log d)$, *then PRIME achieves* $\|\hat{\mu} - \mu\|_2 =$

$O(\alpha\sqrt{\log(1/\alpha)})$ *with probability* 0.9. *The notation* $\widetilde{\Omega}(\cdot)$ *hides logarithmic terms in* d, R, *and* $1/\alpha$.

DPRANGE uses $(0.01\varepsilon, 0.01\delta)$ of the total privacy budget, and DPMMWFILTER uses the rest. The differential privacy guarantee of DPMMWFILTER follows from the interactive version of the algorithm provided in Algorithm 17 in §A.5.1, which explicitly shows how many times we (privately) access the dataset. We interpret the main result in the following remarks.

Remark 1. To achieve an error of $O(\alpha\sqrt{\log(1/\alpha)})$, the first term $\widetilde{\Omega}(d/\alpha^2 \log(1/\alpha))$ is necessary even if there is no corruption. The accuracy of $O(\alpha\sqrt{\log(1/\alpha)})$ matches the lower bound shown in [68] for any polynomial time statistical query algorithm, and it nearly matches the information theoretical lower bound on robust estimation of $O(\alpha)$ up to a logarithmic factor. This is the lowest error one can achieve even with infinite samples.

Remark 2. On the other hand, the second term of $\widetilde{\Omega}(d^{3/2}/(\varepsilon\alpha \log(1/\alpha)))$ in the sample complexity has an extra factor of $d^{1/2}$ compared to the optimal one achieved by exponential time algorithms: private Tukey median (cf. Theorem 5), private hypothesis testing [36], and Algorithm 8. It is an open question if this gap can be closed by a polynomial time algorithm.

The bottleneck is the spectral analysis of the covariance matrix, which is a private PCA in DPFILTER and private matrix multiplicative weights in DPMMWFILTER (lines 14 and 15). Such spectral analyses are crucial in filter-based robust estimators, as captured by Lemma 2.3.2. Even for the simple task of privately computing the top principal component, the best polynomial time algorithm requires $O(d^{3/2})$ samples [80, 46, 209], and this sample complexity is also necessary. A lower bound from [80, Corollary 25] shows that if $n \le cd^{3/2}/(\tilde{\alpha}\sqrt{\log d})$ for some constant c then for any $(1, 1/d^2)$-differentially private estimator $\hat{v}$ of the top principal component v of the covariance matrix Σ, we have $v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} \ge \tilde{\alpha}$ with probability $1 - 1/d$ when each sample is bounded by $\|x_i\|_2 = O(\sqrt{d})$.

Remark 3. Matrix multiplicative weights were utilized in [73] to make robust mean estimation faster using only $O((\log d)^2)$ filtering steps. This is a significant improvement over

$O(d)$ iterations of previous filtering approaches [64]; we provide a detailed explanation in the next section. This speed up is critical in achieving the desired sample efficiency since privacy is leaked every time we access the data.

Remark 4. To boost the success probability to $1 - \zeta$ for some small $\zeta > 0$, we need an extra $\log(1/\zeta)$ factor in the sample complexity to make sure the dataset satisfies the regularity condition with probaibility $\zeta/2$. Then we can run PRIME $\log(1/\zeta)$ times and choose the output of a run that satisfies $n^{(s)} > n - 10\alpha$ and $\lambda^{(s)} \leq C\alpha \log(1/\alpha)$ at termination.

Algorithm 6: Differentially private filtering with matrix multiplicative weights (DPMMWFILTER)

Input: $S = \{x_i \in \bar{x} + [-B/2, B/2]^d\}_{i=1}^n$, $\alpha \in (0, 1/2)$, $T_1 = O(\log(B\sqrt{d})), T_2 = O(\log d)$, privacy (ε, δ)

Initialize $S^{(1)} \leftarrow [n]$, $\varepsilon_1 \leftarrow \varepsilon/(4T_1)$, $\delta_1 \leftarrow \delta/(4T_1)$, $\varepsilon_2 \leftarrow \min\{0.9, \varepsilon\}/(4\sqrt{10T_1T_2\log(4/\delta)})$, $\delta_2 \leftarrow \delta/(20T_1T_2)$, a large enough constant $C > 0$

for epoch $s = 1, 2, \ldots, T_1$ **do**

- $\lambda^{(s)} \leftarrow \|M(S^{(s)}) - \mathbf{I}\|_2 + \mathrm{Lap}(2B^2d/(n\varepsilon_1))$
- $n^{(s)} \leftarrow |S^{(s)}| + \mathrm{Lap}(1/\varepsilon_1)$
- **if** $n^{(s)} \le 3n/4$ **then Output:** $\emptyset$
- **if** $\lambda^{(s)} \le C\,\alpha\log(1/\alpha)$ **then**
  - **Output:** $\mu^{(s)} \leftarrow (1/|S^{(s)}|)\big(\sum_{i\in S^{(s)}} x_i\big) + \mathcal{N}(0, (2B\sqrt{2d\log(1.25/\delta_1)}/(n\,\varepsilon_1))^2\mathbf{I}_{d\times d})$
- $\alpha^{(s)} \leftarrow 1/(100(0.1/C + 1.01)\lambda^{(s)})$
- $S_1^{(s)} \leftarrow S^{(s)}$
- **for** $t = 1, 2, \ldots, T_2$ **do**
  - $\lambda_t^{(s)} \leftarrow \|M(S_t^{(s)}) - \mathbf{I}\|_2 + \mathrm{Lap}(2B^2d/(n\varepsilon_2))$
  - **if** $\lambda_t^{(s)} \le 0.5\lambda_0^{(s)}$ **then**
    - terminate epoch
  - **else**
    - $\Sigma_t^{(s)} \leftarrow M(S_t^{(s)}) + \mathcal{N}(0, (2B^2d\sqrt{2\log(1.25/\delta_2)}/(n\varepsilon_2))^2\mathbf{I}_{d^2\times d^2})$
    - $U_t^{(s)} \leftarrow (1/\operatorname{Tr}(\exp(\alpha^{(s)}\sum_{r=1}^t(\Sigma_r^{(s)} - \mathbf{I})))\exp(\alpha^{(s)}\sum_{r=1}^t(\Sigma_r^{(s)} - \mathbf{I}))$
    - $\psi_t^{(s)} \leftarrow \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)}\right\rangle + \mathrm{Lap}(2B^2d/(n\varepsilon_2))$
    - **if** $\psi_t^{(s)} \le (1/5.5)\lambda_t^{(s)}$ **then**
      - $S_{t+1}^{(s)} \leftarrow S_t^{(s)}$
    - **else**
      - $Z_t^{(s)} \leftarrow \mathrm{Unif}([0,1])$
      - $\mu_t^{(s)} \leftarrow (1/|S_t^{(s)}|)\big(\sum_{i\in S_t} x_i\big) + \mathcal{N}(0, (2B\sqrt{2d\log(1.25/\delta_2)}/(n\,\varepsilon_2)\mathbf{I}_{d\times d})^2)$
      - $\rho_t^{(s)} \leftarrow$ DP-1Dfilter$(\mu_t^{(s)}, U_t^{(s)}, \alpha, \varepsilon_2, \delta_2, S_t^{(s)})$ [Algorithm 7]
      - $S_{t+1}^{(s)} \leftarrow S_t^{(s)} \setminus \{i \mid i \in \mathcal{T}_{2\alpha}$ for $\{\tau_j = (x_j - \mu_t^{(s)})^\top U_t^{(s)}(x_j - \mu_t^{(s)})\}_{j\in S_t^{(s)}}$ and $\tau_i \ge \rho_t^{(s)} Z_t^{(s)}\}$, where $\mathcal{T}_{2\alpha}$ is defined in Definition 2.3.1.
- $S^{(s+1)} \leftarrow S_t^{(s)}$

Output: $\mu^{(T_1)}$

2.3.3.1 *Matrix multiplicative weights and a proof sketch of Theorem 7*

We now provide the intuition for using matrix multiplicative weights in line 15 of Algorithm 6 and a proof sketch of Theorem 7, and we refer to §A.5 for a formal proof. DPMMWFILTER runs T_1 epochs in the outer-loop and T_2 iterations at each epoch in the inner-loop. The inner-loop ensures that the covariance strictly decreases after T_2 iterations. The outer-loop ensures that the covariance decreases sufficiently after T_1 epochs. Lemma 2.3.2 ensures that this is sufficient for robust estimation.

Proof sketch. The next lemma guarantees that (i) we are guaranteed to have more corrupted samples removed than the clean samples (in expectation) at every iteration t, and (ii) we get a decreasing covariance in its spectral norm at every epoch s. We provide a proof of this lemma in §A.5.3. Formally, we define S_{good} as the original set of n clean samples (as defined in Assumption 1) and S_{bad} as the set of corrupted samples that replace αn of the clean samples. The (rescaled) covariance is denoted by $M(S^{(s)}) \triangleq (1/n)\sum_{i\in S^{(s)}}(x_i - \mu(S^{(s)}))(x_i - \mu(S^{(s)}))^\top$, where $\mu(S^{(s)}) \triangleq (1/|S^{(s)}|)\sum_{i\in S^{(s)}} x_i$ denotes the mean.

Lemma 2.3.8 (informal version of Lemma A.5.3)**.** *Under the hypotheses of Lemma A.5.3, if* $n = \widetilde{\Omega}(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\varepsilon\alpha))$ *and* $|S_t^{(s)} \cap S_{\text{good}}| \geq (1-10\alpha)n$ *then there exists a constant* $C > 0$ *such that for each epoch* s *and iteration* t*,*

- *in expectation, more corrupted samples are removed than the uncorrupted samples, i.e.,* $\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| \leq \mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}|$*, and*

- *for each epoch* s*, if* $\|M(S^{(s)}) - \mathbf{I}\|_2 \geq C\,\alpha\log(1/\alpha)$ *then the* s*-th epoch terminates after* $O(\log d)$ *iterations and outputs* $S^{(s+1)}$ *such that* $\|M(S^{(s+1)}) - \mathbf{I}\|_2 \leq 0.98\|M(S^{(s)}) - \mathbf{I}\|_2$ *with probability* $1 - O(1/(\log d)^2)$*.*

In $s = O(\log_{0.98}((C\alpha\log(1/\alpha))/\|M(S^{(1)}) - \mathbf{I}\|_2))$ epochs, this lemma guarantees that we find a candidate set $S^{(s)}$ of samples with $\|M(S^{(s)} - \mathbf{I}\|_2 \leq C\alpha\log(1/\alpha)$. Lemma 2.3.2 ensures that we get the desired bound of $\|\mu(S^{(s)}) - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ as long as $S^{(s)}$ has enough

clean data, i.e., $|S^{(s)} \cap S_{\text{good}}| \geq n(1-\alpha)$. Since Lemma 2.3.8 gets invoked at most $O((\log d)^2)$ times, we can take a union bound, and the following argument conditions on the good events in Lemma 2.3.8 holding, which happens with probability at least 0.99. To turn the average case guarantee of Lemma 2.3.8 into a constant probability guarantee, we apply the optional stopping theorem. Recall that the s-th epoch starts with a set $S^{(s)}$ and outputs a filtered set $S_t^{(s)}$ at the t-th inner iteration. We measure the progress by by summing the number of clean samples removed up to epoch s and iteration t and the number of remaining corrupted samples, defined as $d_t^{(s)} \triangleq |(S_{\text{good}} \cap S^{(1)}) \setminus S_t^{(s)}| + |S_t^{(s)} \setminus (S_{\text{good}} \cap S^{(1)})|$. Note that $d_1^{(1)} = \alpha n$, and $d_t^{(s)} \geq 0$. At each epoch and iteration, we have

$$\mathbb{E}[d_{t+1}^{(s)} - d_t^{(s)} | d_1^{(1)}, d_2^{(1)}, \cdots, d_t^{(s)}] \;=\; \mathbb{E}\left[|S_{\text{good}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})| - |S_{\text{bad}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})|\right] \;\leq\; 0,$$

from part 1 of Lemma 2.3.8. Hence, $d_t^{(s)}$ is a non-negative super-martingale. By the optional stopping theorem, at stopping time, we have $\mathbb{E}[d_t^{(s)}] \leq d_1^{(1)} = \alpha n$. By the Markov inequality, $d_t^{(s)}$ is less than $10\alpha n$ with probability 0.9, i.e., $|S_t^{(s)} \cap S_{\text{good}}| \geq (1-10\alpha)n$. The desired bound in Theorem 7 follows from Lemma 2.3.2.

Matrix multiplicative weights (MMWs). We are left to prove that the MMW filtering (lines 9-23 in DPMMWFILTER) satisfies our main technical result in Lemma 2.3.8. Recall that PCA-based filtering in DPFILTER requires $O(d)$ iterations in the worst case since it checks only one direction at a time. If $O(d)$ samples are corrupted by each taking a clean sample and arbitrarily changing in one distinct coordinate, then it takes $O(d)$ iterations filtering them out one by one.

The MMW-based approach, pioneered in [73] and generalized to covariance estimation [154] and heavy-tailed estimation [105], filters out multiple directions jointly. For simplicity, we present the proof sketch when privacy is not required ($\varepsilon = \infty$) and give the full proof in the general ($\varepsilon < \infty$) setting in §A.5.2. Define $U_t^{(s)}$ via the matrix multiplicative update:

$$U_t^{(s)} \;=\; \frac{1}{\text{Tr}\big(\,\exp(\alpha^{(s)} \sum_{r\in[t]} (\Sigma_r^{(s)} - \mathbf{I}))\,\big)} \exp\left(\alpha^{(s)} \sum_{r\in[t]} (\Sigma_r^{(s)} - \mathbf{I})\right),$$

where $\Sigma_r^{(s)} = M(S_t^{(s)}) = (1/n)\sum_{i\in S}(x_i - \mu(S_t^{(s)}))(x_i - \mu(S_t^{(s)}))^\top$ since $\varepsilon = \infty$. As $U_r^{(s)}$'s are multiplicative weight updates for online constrained linear optimization with objective $\langle \Sigma_r^{(s)}, U\rangle$ at the r-th iteration, it is known from [8], for example, that for the choice of $\alpha^{(s)}$ that satisfies $\alpha^{(s)}(\Sigma_r^{(s)} - \mathbf{I}) \preceq \mathbf{I}$, the following regret bound in Eq. (2.2) is achieved (Lemma A.6.13):

$$\begin{aligned}
\Big\|\sum_{r\in[t]}\big(\Sigma_r^{(s)} - \mathbf{I}\big)\Big\|_2 &= \max_{U,\|U\|_*=1}\Big\langle \sum_{r\in[t]}\Sigma_r^{(s)} - \mathbf{I}, U\Big\rangle \\
&\le \sum_{r\in[t]}\big\langle \Sigma_r^{(s)} - \mathbf{I}, U_r^{(s)}\big\rangle + \sum_{r\in[t]}\alpha^{(s)}\|\Sigma_r^{(s)} - \mathbf{I}\|_2\big\langle U_r^{(s)}, \big|\Sigma_r^{(s)} - \mathbf{I}\big|\big\rangle + \frac{\log d}{\alpha^{(s)}} \\
&\le \sum_{r\in[t]}\big\langle \Sigma_r^{(s)} - \mathbf{I}, U_r^{(s)}\big\rangle + \frac{1}{100}\sum_{r\in[t]}\big\langle \big|\Sigma_r^{(s)} - \mathbf{I}\big|, U_r^{(s)}\big\rangle + 200\log(d)\|M(S^{(s)}) - \mathbf{I}\|_2\,, \quad (2.2)
\end{aligned}$$

where $\|A\|_* = \sum_i \sigma(A)$ is the nuclear norm, $|\cdot|$ of a symmetric matrix is defined by its eigenvalue decomposition as $|U\mathrm{diag}([\lambda_i]_{i=1}^d)V^\top| = U\mathrm{diag}([|\lambda_i|]_{i=1}^d)V^\top$, and we used the fact that $\Sigma_{r+1}^{(s)} \preceq \Sigma_r^{(s)}$ is a decreasing sequence (Lemma A.6.1) and $1/200 \le \alpha^{(s)}\|\Sigma^{(s)} - \mathbf{I}\|_2 \le 1/100$. By carefully designing the private filtering algorithm in DP-1DFILTER, we make sufficient progress in each iteration in reducing the covariance, as shown in Lemma A.5.4. This gives $\langle \Sigma_r^{(s)} - \mathbf{I}, U_r^{(s)}\rangle \le 0.95\|M(S_1^{(s)}) - \mathbf{I}\|_2 + 2c\alpha\log 1/\alpha$ and $\langle |\Sigma_r^{(s)} - \mathbf{I}|, U_r^{(s)}\rangle \le 0.95\|M(S_1^{(s)}) - \mathbf{I}\|_2 + 2c\alpha\log 1/\alpha$. For details of this analysis, we refer to the proof of Lemma A.5.5 in §A.5.3.3.

$$\begin{aligned}
\Big\|\Sigma_{T_2}^{(s)} - \mathbf{I}\Big\|_2 &\le \frac{1}{T_2}\Big\|\sum_{r\in[T_2]}\big(\Sigma_r^{(s)} - \mathbf{I}\big)\Big\|_2 \\
&\le 0.96\,\|\Sigma_1^{(s)} - \mathbf{I}\|_2 + 2\,c\,\alpha\log 1/\alpha + \frac{200\log d}{T_2}\|\Sigma_1^{(s)} - \mathbf{I}\|_2 \\
&\le 0.98\,\|\Sigma_1^{(s)} - \mathbf{I}\|_2\,,
\end{aligned}$$

where the first inequality follows from the monotonicity of $\Sigma_r^{(s)}$ and the last one from the fact that stopping criteria of $\|\Sigma^{(s)} - \mathbf{I}\|_2 \le C\alpha\log(1/\alpha)$ have not been met so far and $T_2 = O(\log d)$. Hence, the MMW approach ensures that $O(\log d)$ steps are sufficient for the spectral norm of the covariance to decrease strictly.

Algorithm 7: Differentially private 1D-filter (DP-1DFILTER)

Input: μ, U, $\alpha \in (0, 1/2)$, target privacy (ε, δ), $S = \{x_i \in \bar{x} + [-B/2, B/2]^d\}$

1 Set $\tau_i \leftarrow (x_i - \mu)^\top U (x_i - \mu)$ for all $i \in S$

2 Set $\tilde{\psi} \leftarrow (1/n) \sum_{i \in S} (\tau_i - 1) + \mathrm{Lap}(B^2 d / n\varepsilon))$

3 Compute a histogram over geometrically sized bins

$$I_1 = [1/4, 1/2), I_2 = [1/2, 1), \ldots, I_{2+\log(B^2 d)} = [2^{\log(B^2 d)-1}, 2^{\log(B^2 d)}]$$

$$h_j \leftarrow \frac{1}{n} \cdot |\{i \in S \,|\, \tau_i \in [2^{-3+j}, 2^{-2+j})\}| \,, \qquad \text{for all } j = 1, \ldots, 2 + \log(B^2 d)$$

4 Compute a privatized histogram $\tilde{h}_j \leftarrow h_j + \mathcal{N}(0, (4\sqrt{2\log(1.25/\delta)}/(n\varepsilon))^2)$, for all $j \in [2 + \log(B^2 d)]$

5 Set $\tilde{\tau}_j \leftarrow 2^{-3+j}$, for all $j \in [2 + \log(B^2 d)]$

6 Find the largest $\ell \in [2 + \log(B^2 d)]$ satisfying $\sum_{j \geq \ell} (\tilde{\tau}_j - \tilde{\tau}_\ell) \tilde{h}_j \geq 0.31 \tilde{\psi}$

Output: $\rho = \tilde{\tau}_\ell$

2.3.3.2 Novel private DP-1DFILTER

Once $U_t^{(s)}$ is obtained from the proposed MMW approach, we run a filter (DPMMWFILTER line 23) to remove suspected corrupted samples. The idea is to remove suspected corrupted samples by their contribution to the covariance matrix as projected onto $U_t^{(s)}$, denoted as τ_i for $i = 1, 2, \ldots, n$. A corresponding non-private filter in [73, Algorithm 9] requires $O(n)$ iterations of 1Dfilter at each inner-loop, a prohibitively large number of accesses to the data under our private setting. Therefore, we introduce a novel private DP-1DFILTER in Algorithm 7 that accesses the data only once.

Lemma 2.3.9 (Private 1-D filter: picking threshold privately)**.** *Algorithm DP-1Dfilter$(\mu, U, \alpha, \varepsilon, \delta, S)$ running on a dataset $\{\tau_i = (x_i - \mu)^\top U (x_i - \mu)\}_{i \in S}$ is (ε, δ)-DP. Define $\psi \triangleq \frac{1}{n} \sum_{i \in S} (\tau_i - 1)$.*

If τ_i's satisfy

$$\begin{aligned} \frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S}\tau_i &\leq \psi/1000 \\ \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S}(\tau_i-1) &\leq \psi/1000\,, \end{aligned}$$

and $n \geq \widetilde{\Omega}\left(\frac{B^2 d\sqrt{\log(1/\delta)}}{\varepsilon\alpha}\right)$*, then DP-1Dfilter outputs a threshold* ρ *such that with probability* $1 - O(1/\log^3 d)$*,*

$$\frac{1}{n}\sum_{\tau_i<\rho}(\tau_i - 1) \leq 0.75\psi \quad \textit{and} \tag{2.3}$$

$$2\Big(\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i\leq\rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i>\rho\}\Big) \leq \sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i\leq\rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i>\rho\}\,. \tag{2.4}$$

DP-1DFILTER finds a private threshold ρ such that when a randomized filter is applied with the scale of ρ, we cut enough samples to make progress in each iteration (Eq. (2.3)) and while ensuring that we do not remove too many uncorrupted samples (Eq. (2.4)). Finding such a threshold is straightforward in a non-private setting; one can choose the largest ρ such that (2.3) holds. The regularity of the uncorrupted samples ensures that the safety condition is also met.

We use a private histogram of the scores to approximate this threshold. However, a standard fixed size binning fails: when evaluating the contribution of the points below (or above) threshold, the error of the Gaussian mechanism accumulates over $O(B^2 d)$ bins. This introduces $O(B^2 d)$ error in approximating $(1/n)\sum_{\tau_i<\rho}(\tau_i - 1)$. Instead, we geometrically increase bin sizes using only $O(\log B^2 d)$ bins; thus, the approximation error is now within $\widetilde{O}(d/\epsilon n)$. This introduces a multiplicative error in our quantization, which luckily fits well with our objective in Eq. (2.3). This, together with our regularity of uncorrupted samples, will satisfy our safety condition in Eq. (2.4).

2.4 Exponential time approaches for sub-Gaussian distributions

The existing result for robust and private mean estimation in [36] strictly requires the uncorrupted samples to be drawn from a Gaussian distribution, and the run-time is exponential in the dimension. The technique heavily relies on covering the parameters of Gaussian distributions with an α-cover, which cannot be extended to any non-parametric family of distributions. To this end, we introduce a new family of (exponential time) algorithms that can provide near optimal sample complexity for both sub-Gaussian distributions and second moment bounded distributions. We provide a proof in §A.7.2.

Theorem 8 (Exponential time algorithm for sub-Gaussian distributions)**.** *Algorithm 8 is* (ε, δ)*-differentially private if* $n = \Omega(d^{1/2}\log(1/\delta)/(\varepsilon\alpha\sqrt{\log(1/\alpha)}))$*. Under Assumption 1, if*

$$n = \Omega\Big(\frac{d + \log(1/\zeta)}{\alpha^2 \log(1/\alpha)} + \frac{d\log(dR/\alpha) + \log(1/\zeta)}{\varepsilon\alpha} + \frac{\sqrt{d\log(1/\delta)}\min\{\log(dR/\zeta), \log(d/\zeta\delta)\}}{\varepsilon} \Big)\,,$$

this algorithm achieves $\|\hat{\mu} - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ *with probability* $1 - \zeta$*.*

The main idea is to use the *resilience* property of the samples to (i) check that the uncorrupted portion of the samples is drawn from the distribution of interest, and (ii) bound the sensitivity of the subsequent exponential mechanism.

Remark 1. In an attempt to design efficient algorithms for robust and private mean estimation, [60] proposed an algorithm with a mis-calculated sensitivity, which can result in violating the privacy guarantee. Our approach for checking the resilience can be used as a pre-processing step to ensure the desired sensitivity bound is met, but at the cost of exponential run-time.

Definition 2.4.1 (Resilience (Definition 1 in [186]))**.** *A set of points* $\{x_i\}_{i\in S}$ *lying in* $\mathbb{R}^d$ *is* (σ, α)*-resilient around a point* μ *if, for all subsets* $T \subset S$ *of size at least* $(1-\alpha)|S|$*,*

$$\Big\| \frac{1}{|T|} \sum_{i\in T} (x_i - \mu) \Big\|_2 \;\leq\; \sigma.$$

We define $R(S)$ as a surrogate for resilience. The intuition is that if the dataset S indeed consists of a $1-\alpha$ fraction of independent samples from the promised class of distributions, the goodness score $R(S)$ will be close to the resilience property of the good data.

Definition 2.4.2 (Goodness of a set)**.** *For $\mu(S) = (1/|S|)\sum_{i\in S} x_i$, let us define*

$$R(S) \triangleq \min_{S'\subset S, |S'|=(1-2\alpha)|S|.} \max_{T\subset S', |T|=(1-\alpha)|S'|.} \|\mu(T) - \mu(S')\|_2 \,. \tag{2.5}$$

Algorithm 8 first checks if the resilience of the dataset matches that of the promised distribution. The data is pre-processed with DPRANGE to ensure we can check $R(S)$ privately. Once resilience is cleared, we can safely use the exponential mechanism based on score function $d(\hat{\mu}, S)$, which is defined in Definition 2.4.3, to select an approximate mean $\hat{\mu}$. The choice of the sensitivity critically relies on the fact that resilient datasets have small sensitivity. As the loss for exponential mechanism, we propose the distance between a robust projected mean and the candidate $\hat{\mu}$ defined as follows.

Definition 2.4.3. *For a set of data $\{x_i\}_{i\in S}$ lying in $\mathbb{R}^d$, for any $v \in \mathbb{S}^{d-1}$, define $\mathcal{T}^v$ to be the $3\alpha|S|$ points with the largest $v^\top x_i$ value, $\mathcal{B}^v$ to be the $3\alpha|S|$ points with the smallest $v^\top x_i$ value, and $\mathcal{M}^v = S \setminus (\mathcal{T}^v \cup \mathcal{B}^v)$. Define*

$$d(\hat{\mu}, S) \triangleq \max_{v\in\mathbb{S}^{d-1}} \left|v^\top (\mu(\mathcal{M}^v) - \hat{\mu})\right| \,.$$

Run-time. Computing $R(S)$ exactly can take $O(de^{\Theta(n)})$ operations. The exponential mechanism implemented with α-covering for $\hat{\mu}$ and a constant covering for v can take $O(nd(R/\alpha)^d)$ operations.

Algorithm 8: Exponential-time private and robust mean estimation

Input: $S = \{x_i\}_{i\in[n]}$, $\alpha \in (0, 1/2)$, R, (ε, δ)

$(\bar{x}, B) \leftarrow$ DPRANGE$(\{x_i\}_{i=1}^n, R, (1/3)\varepsilon, (1/3)\delta)$ [DPRANGE-HT$(\cdot)$ for hevay-tail]

Project the data points onto the ball: $\tilde{x}_i \leftarrow \mathcal{P}_{\mathcal{B}_{\sqrt{d}B/2}(\bar{x})}(x_i)$, for all $i \in [n]$

$\widehat{R}(S) \leftarrow R(S) + \mathrm{Lap}(3Bd^{1/2}/(n\varepsilon))$

if $\widehat{R}(S) > 2\,\alpha\sqrt{\log(1/\alpha)}$ **then Output:** $\emptyset$ [$\widehat{R}(S) > 2c_\zeta\sqrt{\alpha}$ for hevay-tail]

else Output: a randomly drawn point $\hat{\mu} \in [-2R, 2R]^d$ sampled from a density

$$r(\hat{\mu}) \propto e^{-(1/(24\sqrt{\log(1/\alpha)}))\varepsilon\, n\, d(\hat{\mu},S)} \quad [e^{-(\varepsilon n\sqrt{\alpha}/(24c_\zeta))d(\hat{\mu},S)} \text{ for heavy-tail}]$$

where $d(\hat{\mu}, S)$ is defined in Definition 2.4.3

2.5 Heavy-tailed distributions: algorithm and analysis

We consider distributions with bounded covariance as defined in Assumption 2. Under these assumptions, Algorithm 8 achieves near optimal guarantees but takes exponential time. The dominant term in the sample complexity $\widetilde{\Omega}(d/(\varepsilon\alpha))$ cannot be improved as it matches that of the optimal non-robust private estimation [135]. The accuracy $O(\sqrt{\alpha})$ cannot be improved as it matches that of the optimal non-private robust estimation [73]. We provide a proof in §A.7.1.

Theorem 9 (Exponential time algorithm for covariance bounded distributions)**.** *If* $n = \Omega(d^{1/2}\log(1/\delta)/(\varepsilon\alpha))$*, Algorithm 8 is* (ε,δ)*-differentially private. Under Assumption 2, if*

$$n = \Omega((d\log(dR/\alpha))/(\varepsilon\alpha) + (1/\varepsilon)d^{1/2}\log^{3/2}(1/\delta)\min\{\log(dR), \log(d/\delta)\}) \ ,$$

this algorithm achieves $\|\hat{\mu} - \mu\|_2 = O(\sqrt{\alpha})$ *with probability* 0.9*.*

We propose an efficient algorithm PRIME-HT and show that it achieves the same optimal accuracy but at the cost of increased sample complexity of $O(d^{3/2}\log(1/\delta)/(\varepsilon\alpha))$. In the first step, we need increase the radius of the ball to $O(\sqrt{d/\alpha})$ to include a $1-\alpha$ fraction of the clean samples, where DPRANGE-HT returns $B = O(1/\sqrt{\alpha})$ and $\mathcal{B}_{\sqrt{d}B/2}(\bar{x})$ is a ℓ_2-ball of radius $\sqrt{d}B/2$ centered at $\bar{x}$. This is followed by a matrix multiplicative weight filter similar to DPMMWFILTERR but the parameter choices are tailored for covariance bounded distributions. We provide a proof in §A.8.2.

Theorem 10 (Efficient algorithm for covariance bounded distributions)**.** PRIME-HT *is* (ε,δ)*-differentially private if* $n = \tilde{\Omega}((1/\varepsilon)\log(1/\delta))$*. Under Assumption 2 there exists a universal constant* $c \in (0, 0.1)$ *such that if* $\alpha \le c$*, and* $n = \widetilde{\Omega}((d^{3/2}/(\varepsilon\alpha))\log(1/\delta))$*,* $T_1 = \Omega(\log(d/\alpha))$*, and* $T_2 = \Omega(\log d)$*, then* PRIME-HT *achieves* $\|\hat{\mu} - \mu\|_2 = O(\alpha^{1/2})$ *with probability* 0.9*. The notation* $\widetilde{\Omega}(\cdot)$ *hides logarithmic terms in* d*,* R*, and* $1/\alpha$*.*

Remark 1. To boost the success probability to $1-\zeta$ for some small $\zeta > 0$, we will randomly split the data into $O(\log(1/\zeta))$ subsets of equal sizes, and run Algorithm 9 to obtain a mean

estimation from each of the subset. Then we can apply multivariate "mean-of-means" type estimator [162] to get $\|\hat{\mu} - \mu\|_2 = O(\alpha^{1/2})$ with probability $1 - \zeta$.

Algorithm 9: PRIvate and robust Mean Estimation for covariance bounded distributions (PRIME-HT)

Input: $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$, range $[-R, R]^d$, adversarial fraction $\alpha \in (0, 1/2)$, number of iterations $T_1 = O(\log(d/\alpha)), T_2 = O(\log d)$, target privacy (ε, δ)

$(\bar{x}, B) \leftarrow$ DPRANGE-HT$(\{x_i\}_{i=1}^n, R, 0.01\varepsilon, 0.01\delta)$ [Algorithm 18 in §A.8]

Project the data onto the ball: $\tilde{x}_i \leftarrow \mathcal{P}_{\mathcal{B}_{\sqrt{d}B/2}(\bar{x})}(x_i)$, for all $i \in [n]$

$\hat{\mu} \leftarrow$ DPMMWFILTER-HT$(\{\tilde{x}_i\}_{i=1}^n, \alpha, T_1, T_2, 0.99\varepsilon, 0.99\delta)$ [Algorithm 19 in §A.8]

Output: $\hat{\mu}$

2.6 Discussion

Differentially private mean estimation is brittle against a small fraction of the samples being corrupted by an adversary. We show that robustness can be achieves without any increase in the sample complexity by introducing a novel mean estimator. The innovation is in leveraging the resilience property of well-behaved distributions in an innovative way to not only find robust mean (which is the typical use case of resilience) but also bound sensitivity for optimal privacy guarantee. However, this algorithm takes an exponential time. We therefore propose an efficient algorithm that achieves the optimal target accuracy at the cost of an increase of sample complexity. With appropriately chosen parameters, we show that our exponential time approach achieves near-optimal guarantees for both sub-Gaussian and covariance bounded distributions, and our efficient approach achieves target optimal accuracy but at the cost of an extra $d^{1/2}$ factor in the sample complexity.

There are several directions for improving our results further and applying the framework to solve other problems. PRIME provides a new design principle for private and robust estimation. This can be more broadly applied to fundamental statistical analyses such as robust covariance estimation [62, 64, 154] robust PCA [145, 121], and robust linear regression [143, 70].

PRIME could be improved in a few directions. First, the sample complexity of $\widetilde{\Omega}((d/(\alpha^2 \log(1/\alpha)))+ (d^{3/2}/(\varepsilon\alpha \log(1/\alpha))) \log(1/\delta))$ in Theorem 7 is suboptimal in the second term. Improving the $d^{3/2}$ factor requires bypassing differentially private singular value decomposition, which seems to be a challenging task. However, it might be possible to separate the $\log(1/\delta)$ factor from the rest of the terms and get an additive error of the form $\widetilde{\Omega}((d/(\alpha^2 \log(1/\alpha))) + (d^{3/2}/(\varepsilon\alpha \log(1/\alpha))) + (1/\varepsilon) \log(1/\delta))$. This requires using Laplace mechanism in private MMW (line 15 Algortihm 6). Secondly, the time complexity of PRIME is dominated by computation time of the matrix exponential in (line 15 Algortihm 6). Total number of operations scale as $\widetilde{O}(d^3 + nd^2)$. One might hope to achieve $\widetilde{O}(nd)$ time complexity using approximate computations of τ_j's using techniques from [73]. This does not improve the sample complexity, as the number of times the dataset is accessed remains the same. Finally, for (non-robust) private mean estimation, CoinPress provides a practical improvement in the small sample regime by progressively refining the search space [32]. The same principle could be applied to PRIME to design a robust version of CoinPress.

Chapter 3

HPTR: A UNIFYING FRAMEWORK FOR DIFFERENTIALLY PRIVATE AND ROBUST ESTIMATION

3.1 Introduction

Estimating a parameter of a distribution from i.i.d. samples is a canonical problem in statistics. For such problems, characterizing the computational and statistical cost of ensuring differential privacy (DP) has gained significant interest with the rise of DP as the de facto measure of privacy. This is spearheaded by exciting and foundational algorithmic advances, e.g., [25, 139, 129, 135, 40]. However, the computational efficiency of these algorithms often comes at the cost of requiring superfluous assumptions that are not necessary for statistical efficiency, such as known bounds on the parameters or knowledge of higher-order moments. Without such assumptions, the optimal sample complexity remains unknown even for canonical statistical estimation problems under differential privacy. Further, each algorithm needs to be customized to those assumptions or to the problem instances.

We take an alternative route of focusing only on the statistical cost of differential privacy without concerning computational efficiency. Our goal is to introduce a general unifying framework that (1) can be readily applied to each problem instance, (2) provides a tight characterization of the statistical complexity involved, and (3) requires minimal assumptions. Achieving this goal critically relies on three key ingredients: the exponential mechanism introduced in [164], robust statistics, and the Propose-Test-Release mechanism introduced in [77]. We first explain these three components of our approach, and then demonstrate the utility of our proposed framework, called High-dimensional Propose-Test-Release (HPTR), in canonical example problems of mean estimation, linear regression, covariance estimation, and principal component analysis.

Exponential mechanism and sensitivity. Differential privacy (DP) is an agreed upon measure of privacy that provides plausible deniability to the individual entries. Given a dataset S of size n and its empirical distribution $\hat{p}_S = (1/n)\sum_{x_i \in S} \delta_{x_i}$, its *neighborhood* is defined as $\mathcal{N}_S = \{S' : |S'| = |S|, d_{\mathrm{TV}}(\hat{p}_S, \hat{p}_{S'}) \leq 1/n\}$, which is a set of datasets at Hamming distance[1] at most one from S, and $d_{\mathrm{TV}}(\cdot)$ is the total variation. Plausible deniability is achieved by introducing the right amount of randomness. A randomized estimator $\hat{\theta}(S)$ is said to be (ε, δ)-differentially private for some target $\varepsilon \geq 0$ and $\delta \in [0, 1]$ if $\mathbb{P}(\hat{\theta}(S) \in A) \leq e^{\varepsilon}\mathbb{P}(\hat{\theta}(S') \in A) + \delta$ for all neighboring datasets S, S' and all measurable subset $A \subseteq \mathbb{R}^p$ [78]. Consider a binary hypothesis testing on two hypotheses, H_0, where the estimate came from a dataset S, and H_1, where the the estimate came from a dataset S' that is a neighbor of S. The DP condition with a sufficiently small (ε, δ) ensures that an adversary cannot succeed in this test with high confidence [127], which provides plausible deniability.

The *sensitivity* plays a crucial role in designing DP estimators. Consider an example of mean estimation, where the error is measured in the Mahalanobis distance defined as $D_p(\hat{\mu}) = \|\Sigma_p^{-1/2}(\hat{\mu} - \mu_p)\|$, where μ_p and Σ_p are the mean and covariance of the sample-generating distribution p. This is a preferred error metric since it has unit variance in all directions and is invariant to a linear transformation of the samples. A corresponding empirical loss is $D_{\hat{p}_S}(\hat{\mu}) = \|\Sigma_{\hat{p}_S}^{-1/2}(\hat{\mu} - \mu_{\hat{p}_S})\|$. The exponential mechanism from [164] produces an $(\varepsilon, 0)$-DP estimate $\hat{\mu}$ by sampling from a distribution that approximately and stochastically minimizes this empirical loss:

$$\hat{\mu} \sim \frac{1}{Z(S)} e^{-\frac{\varepsilon}{2\Delta} D_{\hat{p}_S}(\hat{\mu})},$$

where $Z(S) = \int \exp\{-(\varepsilon/2\Delta) D_{\hat{p}_S}(\hat{\mu}) d\hat{\mu}$. The sensitivity is defined as $\Delta := \max_{\hat{\mu}, S, S' \in \mathcal{N}_S} |D_{\hat{p}_S}(\hat{\mu}) - D_{\hat{p}_{S'}}(\hat{\mu})|$, which is the influence of one data point on the loss. From this definition, the $(\varepsilon, 0)$-DP guarantee follows immediately (e.g., Lemma 3.2.3).

Using the exponential mechanism is crucial in HPTR for two reasons: adaptivity and

[1] There are two notions of a neighborhood in DP, which are equally popular. We use the one based on exchanging an entry, but all the analyses can seamlessly be applied to the one that allows for insertion and deletion of an entry.

flexibility. First, it naturally adapts to the geometry of the problem, which is encoded in the loss. For example, a more traditional Gaussian mechanism [79] needs to estimate Σ_p privately in order to add a Gaussian noise tailored to that estimated Σ_p, which increases sample complexity significantly. On the other hand, the exponential mechanism seamlessly adapts to Σ_p without explicitly and privately estimating it. Further, the exponential mechanism allows us significant flexibility to design different loss functions, some of which can dramatically reduce the sensitivity. Discovering such a loss function is the main focus of this chapter.

One major challenge is that the sensitivity is unbounded when the support of the distribution is unbounded. A common solution is to privately estimate a bounded domain that the samples lie in and use it to bound the sensitivity (e.g., [139, 129, 160]). We propose a fundamentally different approach using robust statistics.

Robust statistics and resilience. The *resilience* (also known as stability) defined in [186] plays a critical role in robust statistics. For the mean, for example, a dataset S is said to be (α, ρ)-resilient for some $\alpha \in [0, 1]$ and $\rho > 0$ if for all $v \in \mathbb{R}^d$ with $\|v\| = 1$ and all subset $T \subseteq S$ of size at least $|T| \geq \alpha n$,

$$\left| \langle v, \mu_{\hat{p}_T} \rangle - \langle v, \mu_{\hat{p}_S} \rangle \right| \leq \frac{\rho}{\alpha} . \tag{3.1}$$

A more precise statement is in Definition 3.3.2. This measures how resilient the empirical mean is to subsampling or deletion of a fraction of the samples. This resilience is a central concept in robust statistical estimation when a fraction of the dataset is arbitrarily corrupted by an adversary [186, 217]. We show and exploit the fact that resilience is fundamentally related to the sensitivity of robust statistics.

For each direction $v \in \mathbb{R}^d$ with $\|v\| = 1$, we construct a robust mean of a one-dimensional projected dataset, also known as trimmed mean, $S_v = \{\langle v, x_i \rangle \in \mathbb{R}\}_{x_i \in S}$, as follows. For some $\alpha \in [0, 1/2)$, remove αn data points corresponding to the largest entries in S_v and also remove the αn smallest entries. The mean of the remaining $(1 - 2\alpha)n$ points is the robust one-dimensional mean, which we denote by $\langle v, \mu_{\hat{p}_v}^{(robust)} \rangle \in \mathbb{R}$. From the resilience above, we know that the mean of the removed top part is upper bounded by $\langle v, \mu_{\hat{p}_S} \rangle + \rho/\alpha$. The mean

of the removed bottom part is lower bounded by $\langle v, \mu_{\hat{p}_S} \rangle - \rho/\alpha$. Hence, the effective support of this robust one-dimensional mean estimator is upper and lower bounded by the same. This can be readily translated into a bound in sensitivity of the estimate, $\langle v, \mu_{\hat{p}_v}^{(robust)} \rangle$ (e.g., Lemma 3.3.11). A similar sensitivity bound holds for the robust one-dimensional variance estimator, $v^\top \Sigma_{\hat{p}_v}^{(robust)} v$, defined similarly.

We propose an approach that critically relies on this observation that *one-dimensional robust statistics have low sensitivity on resilient datasets, i.e., datasets satisfying the resilience property with small ρ.*

This suggests that if we can design a score function that only depends on one-dimensional robust statistics of the data, it might inherit the low sensitivity of those robust statistics. To this end, we first transform the target error metric into an equivalent expression that only depends on one-dimensional (population) mean, $\langle v, \mu_p \rangle$, and variance, $v^\top \Sigma_p v$, i.e.,

$$\|\Sigma_p^{-1/2}(\hat{\mu} - \mu_p)\| \;=\; \max_{v \in \mathbb{R}^d, \|v\|=1} \frac{\langle v, \hat{\mu} \rangle - \langle v, \mu_p \rangle}{\sqrt{v^\top \Sigma_p \, v}} \;,$$

which follows from Lemma 3.3.1. Next, we replace the population statistics with robust empirical ones to define a new empirical loss, $D_{\hat{p}_S}(\hat{\mu}) = \max_{v \in \mathbb{R}^d, \|v\|=1} (\langle v, \hat{\mu} \rangle - \langle v, \mu_{\hat{p}_v}^{(robust)} \rangle) / \sqrt{v^\top \Sigma_{\hat{p}_v}^{(\text{robust})} \, v}$. Precise definitions of these robust statistics can be found in Eq. (3.5). For resilient datasets, such a score function has a dramatically smaller sensitivity compared to those that rely on high-dimensional robust statistics. For mean estimation under a sub-Gaussian distribution, the sensitivity of the proposed loss is $\tilde{O}(1/n)$, whereas a loss using a high-dimensional robust statistics has $\Omega(\sqrt{d}/n)$ sensitivity.

Such an improved sensitivity immediately leads to a better utility guarantee of the exponential mechanism. We explicitly prescribe such loss functions for the canonical problems of mean estimation, linear regression, covariance estimation, and principal component analysis. This leads to near-optimal utility in most cases and improves upon the state-of-the-art in others, as we demonstrate in Section 3.1.1. Further, this approach can potentially be more generally applied to a much broader class of problems. One remaining challenge is that the tight sensitivity bound we provide holds only for a resilient dataset. To reject bad datasets,

we adopt the Propose-Test-Release (PTR) framework pioneered in the seminal work of [77].

Propose-Test-Release and local sensitivity. The tight sensitivity bound we provide on the proposed exponential mechanism is *local* in the sense that it only holds for resilient datasets. However, differential privacy must be guaranteed for any input, whether it is resilient (with desired level of α and ρ) or not. We adopt Propose-Test-Release introduced in [77] to handle such locality of sensitivity. In the first step, one proposes an upper bound on the sensitivity of the loss $D_S(\hat{\theta})$, determined by the resilience of the dataset, which in turn is determined solely by the distribution family of interest and the target error rate. In the second step, one tests if the combination of the given dataset S, sensitivity bound Δ, and the exponential mechanism with loss $D_S(\hat{\theta})$ satisfy the DP conditions. A part of the privacy budget is used to test this in a differential private manner, such that the subsequent exponential mechanism can depend on the result of this test, i.e., we only proceed to the third step if S passes the test. Otherwise, the process stops and outputs a predefined symbol, $\perp$. In the third step, one releases the DP estimate via the exponential mechanism. This ensures DP for any input S. We are adopting the Propose-Test-Release mechanism pioneered in [77], which we explain in detail in Section 2.1.1. The resulting framework, which we call High-dimensional Propose-Test-Release (HPTR) is provided in Section 3.1.2.

Contributions. We introduce a novel (computationally inefficient) algorithm for differentially private statistical estimation, with the goal of characterizing the achievable sample complexity for various problems with minimal assumptions. The proposed framework, which we call High-dimensional Propose-Test-Release (HPTR), makes a fundamental connection between differential privacy and robust statistics, thus achieving a sample complexity that significantly improves upon other state-of-the-art approaches. HPTR is a generic framework that can be seamlessly applied to various statistical estimation problems, as we demonstrate for mean estimation, linear regression, covariance estimation, and principal component analysis. Further, our analysis technique, which requires minimal assumptions, also seamlessly generalizes to all problem instances of interest.

HPTR uses three crucial components: the exponential mechanism, robust statistics, and the Propose-Test-Release mechanism from [77]. Building upon the inherent adaptivity and flexibility of the exponential mechanism, we propose using a novel loss function (also called a score function in a typical design of exponential mechanisms) that accesses the data only via one-dimensional robust statistics. The use of 1-D robust statistics is critical, because it dramatically reduces the sensitivity. We prove this sensitivity bound using the fundamental concept of resilience, which is central in robust statistics. This novel robust exponential mechanism is incorporated within the PTR framework to ensure that differential privacy is guaranteed on all input datasets, including those that are not necessarily compliant with the statistical assumptions. One byproduct of using robust statistics is that robustness comes for free. HPTR is inherently robust to adversarial corruption of the data and achieves the optimal robust error rate under standard data corruption models.

We present informal version of our main theoretical results in Section 3.1.1. We present HPTR algorithm in detail in Section 3.1.2. We provide a sketch of the proof and the main technical contributions in Section 3.1.3. Detailed explanations of the setting, main results, and the proofs for each instance of the problems are presented in Sections 3.3–3.6 for mean estimation, linear regression, covariance estimation, and principal component analysis, respectively.

Notations. Let $[n] = \{1, 2, \ldots, n\}$. For $x \in \mathbb{R}^d$, we use $\|x\| = (\sum_{i\in[d]}(x_i)^2)^{1/2}$ to denote the Euclidean norm. For $X \in \mathbb{R}^{d_1\times d_2}$, we use $\|X\| = \max_{\|v\|_2=1} \|Xv\|_2$ to denote the spectral norm. The $d \times d$ identity matrix is $\mathbf{I}_{d\times d}$. The Kronecker product is denoted by $x \otimes y$ for $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$, such that $(x \otimes y)_{(i-1)d+j} = x_i y_j$ for $i \in [d_1]$ and $j \in [d_2]$. Whenever it is clear from context, we use S to denote both a set of data points and also the set of indices of those data points. We use $S \sim S'$ to denote that two datasets S, S' of size n are neighbors, such that $d_{\mathrm{TV}}(\hat{p}_S, \hat{p}_{S'}) \le 1/n$ where $d_{\mathrm{TV}}(\cdot)$ is the total variation and $\hat{p}_S$ is the empirical distribution of the data points in S. We use $\mu(S)$ and $\Sigma(S)$ to denote mean and covariance of the data points in a dataset S, respectively. We use μ_p and Σ_p to denote mean

and covariance of the distribution p.

3.1.1 Main results and related work

For each canonical problem of interest in statistical estimation, HPTR can readily be applied to, in most cases, significantly improve upon known achievable sample complexity. Most of the lower bounds we reference are copied in Appendix B.3 for completeness.

3.1.1.1 DP mean estimation

We apply our proposed HPTR framework to the standard DP mean estimation problem, where i.i.d. samples $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ are drawn from a distribution $P_{\mu,\Sigma}$ with an unknown mean μ (which corresponds to θ in the general notation) and an unknown covariance $\Sigma \succ 0$, and we want to produce a DP estimate $\hat{\mu}$ of the mean. The resulting error is measured in Mahalanobis distance, $D_{P_{\mu,\Sigma}}(\hat{\mu}) = \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, which is scale-invariant and naturally captured the uncertainty in all directions.

This problem is especially challenging since we aim for a tight guarantee that adapts to the unknown Σ as measured in the Mahalanobis distance without sufficient samples to directly estimate Σ, as we explain below. Despite being a canonical problem in DP statistics, the optimal sample complexity is not known even for standard distributions: sub-Gaussian and heavy-tailed distributions. We characterize the optimal sample complexity of the two problems by providing the guarantee of HPTR and the matching sample complexity lower bounds. A precise definition of sub-Gaussian distributions is provided in Eq. (3.21).

Theorem 11 (DP sub-Gaussian mean estimation algorithm, Corollary 3.3.13 informal)**.** *Consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ. There exists an (ε, δ)-differentially private algorithm $\hat{\mu}(S)$ that given*

$$n = \tilde{O}_{\xi,\zeta}\Big(\frac{d}{\xi^2} + \frac{d}{\varepsilon\xi}\Big),$$

achieves Mahalanobis error $\|\Sigma^{-1/2}(\hat{\mu}(S)-\mu)\| \le \xi$ *with probability* $1-\zeta$*, where* $\tilde{O}_{\xi,\zeta}$ *hides the logarithmic dependency on* ξ, ζ *and we assume* $\delta = e^{-O(d)}$*.*

HPTR is the first algorithm for sub-Gaussian mean estimation with unknown covariance that matches the best known sample complexity lower bound of $n = \widetilde{\Omega}(d/\xi^2 + d/(\xi\varepsilon))$ from [139, 129] up to logarithmic factors in error ξ and failure probability ζ. Existing algorithms are suboptimal as they require either a larger sample size or strictly Gaussian assumptions.

Advances in DP mean estimation started with computationally efficient approaches of [139, 129, 25]. We discuss the results as follows, and omit the polynomial factors in $\log(1/\delta)$. When the covariance Σ is known, Mahalanobis error ξ can be achieved with $n = \tilde{O}(d/\xi^2 + d/(\xi\varepsilon))$ samples. Under a relaxed assumption that $\mathbf{I}_{d\times d} \preceq \Sigma \preceq \kappa\mathbf{I}_{d\times d}$ with a known κ, $n = \tilde{O}(d/\xi^2 + d/(\xi\varepsilon) + d^{1.5}/\varepsilon)$ samples are required using a specific preconditioning approach tailored for the assumption and the knowledge of κ. For general unknown Σ, $O(d^2/\xi^2 + d^2/(\xi\varepsilon))$ samples are required using an explicit DP estimation of the covariance. Empirically, further gains can be achieved with CoinPress [32].

Computationally inefficient approaches followed with a goal of identifying the fundamental optimal sample complexity with minimal assumptions [36, 4]. For the unknown covariance setting, the best known result under Mahalanobis error is achieved by [34]. Through analyzing the differentially private Tukey median estimator introduced in [160], [34] shows that $n = \tilde{O}(d/\xi^2 + d/(\xi\varepsilon))$ is sufficient even when the covariance is unknown. However, the approach heavily relies on the assumption that the distribution is strictly Gaussian. For sub-Gaussian distributions, [34] proposes a different approach achieving sample complexity of $n = \tilde{O}(d/\xi^2 + d/(\xi\varepsilon^2))$ samples with a sub-optimal $(1/\varepsilon^2)$ dependence.

Beyond the sub-Gaussian setting, it is natural to consider the DP mean estimation for distributions with heavier tails. We apply HPTR framework to the more general mean estimation problems for hypercontractive distributions. A distribution $P_{\mu,\Sigma}$ with mean μ and covariance Σ is (κ, k)-hypercontractive if for all $v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim P_X}[|\langle v, (x-\mu)\rangle|^k] \le \kappa^k (v^\top \Sigma v)^{k/2}$. The assumption of hypercontractivity is similar to the bounded k-th moment assumptions,

except requiring an additional lower bound on the covariance. This additional assumption is necessary for our setting to make sure the Mahalanobis error guarantee is achievable. We state our main result for hypercontractive mean estimation as follows. For simplicity of the statement, we assume k, κ are constants.

Theorem 12 (DP hypercontractive mean estimation algorithm, Corollary 3.3.16 informal)**.** *Consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a (κ, k)-hypercontractive distribution with mean μ and covariance Σ. There exists an (ε, δ)-differentially private algorithm $\hat{\mu}(S)$ that given*

$$n = \tilde{O}_d\Big(\frac{d}{\xi^2} + \frac{d}{\varepsilon\xi^{1+1/(k-1)}}\Big),$$

achieves Mahalanobis error $\|\Sigma^{-1/2}(\hat{\mu}(S) - \mu)\| \le \xi$ with probability at least 0.99, where $\tilde{O}_d$ hides a logarithmic factor on d, and we assumes $\delta = e^{-O(d)}$.

We prove an $n = \Omega(d/\varepsilon\xi^{1+1/(k-1)})$ sample complexity lower bound for hypercontractive DP mean estimation in Proposition 3.3.18 to show the optimality of our upper bound result. Notice that the first term $\tilde{O}_d(d/\xi^2)$ in the upper bound cannot be improved up to logarithmic factors even if we do not require privacy, thus HPTR is the first algorithm that achieves optimal sample complexity for hypercontractive mean estimation under Mahalanobis distance up to logarithmic factors in d. When the covariance is known, an existing DP mean estimator of [135] achieves a stronger $(\varepsilon, 0)$-DP with a similar sample size of $n = \widetilde{O}(d/\xi^2 + d/(\varepsilon\xi^{1+1/(k-1)}))$, and no prior result is known for the unknown covariance case.

3.1.1.2 DP linear regression

We next apply HPTR framework to DP linear regression. Given i.i.d. samples $S = \{(x_i, y_i)\}_{i\in[n]}$ from a distribution $P_{\beta,\Sigma,\gamma^2}$ of a linear model: $y_i = x_i^\top \beta + \eta_i$, where the input $x_i \in \mathbb{R}^d$ has zero mean and covariance Σ and the noise $\eta_i \in \mathbb{R}$ has variance γ^2 satisfying $\mathbb{E}[x_i\eta_i] = 0$, the goal of DP linear regression is to output a DP estimate $\hat{\beta}$ of the unknown model parameter β, without knowledge about the covariance Σ. The resulting error is measured in $D_{P_{\beta,\Sigma,\gamma^2}}(\hat{\beta}) = (1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|$ which is equivalent to the standard *root excess*

risk of the estimated predictor $\hat{\beta}$. Similar to Mahalanobis distance for mean estimation, this is challenging since we aim for a tight guarantee that adapts to the unknown Σ without having enough samples to directly estimate Σ.

Theorem 13 (DP sub-Gaussian linear regression, Corollary 3.4.16 informal)**.** *Consider a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ generated from a linear model $y_i = x_i^\top \beta + \eta_i$ for some $\beta \in \mathbb{R}^d$, where $\{x_i\}_{i\in[n]}$ are i.i.d. sampled from zero-mean d-dimensional sub-Gaussian distribution with unknown covariance Σ, and $\{\eta_i\}_{i\in[n]}$ are i.i.d. sampled from zero mean one-dimensional sub-Gaussian with variance γ^2. We further assume the data x_i and the noise η_i are independent. There exists a (ε, δ)-differentially private algorithm $\hat{\beta}(S)$ that given*

$$n \;=\; \tilde{O}_{\xi,\zeta}\Big(\frac{d}{\xi^2} + \frac{d}{\varepsilon\xi}\Big)\,,$$

achieves error $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}(S)-\beta)\| \le \xi$ with probability $1-\zeta$, where $\tilde{O}_{\xi,\zeta}$ hides the logarithmic dependency on ξ, ζ and we assume $\delta = e^{-O(d)}$.

HPTR is the first algorithm for sub-Gaussian distributions with an unknown covariance Σ that up to logarithmic factors matches the lower bound of $n = \tilde{\Omega}(d/\xi^2 + d/(\xi\varepsilon))$ assuming $\varepsilon < 1$ and $\delta < n^{-1-\omega}$ for some $\omega > 0$ from [40, Theorem 4.1]. An existing algorithm for DP linear regression from [40] is suboptimal as it require Σ to be close to the identity matrix, which is equivalent to assuming that we know Σ. [77] proposes to use PTR and B-robust regression algorithm from [99] for differentially private linear regression under i.i.d. data assumptions (also exponential time), but only asymptotic consistency is proven as $n \to \infty$. Under an alternative setting where the data is deterministically given without any probabilistic assumptions, significant advances in DP linear regression has been made [201, 142, 168, 71, 27, 208, 88, 167, 206, 180]. The state-of-the-art guarantee is achieved in [206, 180] which under our setting translates into a sample complexity of $n = O(d^{1.5}/(\xi\varepsilon))$. The extra $d^{1/2}$ factor is due to the fact that no statistical assumption is made, and cannot be improved under the deterministic setting (not necessarily i.i.d.) that those algorithms are designed for.

Similar to mean estimation, we also consider the DP linear regression for distributions with heavier tails, and apply HPTR framework to the linear regression problem under (k, κ)-hypercontractive distributions (see Definition 3.3.14). HPTR can handle both independent and dependent noise, and we state the independent noise case here the dependent noise case in Section 3.4.3.3. For simplicity of the statement, we assume k, κ are constants.

Theorem 14 (DP hypercontractive linear regression with independent noise, Corollary 3.4.18 informal)**.** *Consider a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ generated from a linear model $y_i = x_i^\top \beta + \eta_i$ for some $\beta \in \mathbb{R}^d$, where $\{x_i\}_{i\in[n]}$ are i.i.d. sampled from zero-mean d-dimensional (κ, k)-hypercontractive distribution with unknown covariance Σ and η_i are i.i.d. sampled from zero mean one-dimensional (κ, k)-hypercontractive distribution with variance γ^2. We further assume the data x_i and the noise $\{\eta_i\}_{i\in[n]}$ are independent. There exists an (ε, δ)-differentially private algorithm $\hat{\beta}(S)$ that given*

$$n \;=\; \tilde{O}_d\Big(\frac{d}{\xi^2} + \frac{d}{\varepsilon \xi^{1+1/(k-1)}} \Big) \,,$$

achieves error $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}(S) - \beta)\| \le \xi$ with probability 0.99, where $\tilde{O}_d$ hides a logarithmic factor on d, and we assume $\delta = e^{-O(d)}$.

The first term in the sample complexity cannot be improved as it matches the classical lower bound of linear regression even without privacy constraint. For the second term, the sub-Gaussian lower bound of $n = \tilde{\Omega}(d/(\varepsilon\xi))$ from [40, Theorem 4.1] continues to hold in the hypercontractive setting. We do not have a matching lower bound for the second term. To the best of our knowledge, HPTR is the first algorithm for linear regression that guarantees (ε, δ)-DP under hypercontractive distributions with independent noise.

When applied to the setting where noise η_i is dependent on the input vector x_i, HPTR is the first algorithm for linear regression that guarantees (ε, δ)-DP. We refer the readers to Section 3.4.3.3 for more detailed description of our result.

3.1.1.3 DP covariance estimation

We present HPTR applied to covariance estimation from i.i.d. samples under a Gaussian distribution $\mathcal{N}(0, \Sigma)$. The main reason for this choice is that the Mahalanobis error $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F$ of the Kronecker product $x_i \otimes x_i$ is proportional to the natural error metric of total variation for Gaussian distributions. The strength of HPTR framework is that it can be seamlessly applied to general distributions, for example sub-Gaussian or heavytailed, but the resulting Mahalanobis error becomes harder to interpret as it involves respective fourth moment tensors.

Theorem 15 (DP Gaussian covariance estimation, Corollary 3.5.9 informal)**.** *Consider a dataset $S = \{x_i\}_{i=1}^n$ of n i.i.d. samples from $\mathcal{N}(0, \Sigma)$. There exists a (ε, δ)-differentially private estimator $\hat{\Sigma}$ that given*

$$n \;=\; \tilde{O}_{\xi,\zeta}\Big(\frac{d^2}{\xi^2} + \frac{d^2}{\xi\varepsilon}\Big)\,,$$

achieves error $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F \le \xi$ with probability $1-\zeta$, where $\tilde{O}_{\xi,\zeta}$ hides the logarithmic dependency on ξ, ζ and we assume $\delta = e^{-O(d)}$.

This Mahalanobis distance guarantee (for the Kronecker product, $\{x_i \otimes x_i\}$, of the samples) implies that the estimated Gaussian distribution is close to the underlying one in total variation distance (see for example [129, Lemma 2.9]): $d_{\mathrm{TV}}(\mathcal{N}(0,\hat{\Sigma}), \mathcal{N}(0,\Sigma)) = O(\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F) = O(\xi)$. The sample complexity is near-optimal, matching a lower bound of $n = \Omega(d^2/\xi^2 + \min\{d^2, \log(1/\delta)\}/(\varepsilon\xi))$ up to a logarithmic factor when $\delta = e^{-\Theta(d)}$. The first term follows from the classical estimation of the covariance without DP. The second term follows from extending the lower bound in [129] constructed for pure differential privacy with $\delta = 0$ and matches the second term in our upper bound when $\delta = e^{-\Theta(d^2)}$. We note that a similar upper bound is achieved by the state-of-the-art (computationally inefficient) algorithm in [4], which improves over HPTR in the lower order terms not explicitly shown in this informal version of our theorem. Both HPTR and [4, 10] improve upon computationally efficient approaches of [139, 129] which require additional assumption that $\mathbf{I}_{d\times d} \preceq \Sigma \preceq \kappa\mathbf{I}_{d\times d}$ with a

known κ. Recently, [133] introduced a novel preconditioning approach that is polynomial time and removes the upper and lower bounds on Σ completely, but requires sample complexity of $n = \tilde{O}(d^2/\xi^2 + d^2\text{polylog}(1/\delta)/(\xi\varepsilon) + d^{5/2}\text{polylog}(1/\delta)/\varepsilon)$.

3.1.1.4 DP principal component analysis

We next apply HPTR to the task of estimating the top PCA direction from i.i.d. sampless

Theorem 16 (DP sub-Gaussian principle component analysis, Corollary 3.6.5)**.** *Consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a zero-mean sub-Gaussian distribution with unknown covariance Σ. There exists an (ε, δ)-differentially private estimator $\hat{u}$ that given*

$$n \;=\; \tilde{O}_{\xi,\zeta}\Big(\frac{d}{\xi^2} + \frac{d}{\varepsilon\xi}\Big)\,,$$

achieves error $1 - \frac{\hat{u}^\top\Sigma\hat{u}}{\|\Sigma\|} \le \xi$ with probability $1-\zeta$, where $\tilde{O}_{\xi,\zeta}$ hides the logarithmic dependency on ξ, ζ and we assume $\delta = e^{-O(d)}$.

HPTR is the first estimator for sub-Gaussian distributions to nearly match the information-theoretic lower bound of $n = \Omega(d/\xi^2 + \min\{d, \log(1/\delta)\}/(\varepsilon\xi))$ as we showed in Proposition 3.6.6. The first term $\Omega(d/\xi^2)$ is unavoidable even without DP (Proposition 3.6.7). The second term in the lower bound follows from Proposition 3.6.6, which matches the second term in the upper bound when $\delta = e^{-\Theta(d)}$. Existing DP PCA approaches from [33, 46, 136, 80, 102, 103, 100] are designed for arbitrary samples not necessarily drawn i.i.d. and hence require a larger samples size of $n = \tilde{O}(d/\xi^2 + d^{1.5}/(\xi\varepsilon))$. This is also unavoidable for more general deterministic data, as it matches an information theoretic lower bound [80] under weaker assumptions on the data than i.i.d. Gaussian.

Theorem 17 (DP hypercontractive principle component analysis, Corollary 3.6.10)**.** *Consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a zero-mean (κ, k)-hypercontractive distribution with unknown covariance Σ. There exists an (ε, δ)-differentially private estimator $\hat{u}$ that given*

$$n \;=\; \tilde{O}_{\xi,d}\Big(\frac{d}{\xi^{(2k-2)/(k-2)}} + \frac{d}{\varepsilon\xi^{1+2/(k-2)}}\Big)\,,$$

achieves error $1 - \frac{\hat{u}^\top \Sigma \hat{u}}{\|\Sigma\|} \le \xi$ *with probability* 0.99*, where* $\tilde{O}_{\xi,d}$ *hides the logarithmic dependency on* ξ, d *and we assume* $\delta = e^{-O(d)}$*.*

HPTR is the first estimator for hypercontractive distributions to guarantee differential privacy for PCA with sample complexity scaling as $O(d)$. As a complement of our algorithm, we proved a $n = \Omega(d/\xi^2 + \min\{d, \log(1/\delta)\}/(\xi^{1+2/(k-2)}\varepsilon))$ information-theoretic lower bound in Proposition 3.6.11. The first term $\Omega(d/\xi^2)$ in the lower bound is needed even without DP, and there is a gap of factor $O(\xi^{-2/(k-2)})$ compared to the first term in our upper bound. The second term in the lower bound matches the last term in the upper bound when $\delta = e^{-\Theta(d)}$.

3.1.2 Algorithm

The proposed High-dimensional Propose-Test-Release (HPTR) is not computationally efficient, as the TEST step requires enumerating over a certain neighborhood of the input dataset and the RELEASE step requires enumerating over all directions in high dimension. The strengths of HPTR is that (*i*) the same framework can be seamlessly applies to many problems as we demonstrate in Sections 3.3–3.6; (*ii*) a unifying recipe can be applied for all those instances to give tight utility guarantees as we explicitly prescribe in Section 3.1.2.1; and (*iii*) the algorithm is simple and the analysis is clear such that it is transparent how the distribution family of interest translates into the utility guarantee (via resilience).

As a byproduct of the simplicity of the algorithm and clarity of the analysis, we achieve the state-of-the-art sample complexity for all those problem instances with minimal assumptions, oftentimes nearly matching the information theoretic lower bounds. As a byproduct of the use of robust statistics, we guarantee robustness against adversarial corruption for free (e.g., Theorems 20, 22, 24).

We describe the framework for general statistical estimation problem where data is drawn i.i.d. from a distribution represented by two unknown parameters $\theta \in \mathbb{R}^p$ and ϕ and is coming from a known family of distributions. An example of a problem instance of this type would be mean estimation from heavy-tailed distributions that are (κ, k)-hypercontractive with

unknown mean μ (which in the general notation is θ) and unknown covariance Σ (which corresponds to ϕ).

The main component is an exponential mechanism in RELEASE step below that uses a loss function $D_S(\hat{\theta})$ and a support $B_{\tau,S}$ defined as

$$B_{\tau,S} = \{\hat{\theta} \in \mathbb{R}^p : D_S(\hat{\theta}) \leq \tau\} .$$

Bounding the support of the exponential mechanism is important since the sensitivity also depends on $\hat{\theta}$ in many problems of interest. We discuss this in detail in the example of mean estimation in Section 3.3.2.2. The specific choices of the threshold τ only depend on the tail of the distribution family of interest and not the parameters θ or ϕ or the data. In particular, we use the resilience property of the distribution family to prescribe the choice of τ for each problem instance that gives us the tight utility guarantees. As explained in Section 3.1, we use one-dimensional robust statistics to design the loss functions, which we elaborate for each problem instances in Sections 3.3–3.6, where we also explain how to choose the sensitivity for each case based on the resilience of the distribution family only.

After we PROPOSE the choice of the sensitivity Δ and threshold τ for the problem instance in hand, we TEST to make sure that the given dataset S is consistent with the assumptions made when selecting $D_S(\hat{\theta})$, Δ, and τ. This is done by testing the safety of the exponential mechanism, by privately checking the margin to safety, i.e., how many data points need to be changed from S for the exponential mechanism to violate differential privacy conditions. If the margin is large enough, HPTR proceeds to RELEASE. Otherwise, it halts and outputs $\perp$. A set of such unsafe datasets is defined as

$$\begin{aligned}\mathsf{UNSAFE}_{(\varepsilon,\delta,\tau)} \;=\; \Big\{ & S' \subseteq \mathbb{R}^{d\times n} \,|\, \exists S'' \sim S' \text{ and } \exists E \subseteq \mathbb{R}^p \text{ such that} \\ & \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E) > e^{\varepsilon}\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) + \delta \text{ or } \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) > e^{\varepsilon}\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E) + \delta \Big\} ,\end{aligned} \tag{3.2}$$

where $r_{(\varepsilon,\Delta,\tau,S)}$ denotes the pdf of the exponential mechanism in Eq. (3.3). Given a loss (or a distance) function, $D_S(\hat{\theta})$, which is a surrogate for the target measure of error, $D_\phi(\hat{\theta}, \theta)$, High-dimensional Propose-Test-Release (HPTR) proceeds as follows:

1. **Propose**: Propose a target bound Δ on local sensitivity and a target threshold τ for safety.

2. **Test**:

   2.1. Compute the safety margin $m_\tau = \min_{S'} d_H(S, S')$ such that $S' \in \mathsf{UNSAFE}_{(\varepsilon/2,\delta/2,\tau)}$.

   2.2. If $\hat{m}_\tau = m_\tau + \mathrm{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta)$, then output $\perp$, and otherwise continue.

3. **Release**: Output $\hat{\theta}$ sampled from a distribution with a pdf:

$$r_{(\varepsilon,\Delta,\tau,S)}(\hat{\theta}) \;=\; \begin{cases} \frac{1}{Z}\,\exp\Big\{-\frac{\varepsilon}{4\Delta}D_S(\hat{\theta})\Big\} & \text{if } \hat{\theta} \in B_{\tau,S}\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{3.3}$$

   where $Z = \int_{B_{\tau,S}} \exp\{-(\varepsilon D_S(\hat{\theta}))/(4\Delta)\}d\hat{\theta}$.

It is straightforward to show that (ε, δ)-differential privacy is satisfied for all input S.

Theorem 18. *For any dataset $S \subset \mathcal{X}^n$, distance function $D_S : \mathbb{R}^p \to \mathbb{R}$ on that dataset, and parameters $\varepsilon, \delta, \Delta$ and τ, HPTR is (ε, δ)-differentially private.*

Proof. The differentially private margin $\hat{m}_\tau$ is $(\varepsilon/2, 0)$-differentially private, because the sensitivity of the margin is one, and we are adding a Laplace noise with parameter $2/\varepsilon$. The Test step (together with the exponential mechanism) is $(0, \delta/2)$-differentially private since there is a probability $\delta/2$ event that a unsafe dataset S with a small margin m_τ is classified as a safe dataset and passes the test. On the complimentary event, namely, that the dataset that passed the Test is indeed safe, the Release step is $(\varepsilon/2, \delta/2)$-differentially private since we use $\mathsf{UNSAFE}_{(\varepsilon/2,\delta/2,\tau)}$ in the Test step. □

3.1.2.1 Utility analysis of HPTR for statistical estimation

We prescribe the following three-step recipe as a guideline for applying HPTR to each specific statistical estimation problem and obtaining a utility guarantee. Consider a problem of estimating an unknown θ from samples from a generative model $P_{\theta,\phi}$, where the error is measured in $D_\phi(\hat{\theta}, \theta)$.

- Step 1: Design a surrogate $D_S(\hat{\theta})$ for the target error metric $D_\phi(\hat{\theta}, \theta)$ using only one-dimensional robust statistics on S.

- Step 2: Assuming *resilience* of the dataset, propose an appropriate sensitivity bound Δ and threshold τ and analyze the utility of HPTR.

- Step 3: For each specific family of generative models $P_{\theta,\phi}$ with a known tail bound, characterize the resulting resilience and substitute it in the utility analysis from the previous step, which gives the final guarantee.

We demonstrate how to apply this recipe and carry out the utility analysis for mean estimation (Section 3.3), linear regression (Section 3.4), covariance estimation (Section 3.5), and PCA (Section 3.6). We explain and justify the use of one-dimensional robust statistics in Step 1 and the assumption on the resilience of the dataset in Step 2 in the next section using the mean estimation problem as a canonical example. It is critical to construct $D_S(\hat{\theta})$ using only one-dimensional and robust statistics; this ensures that $D_S(\hat{\theta})$ has a small sensitivity as demonstrated in Section 3.3.1. We prove error bounds only assuming resilience of the dataset; this relies on a fundamental connection between sensitivity and resilience as explained in Section 3.3.2.

3.1.3 Technical contributions and proof sketch

We use the canonical example of mean estimation to explain our proof sketch. For i.i.d. samples from a sub-Gaussian distribution $P_{\mu,\Sigma}$ with mean μ and covariance Σ, we show in Theorem 19 that HPTR achieves a near optimal sample complexity of $n = \tilde{O}(d/\alpha^2 + d/(\alpha\varepsilon))$ to get Mahalanobis error $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = \tilde{O}(\alpha)$ for some target accuracy $\alpha \in [0, 1]$.

Our proof strategy is to first show that the robust one-dimensional statistics have small sensitivity if the dataset is resilient. Consequently, the loss function $D_S(\hat{\mu})$ has a small *local* sensitivity, i.e. the sensitivity is small if restricted to $\hat{\mu}$ close to μ and if the dataset is resilient. To ensure DP, we run RELEASE only when those two locality conditions are satisfied; we first

PROPOSE the sensitivity Δ and a threshold τ, and then we TEST that DP guarantees are met on the given dataset with those choices. We prove that resilient datasets pass this safety test with a high probability and achieve the desired accuracy. We give a sketch of the main steps below.

One-dimensional robust statistics have small sensitivity on resilient datasets. A set $S = \{x_i \in \mathbb{R}^d\}_{i\in[n]}$ of i.i.d. samples from a sub-Gaussian distribution has the following resilience property with probability $1-\zeta$ if $n = \tilde{\Omega}(d/\alpha^2)$, where $\tilde{\Omega}$ hides polylogarithmic factors α and the failure probability ζ:

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i - \mu\rangle\right| \;\le\; 2\sigma_v\sqrt{\log(1/\alpha)}\,, \text{ and } \left|\frac{1}{|T|}\sum_{x_i\in T}\left(\langle v, x_i-\mu\rangle^2 - \sigma_v^2\right)\right| \;\le\; 2\sigma_v^2\log(1/\alpha)\,,$$

for any subset $T \subset S$ of size at least αn and for any unit norm $v \in \mathbb{R}^d$ where $\sigma_v^2 = v^\top\Sigma v$ (Lemma 3.3.12). This means that the α-tail of the distribution (when projected down to one dimension) cannot be too far from the true one in mean and variance. For mean estimation, we use the loss function of $D_S(\hat{\mu}) = \max_{v\in\mathbb{R}^d,\|v\|=1}\langle v, \hat{\mu} - \mu(\mathcal{M}_{v,\alpha})\rangle/\sqrt{v^\top\Sigma(\mathcal{M}_{v,\alpha})v}$, where $\mu(T)$ and $\Sigma(T)$ are mean and covariance of a dataset T and $\mathcal{M}_{v,\alpha} \subset S$ is defined as follows. For each direction v, we partition S into three sets $\mathcal{T}_{v,\alpha}, \mathcal{M}_{v,\alpha}$, and $\mathcal{B}_{v,\alpha}$. $\mathcal{T}_{v,\alpha} \subset S$ is the subset of datapoints corresponding to the largest αn datapoints in $\{\langle v, x_i\rangle\}_{x_i\in S}$, the projected data points in the direction v. $\mathcal{B}_{v,\alpha} \subset S$ corresponds to the smallest αn values, and $\mathcal{M}_{v,\alpha} \subset S$ is the remaining $(1-2\alpha)n$ data points.

We show that the robust projected mean, $\langle v, \mu(\mathcal{M}_{v,\alpha})\rangle$ has sensitivity $O(\sigma_v\sqrt{\log(1/\alpha)}/n)$. Under the resilience above, the top α-tail, $\mathcal{T}_{v,\alpha}$, has the empirical mean that is no more than $O(\sigma_v\sqrt{\log(1/\alpha)}\,)$ away from the true projected mean $\langle v, \mu\rangle$, and the same is true for $\mathcal{B}_{v,\alpha}$. It follows that there exists at least one data point in $\mathcal{T}_{v,\alpha}$ and one data point in $\mathcal{B}_{v,\alpha}$ that are no more than $O(\sigma_v\sqrt{\log(1/\alpha)})$ away from μ_v. This implies that the range of the middle subset $\mathcal{M}_{v,\alpha}$ is provably bounded by $O(\sigma_v\sqrt{\log(1/\alpha)})$, and the sensitivity of the robust mean $\langle v, \mu(\mathcal{M}_{v,\alpha})\rangle$ is guaranteed to be $O(\sigma_v\sqrt{\log(1/\alpha)}/n)$. We can similarly show that $v^\top\Sigma(\mathcal{M}_{v,\alpha})v$ has sensitivity $O(\sigma_v^2\log(1/\alpha)/n)$.

Under the above sensitivity bounds for the one dimensional statistics, it follows (for example, in Eq. (3.20)) that the sensitivity of the loss function $D_S(\hat{\mu})$ is bounded by $O(\sqrt{\log(1/\alpha)}/n)$ as long as $D_S(\hat{\mu}) \leq \tau := C\alpha\sqrt{\log(1/\alpha)}$ and the dataset is resilient. It is worth noting here that since the sensitivity is only small when $D_S(\hat{\mu}) \leq \tau$, our exponential mechanism only samples from the set $B_{\tau,S}$, which contains only the hypotheses with small scores. We handle this locality with TEST step that checks that the DP conditions are satisfied for the given dataset and the choice of $\Delta := C'\sqrt{\log(1/\alpha)}/n$ and $\tau := C\alpha\sqrt{\log(1/\alpha)}$. It is critical for ensuring DP that these choices only depend on the resilience (which is the property of the distribution family of interest, which in this case is sub-Gaussian) and the target accuracy, and not on the dataset S.

Sample complexity analysis. Assuming the sensitivity is bounded by $\Delta = C'\sqrt{\log(1/\alpha)}/n$, which we ensure with the safety test, we analyze the utility of the exponential mechanism. For a target accuracy of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = O(\alpha\sqrt{\log(1/\alpha)})$, we consider two sets, $B_{\text{out}} = \{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq c_0\alpha\sqrt{\log(1/\alpha)}\}$ and $B_{\text{in}} = \{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq c_1\alpha\sqrt{\log(1/\alpha)}\}$, for some $c_0 > c_1$. The exponential mechanism achieves accuracy $c_0\alpha\sqrt{\log(1/\alpha)}$ with probability $1 - \zeta$ if

$$\mathbb{P}(\hat{\mu} \notin B_{\text{out}}) \leq \frac{\mathbb{P}(\hat{\mu} \notin B_{\text{out}})}{\mathbb{P}(\hat{\mu} \in B_{\text{in}})} \lesssim \frac{\text{Vol}(B_{\tau,S})}{\text{Vol}(B_{\text{in}})} \frac{e^{-\frac{\varepsilon}{4\Delta}c_0\alpha\sqrt{\log(1/\alpha)}}}{e^{-\frac{\varepsilon}{4\Delta}c_1\alpha\sqrt{\log(1/\alpha)}}} \leq e^{O(d)}e^{-\frac{\varepsilon}{4\Delta}(c_0-c_1)\alpha\sqrt{\log(1/\alpha)}} \leq \zeta \,,$$

where the second inequality requires $D_S(\hat{\mu}) \simeq \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, which we show in Lemma 3.3.6. Since $\Delta = O(\sqrt{\log(1/\alpha)}/n)$, it is sufficient to have a large enough c_0 and $n = \tilde{O}((d + \log(1/\zeta))/(\alpha\varepsilon))$ with a large enough constant. Together with the sample size required to ensure resilience, this gives the desired sample complexity of $n = \tilde{O}(d/\alpha^2 + (d+\log(1/\zeta))/(\alpha\varepsilon))$ where $\tilde{O}$ hides polylogarithmic factors in $1/\alpha$ and $1/\delta$.

Safety test. We are left to show that for a resilient dataset, the failure probability of the safety test, $\mathbb{P}(m_\tau + \text{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta))$, is less than ζ. This requires the safety margin to be large enough, i.e. $m_\tau \geq k^* = (2/\varepsilon)\log(4/(\delta\zeta))$. Recall that the safety margin is defined as the Hamming distance to the closest dataset to S where the $(\varepsilon/2, \delta/2)$-DP condition of

the exponential mechanism is violated. We therefore need to show that the DP condition is satisfied for not only S but any dataset S' at Hamming distance at most k^* from S. We treat S' as a *corrupted* version of a resilient S by a fraction k^*/n-corruption. Since we are using robust statistics that are designed to be robust against data corruption, we can show that the corrupted resilient set still has a low sensitivity for $D_{S'}(\hat{\mu})$. Building upon the proof techniques developed in [34] for a safety test for a Tukey median based exponential mechanism, we use the fact that S' is a corrupted version of a resilient dataset S to show that the safety test passes with high probability.

3.2 Preliminaries

We give the backgrounds on differential privacy and the Propose-Test-Release mechanism. We say two datasets S and S' of the same size are neighboring if the Hamming distance between them is at most one. There is another equally popular definition where injecting or deleting one data point to S is considered as a neighboring dataset. All our analysis generalizes to that definition also, but notations get slightly heavier.

Definition 3.2.1 ([78]). *We say a randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if for all neighboring databases $S \sim S' \in \mathcal{X}^n$, and all $Y \subseteq \mathcal{Y}$, we have $\mathbb{P}(M(S) \in Y) \leq e^{\varepsilon}\mathbb{P}(M(S') \in Y) + \delta$.*

HPTR relies on the exponential mechanism for its adaptivity and flexibility.

Definition 3.2.2 (Exponential mechanism [164]). *The exponential mechanism $M_{\text{exp}} : \mathcal{X}^n \to \Theta$ takes database $S \in \mathcal{X}^n$, candidate space Θ, score function $D_S(\hat{\theta})$ and sensitivity Δ as input, and select output with probability proportional to $\exp\{-\varepsilon D_S(\hat{\theta})/2\Delta\}$.*

The exponential mechanism is $(\varepsilon, 0)$-DP if the sensitivity of $D_S(\hat{\theta})$ is bounded by Δ.

Lemma 3.2.3 ([164]). *If $\max_{\hat{\theta}\in\Theta} \max_{S\sim S'} |D_S(\hat{\theta}) - D_{S'}(\hat{\theta})| \leq \Delta$, then the exponential mechanism is $(\varepsilon, 0)$-DP.*

Starting from the seminal paper [77], there are increasing efforts to apply differential privacy to statistical problems, where the dataset consists of i.i.d. samples from a distribution. There are two main challenges. First, the support is typically not bounded, and hence, the sensitivity is unbounded. [77] proposed to resolve this by using robust statistics, such as median, to estimate the mean. The second challenge is that while median is quite insensitive on i.i.d. data, this low sensitivity is only local and holds only for i.i.d. data from a certain class of distributions. This led to the original definition of local sensitivity in the following.

Definition 3.2.4 (Local Sensitivity)**.** *We define local sensitivity of dataset* $S \in \mathcal{X}^n$ *and function* $f : \mathcal{X}^n \to \mathbb{R}$ *as* $\Delta_f(S) := \max_{S' \sim S} |f(S) - f(S')|$.

[77] introduced the Propose-Test-Release mechanism to resolve both issues. First, a certain robust statistic $f(S)$, such as median, mode, Inter-Quantile Range (IQR), or B-robust regression model [99] is chosen as a query. It can be used to approximate a target statistic of interest, such as mean, range, or linear regression model, or the robust statistic itself could be the target. Then, the PTR mechanism proceeds in three steps. In the propose step, a local sensitivity Δ is proposed such that $\Delta_f(S) \leq \Delta$ for all S that belongs to a certain family. In the test step, a safety margin m, which is how many data points have to be changed to violate the local sensitivity, is computed and a private version of the safety margin, $\hat{m}$, is compared with a threshold. If the safety margin is large enough, then the algorithm outputs $f(S)$ via a Laplace mechanism with parameter $2\Delta/\varepsilon$. Otherwise, the algorithm halts and outputs $\perp$.

Definition 3.2.5 (Propose-Test-Release (PTR) [77, 198])**.** *For a query function* $f : \mathcal{X}^n \to \mathbb{R}$*, the* PTR *mechanism* $M_{\text{PTR}} : \mathcal{X}^n \to \mathbb{R}$ *proceeds as follows:*

1. **Propose***: Propose a target bound* $\Delta \geq 0$ *on local sensitivity.*

2. **Test***:*

   2.1. *Compute* $m = \min_{S'} d_H(S, S')$ *such that local sensitivity of* S' *satisfies* $\Delta_f(S') \geq \Delta$.

2.2. If $\hat{m} = m + \mathrm{Lap}(2/\varepsilon) < (2/\varepsilon)\log(1/\delta)$ *then output* $\perp$*, and otherwise continue.*

3. **Release***: Output* $f(S) + \mathrm{Lap}(2\Delta/\varepsilon)$.

It immediately follows that PTR is (ε, δ)-differentially private for any input dataset.

Lemma 3.2.6 ([77, 198])**.** M_{PTR} *is* (ε, δ)*-DP*

Given a robust statistic of interest, the art is in identifying the family of datasets with small local sensitivity and showing that the sensitivity is small enough to provide good utility. For example, for privately releasing the mode, for the family of distributions whose occurrences of the mode is at least $(4/\varepsilon)\log(1/\delta)$ larger than the occurrences of the second most frequent value, the local sensitivity is zero and PTR outputs the true mode with probability at least $1 - \delta$ [198]. Such a specialized PTR mechanism for zero local sensitivity is also called the stability based method.

In general, a naive method of computing m in the TEST step requires enumerating over all possible databases $S \in \mathcal{X}^n$. For typical one-dimensional data/statistics, for example median estimation, this step can be computed efficiently. This led to a fruitful line of research in DP statistics on one-dimensional data. [77, 35] propose PTR mechanisms for the range and the median of of a 1-D smooth distribution and [20, 18, 35] propose PTR mechanisms that can estimating median and mean of a 1-D sub-Gaussian distribution. The stability-based method introduced in [198] can be used to release private histograms, among other things, which can be subsequently used as a black box to solve several important problems including range estimation of a 1-D sub-Gaussian distribution [139, 129, 160] or a 1-D heavy-tailed distribution [135, 160], and general counting queries. PTR and stability-based mechanisms are powerful tools when estimating robust statistics of a distribution from i.i.d. samples.

Even if computational complexity is not concerned, however, directly applying PTR to high dimensional distributions can increase the statistical cost significantly, which has limited the application of PTR. One exception is the recent work of [34]. For the mean estimation problem with Mahalanobis error metric of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, the private Tukey

median mechanism introduced in [160] is studied. One major limitation of the utility analysis is that private Tukey median requires the support to be bounded. In [160], this is circumvented by assuming the covariance Σ is known, in which case one can find a support with, for example, the private histogram of [198]. Instead, [34] proposed using private Tukey median inside the PTR mechanism and designed an advanced safety test for high-dimensional problems. This naturally bounds the support that adapts to the geometry of the problem without explicitly and privately estimating Σ. One notable byproduct of this approach is that the resulting exponential mechanism is no longer pure DP, but rather (ε, δ)-DP. This is because the resulting exponential mechanism has a support that depends on the dataset S, and hence two exponential mechanisms on two neighboring datasets have different supports. The limitations of the private Tukey median are that (i) it requires symmetric distributions, like Gaussian distributions, and do not generalize to even sub-Gaussian distributions, and (ii) it only works for mean estimation. To handle the first limitation, [34] propose another PTR mechanism using Gaussian noise, which works for more general sub-Gaussian distributions but achieves sub-optimal sample complexity.

HPTR builds upon this advanced PTR with the high-dimensional safety test from [34]. However, there are major challenges in applying this safety test to HPTR, which we overcome with the resilience property of the dataset and the robustness of the loss function. For private Tukey median, the sensitivity is always one for any $\hat{\mu}$ and any S, and the only purpose of the safety test is to ensure that the support is not too different between two neighboring datasets. For HPTR, the sensitivity is local in two ways: it requires S to be resilient and the estimate $\hat{\mu}$ to be sufficiently close to μ. To ensure a large enough margin when running the safety test, HPTR requires this local sensitivity to hold not just for the given S but for all S' within some Hamming distance from S. We use the fact that this larger neighborhood is included in an even larger set of databases that are adversarial corruption of the α-fraction of the original resilient dataset S with a certain choice of α. The robustness of our loss function implies that the bounded sensitivity is preserved under such corruption of a resilient dataset. This is critical in proving that a resilient dataset passes the safety test with high probability.

We take a first-principles approach to design a universal framework for DP statistical estimation that blends exponential mechanism, robust statistics, and PTR. The exponential mechanism in HPTR adapts to the geometry of the problem without explicitly estimating any other parameters and also gives us the flexibility to apply to a wide range of problems. The choice of the loss functions that only depend on one-dimensional statistics is critical in achieving the low sensitivity, which directly translates into near optimal utility guarantees for several canonical problems. Ensuring differential privacy is achieved by building upon the advanced PTR framework of [34], with a few critical differences. Notably, the safety analysis uses the resilience of robust statistics in a fundamental way.

On the other hand, there is a different way of handling local sensitivity, which is known as smooth sensitivity. Introduced in [171], smooth sensitivity is a smoothed version of local sensitivity on the neighborhood of the dataset, defined as

$$\Delta_f^{\text{smooth}}(S) \;=\; \max_{S' \in \mathcal{X}^n} \{\Delta_f(S') e^{-\varepsilon d_{\text{H}}(S,S')}\}$$

Note that, in general, computing smooth sensitivity is also computationally inefficient with an exception of [19]. Using smooth sensitivity, [152, 181, 44, 19] leverage robust M-estimators for differentially private estimation and inference. The intuition is based on the fact that the influence function of the M-estimators can be used to bound the smooth sensitivity. The applications include: linear regression, location estimation, generalized linear models, private testing. However, these approaches require restrictive assumptions on the dataset that need to be checked (for example via PTR) and fine-grained analyses on the statistical complexity is challenging; there is no sample complexity analysis comparable to ours. One exception is [39], which proposes a smooth sensitivity based approach and gives an upper bound on the sub-Gaussian mean estimation error for a finite n, but only for one-dimensional data.

3.3 Mean estimation

In a standard mean estimation, we are given i.i.d. samples $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ drawn from a distribution $P_{\mu,\Sigma}$ with an unknown mean μ (which corresponds to θ in the general notation)

and an unknown covariance $\Sigma \succ 0$ (which corresponds to ϕ in the general notation), and we want to produce a DP estimate $\hat{\mu}$ of the mean. The resulting error is best measured in Mahalanobis distance, $D_\Sigma(\hat{\mu}, \mu) = \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, because this is a scale-invariant distance; every direction has unit variance after whitening by Σ.

This problem is especially challenging since we aim for a tight guarantee that adapts to the unknown Σ as measured in the Mahalanobis distance without enough samples to directly estimate Σ (see Section 3.1.1 for a survey). Despite being a canonical problem in DP statistics, the optimal sample complexity is not known even for standard sub-Gaussian and heavy-tailed distributions. We characterize the optimal sample complexity by showing that HPTR matches the known lower bounds in Section 3.3.3. This follows directly from the general three-step strategy outlined in Section 3.1.2.1.

3.3.1 Step 1: Designing the surrogate $D_S(\hat{\mu})$ for the Mahalanobis distance

We want to privately release $\hat{\mu}$ with small Mahalanobis distance $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$. In the exponential mechanism in RELEASE step, we propose using the surrogate distance,

$$D_S(\hat{\mu}) \quad = \quad \max_{v:\|v\|\le 1} \frac{\langle v, \hat{\mu}\rangle - \mu_v(\mathcal{M}_{v,\alpha})}{\sigma_v(\mathcal{M}_{v,\alpha})} \,, \tag{3.4}$$

where the robust one-dimensional mean $\mu_v(\mathcal{M}_{v,\alpha})$ and variance $\sigma_v^2(\mathcal{M}_{v,\alpha})$ are defined as follows. We partition $S = \{x_i\}_{i=1}^n$ into three sets, $\mathcal{B}_{v,\alpha}$, $\mathcal{M}_{v,\alpha}$, and $\mathcal{T}_{v,\alpha}$, by considering a set of projected data points $S_v = \{\langle v, x_i\rangle\}_{x_i \in S}$ and letting $\mathcal{B}_{v,\alpha}$ be the data points corresponding to the subset of bottom $(2/5.5)\alpha n$ data points with the smallest values in S_v, $\mathcal{T}_{v,\alpha}$ be the subset of the top $(2/5.5)\alpha n$ data points with the largest values, and $\mathcal{M}_{v,\alpha}$ be the subset of remaining $(1 - (4/5.5)\alpha)n$ data points. For a fixed direction v, define

$$\mu_v(\mathcal{M}_{v,\alpha}) \;=\; \frac{1}{|\mathcal{M}_{v,\alpha}|} \sum_{x_i \in \mathcal{M}_{v,\alpha}} \langle v, x_i\rangle \;, \text{ and } \; \sigma_v^2(\mathcal{M}_{v,\alpha}) = \frac{1}{|\mathcal{M}_{v,\alpha}|} \sum_{x_i \in \mathcal{M}_{v,\alpha}} (\langle v, x_i\rangle - \mu_v(\mathcal{M}_{v,\alpha}))^2 \tag{3.5}$$

which are robust estimates of the population projected mean $\mu_v = \langle v, \mu\rangle$ and the population projected variance $\sigma_v^2 = v^\top \Sigma v$.

General guiding principles for designing $D_S(\hat{\mu})$. We propose the following three design principles that apply more generally to all problem instances of interest. The first guideline is that it should recover the target error metric $D_\Sigma(\hat{\mu}, \mu) = \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$ when we substitute the population statistics, e.g., μ_v and σ_v for mean estimation, for their robust counterparts: $\mu_v(\mathcal{M}_{v,\alpha})$ and $\sigma_v(\mathcal{M}_{v,\alpha})$. This ensures that minimizing $D_S(\hat{\mu})$ is approximately equivalent to minimizing the target metric $D_\Sigma(\hat{\mu}, \mu) = \|\Sigma^{-1/2}(\hat{\mu}-\mu)\|$ (Lemma 3.3.6). For mean estimation, this equivalence is shown in the following lemma.

Lemma 3.3.1. *For any $\mu \in \mathbb{R}^d$ and $0 \prec \Sigma \in \mathbb{R}^{d\times d}$, let $\mu_v = \langle v, \mu\rangle$ and $\sigma_v^2 = v^\top \Sigma v$. Then, we have*

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \;=\; \max_{v:\|v\|\le 1} \frac{\langle v, \hat{\mu}\rangle - \mu_v}{\sigma_v} \,.$$

Proof. Let $\hat{\mu} - \mu = \sum_{\ell=1}^d a_\ell u_\ell$ with $a_\ell = \langle u_\ell, \hat{\mu} - \mu\rangle$, $\|a\| = \|\hat{\mu} - \mu\|$ and u_ℓ's are the singular vectors of Σ. Similarly, let $v = \sum_{\ell=1}^d b_\ell u_\ell$ with $\|b\| = 1$. Then, we have

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|^2 = \sum (a_\ell^2/\sigma_\ell) \text{ and } \frac{\langle v, (\hat{\mu} - \mu)\rangle}{\sigma_v} = \frac{\langle a, b\rangle}{\sqrt{\sum b_\ell^2 \sigma_\ell}} \,.$$

From Cauchy-Schwarz, we have $\langle a, b\rangle^2 \le (\sum b_\ell^2 \sigma_\ell)(\sum a_\ell^2 \sigma_\ell^{-1})$, which proves that

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \ge \max_{v:\|v\|=1} (1/\sigma_v)\langle v, (\hat{\mu} - \mu)\rangle \,.$$

To show equality, we find v that makes Cauchy-Schwarz inequality tight. Let $v = \sum_{\ell=1}^d b_\ell u_\ell$ with a choice of $b_\ell = (1/Z) a_\ell \sigma_\ell^{-1}$ and $Z = \sqrt{\sum_\ell a_\ell^2 \sigma_\ell^{-2}}$. This implies $\|b\| = 1$ and

$$\langle a, b\rangle \;=\; \frac{1}{Z}\sum_{\ell=1}^d (1/\sigma_\ell) a_\ell^2, \text{ and } \quad \sqrt{\sum b_\ell^2 \sigma_\ell} \;=\; \frac{1}{Z}\sqrt{\sum_{\ell=1}^d (1/\sigma_{u_\ell}) a_\ell^2} \,,$$

which implies that there exists a v such that $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| = (1/\sigma_v)\langle v, \hat{\mu}-\mu\rangle$ and $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| \le \max_{v:\|v\|=1}(1/\sigma_v)\langle v, \hat{\mu} - \mu\rangle$. □

The second guideline is that $D_S(\hat{\mu})$ should depend only on the *one-dimensional* statistics of the data. This is critical since the sensitivity of high-dimensional statistics increases with the

ambient dimension d. For example, consider using the robust mean estimate $\hat{\mu}_{\text{robust}}(S) \in \mathbb{R}^d$ from [73] and using the Euclidean distance $D_S(\hat{\mu}) = \|\hat{\mu} - \hat{\mu}_{\text{robust}}(S)\|$ in the exponential mechanism, where we are assuming $\Sigma = \mathbf{I}$ for simplicity. It can be shown that, even for Gaussian distributions, this requires $n = \widetilde{\Omega}(d^{3/2}/(\varepsilon\alpha) + d/\alpha^2)$ samples to achieve an accuracy of $\|\hat{\mu} - \mu\| = \widetilde{O}(\alpha)$. This is significantly sub-optimal compared to what HPTR achieves in Corollary 3.3.13, which leverages the fact that sensitivity of one-dimensional statistics is dimension-independent.

The last guideline is to use robust statistics. Robust statistics have small sensitivity on *resilient* datasets, which is critical in achieving the near-optimal guarantees. We elaborate on it in Section 3.3.2.2.

3.3.2 Step 2: Utility analysis under resilience

For utility, we prefer smaller Δ and τ to ensure that the exponential mechanism samples $\hat{\mu}$ closer to the minimum of $D_S(\hat{\mu}) \approx \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$. However, aggressive choices can violate the DP condition and hence fail the safety test. Near-optimal utility can be achieved by selecting Δ and τ based on the *resilience* of the dataset, defined as follows.

Definition 3.3.2 (Resilience for mean estimation [186, 217])**.** *For some $\alpha \in (0, 1)$, $\rho_1 \in \mathbb{R}_+$, and $\rho_2 \in \mathbb{R}_+$, we say a set of n data points S_{good} is (α, ρ_1, ρ_2)-resilient with respect to (μ, Σ) if for any $T \subset S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$, the following holds for all $v \in \mathbb{R}^d$ with $\|v\| = 1$:*

$$\left|\frac{1}{|T|}\sum_{x_i \in T}\langle v, x_i\rangle - \mu_v\right| \leq \rho_1\,\sigma_v\,, \quad \textit{and} \tag{3.6}$$

$$\left|\frac{1}{|T|}\sum_{x_i \in T}\big(\langle v, x_i\rangle - \mu_v\big)^2 - \sigma_v^2\right| \leq \rho_2\,\sigma_v^2\,, \tag{3.7}$$

where $\mu_v = \langle v, \mu\rangle$ and $\sigma_v^2 = v^\top \Sigma v$.

Originally, resilience is introduced in the context of robust statistics. Resilience measures how sensitive the sample statistics are to removing an α-fraction of the data points. A dataset from a distribution with a lighter tail has smaller resilience (ρ_1, ρ_2). For example, sub-Gaussian distributions have $\rho_1 = O(\alpha\sqrt{\log(1/\alpha)})$ and $\rho_2 = O(\alpha \log(1/\alpha))$ (Lemma 3.3.12),

which are smaller than the resilience of heavy-tailed distributions with bounded k-th moment, i.e., $\rho_1 = O(\alpha^{1-1/k})$ and $\rho_2 = O(\alpha^{1-2/k})$ (Lemma 3.3.15). Resilience plays a crucial role in robust statistics, where the resilience of a dataset determines the minimax sample complexity of estimating population statistics from adversarially corrupted samples [186, 217].

In the context of differential privacy, our design of HPTR is guided by our analysis showing that the sensitivity of one-dimensional robust statistics is fundamentally governed by resilience. Leveraging this three-way connection between the use of robust statistics in the algorithm, the resilience of the data, and the sensitivity of the distance $D_S(\hat{\mu})$ is crucial in achieving near-optimal utility.

Concretely, we consider α as a free parameter that we can choose depending on the target accuracy. For example, let $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = 32\rho_1$ be our target accuracy. Note that we did not optimize the constants in our analysis, and they can be further tightened. In the case of sub-Gaussian distributions, we have $\rho_1 = C'\alpha\sqrt{\log(1/\alpha)}$ w.h.p. when the sample size is large enough. This determines the value of α that achieves the target accuracy and also the choice of Δ and τ, as follows.

The robust statistics of a resilient dataset (i.e., one with small resilience) cannot change too much when a small fraction of the dataset is changed. This is made precise in Lemma 3.3.11, which shows, for example, that the robust mean $\mu_v(\mathcal{M}_{v,\alpha})$ can change only by $O(\rho_1/(\alpha n))$ when one data point is arbitrarily changed. This implies the sensitivity of $D_S(\hat{\mu})$ is also small: $\Delta = O(\rho_1/(\alpha n))$. Choosing $\tau = 42\rho_1$ to be larger by a constant factor from the target accuracy, we show that a sample size of $n = O(d/(\varepsilon\alpha))$ is sufficient to achieve the desired utility.

Theorem 19 (Utility guarantee for mean estimation)**.** *There exist positive constants c and C such that for any (α, ρ_1, ρ_2)-resilient set S with respect to some $(\mu \in \mathbb{R}^d, \Sigma \succ 0)$ satisfying $\alpha \in (0, c)$, $\rho_1 < c$, $\rho_2 < c$, and $\rho_1^2 \le c\alpha$,* HPTR *with the choices of the distance function in Eq. (3.4), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \le 32\rho_1$ with probability*

$1-\zeta$, *if*

$$n \geq C\frac{d+\log(1/(\delta\zeta))}{\varepsilon\alpha}.$$

This theorem shows how a resilient dataset (which is a deterministic condition) implies small error for HPTR. We make formal connections to standard assumptions on the sample generating distributions and their respective resiliences in Section 3.3.3, where we also discuss the optimality of this utility guarantee. For example, sub-Gaussian distributions have $\rho_1 = O(\alpha\sqrt{\log(1/\alpha)})$ when $n \geq C'd/(\alpha\log(1/\alpha))^2$ for any α smaller than a universal constant. This implies that HPTR achieves a target accuracy of $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| \leq \tilde{\alpha}$ with sample size $\tilde{O}(\frac{d}{\tilde{\alpha}^2}+\frac{d}{\tilde{\alpha}\varepsilon})$, where $\tilde{O}$ hides logarithmic factors in $1/\alpha$, δ, and ζ. We explain the intuition behind our analysis and provide a complete proof in Sections 3.3.2.2–3.3.2.6. One by-product of using robust statistics is that we get robustness for free, as we next show.

3.3.2.1 Robustness of HPTR

One by-product of using robust statistics is that HPTR is also robust to adversarial corruption. We therefore provide a more general guarantee that simultaneously achieves DP and robustness. Suppose we are given a dataset S that is a corrupted version of a resilient dataset S_{good}.

Assumption 3 (α_{corrupt}-corruption)**.** *Given a set $S_{\text{good}} = \{\tilde{x}_i \in \mathbb{R}^d\}_{i=1}^n$ of n data points, an adversary inspects all data points, selects $\alpha_{\text{corrupt}}n$ of the data points, and replaces them with arbitrary dataset S_{bad} of size $\alpha_{\text{corrupt}}n$. The resulting corrupted dataset is called $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$.*

This adaptive adversary is strong since the corruption can adapt to the entire dataset (for example, it covers the Huber contamination model [115] and the non-adaptive adversarial model [151]). This threat model is now standard in robust statistics literature [186]. If the original S_{good} is resilient, we show that the same guarantee as Theorem 19 holds under corruption up to an α_{corrupt} fraction of S_{good} for sufficiently small $\alpha_{\text{corrupt}} \leq (1/5.5)\alpha$. The factor 1/5.5 is due to the fact that the algorithm treats some of the good data points as

outliers (which is at most $4\alpha_{\text{corrupt}}$ due to the top and bottom tails cut in the definition of $\mathcal{M}_{v,(2/5.5)\alpha}$), and we need to handle neighboring datasets up to $(0.5/5.5)\alpha n$ Hamming distance. Hence, we need to ensure resilience for α that is at least 5.5 times larger than the corruption α_{corrupt}.

Definition 3.3.3 (Corrupt good set)**.** *We say a dataset S is $(\alpha_{\text{corrupt}}, \alpha, \rho_1, \rho_2)$-corrupt good with respect to (μ, Σ) if it is an α_{corrupt}-corruption of an (α, ρ_1, ρ_2)-resilient dataset S_{good}.*

We get the following theorem showing that HPTR can tolerate up to $(1/5.5)\alpha$ fraction of the data being arbitrarily corrupted.

Theorem 20 (Robustness)**.** *There exist positive constants c and C such that for any $((2/11)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S with respect to $(\mu \in \mathbb{R}^d, \Sigma \succ 0)$ satisfying $\alpha < c$, $\rho_1 < c$, $\rho_2 < c$, and $\rho_1^2 \le c\alpha$,* HPTR *with the distance function in Eq. (3.4), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \le 32\rho_1$ with probability $1 - \zeta$, if*

$$n \;\ge\; C\frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,.$$

In Sections 3.3.2.2–3.3.2.6, we prove this more general result. When there is no adversarial corruption, Theorem 19 immediately follows as a special case by selecting α as a free parameter depending on the target accuracy. The constants in all the theorems can be improved if we track them more carefully, and we did not attempt to optimize them in this work.

3.3.2.2 Proof strategy for Theorem 20

We show in Section 3.3.2.5 that the robust one-dimensional statistics, $\mu_v(\mathcal{M}_{v,\alpha})$ and $\sigma_v^2(\mathcal{M}_{v,\alpha})$, have small sensitivity if the dataset is resilient. Consequently, $D_S(\hat{\mu})$ has a small *local* sensitivity, i.e., the sensitivity is small if restricted to $\hat{\mu}$ close to μ and if the dataset is resilient. To ensure DP, we run RELEASE only when those two locality conditions are satisfied; we first PROPOSE the sensitivity Δ and a threshold τ, and then we TEST that DP guarantees are met on the given dataset with those choices. Resilient datasets (i) pass this safety test with a high probability and (ii) achieve the desired accuracy, both of which rely on our general

analysis of HPTR with a general distance function (Theorem 28). We give sketches of the main steps below.

One-dimensional robust statistics have small sensitivity on resilient datasets. Consider the robust projected mean $\mu_v(\mathcal{M}_{v,\alpha})$ for some small enough $\alpha > 0$. If S is (α, ρ_1, ρ_2)-resilient, then the following technical lemma shows that the top and bottom $(2/5.5)\alpha$-tails cannot deviate too much from the mean.

Lemma 3.3.4 (Lemma 10 from [186]). *For a (α, ρ_1, ρ_2)-resilient dataset S with respect to (μ, Σ) and any $0 \le \tilde{\alpha} \le \alpha$, the following holds for any subset $T \subset S$ of size at least $\tilde{\alpha}n$ and for any unit norm $v \in \mathbb{R}^d$:*

$$\Big|\frac{1}{|T|}\sum_{x_i \in T} \langle v, x_i - \mu\rangle\Big| \;\le\; \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\,\rho_1\,\sigma_v\;, \;\; \textit{and} \tag{3.8}$$

$$\Big|\frac{1}{|T|}\sum_{x_i \in T} \big(\,\langle v, x_i - \mu\rangle^2 - \sigma_v^2\,\big)\Big| \;\le\; \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\,\rho_2\,\sigma_v^2\;. \tag{3.9}$$

Under the definitions in Eq. (3.4), the top $(2/5.5)\alpha$-tail denoted by $\mathcal{T}_{v,\alpha}$ and bottom $(2/5.5)\alpha$-tail denoted by $\mathcal{B}_{v,\alpha}$ have the empirical means that are no more than $O(\sigma_v\rho_1/\alpha)$ away from the true projected mean μ_v, respectively. It follows that there exists at least one data point in $\mathcal{T}_{v,\alpha}$ and one data point in $\mathcal{B}_{v,\alpha}$ that are no more than $O(\sigma_v\rho_1/\alpha)$ away from μ_v. This implies that the range of the middle subset $\mathcal{M}_{v,\alpha}$ is provably bounded by $O(\sigma_v\rho_1/\alpha)$, and the sensitivity of the robust mean $\mu_v(\mathcal{M}_{v,\alpha})$ is guaranteed to be $O(\sigma_v\rho_1/(\alpha n))$. We can similarly show that $\sigma_v^2(\mathcal{M}_{v,\alpha})$ has sensitivity $O(\sigma_v^2\rho_1^2/(\alpha^2 n))$, as shown in Eq. (3.19). Note that these sensitivity bounds are *local* in the sense that they require the data to be (α, ρ_1, ρ_2)-resilient.

Small local sensitivity of $D_S(\hat{\mu})$. Under the above sensitivity bounds for $\mu_v(\mathcal{M}_{v,\alpha})$ and $\sigma_v^2(\mathcal{M}_{v,\alpha})$, it follows after some calculations as shown in Eq. (3.20) that the sensitivity for a resilient dataset S is bounded by

$$|\,D_S(\hat{\mu}) - D_{S'}(\hat{\mu})\,| \;\le\; C'\frac{\rho_1}{\alpha n}\Big(1 + \frac{\rho_1\|\Sigma^{-1/2}(\hat{\mu}-\mu)\|}{\alpha}\Big)\;, \tag{3.10}$$

for some constant C' and all neighboring datasets S', assuming ρ_2 is sufficiently small. Note that this sensitivity bound is *local* for two reasons; for this sensitivity to be small (i.e. $O(\rho_1/(\alpha n))$), we require S to be resilient and $\hat{\mu}$ to be close to μ. Thus, the meaning of *local* here is two fold, while traditionally local sensitivity in the privacy literature only concerns the sensitivity of a particular dataset S. We handle these two localities with the TEST step, among other things, checks that the DP conditions are satisfied for the given dataset and the choice of Δ and τ, which bounds the support of the exponential mechanism to be within $B_{\tau,S} = \{\hat{\mu} : D_S(\hat{\mu}) \le \tau\}$ with a choice of $\tau = O(\rho_1)$. Consequently, we require $\rho_1^2/\alpha \ll 1$ for the second term in Eq. (3.10) to be dominated by the first. Fortunately, this is indeed true for all scenarios of interests to us. For sub-Gaussian distributions, $\rho_1^2 = \alpha^2 \log(1/\alpha) \ll \alpha$. For k-th moment bounded distributions with $k > 3$, $\rho_1^2 = \alpha^{2-2/k} \ll \alpha$. For covariance bounded distributions, we do not hope to get a Mahalanobis distance guarantee. Instead, we aim for a Euclidean distance guarantee whose sensitivity does not depend on $\hat{\mu}$, and we do not require $\rho_1^2/\alpha \ll 1$ (Section 3.3.3.3).

Sample complexity analysis. Assuming the sensitivity of $D_S(\hat{\mu})$ is bounded by $\Delta = O(\rho_1/(\alpha n))$, which we ensure with the safety test, we analyze the utility of the exponential mechanism. For a target accuracy of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = O(\rho_1)$, we consider two sets $B_{\text{out}} = \{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \le c_0\rho_1\}$ and $B_{\text{in}} = \{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \le c_1\rho_1\}$ for some $c_0 > c_1$. The exponential mechanism achieves accuracy $c_0\rho_1$ with probability $1 - \zeta$ if

$$\mathbb{P}(\hat{\mu} \notin B_{\text{out}}) \le \frac{\mathbb{P}(\hat{\mu} \notin B_{\text{out}})}{\mathbb{P}(\hat{\mu} \in B_{\text{in}})} \lesssim \frac{\text{Vol}(B_{\tau,S})}{\text{Vol}(B_{\text{in}})} \frac{e^{-\frac{\varepsilon}{4\Delta}c_0\rho_1}}{e^{-\frac{\varepsilon}{4\Delta}c_1\rho_1}} \le e^{O(d)} e^{-\frac{\varepsilon}{4\Delta}(c_0 - c_1)\rho_1} \le \zeta \,,$$

where the second inequality requires $D_S(\hat{\mu}) \simeq \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, which we show in Lemma 3.3.6. Since the volume ratio is $\text{Vol}(B_{\tau,S})/\text{Vol}(B_{\text{out}}) = e^{O(d)}$, $\tau = O(\rho_1)$, and $\Delta = O(\rho_1/(\alpha n))$, it is sufficient to have a large enough c_0 and $n = O((d + \log(1/\zeta))/(\alpha\varepsilon))$ with a large enough constant.

Safety test. We are left to show that for a resilient dataset, the failure probability of the safety test, $\mathbb{P}(m_\tau + \text{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta))$, is less than ζ. This requires the safety margin to be large enough, i.e., $m_\tau \ge k^* = (2/\varepsilon)\log(4/(\delta\zeta))$. Recall that the safety margin is defined

as the Hamming distance to the closest dataset to S where the $(\varepsilon/2, \delta/2)$-DP condition of the exponential mechanism is violated. We therefore need to show that the DP condition is satisfied not only for S but for any dataset S' at Hamming distance at most k^* from S.

Consider two exponential mechanisms, $r_{(\varepsilon,\Delta,\tau,S')}$ and $r_{(\varepsilon,\Delta,\tau,S'')}$, on neighboring datasets S' and S''. Since $B_{\tau,S'} \neq B_{\tau,S''}$, we separately analyze the intersection $B_{\tau,S'} \cap B_{\tau,S''}$ and the differences $B_{\tau,S'} \setminus B_{\tau,S''}$ and $B_{\tau,S''} \setminus B_{\tau,S'}$. In the intersection, we show that the two probability distributions are within a multiplicative factor $e^{\varepsilon/2}$ of each other:

$$\mathbb{P}_{r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\mu} \in A) \;\leq\; e^{\varepsilon/2}\mathbb{P}_{r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\mu} \in A)$$

for all $A \subseteq B_{\tau,S'} \cap B_{\tau,S''}$, S' within Hamming distance k^* from a resilient dataset S, and $S'' \sim S'$. The main challenge is that S' is no longer a resilient dataset but a k^*-neighbor of a resilient dataset. Since such S' is $(k^*/n, \alpha, \rho_1, \rho_2)$-corrupt good (Definition 3.3.3), we show that corrupt good sets also inherit the bounded local sensitivity of a resilient dataset seamlessly, as shown in Lemma 3.3.11.

In the set difference, we show that the total probability mass, $\mathbb{P}_{r_{(\varepsilon,\Delta,\tau,S)}}(\hat{\mu} \in B_{\tau,S} \setminus B_{\tau,S'})$ and $\mathbb{P}_{r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\mu} \in B_{\tau,S'} \setminus B_{\tau,S})$, are bounded by δ, respectively, as long as the overlap of the two supports are large enough. This requires $\tau \gg \Delta k^*$, as we show in Appendix B.1.1, which is satisfied for $n \geq (\log(1/(\delta\zeta))/(\alpha\varepsilon))$.

Outline. The analyses for the accuracy and the safety test build upon a universal analysis of HPTR in Theorem 28, which holds more generally for any distance function $D_\phi(\hat{\theta})$ in the estimation problems of interest. For mean estimation, we show in Sections 3.3.2.3-3.3.2.5 that the sufficient conditions of Theorem 28 are met for the choices of constants and parameters: $\rho = \rho_1$, $c_0 = 31.8$, $c_1 = 10.2$, $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, $\tau = 42\rho_1$, and $\Delta = 110\rho_1/(\alpha n)$. We can set c_2 to be a large constant and will only change the constant factor in the sample complexity, which we do not track. A proof of Theorem 20 is provided in Section 3.3.2.6, from which Theorem 19 follows immediately. All the lemmas assume $((1/5.5)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S, $\alpha \leq 0.015$, $\rho_1 \leq 0.013$, and $\rho_2 \leq 0.0005$. We omit this assumption in stating the lemmas for brevity.

3.3.2.3 *Resilience implies robustness*

For the assumption (d) in Theorem 28, we show that $D_S(\hat{\mu})$ is a good approximation of the true distance $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$ in Lemma 3.3.6. We first show that the one-dimensional mean and the variance of the filtered out $\mathcal{M}_{v,\alpha}$ are robust.

Lemma 3.3.5. *For any unit norm* $v \in \mathbb{R}^d$, $|\langle v, \mu - \mu(\mathcal{M}_{v,\alpha})\rangle| \leq 6\rho_1 \sigma_v$ *and* $0.9\sigma_v \leq \sigma_v(\mathcal{M}_{v,\alpha}) \leq 1.1\sigma_v$.

Proof. For the mean bound,

$$\begin{aligned}
&|\langle v, \mu - \mu(\mathcal{M}_{v,\alpha})\rangle| \\
&\leq \frac{|\mathcal{M}_{v,\alpha} \cap S_{\text{bad}}|}{|\mathcal{M}_{v,\alpha}|}|\langle v, \mu(S_{\text{bad}} \cap \mathcal{M}_{v,\alpha}) - \mu\rangle| + \frac{|\mathcal{M}_{v,\alpha} \cap S_{\text{good}}|}{|\mathcal{M}_{v,\alpha}|}|\langle v, \mu(S_{\text{good}} \cap \mathcal{M}_{v,\alpha}) - \mu\rangle| \\
&\leq \frac{(1/5.5)\alpha}{1 - (4/5.5)\alpha}\frac{2\rho_1\sigma_v}{(1/5.5)\alpha} + \frac{1 - (1/5.5)\alpha}{1 - (4/5.5)\alpha}\rho_1\sigma_v \\
&\leq (2\rho_1 + \rho_1)\sigma_v/(1 - (4/5.5)\alpha)\ . \qquad (3.11)
\end{aligned}$$

The second inequality follows from the following. First, $|\langle v, \mu(S_{\text{good}} \cap \mathcal{M}_{v,\alpha}) - \mu\rangle| \leq \sigma_v\rho_1$ by the definition of resilience and that fact that $|S_{\text{good}} \cap \mathcal{M}_{v,\alpha}| \geq (1 - (5/5.5)\alpha)n$. Next, since $|\langle v, \mu(S_{\text{bad}} \cap \mathcal{M}_{v,\alpha}) - \mu\rangle|$ is less than $|\langle v, \mu(S_{\text{good}} \cap \mathcal{T}_{v,\alpha}) - \mu\rangle|$ or $|\langle v, \mu(S_{\text{good}} \cap \mathcal{B}_{v,\alpha}) - \mu\rangle|$, both of which are at most $2\rho_1\sigma_v/(1/5.5)\alpha$, from applying Lemma 3.3.4 with a set size at least $(1/5.5)\alpha n$, we have

$$|\langle v, \mu(S_{\text{bad}} \cap \mathcal{M}_{v,\alpha}) - \mu\rangle| \ \leq\ \frac{2}{(1/5.5)\alpha}\rho_1\sigma_v\ .$$

The mean bound follows from (3.11) and $\alpha \leq 0.1$. For the variance upper bound,

$$\sigma_v(\mathcal{M}_{v,\alpha})^2 \ =\ \frac{1}{(1 - (4/5.5)\alpha)n}\sum_{x_i \in \mathcal{M}_{v,\alpha}} \langle v, x_i - \mu(\mathcal{M}_{v,\alpha})\rangle^2 \ \leq\ \frac{1}{(1 - (4/5.5)\alpha)n}\sum_{x_i \in \mathcal{M}_{v,\alpha}} \langle v, x_i - \mu\rangle^2\ ,$$

where the first inequality follows from the fact that subtracting the empirical mean $\mu(\mathcal{M}_{v,\alpha})$ minimizes the second moment. We can decompose the empirical deviation and show an upper

bound first:

$$\begin{aligned}
&\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}}(\langle v, x_i-\mu\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}\\
=\;&\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}\cap S_{\rm bad}}(\langle v, x_i-\mu\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}+\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}\cap S_{\rm good}}(\langle v, x_i-\mu\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}\\
\le\;&\frac{(1/5.5)\alpha(2\rho_2/(1/5.5)\alpha)\sigma_v^2+(1-(4/5.5)\alpha)\rho_2\sigma_v^2}{1-(4/5.5)\alpha}\;\le\;6\rho_2\sigma_v^2\,, \qquad (3.12)
\end{aligned}$$

where in the second inequality we used resilience on $\mathcal{M}_{v,\alpha}\cap S_{\rm good}$ of size at least $1-(5/5.5)\alpha$. For $x_i\in S_{\rm bad}\cap\mathcal{M}_{v,\alpha}$, we use the fact that

$$\begin{aligned}
\big|\,\langle v, x_i-\mu\rangle^2-\sigma_v^2\,\big| \;&\le\; \max\Big\{\frac{\sum_{j\in S_{\rm good}\cap\mathcal{T}_{v,\alpha}}(\langle v, x_j-\mu\rangle^2-\sigma_v^2)}{|S_{\rm good}\cap\mathcal{T}_{v,\alpha}|},\frac{\sum_{j\in S_{\rm good}\cap\mathcal{B}_{v,\alpha}}(\langle v, x_j-\mu\rangle^2-\sigma_v^2)}{|S_{\rm good}\cap\mathcal{B}_{v,\alpha}|}\Big\}\\
&\le\;\frac{2\rho_2\sigma_v^2}{(1/5.5)\alpha}\,,
\end{aligned}$$

where we used Eq. (3.9) in Lemma 3.3.4 for sets with size at least $(1/5.5)\alpha n$. For the variance deviation lower bound,

$$\begin{aligned}
\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}}(\langle v, x_i-\mu(\mathcal{M}_{v,\alpha})\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}&=\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}}\big(\,\langle v, x_i-\mu\rangle^2-\sigma_v^2-\langle v,\mu-\mu(\mathcal{M}_{v,\alpha})\rangle^2\,\big)}{(1-(4/5.5)\alpha)n}\\
\ge\;&\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}\cap S_{\rm bad}}(\langle v, x_i-\mu\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}+\frac{\sum_{x_i\in\mathcal{M}_{v,\alpha}\cap S_{\rm good}}(\langle v, x_i-\mu\rangle^2-\sigma_v^2)}{(1-(4/5.5)\alpha)n}-36\rho_1^2\sigma_v^2\,,\\
\ge\;&-\frac{2\rho_2\sigma_v^2}{1-(4/5.5)\alpha}-\frac{1-(4/5.5)\alpha}{1-(4/5.5)\alpha}\rho_2\sigma_v^2-36\rho_1^2\sigma_v^2\;\ge\;-(3.2\rho_2+36\rho_1^2)\sigma_v^2\,, \qquad (3.13)
\end{aligned}$$

where we used $\alpha\le 0.1$, the first term only uses the fact that $|S_{\rm bad}|\le(1/5.5)\alpha n$, the second term uses resilience, and the last term uses the mean bound we proved earlier. In (3.12) and (3.13), assuming $\rho_1\le 0.04$, and $\rho_2\le 0.035$, we have $\sqrt{1+6\rho_2}\le 1.1$ and $\sqrt{1-3.2\rho_2-36\rho_1^2}\ge 0.9$. □

We show that resilience implies our estimate of the distance is robust.

Lemma 3.3.6. *If $\hat\mu\in B_{\tau,S}$ and $\tau=42\rho_1$, then $\big|\,\|\Sigma^{-1/2}(\hat\mu-\mu)\|-D_S(\hat\mu)\,\big|\le 6\rho_1+0.1\tau\le 10.2\rho_1$.*

Proof. From Lemma 3.3.5, we know that for all $\hat{\mu} \in B_{t,S}$,

$$D_S(\hat{\mu}) \;=\; \max_{\|v\|=1} \frac{\langle v, \hat{\mu} - \mu(\mathcal{M}_{v,\alpha})\rangle}{\sigma_v(\mathcal{M}_{v,\alpha})} \;\geq\; \max_{\|v\|=1} \frac{\langle v, \hat{\mu} - \mu\rangle - 6\rho_1\sigma_v}{1.1\sigma_v} \,. \tag{3.14}$$

and

$$D_S(\hat{\mu}) \;=\; \max_{\|v\|=1} \frac{\langle v, \hat{\mu} - \mu(\mathcal{M}_{v,\alpha})\rangle}{\sigma_v(\mathcal{M}_{v,\alpha})} \;\leq\; \max_{\|v\|=1} \frac{\langle v, \hat{\mu} - \mu\rangle + 6\rho_1\sigma_v}{0.9\sigma_v} \,. \tag{3.15}$$

Applying Lemma 3.3.1, we get $0.9D_S(\hat{\mu}) - 6\rho_1 \leq \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq 1.1D_S(\hat{\mu}) + 6\rho_1$. Since $D_S(\hat{\mu}) \leq \tau$, we get the desired bound.

□

3.3.2.4 Bounded volume

We show that the assumption (a) in Theorem 28 is satisfied for robust estimate $D_S(\hat{\mu})$.

Lemma 3.3.7. *For $\rho = \rho_1$, $c_1 = 10.2$, $\tau = 42\rho_1$, $\Delta = 110\rho_1/(\alpha n)$, and $c_2 \geq \log(67/12) + \log((c_0 + 2c_1)/c_1)$, we have $(7/8)\tau - (k^* + 1)\Delta > 0$,*

$$\begin{aligned} \frac{\mathrm{Vol}(B_{\tau+(k^*+1)\Delta+c_1\rho,S})}{\mathrm{Vol}(B_{(7/8)\tau-(k^*+1)\Delta-c_1\rho,S})} &\leq e^{c_2 d} \,, \text{ and} \\ \frac{\mathrm{Vol}(\{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq (c_0 + 2c_1)\rho\})}{\mathrm{Vol}(\{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq c_1\rho\})} &\leq e^{c_2 d} \,. \end{aligned}$$

Proof. The second part of assumption (a) follows from the fact that

$$\mathrm{Vol}(\{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq r\}) = c_d|\Sigma|r^d \,,$$

where $|\Sigma| = \prod_{j=1}^{d} \sigma_j(\Sigma)$ is the determinant of Σ and $\sigma_j(\Sigma)$ is the j-th singular value for some constant c_d that depends only on the dimension and selecting $c_2 \geq \log((c_0 + 2c_1)/c_1)$.

The first part is tricky since we do not yet have a handle on the set $B_{t,S}$ for $t > \tau$. In particular, we do not know how $D_S(\hat{\mu})$ relates to $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$ for such a $\hat{\mu}$ outside of $B_{\tau,S}$. To this end, we use the following corollary.

Corollary 3.3.8 (Corollary of Lemma 3.3.6)**.** *If $\hat{\mu} \in B_{2\tau,S}$ and $\tau = 42\rho_1$, then $\big|\, \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| - D_S(\hat{\mu}) \,\big| \le 14.2\rho_1$.*

We will show that $(7/8)\tau - (k^*+1)\Delta > 0$. Since this implies that $\tau + (k^*+1)\Delta \le 2\tau$, we can use the above corollary to show that

$$\begin{aligned}
\frac{\mathrm{Vol}(B_{\tau+(k^*+1)\Delta+c_1\rho,S})}{\mathrm{Vol}(B_{(7/8)\tau-(k^*+1)\Delta-c_1\rho,S})} &\le \frac{\mathrm{Vol}\big(\{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu}-\mu)\| \le \tau + (k^*+1)\Delta + c_1\rho + 14.2\rho_1\}\big)}{\mathrm{Vol}\big(\{\hat{\mu} : \|\Sigma^{-1/2}(\hat{\mu}-\mu)\| \le (7/8)\tau - (k^*+1)\Delta - c_1\rho - 14.2\rho_1\}\big)} \\
&= \Big(\frac{\tau + (k^*+1)\Delta + c_1\rho + 14.2\rho_1}{(7/8)\tau - (k^*+1)\Delta - c_1\rho - 14.2\rho_1}\Big)^d \\
&\le (67/12)^d \;\le\; e^{c_2 d}\,,
\end{aligned}$$

for the choices of $\rho = \rho_1$, $c_1 = 10.2$, $\tau = 42\rho_1$, $\Delta = 110\rho_1/(\alpha n)$, and $c_2 \ge \log(67/12)$, where we used the fact that for $n \ge C\log(1/(\delta\zeta))/(\alpha\varepsilon)$ with a large enough constant C, we have $(k^*+1)\Delta \le 0.3\rho_1$. It follows that the condition $(7/8)\tau - (k^*+1)\Delta > 0$ is also satisfied. □

3.3.2.5 Resilience implies bounded local sensitivity

We show that resilience implies the assumption (b) in Theorem 28 (Lemma 3.3.11). However, since local sensitivity needs to be established first for not just the given set S but also Hamming distance k^*+1 neighborhood of S, we need robustness results for this broader regime. Assuming $(k^*+1)/n \le \alpha/11$, we can extend robustness results analogously, as follows. We consider a set S' with k data points arbitrarily changed from S. This implies that S' is a $((1/5.5)\alpha + (k/n), \alpha, \rho_1, \rho_2)$-corrupt good set with respect to (μ, Σ). We first prove the analogous bounds to Lemma 3.3.5 for this S'.

Lemma 3.3.9. *For an $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2)$-corrupt good set S' with respect to (μ, Σ), $\tilde{\alpha} \le (1/11)\alpha$, and any unit norm $v \in \mathbb{R}^d$, $|\langle v, \mu - \mu(\mathcal{M}_{v,\alpha})\rangle| \le 14\rho_1\,\sigma_v$ and $0.9\sigma_v \le \sigma_v(\mathcal{M}_{v,\alpha}) \le 1.1\sigma_v$.*

Proof. Analogous to (3.11), we have

$$\begin{aligned}
|\langle v, \mu - \mu(\mathcal{M}_{v,\alpha})\rangle| &\le \frac{(1/5.5)\alpha + \tilde{\alpha}}{1 - (4/5.5)\alpha}\,\frac{2\rho_1\sigma_v}{(1/5.5)\alpha - \tilde{\alpha}} + \frac{1 - (1/5.5)\alpha - \tilde{\alpha}}{1 - (4/5.5)\alpha}\rho_1\sigma_v \\
&\le 14\rho_1\sigma_v\,,
\end{aligned}$$

where we used the fact that $(5/5.5)\alpha + \tilde{\alpha} \le \alpha$. Analogous to (3.12), we have

$$\begin{aligned}\frac{\sum_{x_i \in \mathcal{M}_{v,\alpha}} (\langle v, x_i - \mu(\mathcal{M}_{v,\alpha})\rangle^2 - \sigma_v^2)}{(1-(4/5.5)\alpha)n} &\le \frac{((1/5.5)\alpha + \tilde{\alpha})(\frac{2\rho_2}{(1/5.5)\alpha - \tilde{\alpha}})\sigma_v^2 + (1-(1/5.5)\alpha - \tilde{\alpha})\rho_2\sigma_v^2}{1-(4/5.5)\alpha} \\ &\le 14\rho_2\sigma_v^2 \,.\end{aligned}$$

Analogous to (3.13), we have

$$\begin{aligned}\frac{\sum_{x_i \in \mathcal{M}_{v,\alpha}} (\langle v, x_i - \mu(\mathcal{M}_{v,\alpha})\rangle^2 - \sigma_v^2)}{(1-(4/5.5)\alpha)n} &\ge -\frac{((1/5.5)\alpha + \tilde{\alpha})2\rho_2\sigma_v^2}{(1-(4/5.5)\alpha)((1/5.5)\alpha - \tilde{\alpha})} - \rho_2\sigma_v^2 - 14^2\rho_1^2\sigma_v^2 \\ &\ge -(7.3\rho_2 + 196\rho_1^2)\sigma_v^2 \,.\end{aligned}$$

For $\alpha \le 0.045$, $\rho_1 \le 0.013$, and $\rho_2 \le 0.0005$, we have the desired bounds. □

Lemma 3.3.10. *For an $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2)$-corrupt good set S' with respect to (μ, Σ) and $\tilde{\alpha} \le (1/11)\alpha$, if $\hat{\mu} \in B_{t,S'}$ for some $t > 0$ then we have $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \le 14\rho_1 + 1.1t$ and $\left|D(\hat{\mu}, S') - \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|\right| \le 14\rho_1 + 0.1t$.*

Proof. Analogously to the proof of Lemma 3.3.6, we have

$$\begin{aligned}1.1D(\hat{\mu}, S') &\ge -14\rho_1 + \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \,, \text{ and} \\ 0.9D(\hat{\mu}, S') &\le 14\rho_1 + \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \,.\end{aligned}$$

This gives the desired bound. □

The sensitivity of $D_S(\hat{\mu})$ is *local* in two ways. First, we get the desired sensitivity bound for a dataset S that behaves nicely, which is captured by the notion of a $((1/5.5)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S. Second, the sensitivity bound requires the estimate parameter $\hat{\mu}$ to be close to μ in $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$. Both *locality in dataset* and *locality in estimate* are ensured by the safety test (Test step in HPTR). To show that corrupt good datasets pass the safety test, the following lemma establishes that those datasets have small local sensitivity.

Lemma 3.3.11. *For $\Delta = 110\rho_1/(\alpha n)$, $\tau = 42\rho_1$, and an $((1/5.5\alpha), \alpha, \rho_1, \rho_2)$-corrupt good S, if*

$$n = \Omega\Big(\frac{\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big) \,, \tag{3.16}$$

then the local sensitivity in assumption (b) is satisfied.

Remark. Note that to keep $\Delta = O(\rho_1/(\alpha n))$ that we want (and is critical in getting the final utility guarantee), we need the extra corruption to be $k^*/n = O(\alpha)$. This implies $n = \Omega(k^*/\alpha) = \Omega(\log(1/(\delta\zeta))/(\varepsilon\alpha))$. Further, $k^* = \Omega(\log(1/(\delta\zeta))/\varepsilon)$ cannot be improved since it is critical in achieving a small failure probability in the Testing step. Hence, the sample complexity of $\Omega(\log(1/(\delta\zeta))/(\varepsilon\alpha))$ cannot be improved under the current proof strategy.

Proof. Since S is $((1/5.5)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good and $d_H(S, S') \le k^*$, it follows that S' is $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2)$-corrupt good with $\tilde{\alpha} = (k^*/n)$. We further assume that $\tilde{\alpha} \le (1/11)\alpha$, which follows from $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$ and $n = \Omega(\log(1/\delta\zeta)/(\varepsilon\alpha))$ with a large enough constant. We show that this resilience implies that S' is dense around the boundary of $\mathcal{M}_{v,\alpha}$, which in turn implies low sensitivity.

Recall that $\mathcal{T}_{v,\alpha} \subset S$ is the set of data points corresponding to the largest $(2/5.5)\alpha n$ data points in the projected set $S'_{(v)} = \{\langle v, x_i\rangle\}_{x_i \in S'}$, and $\mathcal{B}_{v,\alpha} \subset S$ is the bottom set. Let S_{good} denote the original uncorrupted resilient dataset. Applying Lemma 3.3.4 to $S_{\text{good}} \cap \mathcal{T}_{v,\alpha}$ (and $S_{\text{good}} \cap \mathcal{B}_{v,\alpha}$) of size at least $(1/11)\alpha$ (since the corruption fraction is at most $(1/5.5)\alpha + \tilde{\alpha} \le (1.5/5.5)\alpha$),

$$\left|\, \langle v, \mu(S_{\text{good}} \cap \mathcal{T}_{v,\alpha}) - \mu\rangle \,\right| \;\le\; \frac{2\rho_1\sigma_v}{(1/11)\alpha}\,, \text{ and } \left|\, \langle v, \mu(S_{\text{good}} \cap \mathcal{B}_{v,\alpha}) - \mu\rangle \,\right| \;\le\; \frac{2\rho_1\sigma_v}{(1/11)\alpha}\,.$$

This implies that there is at least one good data point that is closer to the center than the means of the upper tail and the bottom tail:

$$\min_{x_i \in S_{\text{good}} \cap \mathcal{T}_{v,\alpha}} \left|\, \langle v, x_i - \mu\rangle \,\right| \;\le\; \frac{2\rho_1\sigma_v}{(1/11)\alpha}\,, \text{ and } \min_{x_i \in S_{\text{good}} \cap \mathcal{B}_{v,\alpha}} \left|\, \langle v, x_i - \mu\rangle \,\right| \;\le\; \frac{2\rho_1\sigma_v}{(1/11)\alpha}\,.$$

It follows that the distance between two closest points in $\mathcal{T}_{v,\alpha}$ and $\mathcal{B}_{v,\alpha}$ is bounded by

$$\min_{x_i \in S_{\text{good}} \cap \mathcal{T}_{v,\alpha}} \langle v, x_i\rangle - \max_{x_i \in S_{\text{good}} \cap \mathcal{B}_{v,\alpha}} \langle v, x_i\rangle \;\le\; (44/\alpha)\rho_1\sigma_v\,, \tag{3.17}$$

when $\mu \in \mathcal{M}_{v,\alpha}$. When $\mu \in \mathcal{T}_{v,\alpha}$ or $\mu \in \mathcal{B}_{v,\alpha}$, it is straightforward that the above inequality holds. This implies low sensitivity as follows.

Recall that $\mathcal{M}_{v,\alpha}(S')$ denotes the middle part after filtering out the top and bottom $(2/5.5)\alpha$ quantiles from $\{\langle v, x_i\rangle\}_{x_i \in S'}$. For a neighboring dataset S'' and the corresponding $S''_{(v)}$,

consider a scenario where one point x_i in $\mathcal{M}_{v,\alpha}(S')$ is replaced by another point $\tilde{x}_i$. If $\langle v, \tilde{x}_i\rangle \in [\max_{x_i \in S_{\text{good}} \cap \mathcal{B}_{v,\alpha}} \langle v, x_i\rangle , \min_{x_i \in S_{\text{good}} \cap \mathcal{T}_{v,\alpha}} \langle v, x_i\rangle]$, then Eq. (3.17) implies that $|\langle v, x_i - \tilde{x}_i\rangle| \leq (44/\alpha)\rho_1\sigma_v$. Otherwise, $\mathcal{M}_{v,\alpha}(S'')$ will have x_i replaced by either $\arg\min_{j \in S_{\text{good}} \cap \mathcal{T}_{v,\alpha}} \langle v, x_j\rangle$ or $\arg\max_{j \in S_{\text{good}} \cap \mathcal{B}_{v,\alpha}} \langle v, x_j\rangle$. In either case, Eq. (3.17) implies that $|\langle v, x_i - \tilde{x}_i\rangle| \leq (44/\alpha)\rho_1\sigma_v$. The other case of when the replaced sample $x_i \in S$ is not in $\mathcal{M}_{v,\alpha}$ follows similarly.

From this, we get the following bounds on the sensitivity of the robust mean and robust variance. Note that using robust statistics is critical in getting such small sensitivity bounds. Let $\mu' = \mu(\mathcal{M}_{v,\alpha}(S'))$ and $\mu'' = \mu(\mathcal{M}_{v,\alpha}(S''))$, where we write the dataset S' in $\mathcal{M}_{v,\alpha}(S')$ explicitly,

$$\big|\, \langle v, \mu' - \mu''\rangle \,\big| \;\leq\; \frac{44\rho_1\sigma_v}{\alpha(1-(4/5.5)\alpha)n} \,. \tag{3.18}$$

For the variance bound, let $\sigma_v'^2 = \sigma_v^2(\mathcal{M}_{v,\alpha}(S')) = (1/|\mathcal{M}_{v,\alpha}(S')|)\sum_{x_i' \in \mathcal{M}_{v,\alpha}(S')} \langle v, x_i' - \mu'\rangle^2$ and $\sigma_v''^2 = \sigma_v^2(\mathcal{M}_{v,\alpha}(S''))$. Since $(1-(4/5.5)\alpha)n\sigma_v'^2 = \sum_{x_i' \in \mathcal{M}_{v,\alpha}(S')} \langle v, x_i' - \mu'\rangle^2 = \sum_{x_i' \in \mathcal{M}_{v,\alpha}(S')} (\langle v, x_i' - \mu''\rangle^2 - \langle v, \mu'' - \mu'\rangle^2)$, we have $(1-(4/5.5)\alpha)n(\sigma_v'^2 - \sigma_v''^2) = \sum_{x_i' \in \mathcal{M}_{v,\alpha}(S')} \langle v, x_i' - \mu''\rangle^2 - \sum_{x_i'' \in \mathcal{M}_{v,\alpha}(S'')} \langle v, x_i'' - \mu''\rangle^2 - (1-(4/5.5)\alpha)n\langle v, \mu'' - \mu'\rangle^2$. We bound each term separately. Note that $\mathcal{M}_{v,\alpha}(S')$ and $\mathcal{M}_{v,\alpha}(S'')$ only differ in at most one data point. We denote those by x' and x'', respectively. Then,

$$\begin{aligned}
&\Big|\sum_{x_i' \in \mathcal{M}_{v,\alpha}(S')} \langle v, x_i' - \mu''\rangle^2 - \sum_{x_i'' \in \mathcal{M}_{v,\alpha}(S'')} \langle v, x_i'' - \mu''\rangle^2 \Big| = \big|\, \langle v, x' - \mu''\rangle^2 - \langle v, x'' - \mu''\rangle^2 \,\big| \\
&\quad = \big|\, \langle v, x' + x'' - 2\mu''\rangle \langle v, x' - x''\rangle \,\big| \\
&\quad = \big|\, \langle v, x' - \mu'\rangle + \langle v, \mu' - \mu''\rangle + \langle v, x'' - \mu''\rangle \big|\, \big|\langle v, x' - x''\rangle \,\big| \\
&\quad \leq 3\Big(\frac{44\rho_1\sigma_v}{\alpha}\Big)^2 \,,
\end{aligned}$$

and

$$(1-(4/5.5)\alpha)n\langle v, \mu' - \mu''\rangle^2 \;\leq\; (1-(4/5.5)\alpha)n\frac{(44\rho_1\sigma_v)^2}{(\alpha(1-(4/5.5)\alpha)n)^2} \,.$$

This implies that

$$|\sigma_v'^2 - \sigma_v''^2| \;\leq\; \frac{(44\rho_1(\alpha/2)\sigma_v)^2}{(1-(4/5.5)\alpha)n\alpha^2}\Big(3 + \frac{1}{(1-(4/5.5)\alpha)n}\Big) \;\leq\; \frac{4(44\rho_1\sigma_v)^2}{(1-(4/5.5)\alpha)n\alpha^2} \,. \tag{3.19}$$

Together, we get the following bound on the sensitivity of $D(\hat{\mu}, S')$. Since $\max_v a_v - \max_v b_v \leq \max_v |a_v - b_v|$, we have

$$\begin{aligned}
\left| D_{S'}(\hat{\mu}) - D_{S''}(\hat{\mu}) \right| &\leq \max_{v:\|v\|=1} \left| \frac{\langle v, \hat{\mu} - \mu' \rangle}{\sigma'_v} - \frac{\langle v, \hat{\mu} - \mu'' \rangle}{\sigma''_v} \right| \\
&\leq \max_{v:\|v\|=1} \frac{|\langle v, \mu' - \mu'' \rangle|}{\sigma'_v} + \frac{|\langle v, \hat{\mu} - \mu'' \rangle|}{\sigma_v} \left| \frac{\sigma_v}{\sigma'_v} - \frac{\sigma_v}{\sigma''_v} \right| \\
&\leq \frac{44\rho_1}{0.9\alpha(1-(4/5.5)\alpha)n} + \|\Sigma^{-1/2}(\hat{\mu} - \mu'')\| \max_v \frac{\sigma_v}{\sigma'_v \sigma''_v (\sigma'_v + \sigma''_v)} |\sigma'^2_v - \sigma''^2_v| \\
&\leq \frac{44\rho_1}{0.9\alpha(1-(4/5.5)\alpha)n} + \frac{5312\rho_1^2}{\alpha^2(1-(4/5.5)\alpha)n} \|\Sigma^{-1/2}(\hat{\mu} - \mu'')\| \,,
\end{aligned}$$

where we used triangular inequality in the second inequality and the third inequality follows from $\sigma'_v \geq 0.9\sigma_v$ (Lemma 3.3.9), Eqs. (3.18), and Lemma 3.3.1, and the last inequality follows from $\sigma''_v \geq 0.9\sigma_v$ and (3.19).

From Lemma 3.3.10, $\hat{\mu} \in B_{\tau+(k^*+3)\Delta, S}$ implies $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq 14\rho_1 + 1.1(\tau + (k^*+3)\Delta)$. From Lemma 3.3.9, $\|\Sigma^{-1/2}(\mu - \mu'')\| \leq 14\rho_1$. We apply triangular inequality and show that $\|\Sigma^{-1/2}(\hat{\mu} - \mu'')\| \leq c\alpha/\rho_1$ for the choices of Δ, k^*, τ and n, with an arbitrarily small constant c:

$$\begin{aligned}
\|\Sigma^{-1/2}(\hat{\mu} - \mu'')\| &\leq 28\rho_1 + 1.1(\tau + (k^*+3)\Delta) \\
&\leq C\rho_1 + C\frac{\rho_1 \log(1/(\delta\zeta))}{\varepsilon\alpha n} \\
&\leq 2C\rho_1 \,,
\end{aligned}$$

for some constant $C > 0$, where $\Delta = 110\rho_1/(\alpha n)$, $\tau = 42\rho_1$, $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, and $n \geq C' \log(1/(\delta\zeta))/(\varepsilon\,\alpha)$. Under the assumption that $\rho_1^2 \leq c\alpha$ and $\alpha \leq c$ for some small enough c, this implies

$$\begin{aligned}
|D_{S'}(\hat{\mu}) - D_{S''}(\hat{\mu})| &\leq \frac{44\rho_1}{0.9(1-(4/5.5)\alpha)\alpha n}\left(1 + \frac{121\rho_1}{\alpha} 2C\rho_1\right) \\
&\leq \frac{(44/0.9)\rho_1}{\alpha n} \frac{1+44c}{1-(4/5.5)c} \leq \Delta = \frac{110\rho_1}{\alpha n} \,. \qquad (3.20)
\end{aligned}$$

□

3.3.2.6 *Proof of Theorem 20*

We show that the sufficient conditions of Theorem 28 are met for the following choices of constants and parameters: $p = d$, $\rho = \rho_1$, $c_0 = 31.8$, $c_1 = 10.2$, $\tau = 42\rho_1$, and $\Delta = 110\rho_1/(\alpha n)$. We can set c_2 to be a large constant and will only change the constant factor in the sample complexity.

The assumptions (a), (b), and (d) follow from Lemmas 3.3.7, 3.3.11, and 3.3.6, respectively. Assumption (c) follows from

$$\Delta = \frac{110\rho_1}{\alpha n} \leq \frac{1.2\rho_1\varepsilon}{32(c_2 d + (\varepsilon/2) + \log(16/(\delta\zeta)))} = \frac{(c_0 - 3c_1)\rho\varepsilon}{32(c_2 d + (\varepsilon/2) + \log(16/(\delta\zeta)))} ,$$

for a large enough $n \geq C'(d + \log(1/(\delta\zeta)))/(\alpha\varepsilon)$. This finishes the proof of Theorem 20, from which Theorem 19 immediately follows.

3.3.3 *Step 3: Near-optimal guarantees*

We provide utility guarantees for popular families of distributions in private or robust mean estimation literature: sub-Gaussian [25, 150, 186, 217, 139, 129, 40, 36, 32, 4, 34, 62, 64, 73, 106, 66, 114], k-th moment bounded [25, 150, 186, 217, 135], and covariance bounded [25, 150, 186, 217, 135, 73, 105, 57, 58]. We apply known resilience bounds for each family of distributions and substitute them in Theorems 19 and 20. In all cases, the resulting sample complexity is near-optimal, which follows from matching information-theoretic lower bounds.

Since we aim for Mahalanobis distance error bounds, the corresponding mean resilience we need in Definition 3.3.2 scales linearly in the projected standard deviation. For sub-Gaussian distributions, this requires the projected variance $v^\top \Sigma v$ to be lower bounded by how fast the tail is decreasing, as captured by the sub-Gaussian proxy $\Omega(v^\top \Gamma v)$ in Eq. (3.21) (Section 3.3.3.1). For k-th moment bounded distributions with $k > 3$, this requires the projected variance to be lower bounded by $\Omega(\mathbb{E}[|\langle v, x - \mu\rangle|^k]^{2/k})$, a condition known as hypercontractivity (Section 3.3.3.2). When we do not have such lower bounds on the covariance, HPTR can only hope to achieve Euclidean distance error bounds. Under our

design principle, this translates into the choice of $D_S(\hat{\mu}) = \max_{\|v\|\leq 1}\langle v, \hat{\mu}\rangle - \mu_v(\mathcal{M}_{v,\alpha})$. We give an example of this scenario with covariance bounded distributions (Section 3.3.3.3).

3.3.3.1 Sub-Gaussian distributions

We say a distribution P is sub-Gaussian with proxy Γ if for all $\|v\| = 1$ and $t \in \mathbb{R}$,

$$\mathbb{E}_{x\sim P}\big[\, \exp(t\,\langle v, x\rangle)\,\big] \;\leq\; \exp\Big(\frac{t^2\, v^\top \Gamma v}{2}\Big)\,. \tag{3.21}$$

Under this standard sub-Gaussianity, we are only guaranteed mean resilience of Eq. (3.6), for example, with R.H.S scaling as $\rho_1\sqrt{v^\top\Gamma v}$ instead of $\rho_1\sqrt{v^\top\Sigma v}$. This implies that the Mahalanobis distance of any robust estimate can be made arbitrarily large by shrinking the covariance in one direction such that $v^\top\Sigma v \ll v^\top\Gamma v$. To avoid such degeneracy, we add an additional assumption that $\Sigma \succeq c\Gamma$, which is also common in robust statistics literature, e.g., [121]. With this definition, it is known that sub-Gaussian samples are $(\alpha, O(\alpha\sqrt{\log(1/\alpha)}), O(\alpha\log(1/\alpha)))$-resilient.

Lemma 3.3.12 (Resilience of sub-Gaussian samples [217] and [121, Corollary 4])**.** *For any fixed $\alpha \in (0, 1/2)$, consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a sub-Gaussian distribution with mean μ, covariance Σ, and a sub-Gaussian proxy $0 \prec \Gamma \preceq c_1\Sigma$ for a constant c_1. There exist constants c_2 and $c_3 > 0$ such that if $n \geq c_2((d+\log(1/\zeta))/(\alpha\log(1/\alpha))^2)$, then S is $(\alpha, c_3\alpha\sqrt{\log(1/\alpha)}, c_3\alpha\log(1/\alpha))$-resilient with respect to (μ, Σ) with probability $1-\zeta$.*

This lemma and Theorem 19 imply the following utility guarantee. Further, from Theorem 20, the guarantee also holds under α-corruption of the i.i.d. samples from a sub-Gaussian distribution.

Corollary 3.3.13. *Under the hypothesis of Lemma 3.3.12, there exists a constant $c > 0$ such that for any $\alpha \in (0, c)$, a dataset of size*

$$n \;=\; O\Big(\frac{d+\log(1/\zeta)}{(\alpha\log(1/\alpha))^2} + \frac{d+\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big)\,,$$

sensitivity of $\Delta = O((1/n)\sqrt{\log(1/\alpha)})$*, and threshold of* $\tau = O(\alpha\sqrt{\log(1/\alpha)})$*, with large enough constants are sufficient for* HPTR(S) *with the distance function in Eq.* (3.4) *to achieve*

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \;=\; O(\alpha\sqrt{\log(1/\alpha)}) \tag{3.22}$$

with probability $1 - \zeta$*. Further, the same guarantee holds even if* α*-fraction of the samples is arbitrarily corrupted, as shown in Assumption 3.*

This sample complexity is near-optimal up to logarithmic factors in $1/\alpha$ and $1/\zeta$ for $\delta = e^{-O(d)}$. Even for DP mean estimation without corrupted samples, HPTR is the first algorithm for sub-Gaussian distributions with unknown covariance that nearly matches the lower bound of $n = \widetilde{\Omega}(d/\alpha^2 + d/(\alpha\varepsilon) + \log(1/\delta)/\varepsilon)$ from [139, 129], where $\widetilde{\Omega}$ hides polylogarithmic terms in $1/\zeta, 1/\alpha, d, 1/\varepsilon$ and $\log(1/\delta)$. The third term has a gap of $1/\alpha$ factor to our upper bound, but this term is dominated by other terms under the assumption that $\delta = e^{-O(d)}$. For completeness, we state the lower bound in Appendix B.3. Existing algorithms are suboptimal since they require either $n = \widetilde{O}((d/\alpha^2) + (d(\log(1/\delta)^3)/(\alpha\varepsilon^2)))$ samples with $(1/\varepsilon^2)$ dependence to achieve the error rate of Eq. (3.22) [34] or extra conditions, such as strictly Gaussian distributions [34, 36] or known covariance matrices [129, 4, 25].

The error bound is near-optimal in its dependence in α under α-corruption. HPTR is the first estimator that is both (ε, δ)-DP and also achieves the robust error rate of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = O(\alpha\sqrt{\log(1/\alpha)})$, nearly matching the known information-theoretic lower bound of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = \Omega(\alpha)$ [47]. This lower bound holds for any estimator that is not necessarily private and regardless of how many samples are available. In comparison, the existing robust and DP estimator from [160], which runs in polynomial time, requires the knowledge of the covariance matrix Σ and a larger sample complexity of $n = \widetilde{\Omega}((d/\alpha^2) + (d^{3/2}\log(1/\delta))/(\alpha\varepsilon))$. If privacy is not required (i.e., $\varepsilon = \infty$), a robust mean estimator from [217] achieves the same error bound and sample complexity as ours.

3.3.3.2 Hypercontractive distributions

For an integer $k \geq 3$, a distribution $P_{\mu,\Sigma}$ is k-th moment bounded with a mean μ and covariance Σ if for all $\|v\| = 1$, we have $\mathbb{E}_{x\sim P_X}[|\langle v, (x-\mu)\rangle|^k] \leq \kappa^k$ for some $\kappa > 0$. However, similar to the sub-Gaussian case, Mahalanobis distance guarantees require an additional lower bound on the covariance. To this end, we assume hypercontractivity, which is common in robust statistics literature, e.g., [143].

Definition 3.3.14. *A distribution $P_{\mu,\Sigma}$ is (κ, k)-hypercontractive if for all $v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim P_X}[|\langle v, (x-\mu)\rangle|^k] \leq \kappa^k (v^\top \Sigma v)^{k/2}$.*

Although samples from such heavy-tailed distributions are known to be not resilient, it is known that they are $O(\alpha)$-close in total variation distance to an $(\alpha, O(\alpha^{1-1/k}), O(\alpha^{1-2/k}))$-resilient dataset. This means that the resulting dataset is $((1/11)\alpha, \alpha, O(\alpha^{1-1/k}), O(\alpha^{1-2/k}))$-corrupt good, for example. Note that hypercontractivity is invariant under affine transformations, and κ does not depend on the condition number of the covariance.

Lemma 3.3.15 (Resilience of k-th moment bounded samples [217, Lemma G.10])**.** *For any fixed $\alpha \in (0, 1/2)$, consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a (κ, k)-hypercontractive distribution with mean μ and covariance $\Sigma \succ 0$ for some $k \geq 3$. For any $c_3 > 0$, there exist constants c_1 and $c_2 > 0$ that depend only on c_3 such that if*

$$n \;\geq\; c_1\Big(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k}d\log d}{\zeta^{2-4/k}\kappa^2} + \frac{\kappa^2 d\log d}{\alpha^{2/k}}\Big)\,,$$

then S is $(c_3\alpha, \alpha, c_2 k\kappa\alpha^{1-1/k}\zeta^{-1/k}, c_2 k^2\kappa^2\alpha^{1-2/k}\zeta^{-2/k})$-corrupt good with respect to (μ, Σ) with probability $1-\zeta$.

This lemma and Theorem 19 imply the following utility guarantee. Further, from Theorem 20, the guarantee also holds under $(1/5.5 - c_3)\alpha$-corruption of the i.i.d. samples from a (κ, k)-hypercontractive distribution. Choosing appropriate constants, we get the following result.

Corollary 3.3.16. *Under the hypothesis of Lemma 3.3.15, there exists a constant $c_{\kappa,k,\zeta}$ that depends only on k, κ, and ζ such that for any $\alpha \in (0, c_{\kappa,k,\zeta})$, a dataset of size*

$$n = O\Big(\frac{d+\log(1/(\delta\zeta))}{\varepsilon\alpha} + \frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k}d\log d}{\zeta^{2-4/k}\kappa^2} + \frac{\kappa^2 d\log d}{\alpha^{2/k}}\Big),$$

sensitivity of $\Delta = O(1/(n\alpha^{1/k}))$, and threshold of $\tau = O(\alpha^{1-1/k})$ with large enough constants are sufficient for HPTR(S) *with the distance function in Eq. (3.4) to achieve $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| = O(k\kappa\zeta^{-1/k}\alpha^{1-1/k})$ with probability $1-\zeta$. Further, the same guarantee holds even if α-fraction of the samples is arbitrarily corrupted, as shown in Assumption 3.*

This sample complexity is near-optimal in its dependence in d, $1/\varepsilon$, and $1/\alpha$ when $\delta = e^{-\Theta(d)}$. Suppose ζ, k, and κ are $\Theta(1)$. Even for DP mean estimation without robustness, HPTR is the first algorithm that achieves $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| = O(\alpha^{1-1/k})$ with $n = \widetilde{O}(\frac{d}{\alpha^{2(1-1/k)}} + \frac{d+\log(1/\delta)}{\varepsilon\alpha})$ samples, which nearly matches the known lower bounds. The first term $O(d/\alpha^{2(1-1/k)})$ cannot be improved even if we do not require privacy. The second term $O((d+\log(1/\delta))/\varepsilon\alpha)$ nearly matches the lower bound of $n = \Omega(\min\{d, \log((1-e^{-\varepsilon})/\delta)\}/(\varepsilon\alpha))$ for DP mean estimation that we show in Proposition 3.3.18. In typical DP scenarios, we have $0 < \varepsilon \le 1$ and $\delta = e^{-\Theta(d)}$ [25], in which case the upper and lower bounds match. An existing DP mean estimator (without robustness) of [135] achieves a stronger $(\varepsilon, 0)$-DP and similar accuracy but in Euclidean distance with a similar sample size of $n = \widetilde{O}(\frac{d}{\alpha^{2(1-1/k)}} + \frac{d}{\varepsilon\alpha})$. However, it requires a known or identity covariance matrix and a known bound on the unknown mean of the form $\mu \in [-R, R]^d$. Such a bounded search space is critical in achieving a stronger *pure* privacy guarantee with $\delta = 0$.

The error bound is optimal in its dependence in α under α-corruption. The error bound $\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| = O(\alpha^{1-1/k})$ matches the following information-theoretic lower bound in Proposition 3.3.17; no algorithm can distinguish two distributions whose means are at least $O(\alpha^{1-1/k})$ apart from α-fraction of samples corrupted, even with infinite samples. HPTR is the first algorithm that guarantees both differential privacy and robustness (i.e., the error depends only on α and not in d) for k-th moment bounded distributions. If privacy is not

required (i.e., $\varepsilon = \infty$), a robust mean estimator from [217] achieves a similar error bound and sample complexity as ours.

Proposition 3.3.17 (Lower bound for robust mean estimation)**.** *For any $\alpha \in (0, 1/2)$, there exist two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ satisfying the hypotheses of Lemma 3.3.15 such that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \alpha$, and*

$$\|\Sigma^{-1/2}(\mu_1 - \mu_2)\| = \Omega(\alpha^{1-1/k}) \, .$$

Proof. We construct two scalar distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ with $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \alpha$ as follows:

$$\mathcal{D}_1(x) = \begin{cases} (1-\alpha)/2, & \text{if } x \in \{-1, 1\} \\ \alpha & \text{if } x = -\alpha^{1/k} \end{cases}, \quad \text{and} \quad \mathcal{D}_2(x) = \begin{cases} (1-\alpha)/2, & \text{if } x \in \{-1, 1\} \\ \alpha & \text{if } x = \alpha^{1/k} \end{cases}$$

The variance is $\Omega(1)$ for both distributions, and $|\mathbb{E}_{x\sim\mathcal{D}_1}[x] - \mathbb{E}_{x\sim\mathcal{D}_2}[x]| = 2\alpha^{1-1/k}$. Then, it suffices to show that $\mathcal{D}_1$ and $\mathcal{D}_2$ are both $(O(1), k)$-hypercontractive. In fact, we know $\mathbb{E}_{x\sim\mathcal{D}_1}[x] = -\alpha^{1-1/k}$, $\mathbb{E}_{x\sim\mathcal{D}_1}[x^2] = \mathbb{E}_{x\sim\mathcal{D}_2}[x^2] = 1 - \alpha + \alpha^{1-2/k}$, and $\mathbb{E}_{\mathcal{D}_1}[|x|^k] = 2 - \alpha$. Since $\alpha \in (0, 1/2)$, there exists a constant c such that $\mathbb{E}_{x\sim\mathcal{D}_1}[|x - \mu_1|^k] \le c$, which concludes the proof.

□

Proposition 3.3.18 (Lower bound for DP mean estimation)**.** *Let $\mathcal{P}_{\mu,\Sigma,k}$ be the set of $(1, k)$-hypercontractive distributions with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $\mathcal{M}_{\varepsilon,\delta}$ be a class of (ε, δ)-DP estimators using n i.i.d. samples from $P \in \mathcal{P}_{\mu,\Sigma,k}$. Then, for $\varepsilon \in (0, 10)$, there exists a constant c such that*

$$\inf_{\hat{\mu}\in\mathcal{M}_{\varepsilon,\delta}} \sup_{\mu\in\mathbb{R}^d,\Sigma\succ 0, P\in\mathcal{P}_{\mu,\Sigma,k}} \mathbb{E}_{S\sim P^n}[\|\Sigma^{-1/2}(\hat{\mu}(S) - \mu)\|^2] \ge c \min\left\{ \left(\frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon} \right)^{2-2/k}, 1 \right\}.$$

Proof. We extend the proof of [25, Proposition 4] to hypercontractive distributions. Before we prove the lower bound, we first establish the private version of the standard statistical estimation problem. Specifically, let $\mathcal{P}$ denote a family of distributions of interest and $\theta : \mathcal{P} \to \Theta$ denote the population parameter. The goal is to estimate θ from i.i.d. samples

$x_1, x_2, \ldots, x_n \sim \mathcal{P}$. Let $\hat{\theta}$ be an (ε, δ)-differentially private estimator. Furthermore, let $\rho : \Theta \times \Theta \to \mathbb{R}^+$ be a (semi)metric on parameter space Θ and $\ell : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing loss function with $\ell(0) = 0$.

To measure the performance of our (ε, δ)-DP estimator $\hat{\theta}$, we define the *minimax risk* as follows:

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{x_1, x_2, \ldots, x_n \sim P} \left[\ell \left(\rho \left(\hat{\theta} \left(x_1, \ldots, x_n \right), \theta(P) \right) \right) \right] . \tag{3.23}$$

To prove the lower bound of the minimax risk, we construct a well-separated family of distributions and convert the estimation problem into a testing problem. Specifically, let $\mathcal{V}$ be an index set of finite cardinality. Define $\mathcal{P}_{\mathcal{V}} = \{P_v, v \in \mathcal{V}\} \subset \mathcal{P}$ to be an indexed family of distributions. If for all $v \neq v' \in \mathcal{V}$ we have $\rho(P_v, P_{v'}) \geq 2t$, we say $\mathcal{P}_{\mathcal{V}}$ is $2t$-*packing* of Θ.

The proof of [25, Proposition 4] is based on following lemma.

Lemma 3.3.19 ([25, Theorem 3])**.** *Fix $p \in [0, 1]$ and let $\mathcal{P}_{\mathcal{V}}$ be a $2t$-packing of Θ such that $d_{\mathrm{TV}}(P_v, P_{v'}) = p$. Let $\hat{\theta}$ be a (ε, δ) differentially private estimator. Then,*

$$\frac{1}{|\mathcal{V}|} \sum_{\nu \in \mathcal{V}} P_v \left(\rho \left(\widehat{\theta}, \theta(P_v) \right) \geq t \right) \geq \frac{(|\mathcal{V}| - 1) \cdot \left(\frac{1}{2} e^{-\varepsilon \lceil np \rceil} - \delta \frac{1 - e^{-\varepsilon \lceil np \rceil}}{1 - e^{-\varepsilon}} \right)}{1 + (|\mathcal{V}| - 1) \cdot e^{-\varepsilon \lceil np \rceil}} . \tag{3.24}$$

In our problem, we set $\mathcal{P}$ to be $\mathcal{P} = \mathcal{P}_{\mu, \Sigma, k}$. It suffices to construct such an index set $\mathcal{V}$ and indexed family of distributions $\mathcal{P}_{\mathcal{V}}$. We construct a packing set similar to that defined in the proof of [25, Proposition 4]. By [3, Lemma 6], there exists a finite set $\mathcal{V} \subset \mathbb{R}^d$ with cardinality $|\mathcal{V}| = 2^{\Omega(d)}$, $\|v\| = 1$ for all $v \in \mathcal{V}$, and $\|v - v'\| \geq 1/2$ for all $v \neq v' \in \mathcal{V}$. Define Q_0 as $Q_0 = \mathcal{N}(0, \mathbf{I}_{d \times d})$ and Q_v as a point mass on $x = \alpha^{-1/k} c v$, where $v \in \mathcal{V}$. We construct P_v as $P_v = \alpha Q_v + (1 - \alpha) Q_0$.

We first verify that $\mathcal{P}_{\mathcal{V}} \subset \mathcal{P}$. It is easy to see $\mu(P_v) = \mathbb{E}_{x \sim P_v}[x] = \alpha^{1 - 1/k} v$ and $\Sigma(P_v) = \mathbb{E}_{x \sim P_v}[(x - \mu(P_v))(x - \mu(P_v))^\top] = (1 - \alpha)\mathbf{I}_{d \times d} + \alpha(1 - \alpha)\alpha^{-2/k} v v^\top$. This implies $\frac{1}{2}\mathbf{I}_{d \times d} \preceq \Sigma(P_v) \preceq \mathbf{I}_{d \times d}$. Since $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] \leq \mathbb{E}\left[X^k\right]$ for any $X \geq 0$, it suffices to show that $\mathbb{E}_{x \sim P_v}[|\langle u, x \rangle|^k] \leq C^k$ for some constant $C > 0$ and any $\|u\| = 1$. In fact, letting c_k denote the k-th moment of standard Gaussian, we have

$$\mathbb{E}_{x \sim P_v}[|\langle u, x \rangle|^k] = (1 - \alpha) c_k + \alpha \left| \left\langle u, \alpha^{-1/k} v \right\rangle \right|^k = O(1) .$$

It is also easy to see that $d_{\mathrm{TV}}(P_v, P_{v'}) = \alpha$. Let $\rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$. We also have

$$t = \min_{v \neq v' \in \mathcal{V}} \alpha^{1-1/k} \|v - v'\| \geq \frac{1}{2}\alpha^{1-1/k} .$$

Next, we apply the reduction of estimation to testing with this packing $\mathcal{V}$. For (ε, δ)-DP estimator $\hat{\mu}$, using Lemma 3.3.19, we have

$$\begin{aligned}
\sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}[\|\Sigma(P)^{-1/2}(\hat{\mu}(S) - \mu(P))\|^2] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{S \sim P_v^n}[\|\Sigma(P_v)^{-1/2}(\hat{\mu}(S) - \mu(P_v))\|^2] \\
&= t^2 \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(\|\Sigma(P_v)^{-1/2}(\hat{\mu}(S) - \theta(P_v))\| \geq t \right) \\
&\asymp t^2 \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(\|\hat{\mu}(S) - \theta(P_v)\| \geq t \right) \\
&\gtrsim t^2 \frac{e^{d/2} \cdot \left(\frac{1}{2} e^{-\varepsilon \lceil n\alpha \rceil} - \frac{\delta}{1 - e^{-\varepsilon}} \right)}{1 + e^{d/2} e^{-\varepsilon \lceil n\alpha \rceil}} ,
\end{aligned}$$

where the last inequality follows from the fact that $d \geq 2$.

The rest of the proof follows from [25, Proposition 4]. We choose

$$\alpha = \frac{1}{n\varepsilon} \min \left\{ \frac{d}{2} - \varepsilon, \log \left(\frac{1 - e^{-\varepsilon}}{4\delta e^{\varepsilon}} \right) \right\}$$

so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}[\|\Sigma(P)^{-1/2}(\hat{\mu}(S) - \mu(P))\|^2] \gtrsim \alpha^{2-2/k} .$$

This means, for $\varepsilon \in (0, 1)$,

$$\inf_{\hat{\mu} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}[\|\Sigma(P)^{-1/2}(\hat{\mu}(S) - \mu(P))\|^2] \gtrsim \min \left\{ \left(\frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon} \right)^{2-2/k}, 1 \right\} ,$$

which completes the proof. □

3.3.3.3 Covariance bounded distributions

A distribution $P_{\mu,\Sigma}$ is covariance bounded with mean μ and covariance Σ if $\|\Sigma\| \leq 1$. Contrary to the previous cases, the sample variance is not resilient since $\{\langle v, x_i - \mu \rangle^2\}$ do not concentrate.

To get around this issue, we use the Euclidean distance: $D_\phi(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|$. This leads to the surrogate Euclidean distance of

$$D_S(\hat{\mu}) = \max_{\|v\| \le 1} \langle v, \hat{\mu} \rangle - \mu_v(\mathcal{M}_{v,\alpha}) \; . \tag{3.25}$$

Since this does not depend on the robust variance $\sigma_v^2(\mathcal{M}_{v,\alpha})$, we only require the following first order resilience.

Lemma 3.3.20 (Resilience of covariance bounded samples [217, Lemma G.3])**.** *For any fixed $\alpha \in (0, 1/2)$, consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from a covariance bounded distribution with mean μ and covariance $\Sigma \succ 0$. If $n = \Omega(d \log(d/\zeta)/(\alpha))$, then with probability $1 - 3\zeta$, for any subset $T \subset S$ of size $|T| \ge (1-\alpha)n$, there exists a constant $C > 0$ such that the following holds for all $\alpha \in (0, 1/2)$ and for all $v \in \mathbb{R}^d$ with $\|v\| = 1$:*

$$\left| \frac{1}{|T|} \sum_{x_i \in T} \langle v, x_i \rangle - \mu_v \right| \;\; \le \;\; C\alpha^{1/2} \; ,$$

where $\mu_v = \langle v, \mu \rangle$.

This lemma and Theorem 20, adapted for the new $D_S(\hat{\mu}) = \max_{\|v\| \le 1} \langle v, \hat{\mu} \rangle - \mu_v(\mathcal{M}_{v,\alpha})$, imply the following utility guarantee.

Corollary 3.3.21. *Under the hypothesis of Lemma 3.3.20, there exists a constant c_ζ that only depends on ζ such that for $\alpha \in (0, c_\zeta)$, a dataset of size*

$$n \;\; = \;\; O\Big(\frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha} + \frac{d \log(d/\zeta)}{\alpha} \Big) \; ,$$

sensitivity of $\Delta = O(1/(n\sqrt{\alpha}))$, and threshold of $\tau = O(\sqrt{\alpha})$ with large enough constants are sufficient for HPTR(S) *with the distance function in Eq. (3.25) to achieve $\|\hat{\mu} - \mu\| = O(\alpha^{1/2})$ with probability $1 - 3\zeta$. Further, the same guarantee holds even if a α-fraction of the samples is arbitrarily corrupted, as shown in Assumption 3.*

This sample complexity is near-optimal in its dependence on d, $1/\varepsilon$, and $1/\alpha$ for $\delta = e^{-O(d)}$. It matches the information-theoretic lower bound of $n = \Omega(d/\varepsilon\alpha)$ from [135]. For completeness,

we write the lower bound in Appendix B.3. This problem is easier than the sub-Gaussian or k-th moment bounded settings, since the error is measured in Euclidean distance and hence one does not need to adapt to the unknown covariance. Therefore, there exist other algorithms achieving near-optimality and even runs in polynomial time [135].

The error rate is near-optimal under α-corruption, matching the information-theoretic lower bound of $\|\hat{\mu} - \mu\| = \Omega(\alpha^{1/2})$ [73]. Note that there exists an DP and robust algorithm from [160] that achieves near-optimality in both error rate and sample complexity but requires an additional assumption that the spectral norm of the covariance is known and the unknown mean is in a bounded set, $[-R, R]^d$, with a known R.

Remark. Corollary 3.3.21 is suboptimal since (1) the error metric is Euclidean $\|\hat{\mu} - \mu\|$ instead of Mahalanobis $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$, and (2) sample complexity scales as $1/\zeta$ instead of $\log(1/\zeta)$. It remains an open problem if these gaps can be closed. For the former, one could use the Stahel-Donoho outlyingness [184, 74],

$$D_S(\hat{\mu}) \;=\; \sup_{v \in \mathbb{R}^d, \|v\|=1} \frac{|\langle v, \hat{\mu}\rangle - \mathrm{Med}(\langle v, S\rangle)|}{\mathrm{Med}(|\langle v, S\rangle - \mathrm{Med}(\langle v, S\rangle)|)} \;,$$

in the exponential mechanism, which replaces second moment based normalization by a first moment based one that is resilient. Here, $\mathrm{Med}(\langle v, S\rangle)$ is the median of $\{\langle v, x_i\rangle\}_{x_i \in S}$. Further, replacing the median by the median of means can improve the dependence on ζ. Such directions have been fruitful for robust but non-private mean estimation [58].

3.4 Linear regression

In a standard linear regression, we have i.i.d. samples $S = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ from a distribution $P_{\beta,\Sigma,\gamma^2}$ of a linear model:

$$y_i \;=\; x_i^\top \beta + \eta_i \;,$$

where the input $x_i \in \mathbb{R}^d$ has zero mean and covariance Σ and the noise $\eta_i \in \mathbb{R}$ has variance γ^2. We further assume $\mathbb{E}[x_i \eta_i] = 0$, which is equivalent to assuming that the true parameter $\beta = \Sigma^{-1}\mathbb{E}[y_i x_i]$. In DP linear regression, we want to output a DP estimate $\hat{\beta}$ of the unknown

model parameter β (which corresponds to $\theta = \mu$ in the general notation), assuming that both covariance $\Sigma \succ 0$ and noise variance γ^2 (corresponding to $\phi = (\Sigma, \gamma)$ in the general notation) are unknown. The resulting error is measured in $D_{\Sigma,\gamma}(\hat{\beta}, \beta) = (1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|$, which is equivalent to the (re-scaled) root excess prediction risk of the estimated predictor $\hat{\beta}$. Similar to Mahalanobis distance for mean estimation, this is challenging since we aim for a tight guarantee that adapts to the unknown Σ without having enough samples to directly estimate Σ. We follow the three-step strategy of Section 3.1.2.1 and provide utility guarantees.

3.4.1 Step 1: Designing the surrogate $D_S(\hat{\beta})$ for the error metric $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|$

In the RELEASE step of HPTR, we propose the following surrogate error metric for the exponential mechanism:

$$D_S(\hat{\beta}) \;=\; \max_{v:\|v\|\le 1} \frac{\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{x_i \in \mathcal{N}_{v,\hat{\beta},\alpha}} \langle v, x_i(y_i - x_i^\top \hat{\beta})\rangle}{\sigma_v(\mathcal{M}_{v,\alpha})\hat{\gamma}}\,, \tag{3.26}$$

where $\hat{\gamma}^2$ is defined as

$$\hat{\gamma}^2 = \min_{\bar{\beta}} \frac{1}{|\mathcal{B}_{\bar{\beta},\alpha}|} \sum_{i \in \mathcal{B}_{\bar{\beta},\alpha}} (y_i - x_i^\top \bar{\beta})^2\,. \tag{3.27}$$

We define $\mathcal{N}_{v,\hat{\beta},\alpha}$, $\mathcal{M}_{v,\alpha}$ and $\mathcal{B}_{\bar{\beta},\alpha}$ as follows. For a fixed v, $\mathcal{M}_{v,\alpha}$ is defined in Section 3.3.1 as a subset of S with size $(1 - (4/5.5)\alpha)n$ that remains after removing $(4/5.5)\alpha n$ data points corresponding to the top $(2/5.5)\alpha n$ and the bottom $(2/5.5)\alpha n$ of samples when projected down to $S_v = \{\langle v, x_i\rangle\}_{i\in[n]}$. We denote a robust estimate of the variance in direction v as $\sigma_v(\mathcal{M}_{v,\alpha})^2 = (1/|\mathcal{M}_{v,\alpha}|)\sum_{x_i \in \mathcal{M}_{v,\alpha}} \langle v, x_i\rangle^2$ since x_i's are zero mean. Similarly, for a fixed $\hat{\beta}$ and v, we consider a set of projected data points $S_{v,\hat{\beta}} = \{\langle v, x_i(y_i - x_i^\top \hat{\beta})\rangle\}_{i\in[n]}$ and partition S into three disjoint sets, $\mathcal{B}_{v,\hat{\beta},\alpha}$, $\mathcal{N}_{v,\hat{\beta},\alpha}$, and $\mathcal{T}_{v,\hat{\beta},\alpha}$, where $\mathcal{B}_{v,\hat{\beta},\alpha}$ is the subset of S corresponding to the bottom $(2/5.5)\alpha n$ data points with the smallest values in $S_{v,\hat{\beta}}$, $\mathcal{T}_{v,\hat{\beta},\alpha}$ corresponds to the top $(2/5.5)\alpha n$ data points, and $\mathcal{N}_{v,\hat{\beta},\alpha}$ corresponds to the remaining $(1 - (4/5.5)\alpha)n$ middle data points. We use $\mathcal{T}_{v,\hat{\beta},\alpha}$,$\mathcal{N}_{v,\hat{\beta},\alpha}$, and $\mathcal{B}_{v,\hat{\beta},\alpha}$ to denote both the set of paired examples $\{(x_i, y_i)\}$ and the set of indices of those examples, and it should be clear from the context which one we mean.

For a fixed $\bar{\beta}$, $\mathcal{B}_{\bar{\beta},\alpha}$ is defined as a subset of S with size $(1-(3.5/5.5)\alpha)n$ that remains after removing the largest $(2/5.5)\alpha n$ data points in set $S_{\bar{\beta}} = \{(y_i - x_i^\top \bar{\beta})^2\}_{i\in[n]}$.

This choice is justified by Lemma 3.4.1, which shows that if we replace the robust one-dimensional statistics by the true ones, we recover the target error metric. Hence, the exponential mechanism with distance $D_S(\hat{\beta})$ is approximately and stochastically minimizing $\|\Sigma^{1/2}(\hat{\beta}-\beta)\|$. For a more elaborate justification of using $D_S(\hat{\beta})$, we refer to a similar choice for mean estimation in Section 3.3.1.

Lemma 3.4.1. *For any $\beta \in \mathbb{R}^d$, $0 \prec \Sigma \in \mathbb{R}^{d\times d}$, and $\gamma > 0$, let $\sigma_v^2 = v^\top \Sigma v$. If $\mathbb{E}[\eta_i x_i] = 0$, $y_i = x_i^\top \beta + \eta_i$ and $(x_i, y_i) \sim P_{\beta,\Sigma,\gamma^2}$, then we have*

$$\begin{aligned} \|\Sigma^{1/2}(\hat{\beta}-\beta)\| &= \max_{v:\|v\|\le 1} \frac{\mathbb{E}_{P_{\beta,\Sigma,\gamma^2}}[\langle v, x_i(y_i - x_i^\top \hat{\beta})\rangle]}{\sigma_v}\,, \text{ and} \\ \gamma^2 &= \min_{\bar{\beta}\in\mathbb{R}^d} \mathbb{E}[(y_i - x_i^\top \bar{\beta})^2]\,. \end{aligned}$$

Proof. We have

$$\begin{aligned} \max_{v:\|v\|\le 1} \frac{\mathbb{E}_{P_{\beta,\Sigma,\gamma^2}}[\langle v, x_i(y_i - x_i^\top \hat{\beta})\rangle]}{\sigma_v} &= \max_{v:\|v\|\le 1} \frac{\mathbb{E}_{P_{\beta,\Sigma,\gamma^2}}[\langle v, x_i(x_i^\top(\beta-\hat{\beta}) + \eta_i)\rangle]}{\sigma_v} \\ &= \max_{v:\|v\|\le 1} \frac{\langle v, \Sigma(\beta-\hat{\beta})\rangle}{\sigma_v} = \|\Sigma^{1/2}(\beta-\hat{\beta})\|\,, \end{aligned}$$

where the second equality uses the fact that η_i has zero mean and x_i has covariance Σ. The last equality follows from Lemma B.2.1. For the noise, we have $\mathbb{E}[(y_i - x_i^\top \bar{\beta})^2] = \mathbb{E}[(x_i^\top \beta + \eta_i - x_i^\top \bar{\beta})^2] = \mathbb{E}[\eta_i^2] + \mathbb{E}[(\beta-\bar{\beta})x_i x_i^\top (\beta-\bar{\beta})]$, which follows from $\mathbb{E}[\eta_i x_i] = 0$. This is minimized when $\bar{\beta} = \beta$, and the minimum is γ^2. □

3.4.2 Step 2: Utility analysis under resilience

The following resilience is a fundamental property of the dataset that determines the sensitivity of $D_S(\hat{\beta})$. We refer to Section 3.3.2 for a detailed explanation of how resilience relates to sensitivity.

Definition 3.4.2 (Resilience for linear regression)**.** *For some $\alpha \in (0,1)$, $\rho_1 \in \mathbb{R}_+$, $\rho_2 \in \mathbb{R}_+$, and $\rho_3 \in \mathbb{R}_+$, we say a set of n labelled data points $S_{\text{good}} = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ is $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient with respect to (β, Σ, γ) for some $\beta \in \mathbb{R}^d$, positive definite $\Sigma \in \mathbb{R}^{d\times d}$, and $\gamma > 0$ if for any $T \subset S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$, the following holds for all $v \in \mathbb{R}^d$ with $\|v\| = 1$:*

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\langle v, x_i\rangle(y_i - x_i^\top\beta)\right| \;\leq\; \rho_1\,\sigma_v\,\gamma \quad , \tag{3.28}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle^2 - \sigma_v^2\right| \;\leq\; \rho_2\,\sigma_v^2 \quad , \tag{3.29}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle\right| \;\leq\; \rho_3\,\sigma_v \quad , \text{ and} \tag{3.30}$$

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}(y_i - x_i^\top\beta)^2 - \gamma^2\right| \;\leq\; \rho_4\,\gamma^2 \quad , \tag{3.31}$$

where $\sigma_v^2 = v^\top\Sigma v$.

For example, n i.i.d. samples from sub-Gaussian x_i's and sub-Gaussian η_i's (independent of x_i's) is $\big(\alpha, O(\alpha\log(1/\alpha)), O(\alpha\log(1/\alpha)), O(\alpha\sqrt{\log(1/\alpha)}), O(\alpha\log(1/\alpha))\big)$-resilient. A resilient dataset implies a sensitivity of $\Delta = O(\rho_1/(\alpha n)) = O(\log(1/\alpha)/n)$, where α is a free parameter determined by the target accuracy $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(\alpha\log(1/\alpha))$. We show that a sample size of $O((d + \log(1/\delta))/(\varepsilon\alpha))$ is sufficient to achieve the target accuracy for any resilient dataset. In Section 3.4.3, we apply this theorem to resilient datasets from several sampling distributions of interest and characterize the trade-offs.

Theorem 21 (Utility guarantee for linear regression)**.** *There exist positive constants c and C such that for any $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient set S with respect to $(\beta, \Sigma \succ 0, \gamma > 0)$ satisfying $\alpha \in (0, c)$,$\rho_1 < c$, $\rho_2 < c$, $\rho_3^2 \leq c\alpha$ and $\rho_4 < c$,* HPTR *with the distance function in Eq. (3.26), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \leq 32\rho_1$ with probability $1 - \zeta$ if*

$$n \;\geq\; C\,\frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,. \tag{3.32}$$

3.4.2.1 Robustness of HPTR

One by-product of using robust statistics in $D_S(\hat{\beta})$ is that robustness for HPTR comes for free under a standard data corruption model.

Assumption 4 (α_{corrupt}-corruption)**.** *Given a set $S_{\text{good}} = \{(\tilde{x}_i \in \mathbb{R}^d, \tilde{y}_i \in \mathbb{R})\}_{i=1}^n$ of n data points, an adversary inspects all data points, selects $\alpha_{\text{corrupt}} n$ of the data points, and replaces them with arbitrary dataset S_{bad} of size $\alpha_{\text{corrupt}} n$. The resulting corrupted dataset is called $S = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$.*

The same guarantee as Theorem 21 holds under corruption up to a corruption of $\alpha_{\text{corrupt}} < (1/5.5)\alpha$ fraction of a $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient dataset S_{good}. The factor $(1/5.5)$ is due to the fact that the algorithm can remove $(4/5.5)\alpha$ fraction of the good points and a slack of $(0.5/5.5)\alpha$ fraction is needed to resilience of neighboring datasets.

Definition 3.4.3 (Corrupt good set)**.** *We say a dataset S is $(\alpha_{\text{corrupt}}, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good with respect to (β, Σ, γ) if it is an α_{corrupt}-corruption of an $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient dataset S_{good}.*

Theorem 22 (Robustness)**.** *There exist positive constants c and C such that for any $((2/11)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to $(\beta, \Sigma \succ 0, \gamma > 0)$ satisfying $\alpha < c$, $\rho_1 < c$, $\rho_2 < c$, $\rho_3^2 \leq c\alpha$ and $\rho_4 < c$,* HPTR *with the distance function in Eq. (3.26), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \leq 32\rho_1$ with probability $1 - \zeta$, if*

$$n \;\geq\; C\,\frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,. \tag{3.33}$$

We provide a proof in Sections 3.4.2.2-3.4.2.6. When there is no adversarial corruption, Theorem 21 immediately follows by selecting α as a free parameter.

3.4.2.2 Proof strategy for Theorem 22

The overall proof strategy follows that of Section 3.3.2.2 for mean estimation. We highlight the differences here.

Lemma 3.4.4 (Lemma 10 from [186]). *For a $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient set S with respect to (β, Σ, γ) and any $0 \le \tilde{\alpha} \le \alpha$, the following holds for any subset $T \subset S$ of size at least $\tilde{\alpha}n$ and for any unit vector $v \in \mathbb{R}^d$:*

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\langle v, x_i\rangle(y_i - x_i^\top\beta)\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_1\,\sigma_v\,\gamma\,, \tag{3.34}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle^2 - \sigma_v^2\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_2\sigma_v^2\,, \tag{3.35}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_3\sigma_v\,, \text{ and} \tag{3.36}$$

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}(y_i - x_i^\top\beta)^2 - \gamma^2\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_4\,\gamma^2\,. \tag{3.37}$$

This technical lemma is critical in showing that the sensitivity of one-dimensional statistics is bounded by the resilience of the dataset, such that the sensitivity of $D_S(\hat{\beta})$ for a resilient S is bounded by

$$|D_S(\hat{\beta}) - D_{S'}(\hat{\beta})| \le C'\Big(1 + \frac{\rho_3^2}{\alpha}\Big)\frac{\rho_1 + (1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\|}{\alpha n}$$

for some constant C' and for any neighboring dataset S', as shown in Eq (3.47). The desired sensitivity bound is local in two ways: it requires S to be resilient and $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = O(\rho_1)$. Under the assumption that $\rho_3^2/\alpha = O(1)$ with a small enough constant, this achieves the desired bound $\Delta = O(\rho_1/(\alpha n))$ with $\hat{\beta} \in B_{\tau,S}$ and $\tau = O(\rho_1)$. The standard utility analysis of exponential mechanisms shows that the error of $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = O(\rho_1)$ can be achieved when $e^{O(d)-c\frac{\varepsilon}{\Delta}\rho_1} \le \zeta$, which happens if $n = \Omega((d+\log(1/\zeta))/(\varepsilon\alpha))$ with a large enough constant. The TEST step checks the two localities by ensuring that DP conditions are met for the given dataset.

Outline. Analogous to the mean estimation proof, the analyses of utility and the safety test build upon the universal analysis of HPTR in Theorem 28. For linear regression, we show in Sections 3.4.2.3-3.4.2.5 that the assumptions of Theorem 28 are met for a resilient dataset and the choices of constants and parameters: $\rho = \rho_1$, $c_0 = 31.8$, $c_1 = 10.2$, $\tau = 42\rho_1$,

$\Delta = 110\rho_1/(\alpha n)$, $\tau = 42\rho_1$, $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, and a large enough constant c_2. We assume that $\alpha < c$ and $\rho_1 < c$ for a small enough constant c. A proof of Theorem 22 is shown in Section 3.4.2.6, and Theorem 21 immediately follows by selecting α as a free parameter.

The above resilience properties also imply the following useful resilience on the $S_{\bar{\beta}} = \{(y_i - \bar{\beta}^\top x_i)^2\}_{i=[n]}$ for any vector $\bar{\beta}$.

Lemma 3.4.5 (Resilience of residual square)**.** *Let $S_{\text{good}} = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=[n]}$ be $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient with respect to (β, Σ, γ). Let $\rho^* = \max\{\rho_1, \rho_2, \rho_4\}$. Then, we have*

1. *for any $T \in S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$ and any vector $\bar{\beta} \in \mathbb{R}^d$,*

$$\left| \frac{1}{|T|} \sum_{(x_i,y_i)\in T} (y_i - \bar{\beta}^\top x_i)^2 - (\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \right| \leq \rho^*(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \,, \quad (3.38)$$

2. *and for any $0 \leq \tilde{\alpha} \leq \alpha$ and $T \in S_{\text{good}}$ of size $|T| \geq \tilde{\alpha}n$, we have*

$$\left| \frac{1}{|T|} \sum_{(x_i,y_i)\in T} (y_i - \bar{\beta}^\top x_i)^2 - (\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \right| \leq \frac{2-\tilde{\alpha}}{\tilde{\alpha}} \rho^*(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \quad (3.39)$$

Proof. The proof follows directly from resilience properties of Eq. (3.28), (3.29) and (3.31). □

3.4.2.3 Resilience implies robustness

To show that the assumption (d) in Theorem 28 is satisfied, we use the robustness of one-dimensional variance $\sigma_v(\mathcal{M}_{v,\alpha})$ (Lemma 3.4.6) and show that $D_S(\hat{\beta})$ is a good approximation of $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|$ (Lemma 3.4.8).

Lemma 3.4.6. *For an $((2/11)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to (β, Σ, γ) and any unit norm vector $v \in \mathbb{R}^d$, we have $0.9\sigma_v \leq \sigma_v(\mathcal{M}_{v,\alpha}) \leq 1.1\sigma_v$.*

Proof. This follows from Lemma 3.3.5. □

Lemma 3.4.7. *For a $((2/11)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to (β, Σ, γ) and any unit norm vector $v \in \mathbb{R}^d$, we have $0.99\gamma \leq \hat{\gamma} \leq 1.01\gamma$.*

Proof. Analogous to the proof of Lemma 3.4.4, for any fixed $\bar{\beta}$, we have

$$\begin{aligned}
& \left| \frac{1}{|\mathcal{B}_{\bar{\beta},\alpha}|} \sum_{i \in \mathcal{B}_{\bar{\beta},\alpha}} (y_i - x_i^\top \bar{\beta})^2 - (\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \right| \\
\leq & \frac{|\sum_{\mathcal{B}_{\bar{\beta},\alpha} \cap S_{\text{good}}} (y_i - x_i^\top \bar{\beta})^2 - (\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2|}{(1 - (2/5.5)\alpha)n} \\
& + \frac{|\sum_{\mathcal{B}_{\bar{\beta},\alpha} \cap S_{\text{bad}}} (y_i - x_i^\top \bar{\beta})^2 - (\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2|}{(1 - (2/5.5)\alpha)n} \\
\overset{(a)}{\leq} & \frac{(1 - (2/5.5)\alpha)n\rho^*(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2}{(1 - (2/5.5)\alpha)n} + \frac{(2/11)\alpha n \cdot 2\rho^*(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2/((2/11)\alpha)}{(1 - (2/5.5)\alpha)n} \\
\overset{(b)}{\leq} & 4\rho^*(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \,, \qquad (3.40)
\end{aligned}$$

where (a) follows from Lemma 3.4.5, and (b) follows from our assumption that $\alpha \leq c$ for some small enough constant c.

Let $F(\bar{\beta}) = \frac{1}{|\mathcal{B}_{\bar{\beta},\alpha}|} \sum_{i \in \mathcal{B}_{\bar{\beta},\alpha}} (y_i - x_i^\top \bar{\beta})^2$. We know that $\hat{\gamma}^2 = \min_{\bar{\beta}} F(\bar{\beta}) \leq F(\beta)$, which, together with Eq. (3.40), implies

$$\hat{\gamma}^2 \leq (1 + 4\rho^*)\gamma^2 \leq 1.0201\gamma^2 \,,$$

when $\rho^* \leq c$ for some c small enough.

Also, we have

$$\hat{\gamma}^2 \geq (1 - 4\rho^*)(\gamma + \|\Sigma^{1/2}(\beta - \bar{\beta})\|)^2 \geq (1 - 4\rho^*)\gamma^2 \geq 0.9801\gamma^2.$$

when $\rho^* \leq c$ for some c small enough. □

Lemma 3.4.8. *For a $((2/11)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to (β, Σ, γ), if $\hat{\beta} \in B_{\tau,S}$ and $\tau = 42\rho_1$, then $\left| \|\Sigma^{1/2}(\hat{\beta} - \beta)\|/\gamma - D_S(\hat{\beta}) \right| \leq 0.15\tau + 1.1\rho_1 \leq 10.2\rho_1$.*

Proof. By Lemma 3.4.1, Lemma B.2.2 and resilience Eq. (3.28) and Eq. (3.29), we have

$$
\begin{aligned}
& \left| \max_{v:\|v\|\le 1} \frac{\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} \langle v, x_i(y_i - x_i^\top\hat{\beta})\rangle}{\sigma_v} - \left\|\Sigma^{1/2}(\beta-\hat{\beta})\right\| \right| \\
= & \left| \max_{v:\|v\|\le 1} \frac{\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} \left(v^\top x_i x_i^\top(\beta-\hat{\beta}) + v^\top x_i\eta_i\right)}{\sigma_v} - \max_{v:\|v\|\le 1} \frac{v^\top\Sigma(\beta-\hat{\beta})}{\sigma_v} \right| \\
\le & \max_{v:\|v\|\le 1} \left| \frac{v^\top\left(\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} x_i x_i^\top - \Sigma\right)(\beta-\hat{\beta})}{\sigma_v} + \frac{v^\top\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} x_i\eta_i}{\sigma_v} \right| \\
\le & \left\|\Sigma^{-1/2}\left(\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} x_i x_i^\top - \Sigma\right)(\beta-\hat{\beta})\right\| + \left\|\Sigma^{-1/2}\frac{1}{|\mathcal{N}_{v,\hat{\beta},\alpha}|}\sum_{i\in\mathcal{N}_{v,\hat{\beta},\alpha}} x_i\eta_i\right\| \\
\le & \rho_2\|\Sigma^{1/2}(\beta-\hat{\beta})\| + \rho_1\gamma\,.
\end{aligned}
$$

Together with Lemma 3.4.6, this implies

$$
\frac{0.9D_S(\hat{\beta})\hat{\gamma} - \rho_1\gamma}{1+\rho_2} \;\le\; \left\|\Sigma^{1/2}(\beta-\hat{\beta})\right\| \;\le\; \frac{1.1D_S(\hat{\beta})\hat{\gamma} + \rho_1\gamma}{1-\rho_2}\,.
$$

Assuming $\rho_2 \le 0.013$, we have $0.86D_S(\hat{\beta}) - 1.1\rho_1 \le \left\|\Sigma^{1/2}(\beta-\hat{\beta})\right\|/\gamma \le 1.15D_S(\hat{\beta}) + 1.1\rho_1$. Since $D_S(\hat{\beta}) \le \tau$, we get the desired bound. □

3.4.2.4 Bounded volume

We show that the assumption (a) in Theorem 28 is satisfied for robust estimate $D_S(\hat{\beta})$.

Lemma 3.4.9. *For* $\rho = \rho_1$, $c_0 = 31.8$, $c_1 = 10.2$, $\tau = 42\rho_1$, $\Delta = 110\rho_1/(\alpha n)$, *and* $c_2 \ge \log(67/12) + \log((c_0+2c_1)/c_1)$, *we have* $(7/8)\tau - (k^*+1)\Delta > 0$,

$$
\begin{aligned}
\frac{\mathrm{Vol}(B_{\tau+(k^*+1)\Delta+c_1\rho,S})}{\mathrm{Vol}(B_{(7/8)\tau-(k^*+1)\Delta-c_1\rho,S})} &\le e^{c_2 d}\,, \text{ and} \\
\frac{\mathrm{Vol}(\{\hat{\beta} : \|\Sigma^{1/2}(\hat{\beta}-\beta)\|/\gamma \le (c_0+2c_1)\rho\})}{\mathrm{Vol}(\{\hat{\beta} : \|\Sigma^{1/2}(\hat{\beta}-\beta)\|/\gamma \le c_1\rho\})} &\le e^{c_2 d}\,.
\end{aligned}
$$

Proof. The proof is similar to the proof of Lemma 3.3.7. The second part of assumption (a) follows from the fact that

$$\text{Vol}(\{\hat{\beta} : \|\Sigma^{1/2}(\hat{\beta} - \beta)\| \leq r\}) = c_d|\Sigma|r^d \ ,$$

for some constant c_d that depends only on the dimension and selecting $c_2 \geq \log((c_0 + 2c_1)/c_1)$. The first part follows from our choices of c_0, c_1, τ, Δ and the following corollary.

Corollary 3.4.10 (Corollary of Lemma 3.4.8)**.** *If $\hat{\beta} \in B_{2\tau,S}$ and $\tau = 42\rho_1$, then $\left| \|\Sigma^{1/2}(\hat{\beta} - \beta)\|/\gamma - D_S(\hat{\beta}) \right| \leq 14.2\rho_1$.*

□

3.4.2.5 Resilience implies bounded local sensitivity

We show that resilience implies the assumption (b) in Theorem 28 (Lemma 3.4.14). Assuming $(k^* + 1)/n \leq \alpha/2$, we show a set S' with at most k^* data points arbitrarily changed from S has bounded local sensitivity. This implies that S' is a $((1/5.5)\alpha + (k^*/n), \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set with respect to (β, Σ, γ).

Lemma 3.4.11. *For an $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S' with respect to (β, Σ, γ), $\tilde{\alpha} \leq (1/11)\alpha$, and any unit norm $v \in \mathbb{R}^d$, we have $0.9\sigma_v \leq \sigma_v(\mathcal{M}_{v,\alpha}) \leq 1.1\sigma_v$.*

Proof. This follows from Lemma 3.3.9. □

Lemma 3.4.12. *For a $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S' with respect to (β, Σ, γ), and any unit norm vector $v \in \mathbb{R}^d$, we have $0.99\gamma \leq \hat{\gamma} \leq 1.01\gamma$.*

Proof. This proof follows from the proof of Lemma 3.4.7. □

Lemma 3.4.13. *For a $((1/5.5)\alpha + \tilde{\alpha}, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S' with respect to (β, Σ, γ) and $\tilde{\alpha} \leq (1/11)\alpha$, if $\hat{\beta} \in B_{t,S'}$ then we have $\|\Sigma^{1/2}(\hat{\beta} - \beta)\|/\gamma \leq 1.1\rho_1 + 1.15t$ and $\left|D_{S'}(\hat{\beta}) - \|\Sigma^{1/2}(\hat{\beta} - \beta)\|/\gamma\right| \leq 1.1\rho_1 + 0.15t$.*

Proof. This proof follows from the proof of Lemma 3.4.8. □

Lemma 3.4.14. *For $\Delta = 110\rho_1/(\alpha n)$, $\tau = 42\rho_1$, and an $((1/5.5)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good S, if*

$$n \;=\; \Omega\Big(\frac{\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big)$$

with a large enough constant, then the local sensitivity in assumption (b) is satisfied.

Proof. We follow the proof strategy of Lemma 3.3.11 in Section 3.3.2.5. Consider a dataset S' that is at Hamming distance at most $(1/11)\alpha n$ from S and corresponding partition $(\mathcal{T}'_{v,\hat{\beta},\alpha}, \mathcal{N}'_{v,\hat{\beta},\alpha}, \mathcal{B}'_{v,\hat{\beta},\alpha})$ of S' for a specific direction v. By the resilience property of the tails in Eq. (3.34) and Eq. (3.35), Lemma B.2.1, and Lemma B.2.2, we have for any $v \in \mathbb{R}^d$ with unit norm $\|v\| = 1$ and any $\hat{\beta} \in \mathbb{R}^d$,

$$\begin{aligned}
&\frac{v^\top \frac{1}{|\mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in \mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \Big(\big(x_i x_i^\top - \Sigma\big)(\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} \\
\leq\; & \left\| \Sigma^{-1/2} \left(\frac{1}{|\mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in \mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \big(x_i x_i^\top - \Sigma\big)(\beta - \hat{\beta}) \right) \right\| + \qquad (3.41) \\
& \left\| \Sigma^{-1/2} \left(\frac{1}{|\mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in \mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} x_i \eta_i \right) \right\| \\
\leq\; & \frac{2\rho_2}{(1/11)\alpha} \|\Sigma^{1/2}(\beta - \hat{\beta})\| + \frac{2\rho_1}{(1/11)\alpha}\gamma \,, \qquad (3.42)
\end{aligned}$$

where S_{good} is the original uncorrupted resilient dataset. Similarly, we have

$$\frac{v^\top \frac{1}{|\mathcal{B}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in \mathcal{B}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \Big(\big(x_i x_i^\top - \Sigma\big)(\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} \;\leq\; \frac{2\rho_2}{(1/11)\alpha} \|\Sigma^{1/2}(\beta - \hat{\beta})\| + \frac{2\rho_1}{(1/11)\alpha}\gamma \,.$$

This implies

$$\begin{aligned}
\min_{i \in \mathcal{T}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \frac{v^\top \Big(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} - \max_{i \in \mathcal{B}'_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \frac{\tilde{v}^\top \Big(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} & \\
\leq\; \frac{44\rho_1}{\alpha}\gamma + \frac{44\rho_2}{\alpha} \|\Sigma^{1/2}(\beta - \hat{\beta})\| \,. \qquad (3.43) &
\end{aligned}$$

Analogous to Lemma 3.3.11, for a neighboring databases S' and S'', the corresponding middle sets ${\mathcal{N}'}_{v,\hat{\beta},\alpha}$ and ${\mathcal{N}''}_{v,\hat{\beta},\alpha}$ differ by at most one entry. Denote those entries by x'_i and $\eta'_i = y'_i - \langle \beta, x'_i \rangle$ in ${\mathcal{N}'}_{v,\hat{\beta},\alpha}$ and x''_j and η''_j in ${\mathcal{N}''}_{v,\hat{\beta},\alpha}$. Then, from Eq. (3.43), we have

$$\left| v^\top \left(\left(x'_i {x'_i}^\top - x''_j {x''_j}^\top \right) (\beta - \hat{\beta}) + x'_i \eta'_i - x''_j \eta''_j \right) \right| \leq \left(\frac{44\rho_1}{\alpha}\gamma + \frac{44\rho_2}{\alpha} \|\Sigma^{1/2}(\beta - \hat{\beta})\| \right) \sigma_v \ ,$$

which implies that

$$\begin{aligned} &\left| v^\top \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i \in {\mathcal{N}'}_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \right) - v^\top \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i \in {\mathcal{N}''}_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \right) \right| \\ &\qquad \leq \quad \frac{\sigma_v}{(1-(4/5.5)\alpha)n} \left(\frac{44\rho_1}{\alpha}\gamma + \frac{44\rho_2}{\alpha} \|\Sigma^{1/2}(\beta - \hat{\beta})\| \right) \ . \end{aligned} \tag{3.44}$$

By the resilience properties in Eq. (3.28) and Eq. (3.29), and Lemma B.2.2, Lemma 3.4.1, and the fact that ${\mathcal{N}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}$ is at least of size $(1-\alpha)n$, we have for the data points in ${\mathcal{N}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}$,

$$\frac{v^\top \frac{1}{|{\mathcal{N}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in {\mathcal{N}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \left(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \right)}{\sigma_v} \leq (1+\rho_2) \|\Sigma^{1/2}(\hat{\beta} - \beta)\| + \rho_1 \gamma \ .$$

By Eq. (3.42), for any $x''_i \in {\mathcal{N}''}_{v,\hat{\beta},\alpha} \cap S_{\text{bad}}$ (where $S_{\text{bad}} = S'' \setminus S_{\text{good}}$), we have

$$\begin{aligned} \frac{v^\top \left(x''_i {x''_i}^\top (\beta - \hat{\beta}) + x''_i \eta''_i \right)}{\sigma_v} \quad &\leq \quad \frac{v^\top \frac{1}{|{\mathcal{T}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}|} \sum_{i \in {\mathcal{T}''}_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \left(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \right)}{\sigma_v} \\ &\leq \quad \left(\frac{22\rho_2}{\alpha} + 1 \right) \|\Sigma^{1/2}(\hat{\beta} - \beta)\| + \frac{22\rho_1}{\alpha}\gamma \ . \end{aligned}$$

Since $|S_{\text{bad}}| \leq (1.5/5.5)\alpha n$ and $\alpha < c$ for some small enough constant c, we have

$$\begin{aligned}
& \frac{v^\top \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i \in \mathcal{N}''_{v,\hat{\beta},\alpha}} \Big(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} \\
= & \frac{v^\top \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i \in \mathcal{N}''_{v,\hat{\beta},\alpha} \cap S_{\text{bad}}} \Big(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} + \\
& \frac{v^\top \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i \in \mathcal{N}''_{v,\hat{\beta},\alpha} \cap S_{\text{good}}} \Big(x_i x_i^\top (\beta - \hat{\beta}) + x_i \eta_i \Big)}{\sigma_v} \\
\leq & \frac{(6\rho_2 + (1.5/5.5)\alpha)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| + 6\rho_1\gamma}{1 - (4/5.5)\alpha} + \Big((1+\rho_2)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| + \rho_1\gamma \Big) \\
\leq & 7\rho_1\gamma + (1 + \alpha + 7\rho_2)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \, . \qquad (3.45)
\end{aligned}$$

Analogous to Eq. (3.19), by using the resilience properties in Eqs. (3.29) and (3.30), we have

$$\begin{aligned}
|\sigma_v'^2 - \sigma_v''^2| &= \frac{1}{(1-(4/5.5)\alpha)n} \left| \sum_{x_i \in \mathcal{N}'_{v,\hat{\beta},\alpha}} \langle v, x_i \rangle^2 - \sum_{x_i \in \mathcal{N}''_{v,\hat{\beta},\alpha}} \langle v, x_i \rangle^2 \right| \\
&\leq \frac{64 \cdot 11^2 \cdot \rho_3^2 \sigma_v^2}{\alpha^2 (1-(4/5.5)\alpha)n} \, . \qquad (3.46)
\end{aligned}$$

By Eqs. (3.45), (3.44), and (3.46), we have

$$
\begin{aligned}
& \left|D_{S'}(\hat{\beta}) - D_{S''}(\hat{\beta})\right| \\
\leq & \max_{v:\|v\|=1} \left| \frac{v^\top \frac{1}{|\mathcal{N}'_{v,\hat{\beta},\alpha}|} \sum_{i\in\mathcal{N}'_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top(\beta-\hat{\beta}) + x_i\eta_i\right)}{\sigma'_v\hat{\gamma}'} - \frac{v^\top \frac{1}{|\mathcal{N}''_{v,\hat{\beta},\alpha}|} \sum_{i\in\mathcal{N}''_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top(\beta-\hat{\beta}) + x_i\eta_i\right)}{\sigma''_v\hat{\gamma}''} \right| \\
\leq & \max_{v:\|v\|=1} \left| \frac{v^\top \left(\frac{1}{(1-(4/5.5)\alpha)n} \sum_{i\in\mathcal{N}'_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top(\beta-\hat{\beta}) + x_i\eta_i\right) - \frac{1}{(1-(4/5.5)\alpha)n} \sum_{i\in\mathcal{N}''_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top(\beta-\hat{\beta}) + x_i\eta_i\right) \right.}{\sigma'_v\hat{\gamma}'} \right. \\
& + \max_{v:\|v\|=1} \frac{v^\top \frac{1}{|\mathcal{N}''_{v,\hat{\beta},\alpha}|} \sum_{x_i\in\mathcal{N}''_{v,\hat{\beta},\alpha}} \left(x_i x_i^\top(\beta-\hat{\beta}) + x_i\eta_i\right)}{\sigma_v} \left| \frac{\sigma_v}{\sigma'_v\hat{\gamma}'} - \frac{\sigma_v}{\sigma''_v\hat{\gamma}''} \right| \\
\leq & \frac{44\rho_1}{0.9\cdot 0.99(1-(4/5.5)\alpha)n\alpha} + \frac{44\rho_2}{0.9\cdot 0.99(1-(4/5.5)\alpha)n\alpha} \frac{\|\Sigma^{1/2}(\beta-\hat{\beta})\|}{\gamma} \\
& + \frac{64\cdot 11^2\cdot\rho_3^2\cdot 0.02\gamma}{0.9^3\alpha^2(1-(4/5.5)\alpha)n\cdot 0.99^2\gamma^2} \left(7\rho_1\gamma + (1+\alpha+7\rho_2)\|\Sigma^{1/2}(\hat{\beta}-\beta)\|\right) \qquad (3. \\
\leq & \left(\frac{0.12}{\alpha n} + \frac{0.016}{\alpha n}\right) \frac{\|\Sigma^{1/2}(\hat{\beta}-\beta)\|}{\gamma} + \left(\frac{9\rho_1}{\alpha n} + \frac{0.07\rho_1}{\alpha n}\right) \\
\leq & \frac{0.2}{\alpha n} \frac{\|\Sigma^{1/2}(\hat{\beta}-\beta)\|}{\gamma} + \frac{50\rho_1}{\alpha n}
\end{aligned}
$$

where the last three inequalities follow from our assumptions that $\alpha \leq c$ and $\rho_2 \leq c$, $\rho_3^2 \leq c\alpha$, $\rho_4 \leq c$ with a small enough constant c and Lemma 3.4.12. From Lemma 3.4.13, we know that if $\hat{\beta} \in B_{\tau+(k^*+3)\Delta,S}$, we have $\|\Sigma^{1/2}(\hat{\beta}-\beta)\|/\gamma \leq 1.1\rho_1 + 1.15(\tau+(k^*+3)\Delta)$. We show that $\|\Sigma^{1/2}(\hat{\beta}-\beta)\| \leq 50\rho_1\gamma$ for the choices of Δ, k^*, τ and n:

$$
\begin{aligned}
1.1\rho_1 + 1.15(\tau+(k^*+3)\Delta) &\leq 49\rho_1 + \frac{50\rho_1\log(1/(\delta\zeta))}{\varepsilon\alpha n} \\
&\leq 50\rho_1 \,,
\end{aligned}
$$

where $\Delta = 110\rho_1/(\alpha n)$, $\tau = 42\rho_1$, $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, $\varepsilon \leq \log(4/\delta\zeta)$ and $n \geq C'\log(1/(\delta\zeta))/(\varepsilon\,\alpha)$ for some large enough universal constant $C' > 0$. This implies that

$$
|D_{S'}(\hat{\beta}) - D_{S''}(\hat{\beta})| \leq \frac{110\rho_1}{\alpha n} = \Delta \,.
$$

□

3.4.2.6 Proof of Theorem 22

We show that the sufficient conditions of Theorem 28 are met for the following choices of constants and parameters: $p = d$, $\rho = \rho_1$, $c_0 = 31.8$, $c_1 = 10.2$, $\tau = 42\rho_1$, and $\Delta = 110\rho_1/(\alpha n)$. We set c_2 to be a large constant and change only the constant factor in the sample complexity. The assumptions (a), (b), and (d) follow from Lemmas 3.4.9, 3.4.14, and 3.4.8, respectively. The assumption (c) follows from

$$\Delta = \frac{110\rho_1}{\alpha n} \le \frac{1.2\rho_1\varepsilon}{32(c_2 d + (\varepsilon/2) + \log(16/(\delta\zeta)))} = \frac{(c_0 - 3c_1)\rho\varepsilon}{32(c_2 p + (\varepsilon/2) + \log(16/(\delta\zeta)))}$$

for a large enough $n \ge C'(d + \log(1/(\delta\zeta)))/(\alpha\varepsilon)$. This finishes the proof of Theorem 22, from which Theorem 21 follows immediately.

3.4.3 Step 3: Achievability guarantees

We provide utility guarantees for popular families of distributions studied in the private or robust linear regression literature: sub-Gaussian [70, 93, 217, 40, 206] and hypercontractive [217, 143, 51, 120, 22, 174]. Similar to mean estimation, the resilience we need scales with the variance. For sub-Gaussian distributions, this requires a lower bound on the variance of the form $\sigma \preceq c\Gamma$ for the sub-Gaussian proxy Γ. For the k-th moment bounded distributions, we require hypercontractivity.

3.4.3.1 Sub-Gaussian distributions

The most common scenario in linear regression is when both input x_i and noise η_i are sub-Gaussian (as defined in Eq. (3.21)) and independent of each other. The next lemma shows that the resulting dataset is $(O(\alpha\log(1/\alpha)), O(\alpha\log(1/\alpha)), O(\alpha\sqrt{\log(1/\alpha)}), O(\alpha\log(1/\alpha)))$-resilient, which follows from the covariance resilience of sub-Gaussian distributions.

Lemma 3.4.15 (Resilience for sub-Gaussian samples)**.** *Let $\mathcal{D}_1$ be a distribution of $x_i \in \mathbb{R}^d$, which is zero mean sub-Gaussian with covariance Σ and sub-Gaussian proxy $0 \prec \Gamma \preceq c\Sigma$ for some constant c. Let $\mathcal{D}_2$ be a distribution of $\eta_i \in \mathbb{R}$, which is a zero mean one-dimensional*

sub-Gaussian with variance γ^2 *and sub-Gaussian proxy* $\gamma_0^2 \le c\gamma^2$ *for some constant* c. *A multiset of i.i.d. labeled samples* $S = \{(x_i, y_i)\}_{i=1}^n$ *is generated from a linear model with noise* η_i *independent of* x_i: $y_i = x_i^\top \beta + \eta_i$, *where the input* x_i *and the independent noise* η_i *are i.i.d. samples from* $\mathcal{D}_1$ *and* $\mathcal{D}_2$. *There exist constants* c_1 *and* $c_2 > 0$ *such that, for any* $\alpha \in (0, 1/2)$, *if* $n \ge c_1((d + \log(1/\zeta))/(\alpha \log(1/\alpha))^2)$, *then, with probability* $1 - \zeta$, S *is* $(\alpha, c_2\alpha\log(1/\alpha), c_2\alpha\log(1/\alpha), c_2\alpha\sqrt{\log(1/\alpha)}, c_2\alpha\log(1/\alpha))$*-resilient with respect to* (β, Σ, γ).

Proof. This follows from [121, Corollary 4]. Let $\tilde{x}_i := \begin{bmatrix} \Sigma^{-1/2} x_i \\ \eta_i/\gamma \end{bmatrix} \in \mathbb{R}^{d+1}$. By definition, we know that $\tilde{x}_i$ can be seen as samples from a zero mean sub-Gaussian distribution with covariance $\mathbf{I}_{(d+1)\times(d+1)}$. By [121, Corollary 4] and a union bound, we know that if $n = \Omega(d + \log(1/\zeta))/(\alpha\log(1/\alpha))^2$, then there exists a constant C_1 such that with probability $1 - \zeta$, for any $T \subset S$ and $|T| \ge (1-\alpha)n$ and any unit vector $u \in \mathbb{R}^{d+1}$, $v \in \mathbb{R}^d$, we have

$$\left| u^\top \left(\frac{1}{|T|} \sum_{x_i \in T} \tilde{x}_i \tilde{x}_i^\top - \mathbf{I}_{(d+1)\times(d+1)} \right) u \right| \le C_1 \alpha \log(1/\alpha) \ , \tag{3.48}$$

$$\left| v^\top \left(\frac{1}{|T|} \sum_{x_i \in T} \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_{d\times d} \right) v \right| \le C_1 \alpha \log(1/\alpha) \ , \text{ and} \tag{3.49}$$

$$\left| \frac{1}{|T|} \sum_{\eta_i \in T} \frac{\eta_i^2}{\gamma^2} - 1 \right| \le C_1 \alpha \log(1/\alpha) \ . \tag{3.50}$$

Let $u := \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$, where $u_1 \in \mathbb{R}^d$ and $u_2 \in \mathbb{R}$ and $\|u_1\|^2 + u_2^2 = 1$. Then, Eq. (3.48) is equivalent to

$$\left| u_1^\top \left(\frac{1}{|T|} \sum_{i \in T} \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_{d\times d} \right) u_1 + \frac{2u_2}{\gamma} u_1^\top \frac{1}{|T|} \sum_{i \in T} \Sigma^{-1/2} x_i \eta_i + \frac{u_2^2}{\gamma^2} \frac{1}{|T|} \sum_{i \in T} (\eta_i^2 - \gamma^2) \right|$$
$$\le \quad C_1 \alpha \log(1/\alpha) \ . \tag{3.51}$$

By Eq. (3.49) and (3.50), we know

$$\left| u_1^\top \left(\frac{1}{|T|} \sum_{i \in T} \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_{d\times d} \right) u_1 \right| \le C_1 \alpha \log(1/\alpha) \|u_1\|^2$$

$$\left| \frac{u_2^2}{\gamma^2} \frac{1}{|T|} \sum_{i \in T} (\eta_i^2 - \gamma^2) \right| \le C_1 \alpha \log(1/\alpha) u_2^2 \ .$$

This means that

$$-C_1\alpha\log(1/\alpha)(1+\|u_1\|^2+u_2^2) \le \frac{2u_2}{\gamma}u_1^\top\frac{1}{|T|}\sum_{i\in T}\Sigma^{-1/2}x_i\eta_i \le C_1\alpha\log(1/\alpha)(1+\|u_1\|^2+u_2^2) \quad (3.52)$$

For any unit vector $w\in\mathbb{R}^d$, let $u_1 = 0.5w$. Thus, we have $u_2^2 = 0.75$. Eq. (3.52) implies

$$\left|\frac{1}{\gamma}w^\top\frac{1}{|T|}\sum_{i\in T}\Sigma^{-1/2}x_i\eta_i\right| \le C_2\alpha\log(1/\alpha) \quad (3.53)$$

for some constant C_2. This proves the first resilience property in Eq. (3.28). The second, third and fourth resilience properties in Eqs. (3.29), (3.30) and (3.31) follow from [73, Lemma 4.1], [121, Corollary 4] and a union bound.

□

The preceding resilience lemma and Theorem 22 imply the following optimal utility guarantee.

Corollary 3.4.16. *Under the hypothesis of Lemma 3.4.15, there exists a constant $c>0$ such that for any $\alpha\in(0,c)$, a sample size of*

$$n = O\Big(\frac{d+\log(1/\zeta)}{(\alpha\log(1/\alpha))^2} + \frac{d+\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big),$$

a sensitivity of $\Delta = O(\log(1/\alpha)/n)$, and a threshold of $\tau = O(\alpha\log(1/\alpha))$ with large enough constants are sufficient for HPTR(S) *with the distance function in Eq. (3.26) to achieve*

$$\frac{1}{\gamma}\|\Sigma^{1/2}(\hat\beta-\beta)\| \;=\; O(\alpha\log(1/\alpha)) \quad (3.54)$$

with probability $1-\zeta$. Further, the same guarantee holds even if α-fraction of the samples is arbitrarily corrupted, as in Assumption 4.

The sample complexity is nearly optimal. Even for DP linear regression without robustness, HPTR is the first algorithm for sub-Gaussian distributions with an unknown covariance Σ that up to log factors matches the lower bound of $n=\tilde\Omega(d/\alpha^2+d/(\alpha\varepsilon))$ assuming $\varepsilon<1$ and $\delta<n^{-1-\omega}$ for some $\omega>0$ from [40, Theorem 4.1]. For completeness, we provide the

lower bound in Appendix B.3. An existing algorithm for DP linear regression from [40] is suboptimal since it require Σ to be close to the identity matrix, which is equivalent to assuming that we know Σ.

The error bound is nearly optimal under α-corruption, i.e., HPTR is the first robust estimator that is both differentially private and also achieves the near-optimal error rate of $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = O(\alpha\log(1/\alpha))$, matching the known information-theoretic lower bound of $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = \Omega(\alpha)$ [93] up to a log factor. This lower bound holds for any robust estimator that is not necessarily private and regardless of how many samples are available. If privacy is not required (i.e., $\varepsilon = \infty$), a similar guarantee can be achieved by, for example, [70].

3.4.3.2 Hypercontractive distributions with independent noise

We assume that x_i and η_i are independent and (κ, k)-hypercontractive and $(\tilde{\kappa}, k)$-hypercontractive, respectively, as in Definition 3.3.14. For the necessity of hypercontractive conditions for robust linear regression, we refer to [217, Section F.5]. The next lemma shows that the resulting dataset has a subset of size at least $(1-\alpha)n$ that is $(O(\alpha), O(\alpha^{1-1/k}), O(\alpha^{1-2/k}), O(\alpha^{1-1/k}), O(\alpha^{1-2/k}))$-resilient.

Lemma 3.4.17 (Resilience for hypercontractive samples)**.** *For some integer $k \geq 4$ and positive scalar parameters κ and $\tilde{\kappa}$, let $\mathcal{D}_1$ be a (κ, k)-hypercontractive distribution on $x_i \in \mathbb{R}^d$ with zero mean and covariance $\Sigma \succ 0$. Let $\mathcal{D}_2$ be a $(\tilde{\kappa}, k)$-hypercontractive distribution on $\eta_i \in \mathbb{R}$ with zero mean and variance γ^2. A multiset of labeled samples $S = \{(x_i, y_i)\}_{i=1}^n$ is generated from the linear model $y_i = x_i^\top \beta + \eta_i$, where the input x_i and the independent noise η_i are i.i.d. samples from $\mathcal{D}_1$ and $\mathcal{D}_2$. For any $\alpha \in (0, 1/2)$ and any constant $c_3 > 0$, there exist constants c_1 and $c_2 > 0$ that depend only on c_3 such that if*

$$n \;\geq\; c_1\Big(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k}(1+1/\tilde{\kappa}^2)d\log d}{\zeta^{2-4/k}\kappa^2} + \frac{\kappa^2(1+\tilde{\kappa}^2)d\log d}{\alpha^{2/k}}\Big)\,, \tag{3.55}$$

then S is $(c_3\alpha, \alpha, c_2 k\kappa\tilde{\kappa}\alpha^{1-1/k}\zeta^{-1/k}, c_2k^2\kappa^2\alpha^{1-2/k}\zeta^{-2/k}, c_2k\kappa\alpha^{1-1/k}\zeta^{-1/k}, c_2k^2\tilde{\kappa}^2\alpha^{1-2/k}\zeta^{-2/k})$-corrupt good with respect to (β, Σ, γ) with probability $1-\zeta$.

Proof. Since x_i and η_i are independent, we know

$$\mathbb{E}\left[\left|\left\langle v, \gamma^{-1}\Sigma^{-1/2}x\eta\right\rangle\right|^k\right] = \mathbb{E}\left[\left|\left\langle v, \Sigma^{-1/2}x\right\rangle\right|^k\right]\mathbb{E}\left[|\gamma^{-1}\eta|^k\right] \le \kappa^k\tilde{\kappa}^k \ .$$

This implies that $\gamma^{-1}\Sigma^{-1/2}x\eta$ is a k-th moment bounded distribution with covariance $\mathbf{I}_{d\times d}$. By Lemma 3.3.15, under the sample complexity of (3.55), with probability $1-8\zeta$, there exists a subset $S_{\text{good}} \subset S$ such that $|S_{\text{good}}| \ge (1-\alpha)n$, and there exists a constant C such that for any subset $T \subset S_{\text{good}}$ and $|T| \ge (1-10\alpha)|S_{\text{good}}|$, we have

$$\left\|\frac{1}{|T|}\sum_{i\in T}\frac{1}{\gamma}\Sigma^{-1/2}x_i\eta_i\right\| \ \le \ Ck\kappa\tilde{\kappa}\gamma\alpha^{1-1/k}\zeta^{-1/k} \ . \tag{3.56}$$

This proves the first resilience property in Eq. (3.28). The second resilience property in Eq. (3.29), the third in Eq. (3.30) and the fourth in Eq. (3.31) follow directly from Lemma 3.3.15. □

The preceding resilience lemma and Theorem 22 imply the following utility guarantee. HPTR is naturally robust against $(1/5.5 - c_3)\alpha$-corruption of the data. Choosing appropriate constants, we get the following result.

Corollary 3.4.18. *Under the hypothesis of Lemma 3.4.17, there exists a constant $c > 0$ such that for any $\alpha \le c$ and $k^2\kappa^2\alpha^{1-2/k} \le c$, it is sufficient to have a dataset of size*

$$n = O\Big(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k}(1+1/\tilde{\kappa}^2)d\log d}{\zeta^{2-4/k}\kappa^2} + \frac{\kappa^2(1+\tilde{\kappa}^2)d\log d}{\alpha^{2/k}} + \frac{d+\log(1/\delta)}{\alpha\varepsilon}\Big) \tag{3.57}$$

a sensitivity of $\Delta = O(1/(n\alpha^{1/k}))$, and a threshold of $\tau = O(\alpha^{1-1/k})$ with large enough constants for HPTR(S) *with the distance function in Eq. (3.26) to achieve $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(k\kappa\tilde{\kappa}\alpha^{1-1/k}\zeta^{-1/k})$ with probability $1-\zeta$. Further, the same guarantee holds even if α-fraction of the samples is arbitrarily corrupted, as in Assumption 4.*

The error bound is optimal under α-corruption; namely, the error bound $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(\alpha^{1-1/k})$ matches the lower bound $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = \Omega(\alpha^{1-1/k})$ by [22], where the noise η_i is $(1,k)$-hypercontractive and independent of x_i, which is also $(1,k)$-hypercontractive.

For completeness, we provide the lower bound in Appendix B.3. HPTR is the first algorithm that guarantees both differential privacy and an optimal robust error bound of $O(\alpha^{1-1/k})$ for hypercontractive distributions. If only robust error bound under α-corruption is at issue, [217] also achieves the same optimal error bound but does not provide differential privacy. Further, in this robust but not private case with $\varepsilon = \infty$, our sample complexity improves by a factor of $\alpha^{2/k}$ upon the state-of-the-art sample complexity of [217, Theorem 3.3], which shows that $n = O(d/\alpha^2)$ is sufficient to achieve $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(\alpha^{1-1/k})$.

Remark. Suppose $k, \kappa, \tilde{\kappa}$, and ζ are $\Theta(1)$. HPTR achieves $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(\alpha^{1-1/k})$ with $n = \tilde{O}(d/(\alpha^{2-2/k}) + (d + \log(1/\delta))/(\alpha\varepsilon))$ samples, where $\tilde{O}$ hides logarithmic factors in d. The first term cannot be improved upon since it matches the first term of a lower bound of $n = \tilde{\Omega}(d/\alpha^{2-2/k} + d/(\alpha^{1-1/k}\varepsilon))$ from [40, Theorem 4.1], which holds even for a standard, non-robust sub-Gaussian (which is (c_k, k)-hypercontractive for any $k \in \mathbb{Z}_+$ and a constant c_k that depends only on k) linear regression with independent noise (see Appendix B.3 for a precise statement). However, we do not have a matching lower bound for the second term. To the best of our knowledge, HPTR is the first algorithm for linear regression that guarantees (ε, δ)-DP under hypercontractive distributions with independent noise.

3.4.3.3 Hypercontractive distributions with dependent noise

We assume x_i and η_i may be dependent and marginally (κ, k)-hypercontractive and $(\tilde{\kappa}, k)$-hypercontractive, respectively, as defined in Definition 3.3.14. In this case, the first resilience ρ_1 that determines the error rate increases from $O(\alpha^{1-1/k})$ to $O(\alpha^{1-2/k})$ as a result of the potential correlation between input and noise. The next lemma shows that the the resulting dataset has a subset of size at least $(1-\alpha)n$ that is $(O(\alpha), O(\alpha^{1-2/k}), O(\alpha^{1-2/k}), O(\alpha^{1-1/k}), O(\alpha^{1-2/k}))$-resilient.

Lemma 3.4.19 (Resilience for hypercontractive samples with dependent noise)**.** *For some integer $k \geq 4$ and positive scalar parameters κ and $\tilde{\kappa}$, let $\mathcal{D}_1$ be a (κ, k)-hypercontractive distribution on $x_i \in \mathbb{R}^d$ with zero mean and covariance $\Sigma \succ 0$. Let $\mathcal{D}_2$ be a $(\tilde{\kappa}, k)$-hypercontractive*

distribution on $\eta_i \in \mathbb{R}$ *with variance* γ^2. *A multiset of labeled samples* $S = \{(x_i, y_i)\}_{i=1}^n$ *is generated from a linear model as follows:* $y_i = x_i^\top \beta + \eta_i$, *where* $\{(x_i, \eta_i)\}_{i\in[n]}$ *are i.i.d. samples from some distribution* $\mathcal{D}$ *whose marginal distribution for* x_i *is* $\mathcal{D}_1$, *the marginal distribution for* η_i *is* $\mathcal{D}_2$, *and* $\mathbb{E}[x_i\eta_i] = 0$. *For any* $\alpha \in (0, 1/2)$ *and* $c_3 > 0$, *there exist constants* c_1 *and* $c_2 > 0$ *that depend only on* c_3 *such that if*

$$n \;\geq\; c_1\Big(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-4/k}(1+1/\tilde{\kappa}^2)d\log d}{\zeta^{2-4/k}\kappa^2\tilde{\kappa}^2} + \frac{\kappa^2(\tilde{\kappa}^2+1)d\log d}{\alpha^{4/k}}\Big)\,, \tag{3.58}$$

then S *is* $(c_3\alpha, \alpha, c_2 k\kappa\tilde{\kappa}\alpha^{1-2/k}\zeta^{-2/k}, c_2k^2\kappa^2\alpha^{1-2/k}\zeta^{-2/k}, c_2k\kappa\alpha^{1-1/k}\zeta^{-1/k}, c_2k^2\tilde{\kappa}^2\alpha^{1-2/k}\zeta^{-2/k})$-*corrupt good with respect to* (β, Σ, γ) *with probability* $1-\zeta$.

Proof. Since η_i and x_i are dependent, we can bound only the $k/2$-th moment of $\gamma^{-1}\Sigma^{-1/2}x\eta$. By the Holder inequality, we have

$$\mathbb{E}\left[\left|\left\langle v, \Sigma^{-1/2}\gamma^{-1}x\eta\right\rangle\right|^{k/2}\right] \leq \sqrt{\mathbb{E}\left[\left|\langle v, \Sigma^{-1/2}x\rangle\right|^k\right]\mathbb{E}\left[|\gamma^{-1}\eta|^k\right]} \leq \kappa^{k/2}\tilde{\kappa}^{k/2}\,.$$

The rest of the proof follows similarly to the proof of Lemma 3.4.17.

□

The preceding resilience lemma and Theorem 22 imply the following optimal utility guarantee, which achieves an error rate of $O(\alpha^{1-2/k})$.

Corollary 3.4.20. *Under the hypothesis of Lemma 3.4.19, there exists a constant* $c > 0$ *such that for any* $\alpha \leq c$ *and* $k^2\kappa^2\alpha^{1-2/k} \leq c$, *it is sufficient to have a dataset of size*

$$n = O\Big(\frac{d+\log(1/\delta)}{\alpha\varepsilon} + \frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-4/k}(1+1/\tilde{\kappa}^2)d\log d}{\zeta^{2-4/k}\kappa^2\tilde{\kappa}^2} + \frac{\kappa^2(\tilde{\kappa}^2+1)d\log d}{\alpha^{4/k}}\Big)\,,$$

a sensitivity $\Delta = O(1/(n\alpha^{2/k}))$, *and a threshold* $\tau = O(\alpha^{1-2/k})$, *with large enough constants for* HPTR(S) *with the distance function in Eq. (3.26) to achieve* $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = O(k\kappa\tilde{\kappa}\alpha^{1-2/k}\zeta^{-2/k})$ *with probability* $1-\zeta$. *Further, the same guarantee holds even if an* α*-fraction of the samples is arbitrarily corrupted, as in Assumption 4.*

This error rate is optimal in its dependence on α under α-corruption. When η_i and x_i are dependent, [22] gives a lower bound of error rate $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta}-\beta)\| = \Omega(\tilde{\kappa}\alpha^{1-2/k})$

that holds regardless of how many samples we have and without the privacy constraints. For completeness, we provide the lower bound in Appendix B.3. If only a robust error bound under α-corruption is at issue, [217] also achieves the same optimal error bound but does not provide differential privacy. Further, in this robust but not private case with $\varepsilon = \infty$, our sample complexity improves by a factor of $\alpha^{2/k}$ upon the state-of-the-art sample complexity of [217, Theorem 3.3], which shows that $n = O(d/\alpha^2)$ is sufficient to achieve $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O(\alpha^{1-2/k})$.

Remark. Suppose $\zeta, \kappa, \tilde{\kappa}$, and k are $\Theta(1)$. The sample complexity of HPTR is $n = \tilde{O}((d + \log(1/\delta))/\alpha^{2(1-1/k)} + d/(\alpha\varepsilon))$. The first term has a gap of a $\alpha^{-2/k}$ factor compared to the first term of a lower bound of $n = \tilde{\Omega}(d/\alpha^{2(1-2/k)} + d/(\alpha^{1-2/k}\varepsilon))$ from [40, Theorem 4.1], which holds even for standard, non-robust sub-Gaussian DP linear regression. It remains an open question whether this gap can be closed, either by a tighter analysis of the resilience for HPTR or a tighter analysis for a lower bound.

On the upper bound, the gap comes from ensuring stronger resilience than we need. From Theorem 21, we know that we require $\rho_1 \leq c$ and $\rho_3^2 \leq c\alpha$, and from the optimal error rate, we want $\rho_1 \leq c\alpha^{1-2/k}$. The resilience we ensure in Lemma 3.4.19 is $(\alpha, \rho_1 = \alpha^{1-2/k}, \rho_2 = \alpha^{1-2/k}, \rho_3 = \alpha^{1-1/k})$, which guarantees an unnecessarily small ρ_2 and ρ_3. A similar slack was also in the mean estimation, which did not affect the final sample complexity. In this case, i.e., with linear regression and hypercontractive distributions, it enlarges sample complexity. Tighter analysis of the resilience, which guarantees a larger ρ_2 and ρ_3, can improve the first term in the sample complexity in its dependence on α, but it cannot close the $\alpha^{-2/k}$ gap. On the lower bound, we apply a construction of [40, Theorem 4.1], which uses Gaussian distributions and an independent noise. One could potentially tighten the lower bound with a construction that uses hypercontractive distributions and a dependent noise.

For the second term, we provide a nearly matching lower bound of $n = \Omega(\min\{d, \log(1/\delta)\}/\alpha\varepsilon)$ to achieve $(1/\gamma)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|^2 \leq O(\alpha^{2-4/k})$ in Proposition 3.4.21, proving that it is tight when $\delta = \exp(-\Theta(d))$. To the best of our knowledge, HPTR is the first algorithm for linear

regression that guarantees (ε, δ)-DP under hypercontractive distributions with dependent noise.

Proposition 3.4.21 (Lower bound of hypercontractive linear regression with dependent noise)**.** *For any $k \geq 4$, let $\mathcal{P}_{\kappa,k,\Sigma,\gamma^2}$ be a distribution over $(x_i, \eta_i) \in \mathbb{R}^d \times \mathbb{R}$, where x_i is (κ, k)-hypercontractive with zero mean and covariance Σ, and η_i is (κ, k)-hypercontractive with zero mean and variance γ^2. We observe labelled examples a linear model $y_i = x_i^\top \beta + \eta_i$ with $\mathbb{E}[x_i \eta_i] = 0$ such that $\beta = \Sigma^{-1}\mathbb{E}[y_i x_i]$. Let $\mathcal{M}_{\varepsilon,\delta}$ denote a class of (ε, δ)-DP estimators that are measurable functions over n i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^n$ from a distribution. There exist positive constants $c, \gamma, \kappa = O(1)$ such that, for $\varepsilon \in (0, 10)$,*

$$\inf_{\hat{\beta} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma \succ 0, P \in \mathcal{P}_{\kappa,k,\Sigma,\gamma^2}} \frac{1}{\gamma} \mathbb{E}_{P^n}[\|\Sigma^{1/2}(\hat{\beta}(S) - \beta)\|^2] \;\geq\; c \min\left\{ \left(\frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon} \right)^{2-4/k}, 1 \right\} .$$

Proof. We adopt the same framework as used in the proof of Proposition 3.3.18. We choose $\mathcal{P}$ to be $\mathcal{P} = \mathcal{P}_{\Sigma,k}$. It suffices to construct index set $\mathcal{V}$ and indexed family of distributions $\mathcal{P}_{\mathcal{V}}$ such that $d_{\mathrm{TV}}(P_v, P_{v'}) = \alpha$ and $\rho(\beta_v, \beta_{v'}) \geq t$, where β_v is the least square solution of P_v. By [3, Lemma 6], there exists a finite set $\mathcal{V} \subset \mathbb{R}^d$ with cardinality $|\mathcal{V}| = 2^{\Omega(d)}$, $\|v\| = 1$ for all $v \in \mathcal{V}$, and $\|v - v'\| \geq 1/2$ for all $v \neq v' \in \mathcal{V}$. Let $f_{\mu,\Sigma}(x)$ be a density function of $\mathcal{N}(\mu, \Sigma)$. We construct a marginal distribution over $\mathbb{R}^d$ as follows

$$D_x^v(x) = \begin{cases} \alpha/2, & \text{if } x = -\alpha^{-1/k} v \;, \\ \alpha/2, & \text{if } x = \alpha^{-1/k} v \;, \\ (1-\alpha) f_{0, \mathbf{I}_{d\times d}}(x) & \text{otherwise} \;, \end{cases} \quad . \tag{3.59}$$

It is straightforward to verify that $\mathbb{E}_{P_x^v}[x] = 0$, $\mathbb{E}_{P_x^v}[xx^\top] = (1-\alpha)\mathbf{I}_{d\times d} + \alpha^{1-2/k} vv^\top$ and thus $\frac{1}{2}\mathbf{I}_{d\times d} \preceq \mathbb{E}_{P_x^v}[xx^\top] \preceq 2\mathbf{I}_{d\times d}$ for $\alpha \leq 1/2$. Furthermore, we have

$$\mathbb{E}_{x \sim P_x^v}[\,|\langle u, x\rangle|^k\,] \;\;\leq\;\; \langle u, v\rangle^k + (1-\alpha) c_k^k = O(1) \;,$$

where we use the fact that there exists a constant $c_k > 0$ such that the k-th moment of Gaussian distribution is bounded by c_k^k. Since $\frac{1}{2}\mathbf{I}_{d\times d} \preceq \mathbb{E}_{P_x^v}[xx^\top] \preceq 2\mathbf{I}_{d\times d}$, we know that x is

$(O(1), k)$-hypercontractive. We construct conditional distribution $D^v(y|x)$ as follows

$$y|x = \begin{cases} -\alpha^{-1/k} & \text{if } x = -\alpha^{-1/k} v \\ \alpha^{-1/k} & \text{if } x = \alpha^{-1/k} v \\ \mathcal{N}(0,1) & \text{otherwise} \end{cases} .$$

Then, we have

$$\begin{aligned} \beta_v &= \mathbb{E}_{x \sim P_x^v}[xx^\top]^{-1} \mathbb{E}_{x,y \sim P_{x,y}^v}[xy] \\ &= \mathbb{E}_{x \sim P_x^v}[xx^\top]^{-1} \alpha^{1-2/k} v \ . \end{aligned}$$

This implies that $t = \min_{v \neq v' \in \mathcal{V}} \|\beta_v - \beta_{v'}\| \geq 1/2 \alpha^{1-2/k} \min_{v \neq v' \in \mathcal{V}} \|v - v'\| = \Omega(\alpha^{1-2/k})$. We are left to verify that $\eta = y - \langle \beta_v, x \rangle$ is also hypercontractive:

$$\mathbb{E}[|\eta|^k] = \alpha \left| \alpha^{-1/k} - v^\top \mathbb{E}_{x \sim P_x^v}[xx^\top]^{-1} v \alpha^{1-3/k} \right|^k + (1-\alpha) \mathbb{E}_{x \sim \mathcal{N}(0, 2\mathbf{I}_{d \times d})}[|x|^k] = O(1) \ ,$$

where we use the fact that the k-th moment of standard Gaussian is bounded by some constants $C_k > 0$ and $k = O(1)$. It is straightforward to see that the total variation distance $d_{\mathrm{TV}}(P_{x,y}^v, P_{x,y}^{v'}) = \alpha$.

Next, we apply a reduction of estimation to testing with this packing $\mathcal{V}$ similar to that we used in the proof of Proposition 3.3.18. For (ε, δ)-DP estimator $\hat{\beta}$, using Theorem 3.3.19, we have

$$\begin{aligned} & \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}[\|\Sigma(P)^{1/2}(\hat{\beta}(S) - \beta(P))\|^2] \\ \geq \ & \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v^n}[\|\Sigma(P_v)^{1/2}(\hat{\beta}(S) - \beta(P_v))\|^2] \\ = \ & t^2 \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(\|\Sigma(P_v)^{1/2}(\hat{\beta}(S) - \beta(P_v))\| \geq t \right) \\ \asymp \ & t^2 \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(\|\hat{\beta}(S) - \beta(P_v)\| \geq t \right) \\ \gtrsim \ & t^2 \frac{e^{d/2} \cdot \left(\frac{1}{2} e^{-\varepsilon \lceil n\alpha \rceil} - \frac{\delta}{1 - e^{-\varepsilon}} \right)}{1 + e^{d/2} e^{-\varepsilon \lceil n\alpha \rceil}} \ , \end{aligned}$$

where $\beta(P)$ is the least squares solution of the distribution P, $\Sigma(P)$ is the covariance of x from P, and the last inequality follows from the fact that $d \geq 2$. The rest of the proof follows from [25, Proposition 4]. We choose

$$\alpha = \frac{1}{n\varepsilon} \min \left\{ \frac{d}{2} - \varepsilon, \log \left(\frac{1 - e^{-\varepsilon}}{4\delta e^{\varepsilon}} \right) \right\}$$

and $t = \Omega(\alpha^{1-2/k})$ for $\varepsilon \in (0, 10)$ so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}[\|\Sigma(P)(\hat{\beta}(S) - \beta(P))\|^2] \gtrsim \alpha^{2-4/k} .$$

This means that for all $k \geq 4$, there exist some $\kappa, \gamma = O(1)$ such that

$$\inf_{\hat{\beta} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma \succ 0, P \in \mathcal{P}_{\kappa,k,\Sigma,\gamma^2}} \mathbb{E}_{P^n}[\|\Sigma^{1/2}(\hat{\beta}(S) - \beta(P))\|^2] \gtrsim \min \left\{ \left(\frac{d \wedge \log(1 - e^{-\varepsilon}/\delta)}{n\varepsilon} \right)^{2-4/k}, 1 \right\} ,$$

which completes the proof by noting that $\gamma = \Theta(1)$.

□

3.5 Covariance estimation

In a standard covariance estimation, we are given i.i.d. samples $S = \{x_i \in \mathbb{R}^d\}_{i \in [n]}$ drawn from a distribution $P_{\Sigma,\Psi}$ with zero mean, an unknown covariance matrix $0 \prec \Sigma \in \mathbb{R}^{d \times d}$, and an unknown positive semidefinite matrix $\Psi := \mathbb{E}[(x_i \otimes x_i - \Sigma^\flat)(x_i \otimes x_i - \Sigma^\flat)^\top] \in \mathbb{R}^{d^2 \times d^2}$, where $\otimes$ denotes the Kronecker product. We treat the fourth moment matrix Ψ as a linear operator on a subspace $\mathcal{S}_{\text{sym}} \subset \mathbb{R}^{d^2}$, defined as $\mathcal{S}_{\text{sym}} := \{M^\flat \in \mathbb{R}^{d^2} : M \text{ is symmetric}\}$ following the definitions and notations from [66].

Definition 3.5.1. *For any matrix $M \in \mathbb{R}^{d \times d}$, let $M^\flat \in \mathbb{R}^{d^2}$ denote its canonical flattening into a vector in $\mathbb{R}^{d^2}$, and for any vector $v \in \mathbb{R}^{d^2}$, let $v^\sharp$ denote the unique matrix $M \in \mathbb{R}^{d \times d}$ such that $M^\flat = v$.*

This definition of Ψ as an operator on $\mathcal{S}_{\text{sym}}$ is without loss of generality since here we apply Ψ only to flattened symmetric matrices, which significantly lightens the notations, for example, for Gaussian distributions. We consider all $d^2 \times d^2$ matrices in this section to

be linear operators on $\mathcal{S}_{\text{sym}}$, and we restrict our support of the exponential mechanism in RELEASE to be the set of positive definite matrices $\{\hat{\Sigma} \in \mathbb{R}^{d\times d} : \hat{\Sigma} \succ 0\}$.

Lemma 3.5.2 ([66, Theorem 4.12]). *If $P_{\Sigma,\Psi} = \mathcal{N}(0,\Sigma)$, then $\mathbb{E}[x_i \otimes x_i] = \Sigma^\flat$, and, as a matrix in $\mathbb{R}^{d^2\times d^2}$, we have $\Psi_{n(i-1)+j,n(k-1)+\ell} = \Sigma_{i,k}\Sigma_{j,\ell} + \Sigma_{i,\ell}\Sigma_{j,k}$ for all $(i,j,k,\ell) \in [d]^4$; as an operator on $\mathcal{S}_{\text{sym}}$, we can equivalently write it as $\Psi = 2(\Sigma \otimes \Sigma)$.*

Further, we can assume an invertible operator Ψ and define the Mahalanobis distance for $x_i \otimes x_i$, which is $D_\Psi(\hat{\Sigma},\Sigma) = \|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\|$. For Gaussian distributions, for example, we have $D_\Psi(\hat{\Sigma},\Sigma) = (1/\sqrt{2})\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. This is a natural choice of a distance because the total variation distance between two Gaussian distributions is $d_{\text{TV}}(\mathcal{N}(0,\Sigma),\mathcal{N}(0,\Sigma')) = O(\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F)$ (see, for example, [129, Lemma 2.9]). We want a DP estimate of the covariance Σ with a small Mahalanobis distance $D_\Psi(\hat{\Sigma},\Sigma)$. If the sample-generating distribution is not zero-mean, we can either apply a robust mean estimation with a subset of samples to estimate the mean or estimate the covariance using zero mean samples of the form $\{x_i - x_{i+\lceil n/2\rceil}\}_{i\in[n/2]}$.

3.5.1 Step 1: Designing the surrogate $D_S(\hat{\Sigma})$ for the Mahalanobis distance

To sample only positive definite matrices, we restrict the domain of our score function to be $D_\Sigma : \{\hat{\Sigma} \in \mathbb{R}^{d\times d} : \hat{\Sigma} \succ 0\} \to \mathbb{R}_+$ and assume $D_\Sigma(\hat{\Sigma}) = \infty$ for non positive definite $\hat{\Sigma}$:

$$D_S(\hat{\Sigma}) = \max_{V\in\mathbb{R}^{d\times d}: V^\top = V, \|V\|_F = 1} \frac{\langle V, \hat{\Sigma}\rangle - \Sigma_V(\mathcal{M}_{V,\alpha})}{\psi_V(\mathcal{M}_{V,\alpha})}\,, \tag{3.60}$$

where we define the set $\mathcal{M}_{V,\alpha}$ similarly to the definition in Section 3.3.1. We consider a projected dataset $\{\langle V, x_i x_i^\top\rangle\}_{i\in S}$ and partition S into three sets, $\mathcal{B}_{V,\alpha}$, $\mathcal{M}_{V,\alpha}$ and $\mathcal{T}_{V,\alpha}$, where $\mathcal{B}_{V,\alpha}$ corresponds to the subset of $(2/5.5)\alpha n$ data points with smallest values in $\{\langle V, x_i x_i^\top\rangle\}_{i\in S}$, $\mathcal{T}_{V,\alpha}$ is the subset of top $(2/5.5)\alpha n$ data points with the largest values, and $\mathcal{M}_{V,\alpha}$ is the subset of remaining $1 - (4/5.5)\alpha n$ data points. For a fixed symmetric matrix $V \in \mathbb{R}^{d\times d}$ with $\|V\|_F = 1$, we define $\Sigma_V(\mathcal{M}_{V,\alpha}) = \frac{1}{|\mathcal{M}_{V,\alpha}|}\sum_{x_i\in\mathcal{M}_{V,\alpha}} \langle V, x_i x_i^\top\rangle$ and $\psi_V(\mathcal{M}_{V,\alpha})^2 = \frac{1}{|\mathcal{M}_{V,\alpha}|}\sum_{x_i\in\mathcal{M}_{V,\alpha}} \left(\langle V, x_i x_i^\top\rangle - \Sigma_V(\mathcal{M}_{V,\alpha})\right)^2$, which are robust estimates of the

population projected covariance $\Sigma_V = \langle V, \Sigma \rangle$ and projected fourth moment $\psi_V^2 = (V^\flat)^\top \Psi V^\flat$. Next, we show that this score function $D_S(\hat{\Sigma})$ recovers our target error metric $D_\Psi(\hat{\Sigma}, \Sigma) = \|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\|$ when we substitute $\Sigma_V(\mathcal{M}_{V,\alpha})$ and $\psi_V(\mathcal{M}_{V,\alpha})$ with population statistics Σ_V and ψ_V, respectively. This justifies the choice of $D_S(\hat{\Sigma})$, as discussed in Section 3.3.1.

Lemma 3.5.3. *For any $0 \prec \Sigma \in \mathbb{R}^{d\times d}$, $0 \prec \hat{\Sigma}$ and any invertible linear operator $\Psi \in \mathbb{R}^{d^2\times d^2}$ on $\mathcal{S}_{\mathrm{sym}}$, we have*

$$\max_{V\in\mathbb{R}^{d\times d}: V^\top = V, \|V\|_F = 1} \frac{\langle V, \hat{\Sigma}\rangle - \Sigma_V}{\psi_V} = \left\|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\right\| \ , \tag{3.61}$$

where $\Sigma_V = \langle V, \Sigma\rangle$ and $\psi_V^2 = (V^\flat)^\top \Psi V^\flat$.

This follows immediately from Lemma 3.3.1.

3.5.2 Step 2: Utility analysis under resilience

The following resilience property of the dataset is critical in selecting Δ and τ and analyzing utility.

Definition 3.5.4 (Resilience)**.** *For some $\alpha \in (0,1)$, $\rho_1 \in \mathbb{R}_+$, and $\rho_2 \in \mathbb{R}_+$, we say a set of n data points S_{good} is (α, ρ_1, ρ_2)-resilient with respect to (Σ, Ψ) if for any $T \subset S_{\mathrm{good}}$ of size $|T| \geq (1-\alpha)n$, the following holds for all symmetric matrices $V \in \mathbb{R}^{d\times d}$ with $\|V\|_F = 1$:*

$$\left|\frac{1}{|T|}\sum_{x_i\in T} \left\langle V, x_i x_i^\top \right\rangle - \langle V, \Sigma\rangle\right| \ \leq \ \rho_1 \psi_V \ , \ \textit{and} \tag{3.62}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T} \left(\left\langle V, x_i x_i^\top \right\rangle - \langle V, \Sigma\rangle\right)^2 - \psi_V^2\right| \ \leq \ \rho_2 \psi_V \ . \tag{3.63}$$

Note that covariance estimation for $\{x_i\}$ is equivalent to mean estimation for $\{x_i \otimes x_i\}$. We can immediately apply the mean estimation utility guarantee in Theorem 19 to show that $\|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\| = O(\rho_1)$ can be achieved with $n = O(d^2/\varepsilon\alpha)$ samples.

Corollary 3.5.5 (Corollary of Theorem 19)**.** *There exist positive constants c and $C > 0$ such that for any (α, ρ_1, ρ_2)-resilient dataset S with respect to (Σ, Ψ) satisfying $\alpha < c$, $\rho_1 < c$ and*

$\rho_2 < c$, and $\rho_1^2 \le c\alpha$, HPTR *with the distance function in Eq. (3.60), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $\|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\| \le 32\rho_1$ with probability $1-\zeta$ if*

$$n \;\ge\; C\frac{d^2 + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,. \tag{3.64}$$

Under Assumption 3 on α_{corrupt}-corruption and Definition 3.3.3 on corrupt good sets extended to $\{x_i \otimes x_i\}_{i=1}^n$, it follows from Theorem 20 that the same guarantee holds under an adversarial corruption.

Corollary 3.5.6 (Corollary of Theorem 20)**.** *There exist positive constants c and $C > 0$ such that for any $((1/11)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S with respect to (Σ, Ψ) satisfying $\alpha < c$, $\rho_1 < c$ and $\rho_2 < c$, and $\rho_1^2 \le c\alpha$,* HPTR *with the distance function in Eq. (3.60), $\Delta = 110\rho_1/(\alpha n)$, and $\tau = 42\rho_1$ achieves $\|\Psi^{-1/2}(\hat{\Sigma}^\flat - \Sigma^\flat)\| \le 32\rho_1$ with probability $1-\zeta$ if*

$$n \;\ge\; C\frac{d^2 + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,. \tag{3.65}$$

3.5.3 Step 3: Near-optimal guarantees

Covariance estimation has been studied for Gaussian distributions under differential privacy [139, 129, 4] and robust estimation under α-corruption [154, 62, 47, 177, 217]. Note that from Lemma 3.5.2, we know that $\Psi = 2(\Sigma \otimes \Sigma)$ and the Mahalanobis distance simplifies to $D_\Psi(\hat{\Sigma}, \Sigma) = \|\Sigma^{1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F$ for Gaussian distributions.

3.5.3.1 Gaussian distributions

For Gaussian distributions, the second moment resilience in Eq. (3.62) is satisfied with $\rho_1 = O(\alpha\log(1/\alpha))$, and the 4th moment resilience in Eq. (3.63) is satisfied with $\rho_2 = O(\alpha\log^2(1/\alpha))$.

Lemma 3.5.7 (Resilience for Gaussian)**.** *Consider a dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ of n i.i.d. samples from $\mathcal{N}(0, \Sigma)$. If $n = \Omega\left((d^2 + \log(1/\zeta))/(\alpha^2\log(1/\alpha))\right)$ with a large enough constant, then there exists a constant $C > 0$ such that S is $(\alpha, C\alpha\log(1/\alpha), C\alpha\log^2(1/\alpha))$-corrupt good with respect to $(\Sigma, \Psi = 2\Sigma\otimes\Sigma)$ with probability $1-\zeta$.*

Proof. Since x is Gaussian, by Lemma 3.5.2, we have $\Psi = \mathbb{E}[(x\otimes x-\Sigma^\flat)(x\otimes x-\Sigma^\flat)^\top] = 2\Sigma\otimes\Sigma$. We can write $\psi_V^2 = 2\operatorname{Tr}(V^\top\Sigma V\Sigma) = 2\langle V, \Sigma V\Sigma\rangle$.

Lemma 3.5.8 ([154, Lemma B.1] and [73, Fact 4.2])**.** *Let $\delta > 0$ and $\alpha \in (0, 0.5)$. A dataset $S = \{x_1, x_2, \cdots, x_n\}$ consists of n i.i.d. samples from $\mathcal{N}(0, \mathbf{I}_{d\times d})$. If $n = \Omega\left((d^2 + \log(1/\zeta))/(\alpha^2\log(1/\alpha))\right)$ with a large enough constant, then there exists a universal constant $C_1 > 0$ and $C_2 > 0$ such that with probability $1-\zeta$, for any subset $T \subset S$ and $|T| \geq (1-\alpha)n$, we have*

$$\left\|\frac{1}{|T|}\sum_{x_i\in T} x_i\otimes x_i - \mathbf{I}_{d\times d}^\flat\right\| \leq C_1\alpha\log(1/\alpha)\ , \text{ and}$$

$$\left\|\frac{1}{|T|}\sum_{x_i\in T}\left(x_i\otimes x_i - \mathbf{I}_{d\times d}^\flat\right)\left(x_i\otimes x_i - \mathbf{I}_{d\times d}^\flat\right)^\top - 2\mathbf{I}_{d\times d}\otimes\mathbf{I}_{d\times d}\right\| \leq C_2\alpha\log(1/\alpha)^2\ .$$

By Lemma 3.5.8, we know with probability $1-\zeta$ that for any subset $T \subset S$ and $|T| \geq (1-\alpha)n$, we have

$$\left\|\frac{1}{|T|}\sum_{x_i\in T}(\Sigma^{-1/2}x_i)\otimes(\Sigma^{-1/2}x_i) - \mathbf{I}_{d\times d}^\flat\right\| \leq C_1\alpha\log(1/\alpha)\ .$$

This is equivalent to

$$\left|(V^\flat)^\top\frac{1}{|T|}\sum_{x_i\in T}(\Sigma^{-1/2}\otimes\Sigma^{-1/2})(x_i\otimes x_i) - (V^\flat)^\top\mathbf{I}_{d\times d}^\flat\right| \leq C_1\alpha\log(1/\alpha)\ ,$$

for any $\|V\|_F = 1$. This implies that

$$\left|(V^\flat)^\top\frac{1}{|T|}\sum_{x_i\in T}(x_i\otimes x_i) - (V^\flat)^\top(\Sigma\otimes\Sigma)^{1/2}\mathbf{I}_{d\times d}^\flat\right| \leq C_1\alpha\log(1/\alpha)\sqrt{(V^\flat)^\top(\Sigma\otimes\Sigma)V^\flat}\ ,$$

which is also equivalent to, for some constant C,

$$\left|\left\langle V, \frac{1}{|T|}\sum_{x_i\in T}x_ix_i^\top\right\rangle - \langle V,\Sigma\rangle\right| \leq C\alpha\log(1/\alpha)\sqrt{2\langle V,\Sigma V\Sigma\rangle}\ ,$$

which proves the first resilience Eq. (3.62) in Definition 3.5.4.

Similarly, by Lemma 3.5.8, we have

$$\left\|\frac{1}{|T|}\sum_{x_i\in T}\left(\Sigma^{-1/2}x_i\otimes\Sigma^{-1/2}x_i - \mathbf{I}_{d\times d}^\flat\right)\left(\Sigma^{-1/2}x_i\otimes\Sigma^{-1/2}x_i - \mathbf{I}_{d\times d}^\flat\right)^\top - 2\mathbf{I}_{d\times d}\otimes\mathbf{I}_{d\times d}\right\| \leq C_2\alpha\log(1/\alpha)^2\ .$$

This is equivalent, for any $\|V\|_F = 1$, to

$$\Big|\frac{1}{|T|}\sum_{x_i\in T}\big\langle V^\flat, \Sigma^{-1/2}x_i\otimes\Sigma^{-1/2}x_i - \mathbf{I}_{d\times d}^\flat\big\rangle^2 - 2\Big| \;\le\; C_2\alpha\log(1/\alpha)^2\,.$$

This implies

$$\Big|\frac{1}{|T|}\sum_{x_i\in T}\big\langle V^\flat, x_i\otimes x_i - \Sigma^\flat\big\rangle^2 - 2(V^\flat)^\top(\Sigma\otimes\Sigma)V^\flat\Big| \;\le\; C_2\alpha\log(1/\alpha)^2\,\langle V, \Sigma V\Sigma\rangle\;,$$

which is also equivalent, for some constant C, to

$$\Big|\frac{1}{|T|}\sum_{x_i\in T}\big(\langle V, x_ix_i^\top\rangle - \langle V,\Sigma\rangle\big)^2 - 2\,\mathrm{Tr}(V^\top\Sigma V\Sigma)\Big| \;\le\; 2C\alpha\log(1/\alpha)^2\,\langle V, \Sigma V\Sigma\rangle\;,$$

which proves the second resilience Eq. (3.63) in Definition 3.5.4.

□

The second and fourth moment resilience properties of Gaussian distributions in Lemma 3.5.7, together with the utility analysis of HPTR in Corollary. 3.5.6, imply the following utility guarantee.

Corollary 3.5.9. *Under the hypotheses of Lemma 3.5.7, there exists a constant $c > 0$ such that for any $\alpha \in (0, c)$, a dataset of size*

$$n = O\Big(\frac{d^2+\log(1/\zeta)}{\alpha^2\log(1/\alpha)} + \frac{d^2+\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big)\,,$$

a sensitivity of $\Delta = O(\log(1/\alpha)/n)$, and a threshold $\tau = O(\alpha\log(1/\alpha))$ with large enough constants are sufficient for HPTR(S) *with a choice of distance function in Eq. (3.60) to achieve*

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F \;=\; O(\alpha\log(1/\alpha))\,, \tag{3.66}$$

with probability $1-\zeta$. Further, the same guarantee holds even if an α-fraction of the samples is arbitrarily corrupted, as in Assumption 3.

This Mahalanobis distance guarantee (for the Kronecker product, $\{x_i \otimes x_i\}$, of the samples) implies that the predicted Gaussian distribution is close to the sample generating one in total variation distance (see, for example, [129, Lemma 2.9]) $d_{\mathrm{TV}}(\mathcal{N}(0,\hat{\Sigma}),\mathcal{N}(0,\Sigma)) = O(\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\|_F) = O(\alpha\log(1/\alpha))$. This relation also implies that the error bound is near-optimal under α-corruption, matching a lower bound up to a factor of $O(\log(1/\alpha))$. Even if DP is not required and we are given infinite samples, an adversary can move an α fraction of the probability mass to switch a Gaussian distribution into another one at Mahalanobis distance $\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbf{I}_{d\times d}\|_F = \Omega(\alpha)$. Hence, we cannot tell which of the two distributions the (potentially infinite) samples came from.

The sample complexity is near-optimal, matching a lower bound up to a factor of $O(\log(1/\alpha))$ when $\delta = e^{-\Theta(d^2)}$. For a constant ζ, HPTR requires $n = O(d^2/(\alpha^2\log(1/\alpha)) + d^2/(\alpha\varepsilon) + \log(1/\delta)/(\alpha\varepsilon))$. This nearly matches a lower bound (that holds even if there is no corruption) on n to achieve the guarantee of Eq. (3.66) of $n = \Omega(d^2/(\alpha\log(1/\alpha))^2 + \min\{d^2,\log(1/\delta)\}/(\varepsilon\alpha\log(1/\alpha)) + \log(1/\delta)/\varepsilon)$. The first term follows from the classical estimation of the covariance without DP and matches the first term in our upper bound up to a $O(\log(1/\alpha))$ factor. The second term follows from extending the lower bound in [129], constructed for pure differential privacy with $\delta = 0$, and matches the second term in our upper bound up to a $O(\log(1/\alpha))$ factor when $\delta = e^{-\Theta(d^2)}$. The last term, from [139], has a gap of $O(1/\alpha)$ factor compared to the third term in our upper bound, but this term is typically not dominating when δ is large enough, i.e., $\delta = e^{-O(d^2)}$. We note that a slightly tighter upper bound is achieved by the state-of-the-art algorithm in [4] that requires only $O(d^2/(\alpha\log(1/\alpha))^2 + d^2/(\varepsilon\alpha\log(1/\alpha)) + \log(1/\delta)/\varepsilon)$. The state-of-the-art polynomial time algorithm in [133] requires no assumptions on Σ, but the sample complexity is larger: $n = \tilde{O}(d^2/(\alpha\log(1/\alpha))^2 + d^2\mathrm{polylog}(1/\delta)/(\varepsilon\alpha\log(1/\alpha)) + d^{5/2}\mathrm{polylog}(1/\delta)/\varepsilon)$.

If privacy is not an issue (i.e., $\varepsilon = \infty$), HPTR achieves the error in Eq. (3.66) with $n = O(d^2/\alpha^2\log(1/\alpha))$ samples. There are polynomial time estimators that achieve the same guarantee [154, 62]. The gap of $\log(1/\alpha)$ to the lower bound in the error can be tightened using algorithms that are not computationally efficient, as shown in [47, 177].

Remark. When we have a sample size of only $n = O(d/\alpha^2)$, our analysis provides no guarantees. However, for robust covariance estimation under α-corruption, one can still guarantee a bound on a weaker error metric in the spectral norm: $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_{d\times d}\| = O(\alpha \log(1/\alpha))$ [217, Theorem 3.4]. There is no corresponding DP covariance estimator in that small sample regime. A promising direction is to apply the HPTR framework, but it remains challenging to design a score function for this spectral norm distance that depends only on one-dimensional robust statistics.

3.6 Principal component analysis

In Principal Component Analysis (PCA), we are given i.i.d. samples $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ drawn from a zero mean distribution P_Σ with an unknown covariance matrix Σ. We want to find a top eigenvector of Σ, $u \in \arg\max_{\|v\|=1} v^\top \Sigma v$, privately. The performance of our estimate $\hat{u}$ is measured by how much of the covariance is captured in the direction $\hat{u}$ relative to that of u, $D_\Sigma(\hat{u}) = 1 - (\hat{u}^\top \Sigma \hat{u}/u^\top \Sigma u)$, where u is one of the top eigenvectors of Σ. When the mean is not zero, this can be handled similarly to covariance estimation in Section 3.5.

3.6.1 Step 1: Designing the surrogate score function $D_S(\hat{u})$

It is straightforward to design a score function of $D_S : \mathbb{S}^{(d-1)} \to \mathbb{R}_+$, where $\mathbb{S}^{(d-1)}$ is the unit sphere in $\mathbb{R}^d$

$$D_S(\hat{u}) \;=\; 1 - \frac{\hat{u}^\top \Sigma(\mathcal{M}_{\hat{u},\alpha})\hat{u}}{\max_{v\in\mathbb{R}^d:\|v\|=1} v^\top \Sigma(\mathcal{M}_{v,\alpha})v}\,, \tag{3.67}$$

where $\mathcal{M}_{\hat{u},\alpha} \subset S$ is the subset of data points corresponding to the smallest $(1-(2/3.5)\alpha)n$ values in the projected set $S_{\hat{u}} = \{\langle \hat{u}, x_i\rangle^2\}_{x_i\in S}$ and $\Sigma(\mathcal{M}_{\hat{u},\alpha}) = (1/|\mathcal{M}_{\hat{u},\alpha}|)\sum_{x_i\in\mathcal{M}_{\hat{u},\alpha}} x_i x_i^\top$. Note that when we replace $\Sigma(\mathcal{M}_{\hat{u},\alpha})$ with the population covariance matrix Σ, we recover the target error metric of $D_\Sigma(\hat{u}) = 1 - (\hat{u}^\top \Sigma \hat{u}/\max_{\|v\|=1} v^\top \Sigma v)$. For this choice of $D_S(\hat{u})$, the support of the exponential mechanism is already compact, and we do not restrict it any further, say, to be in $B_{\tau,S}$. This simplifies the HPTR algorithm and also the analysis, as

follows. We define

$$\mathsf{UNSAFE}_{\varepsilon} = \Big\{ S' \subset \mathbb{R}^{d\times n} \,|\, \exists S'' \sim S' \text{ and } \exists E \text{ such that } \mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S'')}}(\hat{u}\in E) > e^{\varepsilon}\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S')}}(\hat{u}\in E) \text{ or } \mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S')}}(\hat{u}\in E) > e^{\varepsilon}\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S'')}}(\hat{u}\in E) \Big\} .$$

Note that since the support is the same for all S, we can achieve a stronger pure DP with $\delta = 0$ in the exponential mechanism. However, we still need $\delta > 0$ in the TEST step. HPTR for PCA proceeds as follows.

1. PROPOSE: Propose a target sensitivity bound $\Delta = 80\rho_2/(\alpha n)$.

2. TEST:

   2.1. Compute the safety margin $m = \min_{S'} d_H(S, S')$ such that $S' \in \mathsf{UNSAFE}_{\varepsilon/2}$.

   2.2. If $\hat{m} = m + \mathrm{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta)$, then output $\perp$; otherwise, continue.

3. RELEASE: Output $\hat{u}$ sampled from a distribution with a pdf:

$$r_{(\varepsilon,\Delta,S)}(\hat{u}) = \frac{1}{Z}\exp\left(-\frac{\varepsilon}{4\Delta}D_S(\hat{u})\right) ,$$

   from $\mathbb{S}^{(d-1)} = \{\hat{u}\in\mathbb{R}^d : \|\hat{u}\| = 1\}$ where $Z = \int_{\mathbb{S}^{(d-1)}} \exp\{-(\varepsilon D_S(\hat{u}))/(4\Delta)\}\, d\hat{u}$.

The choice of ρ_2 depends on the hypothesis on the tail of the sample-generating distribution, and α depends on the target accuracy as guided by Theorem 23 (or the fraction of adversarial corruption in the case of the outlier robust PCA setting in Theorem 24). The target privacy guarantee determines (ε, δ).

3.6.2 Step 2: Utility analysis under resilience

The following resilience properties are critical in selecting the sensitivity Δ and in analyzing the utility.

Definition 3.6.1 (Resilience for PCA). *For some $\rho_1 \in \mathbb{R}_+, \rho_2 \in \mathbb{R}_+$, we say a set of n data points $S_{\text{good}} = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ is (α, ρ_1, ρ_2)-resilient with respect to Σ for some positive semidefinite $\Sigma \in \mathbb{R}^{d\times d}$ if for any $T \subset S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$, the following holds for all $v \in \mathbb{R}^d$ with $\|v\| = 1$:*

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle\right| \;\leq\; \rho_1\,\sigma_v \text{ and} \tag{3.68}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle^2 - \sigma_v^2\right| \;\leq\; \rho_2\,\sigma_v^2\,. \tag{3.69}$$

where $\sigma_v^2 = v^\top \Sigma v$.

We refer to Section 3.3.2 for the explanation of how resilience is fundamentally connected to sensitivity. For an example of a Gaussian distribution, the samples are $(\alpha, O(\alpha\sqrt{\log(1/\alpha)}), O(\alpha\log(1/\alpha)))$-resilient (with a large enough n). We show next how resilience implies an error bound for HPTR, which is $O(\alpha\log(1/\alpha))$ for Gaussian distributions.

Theorem 23. *There exist positive constants c and C such that for any (α, ρ_1, ρ_2)-resilient set S with respect to some Σ and satisfying $\alpha < \rho_2 < c$,* HPTR *Section 3.6.1 for PCA with the choices of the distance function in Eq. (3.67) and $\Delta = 80\rho_2/(\alpha n)$ achieves $1 - (\hat{u}^\top\Sigma\hat{u}/\|\Sigma\|) \leq 20\rho_2$ with probability $1-\zeta$ if*

$$n \;\geq\; C\left(\frac{\log(1/(\delta\zeta)) + d\log(1/\rho_2)}{\varepsilon\alpha}\right)\,. \tag{3.70}$$

We discuss the implications of this result in Section 3.6.3 for specific instances of the problem. Under Assumption 3 on α_{corrupt}-corruption of the data and Definition 3.3.3 on the corrupt good sets, we show that HPTR is also robust against corruption.

Theorem 24. *There exist positive constants c and C such that for any $((2/7)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S with respect to some Σ satisfying $\alpha < \rho_2 < c$,* HPTR *in Section 3.6.1 for PCA with the choices of the distance function in Eq. (3.67) and $\Delta = 80\rho_2/(\alpha n)$ achieves $1 - (\hat{u}^\top\Sigma\hat{u}/\|\Sigma\|) \leq 20\rho_2$ with probability $1-\zeta$ if*

$$n \;\geq\; C\left(\frac{\log(1/(\delta\zeta)) + d\log(1/\rho_2)}{\varepsilon\alpha}\right)\,. \tag{3.71}$$

We provide a proof of the robust and DP PCA in Section 3.6.2.2, where Theorem 23 follows immediately by selecting α as a free parameter. As the HPTR Section 3.6.1 for PCA is significantly simpler, we do not apply the general analysis in Theorem 28; instead, we prove the preceding theorem directly. To this end, we first show a bound on sensitivity and next show that the safety test succeeds with high probability in Section 3.6.2.1.

3.6.2.1 Resilience implies bounded local sensitivity

Given the resilience properties of a corrupt good set S, we show that the sensitivity of $D_S(\hat{u})$ is bounded by Δ.

Lemma 3.6.2. *Suppose $\alpha \le c$ for some small enough constant c. For $\Delta = 80\rho_2/(\alpha n)$ and a $((2/7)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good S, if*

$$n \;=\; \Omega\Big(\frac{\log(1/(\delta\zeta))}{\alpha\varepsilon}\Big)\,,$$

with a large enough constant, then for all S' within a Hamming distance $k^* = (2/\varepsilon)\log(4/(\zeta\delta))$ from S, we have

$$\max_{S'' \sim S'} |D_{S''}(\hat{u}) - D_{S'}(\hat{u})| \le \Delta\,, \tag{3.72}$$

for all unit vectors $\hat{u}$ and all neighboring datasets S''.

Proof. The proof is similar to the proof of Lemma 3.3.11. We first assume $(k^*+1)/n \le \alpha/7$, which requires $n = \Omega(\log(1/\delta\zeta))/(\alpha\varepsilon)$ with a large enough constant. This implies that S' is a $((3/7)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set. The rest of this proof uses this assumption. Let $\mathcal{T}_{\hat{u},\alpha}(S') \subset S$ be the subset of data points corresponding to the largest $(2/3.5)\alpha n$ values in the projected set $S'_{\hat{u}} = \{\langle \hat{u}, x_i\rangle^2\}_{x_i \in S'}$. Recall that S_{good} is the original resilient dataset before corruption by an adversary. From Lemma 3.3.4 and the fact that $|S_{\text{good}} \cap \mathcal{T}_{\hat{u},\alpha}(S')| \ge (1/7)\alpha n$, it follows that $(1/|S_{\text{good}} \cap \mathcal{T}_{\hat{u},\alpha}(S')|)\sum_{x_i \in S_{\text{good}} \cap \mathcal{T}_{\hat{u},\alpha}} \langle \hat{u}, x_i\rangle^2 \le (1 + (2\rho_2)/((1/7)\alpha))\sigma^2_{\hat{u}}$, where $\sigma_{\hat{u}} = \sqrt{\hat{u}^\top \Sigma \hat{u}}$. This implies that

$$\min_{x_i \in S_{\text{good}} \cap \mathcal{T}_{\hat{u},\alpha}} \langle \hat{u}, x_i\rangle^2 \;\le\; \Big(1 + \frac{2\rho_2}{(1/7)\alpha}\Big)\sigma^2_{\hat{u}}\,. \tag{3.73}$$

Let $\mathcal{M}_{\hat{u},\alpha}(S')$ be the remaining subset of S', with $(1-(2/3.5)\alpha)n$ smallest values in $\{(\langle \hat{u}, x_i\rangle)^2\}_{i\in[n]}$. $\mathcal{M}_{\hat{u},\alpha}(S')$ and $\mathcal{M}_{\hat{u},\alpha}(S'')$ can differ by at most one data point. Let x' and x'' be the unique pair of data points that are in $\mathcal{M}_{\hat{u},\alpha}(S')$ and $\mathcal{M}_{\hat{u},\alpha}(S'')$, respectively. If there is no such pair, then the two filtered subsets are the same, and the following claims are trivially true.

If $\langle \hat{u}, x''\rangle^2 \le \max_{x_i\in\mathcal{M}_{\hat{u},\alpha}(S')} \langle \hat{u}, x_i\rangle^2 \le \min_{x_i\in S_{\text{good}}\cap\mathcal{T}_{\hat{u},\alpha}(S')} \langle \hat{u}, x_i\rangle^2$, we have $|\langle \hat{u}, x'\rangle^2 - \langle \hat{u}, x''\rangle^2| \le (1+14\rho_2/\alpha)\sigma_{\hat{u}}^2$, where $\sigma_{\hat{u}}^2 = \hat{u}^\top\Sigma\hat{u}$. If $\langle \hat{u}, x''\rangle^2 > \max_{x_i\in\mathcal{M}_{\hat{u},\alpha}(S')} \langle \hat{u}, x_i\rangle^2$, then x'' is at most $\langle \hat{u}, x''\rangle^2 \le \min_{x_i\in S_{\text{good}}\cap\mathcal{T}_{\hat{u},\alpha}(S')}\langle \hat{u}, x_i\rangle^2$, where equality holds if the smallest point in the top subset enters $\mathcal{M}_{\hat{u},\alpha}(S'')$. This also implies $|\langle \hat{u}, x'\rangle^2 - \langle \hat{u}, x''\rangle^2| \le (1+14\rho_2/\alpha)\sigma_{\hat{u}}^2$. Let $\sigma_v'^2 = v^\top\Sigma(\mathcal{M}_{v,\alpha}(S'))v$ and $\sigma_v''^2 = v^\top\Sigma(\mathcal{M}_{v,\alpha}(S''))v$. Then, for any $\|v\|=1$,

$$\begin{aligned}
\left|\sigma_v'^2 - \sigma_v''^2\right| &= \left| v^\top \left(\frac{1}{(1-(2/3.5)\alpha)n} \sum_{x_i\in\mathcal{M}_{v,2\alpha}(S')} x_i x_i^\top - \frac{1}{(1-(2/3.5)\alpha)n} \sum_{x_i\in\mathcal{M}_{v,2\alpha}(S'')} x_i x_i^\top \right) v \right| \\
&\le \frac{2}{n}|\langle v, x'\rangle^2 - \langle v, x''\rangle^2| \le \frac{2}{n}\Big(1+\frac{14\rho_2}{\alpha}\Big) v^\top\Sigma v \,,
\end{aligned}$$

for $\alpha \le c$ small enough. Then, for the local sensitivity, we have

$$\begin{aligned}
|D_{S'}(\hat{u}) - D_{S''}(\hat{u})| &\le \left|\frac{\sigma_{\hat{u}}'^2 - \sigma_{\hat{u}}''^2}{\max_{\|v\|=1}\sigma_v'^2}\right| + \left|\frac{\sigma_{\hat{u}}''^2}{\max_{\|v\|=1}\sigma_v'^2} - \frac{\sigma_{\hat{u}}''^2}{\max_{\|v\|=1}\sigma_v''^2}\right| \\
&\le \frac{2}{n}\Big(1+\frac{14\rho_2}{\alpha}\Big)\frac{\hat{u}^\top\Sigma\hat{u}}{0.9\|\Sigma\|} + \frac{1.1\hat{u}^\top\Sigma\hat{u}}{0.9^2\|\Sigma\|^2}\frac{2}{n}\Big(1+\frac{14\rho_2}{\alpha}\Big)\|\Sigma\| \,,
\end{aligned}$$

where we used the resilience in Eq. (3.69) with a small enough $\rho_2 \le c$ such that $0.9v^\top\Sigma v \le \sigma_v'^2 \le 1.1v^\top\Sigma v$ and $0.9v^\top\Sigma v \le \sigma_v''^2 \le 1.1v^\top\Sigma v$ (which follow from Lemma 3.6.4). When $\rho_2 \le \alpha$, this is bounded by $|D_{S'}\hat{u}) - D_{S''}(\hat{u})| \le 80\rho_2/(\alpha n) = \Delta$.

□

Since the support is the same for all exponential mechanisms regardless of the dataset, the sensitivity bound immediately implies safety. The following lemma shows that we have a sufficient safety margin to succeed with probability of at least $1-\zeta$ since $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$ and the threshold is $(2/\varepsilon)\log(2/\delta)$.

Lemma 3.6.3. *Under the hypothesis of Lemma 3.6.2, for any S' at Hamming distance at most k^* from S, we have $S' \in \mathsf{SAFE}_{\varepsilon/2}$.*

3.6.2.2 Proof of Theorem 24

This proof is similar to the proof of a universal utility analysis in Theorem 28. First, we show that we pass the safety test with high probability. By Lemma 3.6.3, we know $m > k^* = 2/\varepsilon \log(4/(\zeta\delta))$. Then, we have

$$\mathbb{P}\left(\text{output } \perp\right) = \mathbb{P}\left(m + \text{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta)\right) \;\leq\; \frac{\zeta}{2}\,.$$

Next, we assume the dataset passed the safety test and show that $\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}(\hat{u}^\top\Sigma\hat{u} \geq (1-4\rho_2)\|\Sigma\|) \geq 1-\zeta/2$.

Lemma 3.6.4. *For an $((2/7)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S with respect to Σ, then $|\hat{u}^\top\Sigma\hat{u} - \hat{u}^\top\Sigma(\mathcal{M}_{\hat{u},\alpha})\hat{u}| \;\leq\; 4\rho_2\hat{u}^\top\Sigma\hat{u}$.*

Proof. We have

$$\begin{aligned}
|\hat{u}^\top\Sigma\hat{u} - \hat{u}^\top\Sigma(\mathcal{M}_{\hat{u},\alpha})\hat{u}| \;&=\; \frac{|\sum_{i\in\mathcal{M}_{\hat{u},\alpha}}(\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2)|}{(1-(2/3.5)\alpha)n} \\
&\leq \frac{|\sum_{i\in\mathcal{M}_{\hat{u},\alpha}\cap S_{\text{good}}}(\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2)|}{(1-(2/3.5)\alpha)n} + \frac{|\sum_{i\in\mathcal{M}_{\hat{u},\alpha}\cap S_{\text{good}}}(\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2)|}{(1-(2/3.5)\alpha)n} \qquad (3.74)
\end{aligned}$$

For $i \in \mathcal{M}_{\hat{u},\alpha} \cap S_{\text{bad}}$, by Lemma 3.3.4, we have

$$\begin{aligned}
|\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2| \;&\leq\; \max\left\{\frac{\sum_{i\in\mathcal{T}_{\hat{u},\alpha}\cap S_{\text{good}}}(\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2)}{|\mathcal{T}_{\hat{u},\alpha}\cap S_{\text{good}}|}, \sigma_{\hat{u}}^2\right\} \\
&\leq\; \frac{2\rho_2\sigma_{\hat{u}}^2}{(1/3.5)\alpha}\,, \qquad (3.75)
\end{aligned}$$

where in the last inequality, we applied our assumption that $\rho_2 \geq \alpha$.

By the resilience property in Eq. (3.69) on $\mathcal{M}_{\hat{u},\alpha} \cap S_{\text{good}}$, we also have

$$\frac{|\sum_{i\in\mathcal{M}_{\hat{u},\alpha}\cap S_{\text{good}}}(\langle\hat{u},x_i\rangle^2 - \sigma_{\hat{u}}^2)|}{|\mathcal{M}_{\hat{u},\alpha}\cap S_{\text{good}}|} \leq \rho_2\sigma_{\hat{u}}^2\,. \qquad (3.76)$$

Plugging Eq. (3.75) and (3.76) into (3.74), we have

$$|\hat{u}^\top\Sigma\hat{u} - \hat{u}^\top\Sigma(\mathcal{M}_{\hat{u},\alpha})\hat{u}| \leq \frac{2\rho_2\sigma_{\hat{u}}^2 + (1-(2/3.5)\alpha)\rho_2\sigma_{\hat{u}}^2}{1-(2/3.5)\alpha} \leq 4\rho_2\sigma_{\hat{u}}^2\,,$$

for $\alpha \leq c$ small enough. □

This implies that $|D_\Sigma(\hat{u}) - D_S(\hat{u})| \le 4\rho_2$ for an $((2/7)\alpha, \alpha, \rho_1, \rho_2)$-corrupt good set S.

Let $\mu(\cdot)$ denote the uniform measure on the unit sphere. By the fact that for any $0 < r < 2$, a cap of radius r on the $(d-1)$-dimensional unit sphere $\mathbb{S}^{(d-1)}$ has a measure of at least $(1/2)(r/2)^{d-1}$ from, for example, [136, Fact 3.1], we have for some constant $c_2 > 0$ and $\rho_2 \le 1/8$,

$$\mu(\{v \in \mathbb{R}^d : v^\top \Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\| = 1\}) \ge \left(\cos^{-1}(1-4\rho_2)/2\right)^{d-1} \ge e^{-c_2 d \log(1/\rho_2)} \quad (3.77)$$

By Lemma 3.6.4, the choice of $\Delta = 80\rho_2/(\alpha n)$, we have

$$\begin{aligned}
& \mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\|\Sigma\| - \hat{u}^\top \Sigma \hat{u} \le 4\rho_2\|\Sigma\|\right) \\
= & \int_{\{v\in\mathbb{R}^d : v^\top\Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\|=1\}} r_{(\varepsilon,\Delta,S)}(\hat{u})\, d\hat{\mu} \\
\ge & \mathrm{Vol}(\{v \in \mathbb{R}^d : v^\top \Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\| = 1\}) \min_{\hat{\mu}\in\{v\in\mathbb{R}^d : v^\top\Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\|=1\}} r_{(\varepsilon,\Delta,S)}(\hat{u}) \\
\ge & \mathrm{Vol}(\mathbb{S}^{(d-1)})\, \mu(\{v \in \mathbb{R}^d : v^\top \Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\| = 1\}) \min_{\hat{u}\in\{v\in\mathbb{R}^d : v^\top\Sigma v \ge (1-4\rho_2)\|\Sigma\|, \|v\|=1\}} r_{(\varepsilon,\Delta,S)}(\hat{u}) \\
\ge & \mathrm{Vol}(\mathbb{S}^{(d-1)})\, e^{-c_2 d\log(1/\rho_2)} \frac{1}{Z} \exp\left\{ -\frac{\varepsilon}{4\Delta} \max_{\|\hat{u}\|=1, 4\rho_2 \ge 1 - \frac{\hat{u}^\top\Sigma\hat{u}}{\|\Sigma\|}} 1 - \frac{\hat{u}^\top \Sigma(\mathcal{M}_{\hat{u},\alpha})\hat{u}}{\|\Sigma\|} \right\} \\
\ge & \mathrm{Vol}(\mathbb{S}^{(d-1)})\, e^{-c_2 d\log(1/\rho_2)} \frac{1}{Z} \exp\left\{ -\frac{\alpha\varepsilon n}{40} \right\},
\end{aligned}$$

and similarly,

$$\begin{aligned}
\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\|\Sigma\| - \hat{u}^\top\Sigma\hat{u} \ge 20\rho_2\|\Sigma\|\right) &\le \mathrm{Vol}(\mathbb{S}^{(d-1)}) \max_{\hat{u}\in\{v\in\mathbb{R}^d : v^\top\Sigma v \le (1-20\rho_2)\|\Sigma\|, \|v\|=1\}} r_{(\varepsilon,\Delta,S)}(\hat{u}) \\
&\le \mathrm{Vol}(\mathbb{S}^{(d-1)}) \frac{1}{Z} e^{-\varepsilon\alpha n(20\rho_2 - 4\rho_2)\|\Sigma\|/(320\rho_2\|\Sigma\|)} \\
&\le \mathrm{Vol}(\mathbb{S}^{(d-1)}) \frac{1}{Z} \exp\left\{ -\frac{\alpha\varepsilon n}{20} \right\}
\end{aligned}$$

This implies that

$$\log\left(\frac{\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\lambda_1 - \hat{u}^\top\Sigma\hat{u} \le 4\rho_2\|\Sigma\|\right)}{\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\lambda_1 - \hat{u}^\top\Sigma\hat{u} \ge 20\rho_2\|\Sigma\|\right)} \right) \ge \frac{\varepsilon\alpha n}{40} - c_2 d\log(1/\rho_2)\,.$$

If we set $n = \Omega\left(\frac{\log(1/\zeta) + d\log(1/\rho_2)}{\varepsilon\alpha}\right)$, we get

$$\frac{\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\lambda_1 - \hat{u}^\top\Sigma\hat{u} \le 4\rho_2\lambda_1\right)}{\mathbb{P}_{\hat{u}\sim r_{(\varepsilon,\Delta,S)}}\left(\lambda_1 - \hat{u}^\top\Sigma\hat{u} \ge 20\rho_2\lambda_1\right)} \ge \frac{2}{\zeta}\,,$$

which completes the proof.

3.6.3 Step 3: Achievability guarantees

We provide utility guarantees for a private PCA for sub-Gaussian and hypercontractive distributions.

3.6.3.1 Sub-Gaussian distributions

Using the resilience of sub-Gaussian distributions with respect to $(\mu = 0, \Sigma)$ in Lemma 3.3.12, which is the same as the resilience properties we need for the PCA in Definition 3.6.1, Theorem 24 implies the following corollary.

Corollary 3.6.5. *Under the hypothesis of Lemma 3.3.12 with $\mu = 0$ and any PSD matrix $\Sigma \in \mathbb{R}^{d\times d}$, there exist universal constants c and $C > 0$ such that for any $\alpha \in (0, c)$, a dataset of size*

$$n \;=\; O\left(\frac{d + \log(1/\zeta)}{(\alpha \log(1/\alpha))^2} + \frac{\log(1/(\delta\zeta)) + d\log(1/(\alpha\log(1/\alpha)))}{\varepsilon\alpha}\right) ,$$

and sensitivity of $\Delta = O(\log(1/\alpha)/n)$ with large enough constants are sufficient for HPTR(S) *in Section 3.6.1 for a PCA with the choices of the distance function in Eq. (3.67) to achieve*

$$1 - \frac{\hat{u}^\top \Sigma \hat{u}}{\|\Sigma\|} \;\le\; C\alpha\log(1/\alpha) \;, \tag{3.78}$$

with probability $1 - \zeta$. Further, the same guarantee holds even if an α-fraction of the samples is arbitrarily corrupted, as in Assumption 3.

The error bound is near-optimal under α-corruption, matching a lower bound up to a factor of $O(\log(1/\alpha))$. HPTR is the first estimator that guarantees (ε, δ)-DP and also achieves the robust error rate of $1 - \hat{u}^\top\Sigma\hat{u}/\|\Sigma\| = O(\alpha\log(1/\alpha))$, nearly matching the information-theoretic lower bound of $1 - \hat{u}^\top\Sigma\hat{u}/\|\Sigma\| = \Omega(\alpha)$. This lower bound, which can be easily constructed using $\mathcal{N}(0, \mathbf{I} + \alpha e_1 e_1^\top)$ and $\mathcal{N}(0, \mathbf{I} + \alpha e_2 e_2^\top)$, holds for any estimator that is not necessarily private and regardless of how many samples are available. If privacy is not required, a near-optimal robust error rate can be achieved by outlier-robust PCA approaches in [145, 121].

The sample complexity is near-optimal, matching a lower bound up to a factor of $O(\log(1/\alpha))$ when $\delta = e^{-\Theta(d)}$. Even for a DP PCA without corrupted samples, HPTR is the first estimator for sub-Gaussian distributions to nearly match the information-theoretic lower bound of $n = \Omega(d/(\alpha\log(1/\alpha))^2 + \min\{d, \log((1-e^{-\varepsilon})/\delta)\}/(\varepsilon\alpha\log(1/\alpha)))$ to achieve the error in Eq. (3.78). The first term is unavoidable since even without DP and robustness, when the data comes from a Gaussian distribution, estimating the principal component up to error $\alpha\log(1/\alpha)$ requires $\Omega(d/(\alpha\log(1/\alpha))^2)$ samples (Proposition 3.6.7). The second term in the lower bound follows from Proposition 3.6.6, which matches the second term in the upper bound up to a factor of $O(\log(1/\alpha))$ when $\delta = e^{-\Theta(d)}$ and $\varepsilon > 0$. Existing DP PCA approaches from [46, 136, 80] are designed for arbitrary samples not necessarily drawn i.i.d., and hence they require a larger sample size of $n = \tilde{O}(d/\alpha^2 + d^{1.5}\sqrt{\log(1/\delta)}/(\alpha\varepsilon))$ i.i.d. samples from a Gaussian distribution to achieve the guarantee in Eq. (3.78), where $\tilde{O}$ hides polylogarithmic terms in $1/\alpha$ and $1/\zeta$.

Remark. Rank-k PCA under α-corruption from a Gaussian dataset is of great practical interest. An outlier-robust PCA algorithm in [145, Appendix D] outputs an orthonormal matrix $\hat{U} \in \mathbb{R}^{d\times k}$ achieving

$$\mathrm{Tr}(U_k^\top \Sigma U_k) - \mathrm{Tr}(\hat{U}^\top \Sigma \hat{U}) \;=\; O\big(\alpha \mathrm{Tr}(U_k^\top \Sigma U_k) + \nu k^{1/2}\alpha\log(1/\alpha)\big)\,,$$

where $U_k \in \arg\max_{U^\top U = \mathbf{I}_{k\times k}} U^\top \Sigma U$ and $\nu^2 = \max_{V\in\mathbb{R}^{d\times d}, \|V\|_F=1, V=V^\top, \mathrm{rank}(V)\le k}\langle V, \Sigma V \Sigma\rangle$. It is a promising direction to design a DP rank-k PCA algorithm by applying the HPTR framework that can achieve a similar error rate. It is not immediately clear how to design an appropriate score function for general rank k, and a simple technique of peeling off rank-one components one-by-one (using the rank-one PCA with HPTR) will not achieve the target error bound.

Proposition 3.6.6 (Lower bound for private sub-Gaussian PCA)**.** *Let $\mathcal{P}_\Sigma$ be the set of zero-mean sub-Gaussian distributions with covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $\mathcal{M}_{\varepsilon,\delta}$ be a class of (ε,δ)-DP, d-dimensional estimators of the top principal component of Σ using n i.i.d. samples*

from $P \in \mathcal{P}_\Sigma$. Then, for $\varepsilon \in (0, 10)$, there exists a universal constant $c > 0$ such that

$$\inf_{\hat{u} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma \succ 0, P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n} \left[1 - \frac{\hat{u}(S)^\top \Sigma \hat{u}(S)}{\|\Sigma\|} \right] \geq c \cdot \min \left\{ \frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon}, 1 \right\} .$$

Proof. We adopt the same proof strategy as the proof of Proposition 3.3.18 for mean estimation. By [3, Lemma 6], there exists a finite index set $\mathcal{V} \subset \mathbb{R}^d$ with cardinality $|\mathcal{V}| = 2^{\Omega(d)}$, $\|v\| = 1$ for all $v \in \mathcal{V}$ and $\|v - v'\| \geq 1/2$ for all $v \neq v' \in \mathcal{V}$. For each $v \in \mathcal{V}$, we define $\Sigma_v := \mathbf{I}_{d\times d} + \alpha v v^\top$ and $P_v := \mathcal{N}(0, \Sigma_v)$ for some $\alpha \in (0, 1/2)$. It is straightforward to see that $\mathbf{I}_{d\times d} \preceq \Sigma_v \preceq 3\mathbf{I}_{d\times d}/2$ and the top eigenvector of Σ_v is v. For $v \neq v' \in \mathcal{V}$, we know $\|\Sigma_{v'}^{-1/2} \Sigma_v \Sigma_{v'}^{-1/2} - \mathbf{I}_{d\times d}\|_F = O(\alpha)$. By [129, Lemma 2.9], this implies $d_{\mathrm{TV}}(\mathcal{N}(0, \Sigma_v), \mathcal{N}(0, \Sigma'_v)) = O(\alpha)$.

Since $\|v - v'\| \geq 1/2$, we have

$$D_{\Sigma_{v'}}(v) = 1 - \frac{v^\top \Sigma'_v v}{\|\Sigma_{v'}\|} = 1 - \frac{1 + \alpha \langle v, v' \rangle^2}{1 + \alpha} \geq \frac{\alpha}{8(1 + \alpha)} > \frac{\alpha}{12} .$$

The principal component estimation problem can be reduced to a testing problem with this packing $\mathcal{V}$. For the (ε, δ)-DP estimator $\hat{u}$, using Lemma 3.3.19, let $t = \frac{\alpha}{12}$, we have

$$\begin{aligned} \sup_{P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n}[D_\Sigma(\hat{u})] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{S \sim P_v^n}[D_{\Sigma_v}(\hat{u})] \\ &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(D_{\Sigma_v}(\hat{u}) \geq t \right) \\ &\gtrsim t \frac{e^{d/2} \cdot \left(\frac{1}{2} e^{-\varepsilon \lceil n\alpha \rceil} - \frac{\delta}{1 - e^{-\varepsilon}} \right)}{1 + e^{d/2} e^{-\varepsilon \lceil n\alpha \rceil}} , \end{aligned}$$

where the last inequality follows from the fact that $d \geq 2$. The rest of the proof follows from [25, Proposition 4]. We choose

$$\alpha = \frac{1}{n\varepsilon} \min \left\{ \frac{d}{2} - \varepsilon, \log \left(\frac{1 - e^{-\varepsilon}}{4\delta e^{\varepsilon}} \right) \right\}$$

so that

$$\sup_{P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n}[D_{\Sigma_v}(\hat{u})] \gtrsim \alpha .$$

This implies, for $t = \alpha/12$ and $\varepsilon \in (0, 10)$, that

$$\inf_{\hat{u} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma \succ 0, P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n}[D_\Sigma(\hat{u})] \gtrsim \min \left\{ \frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon}, 1 \right\} ,$$

which completes the proof. □

It is well known that even for Gaussian distributions, learning the principal component up to error α requires $\Omega(d/\alpha^2)$. We provide a lower bound proof here for completeness.

Proposition 3.6.7 (Sample Complexity Lower bound for PCA)**.** *Let $\mathcal{P}_\Sigma$ be the set of zero-mean Gaussian distributions with covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $\mathcal{M}_d$ be the class of estimators of the d-dimensional top principal component of Σ using n i.i.d. samples from $P \in \mathcal{P}_\Sigma$. There exists a universal constant $c > 0$ such that*

$$\inf_{\hat{u}\in\mathcal{M}_d} \sup_{\Sigma\succ 0, P\in\mathcal{P}_\Sigma} \mathbb{E}_{S\sim P^n}\left[1 - \frac{\hat{u}(S)^\top \Sigma \hat{u}(S)}{\|\Sigma\|}\right] \;\geq\; c\cdot \min\left\{\sqrt{\frac{d}{n}}, 1\right\}\,.$$

Proof. The following proposition helps to prove a minimax lower bound on estimating $\|\Sigma\|$. We first define some notations.

Definition 3.6.8 (Definition 3.1 in [68])**.** *For a distribution A on the real line with probability density function $A(x)$ and a unit vector $v \in \mathbb{R}^d$, consider the distribution over $\mathbb{R}^n$ with probability density function $P_v(x) = A(v^\top x)\exp(-\|x - (v^\top x)v\|_2^2/2)\cdot(2\pi)^{-(d-1)/2}$.*

Proposition 3.6.9 (Proposition 7.1 in [68])**.** *Let A be a distribution on $\mathbb{R}$ such that A has a mean 0 and $\chi^2(A, N(0,1))$ is finite. Then, there is no algorithm for any d, given $n < d/(8\chi^2(A, N(0,1)))$ samples from a distribution D over $\mathbb{R}^d$ which is either $N(0, I)$ or P_v for some unit vector $v \in \mathbb{R}^d$, that correctly distinguishes between the two cases with probability at least $2/3$.*

To apply Proposition 3.6.9, let A be Gaussian distribution $\mathcal{N}(0, 1+\alpha)$. Through simple calculation, it can be shown that $\chi^2(\mathcal{N}(0,1), \mathcal{N}(0,1+\alpha)) = \frac{1}{\sqrt{1-\alpha^2}} - 1 \leq \alpha^2$ whenever $\alpha^2 \leq 1/2$. Then, for the first case in Proposition 3.6.9, $\|\Sigma\| = \|I\| = 1$, the second case has $\|\Sigma\| = 1+\alpha$, and Proposition 3.6.9 implies that there exists absolute constant c such that

$$\inf_{\hat{\lambda}} \sup_{\Sigma\succ 0, P\in\mathcal{P}_\Sigma} \mathbb{E}_{S\sim P^n}\left[1 - \frac{\hat{\lambda}(S)}{\|\Sigma\|}\right] \;\geq\; c\cdot \min\left\{\sqrt{\frac{d}{n}}, 1\right\}\,.$$

Since we can turn a principal component estimator $u(S)$ into an estimator of $\|\Sigma\|$ through n additional fresh samples to estimate $u(S)^\top \Sigma u(S)$ up to a minor multiplicative error $O(1/\sqrt{n})$.

This implies there exists a universal constant $c > 0$ such that

$$\inf_{\hat{u}\in\mathcal{M}_d} \sup_{\Sigma\succ 0, P\in\mathcal{P}_\Sigma} \mathbb{E}_{S\sim P^n}\left[1 - \frac{\hat{u}(S)^\top \Sigma \hat{u}(S)}{\|\Sigma\|}\right] \;\geq\; c\cdot\min\left\{\sqrt{\frac{d}{n}}, 1\right\} .$$

□

3.6.3.2 Hypercontractive distributions

In this section, we apply our results on hypercontractive distributions in Definition 3.3.14. Using the resilience of hypercontractive distributions with respect to $(\mu = 0, \Sigma)$ in Lemma 3.3.15, which is the same as the resilience properties we need for PCA in Definition 3.6.1, Theorem 24 implies the following corollary.

Corollary 3.6.10. *Under the hypothesis of Lemma 3.3.15 with $k \geq 3$, $\mu = 0$ and any PSD matrix $\Sigma \in \mathbb{R}^{d\times d}$, there exist universal constants c and $C > 0$ such that for any $\alpha \in (0, c)$, a dataset of size*

$$n \;=\; O\left(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k} d\log d}{\zeta^{2-4/k}\kappa^2} + \frac{\kappa^2 d\log d}{\alpha^{2/k}} + \frac{\log(1/(\delta\zeta)) + d\log(1/\alpha^{1-2/k})}{\varepsilon\alpha}\right) ,$$

and sensitivity of $\Delta = O(\alpha^{1-2/k}/n)$ with large enough constants are sufficient for HPTR(S) *in Section 3.6.1 for PCA with the choices of the distance function in Eq. (3.67) to achieve*

$$1 - \frac{\hat{u}^\top\Sigma\hat{u}}{\|\Sigma\|} \;\leq\; C\alpha^{1-2/k} , \tag{3.79}$$

with probability $1-\zeta$. Further, the same guarantee holds even if an α-fraction of the samples is arbitrarily corrupted, as in Assumption 3.

The error bound is optimal under α-corruption up to a constant factor. HPTR is the first estimator that guarantees (ε, δ)-DP and also achieves the robust error rate of $1-\hat{u}^\top\Sigma\hat{u}/\|\Sigma\| = O(\alpha^{1-2/k})$, matching the information-theoretic lower bound of $1 - \hat{u}^\top\Sigma\hat{u}/\|\Sigma\| = \Omega(\alpha^{1-2/k})$. This lower bound can be easily constructed using Eq. (3.59), where two hypercontractive distributions are at total variation distance $O(\alpha)$ and the top principal component of one distribution achieves an error lower bounded by $1 - \hat{u}^\top\Sigma\hat{u}/\|\Sigma\| = \Omega(\alpha^{1-2/k})$. Even if privacy

is not required, there is no outlier-robust PCA estimator matching this optimal error rate for a general k.

The sample complexity is $n = \tilde{O}(d/\alpha^{2(1-1/k)} + (d+\log(1/\delta))/(\varepsilon\alpha))$ for a constant ζ, k, and κ, where $\tilde{O}$ hides logarithmic factors in $1/\alpha$ and d. Even for DP PCA without corrupted samples, HPTR is the first estimator for hypercontractive distributions to guarantee differential privacy. The information-theoretic lower bound is $n = \Omega(d/\alpha^{2(1-2/k)} + \min\{d, \log((1-e^{-\varepsilon})/\delta)\}/(\alpha\varepsilon))$ to achieve the error in Eq. (3.79). The first term is unavoidable, even without DP and robustness, when the data comes from a Gaussian distribution because estimating the principal component up to error $\alpha^{1-2/k}$ requires $\Omega(d/\alpha^{2(1-2/k)})$ samples (Proposition 3.6.7). There is a gap of factor $O(\alpha^{-2/k})$ compared to the first term in our upper bound. Since the sample complexity lower bound in Proposition 3.6.7 is constructed using Gaussian distributions, it might be possible to tighten it further using hypercontractive distributions. The second term in the lower bound follows from Proposition 3.6.11, which matches the last term in the upper bound up to a factor of $O(\log(1/\alpha))$ when $\delta = e^{-\Theta(d)}$ and $\varepsilon > 0$. To the best of our knowledge, HPTR is the first algorithm for PCA that guarantees (ε, δ)-DP under hypercontractive distributions.

Proposition 3.6.11 (Lower bound for hypercontractive private PCA)**.** *Let $\mathcal{P}_\Sigma$ be the set of zero-mean hypercontractive distributions with covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $\mathcal{M}_{\varepsilon,\delta}$ be a class of (ε, δ)-DP estimators using n i.i.d. samples from $P \in \mathcal{P}_\Sigma$. Then, for $\varepsilon \in (0, 10)$, there exists a constant c such that*

$$\inf_{\hat{u}\in\mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma\succ 0, P\in\mathcal{P}_\Sigma} \mathbb{E}_{S\sim P^n}\left[1 - \frac{\hat{u}^\top\Sigma\hat{u}}{\|\Sigma\|}\right] \geq c\min\left\{\left(\frac{d \wedge \log((1-e^{-\varepsilon})/\delta)}{n\varepsilon}\right)^{1-2/k}, 1\right\}. \quad (3.80)$$

Proof. We use the same construction as used in the distribution of x in the proof of Proposition 3.4.21. By [3, Lemma 6], there exists a finite index set $\mathcal{V} \subset \mathbb{R}^d$ with cardinality $|\mathcal{V}| = 2^{\Omega(d)}$, $\|v\| = 1$ for all $v \in \mathcal{V}$ and $\|v - v'\| \geq 1/2$ for all $v \neq v' \in \mathcal{V}$. For each $v \in \mathcal{V}$ and $\alpha \in (0, 1/2)$, we construct the density function of distribution P_v as defined in Eq. (3.59). Let Σ_v denote the covariance matrix of P_v. The proof of Proposition 3.4.21 shows that $\Sigma_v = (1-\alpha)\mathbf{I}_{d\times d} + \alpha^{1-2/k}vv^\top$, $d_{\mathrm{TV}}(P_v, P_v') = \alpha$ and that P_v is $(O(1), k)$-hypercontractive.

Since $\|v - v'\| \geq 1/2$, we know that $\langle v, v' \rangle \leq 7/8$, and we have

$$D_{\Sigma_{v'}}(v) = 1 - \frac{v^\top \Sigma'_v v}{\|\Sigma_{v'}\|} = 1 - \frac{1 - \alpha + \alpha^{1-2/k} \langle v, v' \rangle^2}{1 - \alpha + \alpha^{1-2/k}} \geq \frac{\alpha^{1-2/k}}{8(1 - \alpha + \alpha^{1-2/k})} > \frac{\alpha^{1-2/k}}{12} ,$$

for $\alpha < c$ small enough.

Next, we apply the reduction of estimation to testing with this packing $\mathcal{V}$. For a (ε, δ)-DP estimator $\hat{u}$, using Lemma 3.3.19, let $t = \frac{\alpha^{1-2/k}}{12}$. Then, we have

$$\begin{aligned}
\sup_{P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n}[D_\Sigma(\hat{u})] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{S \sim P_v^n}[D_{\Sigma_v}(\hat{u})] \\
&= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left(D_{\Sigma_v}(\hat{u}) \geq t \right) \\
&\gtrsim t \frac{e^{d/2} \cdot \left(\frac{1}{2} e^{-\varepsilon \lceil n\alpha \rceil} - \frac{\delta}{1 - e^{-\varepsilon}} \right)}{1 + e^{d/2} e^{-\varepsilon \lceil n\alpha \rceil}} ,
\end{aligned}$$

where the last inequality follows from the fact that $d \geq 2$.

The rest of the proof follows from [25, Proposition 4]. We choose

$$\alpha = \frac{1}{n\varepsilon} \min \left\{ \frac{d}{2} - \varepsilon, \log \left(\frac{1 - e^{-\varepsilon}}{4\delta e^\varepsilon} \right) \right\}$$

so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}[D_{\Sigma_v}(\hat{u})] \gtrsim \alpha^{1-2/k} .$$

This means, for $t = (1/12)\alpha^{1-2/k}$ and $\varepsilon \in (0, 10)$, that

$$\inf_{\hat{u} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}[D_\Sigma(\hat{u})] \gtrsim \min \left\{ \left(\frac{d \wedge \log((1 - e^{-\varepsilon})/\delta)}{n\varepsilon} \right)^{1-2/k}, 1 \right\} ,$$

which completes the proof. □

3.7 Discussion

We provided a universal framework for characterizing the statistical efficiency of statistical estimation problems with differential privacy guarantees. Our framework, High-dimensional Propose-Test-Release (HPTR), is computationally inefficient and builds upon three key

components: the exponential mechanism, robust statistics, and the Propose-Test-Release mechanism. Our key insight is that designing an exponential mechanism that accesses the data via only one-dimensional robust statistics can dramatically reduce the resulting local sensitivity. Using resilience, a central concept in robust statistics, we can provide tight local sensitivity bounds. These tight bounds readily translate into near-optimal utility guarantees in several statistical estimation problems of interest: mean estimation, linear regression, covariance estimation, and principal component analysis. Although our framework is written as a conceptual algorithm without a specific implementation, it is possible to implement it with exponential computational complexity following the guidelines of [34], where a similar exponential mechanism with PTR was proposed and an implementation was explicitly provided.

To protect against membership inference attacks, significant progress has been made in training DP models that are practical [1, 213, 13]. To protect against data poisoning attacks, a recent work utilizes robust statistics with great success [104]. In practice, however, we need to protect against both types of attacks to facilitate learning and analysis from shared data. Currently, there is an algorithmic deficiency in this space. Efficient algorithms achieving both DP and robustness against adversarial corruption are known only for mean estimation [160]. We make a valuable contribution to the design of such algorithms for a broad class of problems, including covariance estimation, principal component analysis, and linear regression.

Further, these computationally efficient algorithms typically require more samples. For sub-Gaussian mean estimation with known covariance Σ, an efficient approach of [160] requires $\tilde{O}(d/\alpha^2 + d^{3/2}/(\varepsilon\alpha))$ samples under α-corruption and (ε, δ)-DP to achieve an error of $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = \tilde{O}(\alpha)$. HPTR requires only $O(d/\alpha^2 + d/(\varepsilon\alpha))$ samples. A significant open question is whether this $d^{1/2}$ gap is fundamental and cannot be improved.

Chapter 4

DIFFERENTIALLY PRIVATE PCA

4.1 Introduction

Principal Component Analysis (PCA) is a fundamental statistical tool with multiple applications including dimensionality reduction, data visualization, and noise reduction. Naturally, it is a key part of most standard data analysis and ML pipelines. However, when applied to data collected from numerous individuals, such as the U.S. Census data, outcome of PCA might reveal highly sensitive personal information. We investigate the design of privacy preserving PCA algorithms and the involved privacy/utility tradeoffs, for computing the first principal component, that should serve as the building block of more general rank-k PCA.

Differential privacy (DP) is a widely accepted mathematical notion of privacy introduced in [78], which is a standard in releasing the U.S. Census data [2] and also deployed in commercial systems [189, 82, 84]. A query to a database is said to be (ε, δ)-differentialy private if a strong adversary who knows all other entries but one cannot infer that one entry from the query output, with high confidence. The parameters ε and δ restricts the confidence as measured by the Type-I and II errors [128]. Smaller values of $\varepsilon \in [0, \infty)$ and $\delta \in [0, 1]$ imply stronger privacy and plausible deniability for the participants.

For non-private PCA with n i.i.d. samples in d dimensions, the popular Oja's algorithm (provided in Algorithm 10) achieves the optimal error of $\sin(\hat{v}, v_1) = \tilde{\Theta}(\sqrt{d/n})$, where the error is measured by the sine function of the angle between the estimate, $\hat{v}$, and the principal component, v_1, [119]. For differentially private PCA, there is a natural fundamental question: *what is the extra cost we pay in the error rate for ensuring* (ε, δ)*-DP?*

We introduce a novel approach we call DP-PCA (Algorithm 12) and show that it achieves an error bounded by $\sin(\hat{v}, v) = \tilde{O}(\sqrt{d/n} + d/(\varepsilon n))$ for *sub-Gaussian-like* data defined in

Assumption 5, which is a broad class of light-tailed distributions that includes Gaussian data as a special case. The second term characterizes the cost of privacy and this is tight; we prove a nearly matching information theoretic lower bound showing that no (ε, δ)-DP algorithm can achieve a smaller error. This significantly improves upon a long line of existing private algorithms for PCA, e.g., [46, 33, 103, 101, 80]. These existing algorithms are analyzed for fixed and non-stochastic data and achieve sub-optimal error rates of $O(\sqrt{d/n} + d^{3/2}/(\varepsilon n))$ even in the stochastic setting we consider.

A remaining question is whether the sub-Gaussian-like assumption, namely Assumption A.4, is necessary or if it is an artifact of our analysis and our algorithm. It turns out that such an assumption on the lightness of the tail is critical; we prove an algorithmic independent and information theoretic lower bound (Theorem 4.5.4) to show that, without such an assumption, the cost of privacy is lower bounded by $\Omega(\sqrt{d/(\varepsilon n)})$. This proves a separation of the error depending on the lightness of the tail.

We start with the formal description of the stochastic setting in Section 4.2 and present Oja's algorithm for non-private PCA. Our first attempt in making this algorithm private in Section 4.3 already achieves near-optimal error, if the data is strictly from a Gaussian distribution. However, there are two remaining challenges that we describe in detail in Section 4.4: (i) the excessive number of iterations of Private Oja's Algorithm (Algorithm 11) prevents using typical values of ε used in practice, and (ii) for general sub-Gaussian-like distributions, the error does not vanish even when the noise in the data (as measured by a certain fourth moment of a function of the data) vanishes. The first challenge is due to the analysis that requires amplification by shuffling [81] that is restrictive. The second is due to its reliance on gradient norm clipping [1] that does not adapt to the geometry of the current gradients. This drives the design of DP-PCA in Section 4.5 that critically relies on two techniques to overcome each challenge, respectively. First, minibatch SGD (instead of single sample SGD) significantly reduces the number iterations, thus obviating the need for amplification by shuffling. Next, private mean estimation (instead of gradient norm clipping and noise adding) adapts to the geometry of the problem and adds the minimal noise

necessary to achieve privacy. The main idea of this geometry adaptive stochastic gradient update is explained in detail in Section 4.6, along with a sketch of a proof.

Notations. For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote the Euclidean norm. For a matrix $X \in \mathbb{R}^{d\times d}$, we use $\|X\|_2 = \max_{\|v\|=1} \|Xv\|_2$ to denote the spectral norm. We use $\mathbf{I}_d$ to denote $d \times d$ identity matrix. For $n \in \mathbb{Z}^+$, let $[n] := \{1, 2, \ldots, n\}$. Let $\mathbb{S}_2^{d-1}$ denote the unit d-sphere of ℓ_2, i.e., $\mathbb{S}_2^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$. $\tilde{O}()$ hides logarithmic factors in n, d, and the failure probability ζ.

4.2 Problem formulation and background on DP

Typical PCA assumes i.i.d. data $\{x_i \in \mathbb{R}^d\}$ from a distribution and finds the first eigenvector of $\Sigma = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_i - \mathbb{E}[x_i])^\top] \in \mathbb{R}^{d\times d}$. Our approach allows for a more general class of data $\{A_i \in \mathbb{R}^{d\times d}\}$ that recovers the standard case when $A_i = (x_i - \mathbb{E}[x_i])(x_i - \mathbb{E}[x_i])^\top$.

Assumption 5 ($(\Sigma, \{\lambda_i\}_{i=1}^d, M, V, K, \kappa, a, \gamma^2)$-model)**.** *Let $A_1, A_2, \ldots, A_n \in \mathbb{R}^{d\times d}$ be a sequence of (not necessarily symmetric) matrices sampled independently from the same distribution that satisfy the following with PSD matrices $\Sigma \in \mathbb{R}^{d\times d}$ and $H_u \in \mathbb{R}^{d\times d}$, and positive scalar parameters M, V, K, κ, a, and γ^2:*

- ***A.1.*** *Let $\Sigma := \mathbb{E}[A_i]$, for a symmetric positive semidefinite (PSD) matrix $\Sigma \in \mathbb{R}^{d\times d}$, λ_i denote the i-th largest eigenvalue of Σ, and $\kappa := \lambda_1/(\lambda_1 - \lambda_2)$,*
- ***A.2.*** *$\|A_i - \Sigma\|_2 \le \lambda_1 M$ almost surely,*
- ***A.3.*** *$\max\left\{\left\|\mathbb{E}\left[(A_i - \Sigma)(A_i - \Sigma)^\top\right]\right\|_2, \left\|\mathbb{E}\left[(A_i - \Sigma)^\top(A_i - \Sigma)\right]\right\|_2\right\} \le \lambda_1^2 V$,*
- ***A.4.*** *$\max_{\|u\|=1,\|v\|=1} \mathbb{E}\left[\exp\left(\left(\frac{|u^\top(A_i^\top - \Sigma)v|^2}{K^2\lambda_1^2\|H_u\|_2}\right)^{1/(2a)}\right)\right] \le 2$, where $H_u := (1/\lambda_1^2)\mathbb{E}[(A_i - \Sigma)uu^\top(A_i - \Sigma)^\top]$. We denote $\gamma^2 := \max_{\|u\|=1} \|H_u\|_2$.*

The first three assumptions are required for PCA even if privacy is not needed. The last assumption provides a Gaussian-like tail bound that determines how much noise we need to introduce in the algorithm for (ε, δ)-DP. The following lemma is useful in the analyses.

Lemma 4.2.1. *Under A.1 and A.4 in Assumption 5, for any unit vector u, v, with probability $1-\zeta$,*

$$|u^\top (A_i^\top - \Sigma) v|^2 \;\le\; K^2 \lambda_1^2 \|H_u\|_2 \log^{2a}(2/\zeta)\,. \tag{4.1}$$

4.2.1 Oja's algorithm

In a non-private setting, the following streaming algorithm introduced in [172] achieves optimal sample complexity as analyzed in [119]. It is a projected stochastic gradient ascent on the objective defined on the empirical covariance: $\max_{\|w\|=1}(1/n)\sum_{i=1}^n w^\top A_i w$.

Algorithm 10: (Non-private) Oja's Algorithm

1 Choose w_0 uniformly at random from the unit sphere

2 **for** $t = 1, 2, \ldots, T$ **do** $w_t' \leftarrow w_{t-1} + \eta_t A_t w_{t-1}$, $w_t \leftarrow w_t'/\|w_t'\|$

3 Return w_T

Central to our analysis is the following error bound on Oja's Algorithm from [119].

Theorem 4.2.2 ([119, Theorem 4.1])**.** *Under Assumptions A.1-A.3, suppose the step size $\eta_t = \frac{\alpha}{(\lambda_1-\lambda_2)(\xi+t)}$ for some $\alpha > 1/2$ and $\xi := 20\max\left(\kappa M\alpha, \kappa^2(V+1)\alpha^2/\log(1+(\zeta/100))\right)$. If $T > \xi$ then there exists a constant $C > 0$ such that Algorithm 10 outputs w_T achieving w.p. $1-\zeta$,*

$$\sin^2(w_T, v_1) \le \frac{C\log(1/\zeta)}{\zeta^2}\left(\frac{\alpha^2\kappa^2 V}{(2\alpha-1)T} + d\left(\frac{\xi}{T}\right)^{2\alpha}\right)\,. \tag{4.2}$$

4.2.2 Background on Differential Privacy

Differential privacy (DP), introduced in [78], is a de facto mathematical measure for privacy leakage of a database accessed via queries. It ensures that even an adversary who knows all other entries cannot identify with a high confidence whether a person of interest participated in a database or not.

Definition 4.2.3 (Differential privacy [78])**.** *Given two multisets S and S', we say the pair (S, S') is* neighboring *if $|S \setminus S'| + |S' \setminus S| \le 1$. We say a stochastic query q over a dataset S*

satisfies (ε, δ)-differential privacy *for some* $\varepsilon > 0$ *and* $\delta \in (0,1)$ *if* $\mathbb{P}(q(S) \in A) \leq e^{\varepsilon}\mathbb{P}(q(S') \in A) + \delta$ *for all neighboring* (S, S') *and all subset* A *of the range of* q.

Small values of ε and δ ensures that the adversary cannot identify any single data point with high confidence, thus providing plausible deniability. We provide useful DP lemmas in Appendix 2.1.1. Within our stochastic gradient descent approach to PCA, we rely on the Gaussian mechanism to privatize each update. The *sensitivity* of a query q is defined as $\Delta_q := \sup_{\text{neighboring } (S,S')} \|q(S) - q(S')\|$.

Lemma 4.2.4 (Gaussian mechanism [79])**.** *For a query* q *with sensitivity* Δ_q, $\varepsilon \in (0,1)$, *and* $\delta \in (0,1)$, *the Gaussian mechanism outputs* $q(S) + \mathcal{N}(0, (\Delta_q(\sqrt{2\log(1.25/\delta)})/\varepsilon)^2 \mathbf{I}_d)$ *and achieves* (ε, δ)*-DP.*

This is a special case of a family of output perturbation mechanisms which includes the Laplace mechanism [78] and stair-case mechanisms [94]. The latter is shown to be optimal in one-dimension [95] and for hypothesis testing under local DP [126]. Another mechanism we frequently use is the private histogram learner of [140], whose analysis is provide in Appendix 2.1.1, along with various composition theorems to provide end-to-end guarantees.

4.2.3 Comparisons with existing results in private PCA

We briefly discuss the most closely related work and provide more previous work in Appendix C.1. Most existing results assume a fixed data under a deterministic setting where each sample has a bounded norm, $\|x_i\| \leq \beta$, and the goal is to find the top eigenvector of $\hat{\Sigma} := (1/n)\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top$ for the empirical mean $\hat{\mu}$. For the purpose of comparisons, consider Gaussian $x_i \sim \mathcal{N}(0, \Sigma)$ with $\|x_i\| \leq \beta = O(\sqrt{\lambda_1 d \log(n/\zeta)})$ for all $i \in [n]$ with probability $1 - \zeta$. The first line of approaches in [33, 46, 80] is a Gaussian mechanism that outputs $\text{PCA}(\widehat{\Sigma} + Z)$, where Z is a symmetric matrix with i.i.d. Gaussian entries with a variance $((\beta^2/n\varepsilon)\sqrt{2\log(1.25/\delta)})^2$ to ensure (ε, δ)-DP. The tightest result in [80, Theorem 7] achieves

$$\sin(\hat{v}, v_1) \quad = \quad \tilde{O}\Big(\kappa\Big(\sqrt{\frac{d}{n}} + \frac{d^{3/2}\sqrt{\log(1/\delta)}}{\varepsilon n}\Big)\Big)\,, \tag{4.3}$$

Algorithm 11: Private Oja's Algorithm

Input: $S = \{A_i \in \mathbb{R}^{d\times d}\}_{i=1}^n$, privacy (ε, δ), learning rates $\{\eta_t\}_{t=1}^n$

Randomly permute S and choose w_0 uniformly at random from the unit sphere

Set DP noise multiplier: $\alpha \leftarrow C' \log(n/\delta)/(\varepsilon\sqrt{n})$

Set clipping threshold: $\beta \leftarrow C\lambda_1\sqrt{d}(K\gamma \log^a(nd/\zeta) + 1)$

for *t=1, 2, . . . , n* **do**

  Sample $z_t \sim \mathcal{N}(0, \mathbf{I}_d)$

  $w_t' \leftarrow w_{t-1} + \eta_t \operatorname{clip}_\beta(A_t w_{t-1}) + 2\eta_t\beta\alpha z_t$ where $\operatorname{clip}_\beta(x) = x \cdot \min\{1, \frac{\beta}{\|x\|_2}\}$

  $w_t \leftarrow w_t'/\|w_t'\|$

Return w_n

with high probability, under a strong assumption that the spectral gap is very large: $\lambda_1 - \lambda_2 = \omega(d^{3/2}\sqrt{\log(1/\delta)}/(\varepsilon n))$. In a typical scenario with $\lambda_1 = O(1)$, this requires a large sample size of $n = \omega(d^{3/2}/\varepsilon)$. Since this Gaussian mechanism does not exploit the statistical properties of i.i.d. samples, the second term in this upper bound is larger by a factor of $d^{1/2}$ compared to the proposed DP-PCA (Corollary 4.5.2). The error rate of Eq. (4.3) is also achieved in [103, 101] by adding Gaussian noise to the standard power method for computing the principal components. When the spectral gap, $\lambda_1 - \lambda_2$, is smaller, it is possible to trade-off the dependence in κ and the sampling ratio d/n, which we do not address in this work but is surveyed in Appendix C.1.

4.3 First attempt: making Oja's Algorithm private

Following the standard recipe in training with DP-SGD, e.g., [1] and a recent work [199], we introduce Private Oja's Algorithm in Algorithm 11. At each gradient update, we first apply gradient norm clipping to limit the contribution of a single data point and next add an appropriately chosen Gaussian noise from Lemma D.2.1 to achieve (ε, δ)-DP, end-to-end. The choice of clipping threshold β ensures that, with high probability under our assumption, we do not clip any gradients. The choice of noise multiplier α ensures (ε, δ)-DP.

One caveat in streaming algorithms is that we access data n times, each with a private mechanism, but accessing only a single data point at a time. To prevent excessive privacy loss due to such a large number of data accesses, we apply a random shuffling in line 1 Algorithm 11, in order to benefit from a standard amplification by shuffling [81, 86]. This gives an amplified privacy guarantee that allows us to add a small noise proportional to $\alpha = O(\log(n/\delta)/(\varepsilon\sqrt{n}))$. Without the shuffle amplification, we will instead need a larger noise scaling as $\alpha = O(\log(n/\delta)/\varepsilon)$, resulting in a suboptimal utility guarantee. However, this comes with a restriction that the amplification holds only for small values of $\varepsilon = O(\sqrt{\log(n/\delta)/n})$. Our first contribution in the proposed DP-PCA (Algorithm 12) is to expand this range to $\varepsilon = O(1)$, which includes the practical regime of interest $\varepsilon \in [1/2, 5]$.

Lemma 4.3.1 (Privacy). *If $\varepsilon = O(\sqrt{\log(n/\delta)/n})$ and the noise multiplier is chosen to be $\alpha = \Omega\left(\log(n/\delta)/(\varepsilon\sqrt{n})\right)$, then Algorithm 11 is (ε, δ)-DP.*

Under Assumption 5, we select gradient norm clipping threshold β such that no gradient exceeds β.

Lemma 4.3.2 (Gradient clipping). *Let $\beta = C\lambda_1\sqrt{d}(K\gamma\log^a(nd/\zeta) + 1)$ for some constant $C > 0$. Then with probability $1 - \zeta$, $\|A_t w_{t-1}\| \leq \beta$ for any fixed w_{t-1} independent of A_t, for all $t \in [n]$.*

We provide proofs of both lemmas and the next theorem in Appendix C.4. When no clipping is applied, we can use the standard analysis of Oja's Algorithm from [119] to prove the following utility guarantee.

Theorem 4.3.3 (Utility). *Given n i.i.d. samples $\{A_i \in \mathbb{R}^{d\times d}\}_{i=1}^n$ satisfying Assumption 5 with parameters $(\Sigma, M, V, K, \kappa, a, \gamma^2)$, if*

$$n \;=\; \tilde{O}\Big(\kappa^2 + \kappa M + \kappa^2 V + \frac{d\,\kappa\,(\gamma+1)\,\log(1/\delta)}{\varepsilon}\Big)\,, \tag{4.4}$$

with a large enough constant, then there exists a positive universal constant c_1 and a choice of learning rate η_t that depends on $(t, M, V, K, a, \lambda_1, \lambda_1 - \lambda_2, n, d, \varepsilon, \delta)$ such that Algorithm 11

with a choice of $\zeta = 0.01$ *outputs* w_n *that achieves with probability* 0.99,

$$\sin^2(w_n, v_1) \;=\; \widetilde{O}\left(\kappa^2\Big(\frac{V}{n} + \frac{(\gamma+1)^2 d^2 \log^2(1/\delta)}{\varepsilon^2 n^2}\Big)\right), \tag{4.5}$$

where $\widetilde{O}(\cdot)$ *hides poly-logarithmic factors in* n, d, $1/\varepsilon$, *and* $\log(1/\delta)$ *and polynomial factors in* K.

The first term in Eq. (4.5) matches the non-private error rate for Oja's algorithm in Eq. (4.2) with $\alpha = O(\log n)$ and $T = n$, and the second term is the price we pay for ensuring (ε, δ)-DP.

Remark 4.3.4. *For a canonical setting of a Gaussian data with* $A_i = x_i x_i^\top$ *and* $x_i \sim \mathcal{N}(0, \Sigma)$, *we have, for example from [176, Lemma 1.12], that* $M = O(d\log(n))$, $V = O(d)$, $K = 4$, $a = 1$, *and* $\gamma^2 = O(1)$. *Theorem 4.3.3 implies the following error rate:*

$$\sin^2(w_n, v_1) \;=\; \tilde{O}\Big(\kappa^2\Big(\frac{d}{n} + \frac{d^2 \log^2(1/\delta)}{\varepsilon^2 n^2}\Big)\Big). \tag{4.6}$$

Comparing to a lower bound in Theorem 4.5.3, this is already near optimal. However, for general distributions satisfying Assumption 5, Algorithm 11 (in particular the second term in Eq. (4.5)) can be significantly sub-optimal. We explain this second weakness of Private Oja's Algorithm in the following section (the first weakness is the restriction on $\varepsilon = O(\sqrt{\log(n/\delta)/n})$).

4.4 Two remaining challenges

We explain the two remaining challenges in Private Oja's Algorithm and propose techniques to overcomes these challenges that lead to the design of DP-PCA (Algorithm 12).

First challenge: restricted range of $\varepsilon = O(\sqrt{\log(n/\delta)/n})$**.** This is due to the large number, n, of iterations that necessitates the use of the amplification by shuffling in Theorem C.4.1. We reduce the number of iterations with minibatch SGD. For $T = O(\log^2 n)$

and $t = 1, 2, \ldots, T$, we repeat

$$w'_t \leftarrow w_{t-1} + \frac{\eta_t}{B} \sum_{i=1+B(t-1)}^{Bt-1} \mathrm{clip}_\beta(A_i w_{t-1}) + \frac{w\eta_t\beta\alpha}{B} z_t \ , \text{ and } \quad w_t \leftarrow w'_t/\|w'_t\| \ , \tag{4.7}$$

where $z_t \sim \mathcal{N}(0, \mathbf{I}_d)$ and the minibatch size is $B = \lfloor n/T \rfloor$. Since the dataset is accessed only $T = O(\log^2 n)$ times, the end-to-end privacy is analyzed with the serial composition (Lemma C.2.2) instead of the amplification by shuffling. This ensures (ε, δ)-DP for any $\varepsilon = O(1)$, resolving the first challenge, and still achieves the utility guarantee of Eq. (4.5).

Second challenge: excessive noise for privacy. This is best explained with an example.

Example 4.4.1 (Signal and noise separation)**.** *Consider a setting with $A_i = x_i x_i^\top$ and $x_i = s_i + n_i$ where $s_i = v$ with probability half and $s_i = -v$ otherwise for a unit norm vector v and $n_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We want to find the principal component of $\Sigma = \mathbb{E}[x_i x_i^\top] = vv^\top + \sigma^2\mathbf{I}$, which is v. This construction decomposes the signal and the noise. For $A_i = vv^\top + s_i n_i^\top + n_i s_i^\top + n_i n_i^\top$, the signal component is determined by $vv^\top$ that is deterministic due to the sign cancelling. The noise component is $s_i n_i^\top + n_i s_i^\top + n_i n_i^\top$ which is random. We can control the Signal-to-Noise Ratio (SNR), $1/\sigma^2$, by changing σ^2, and we are particularly interested in the regime where σ^2 is small. As we are interested in $\sigma^2 < 1$, this satisfies Assumption 5 with $\lambda_1 = 1 + \sigma^2$, $\lambda_2 = \sigma^2$, $V = O(d\sigma^2)$, $K = O(1)$, $a = 1$, and $\gamma^2 = \sigma^2$. Substituting this into Eq. (4.5), Private Oja's Algorithm achieves*

$$\sin^2(w_n, v_1) \ = \ \tilde{O}\Big(\frac{\sigma^2 d}{n} + \frac{d^2 \log(1/\delta)}{\varepsilon^2 n^2}\Big) \ , \tag{4.8}$$

where we are interested in $\sigma^2 < 1$.

This is problematic since the second term, due to the DP noise, does not vanish as the randomness σ^2 in the data decreases. We do not observe this for Gaussian data where signal and noise scale proportionally as shown below. We reduce the noise we add for privacy, by switching from a simple norm clipping, that adds noise proportional to the norm of the gradients, to private estimation, that only requires the noise to scale as the *range* of the

Figure 4.1: 2-d PCA under the Gaussian data from Remark 4.3.4 (left) shows that the average gradient (red arrow) is smaller than the range of the minibatch of 400 gradients (blue dots). Under Example 4.4.1 (right), the range can be made arbitrarily smaller than the average gradient.

gradients, i.e. the maximum distance between two gradients in the minibatch. The toy example above showcases that the range can be arbitrarily smaller than the maximum norm (Fig. 4.1). We want to emphasize that although the idea of using private estimation within an optimization has been conceptually proposed in abstract settings, e.g., in [131], DP-PCA is the first setting where (i) such separation between the norm and the range of the gradients holds under any statistical model, and hence (ii) the long line of recent advances in private estimation provides significant gain over the simple DP-SGD [1].

4.5 Differentially Private Principal Component Analysis (DP-PCA)

Combining the two ideas of minibatch SGD and private mean estimation, we propose DP-SGD. We use minibatch SGD of minibatch size $B = O(n/\log^2 n)$ to allow for larger range of $\varepsilon = O(1)$. We use Private Mean Estimation to add an appropriate level of noise chosen adaptively according to Private Eigenvalue Estimation. We describe details of both sub-routines in Section 4.6.

We show an upper bound on the error achieved by DP-PCA under an appropriate choice of the learning rate. We provide a complete proof in Appendix C.5.1 that includes the explicit choice of the learning rate η_t in Eq. (C.52), and a proof sketch is provided in Section 4.6.1.

Theorem 4.5.1. *For $\varepsilon \in (0, 0.9)$, DP-PCA guarantees (ε, δ)-DP for all S, B, ζ, and δ. Given*

Algorithm 12: Differentially Private Principal Component Analysis (DP-PCA)

Input: $S = \{A_i\}_{i=1}^n$, (ε, δ), batch size $B \in \mathbb{Z}_+$, learning rates $\{\eta_t\}_{t=1}^{\lfloor n/B \rfloor}$, probability $\zeta \in (0,1)$

Choose w_0 uniformly at random from the unit sphere

for $t = 1, 2, \ldots, T = \lfloor n/B \rfloor$ **do**

  Run Private Top Eigenvalue Estimation (Algorithm 21) with $(\varepsilon/2, \delta/2)$-DP and failure probability $\zeta/(2T)$ on $\{A_{B(t-1)+i} w_{t-1}\}_{i=1}^{\lfloor B/2 \rfloor}$. Let the returned estimation be $\hat{\Lambda}_t > 0$.

  Run Private Mean Estimation (Algorithm 22) with $(\varepsilon/2, \delta/2)$-DP, failure probability $\zeta/(2T)$, and the estimated eigenvalue $2\hat{\Lambda}_t$ on $\left\{A_{B(t-1)+\lfloor B/2 \rfloor + i} w_{t-1}\right\}_{i \in \lfloor B/2 \rfloor}$. Let the returned mean gradient estimate be $\hat{g}_t \in \mathbb{R}^d$.

  $w_t' \leftarrow w_{t-1} + \eta_t \hat{g}_t \quad , \quad w_t \leftarrow w_t' / \|w_t'\|$

Return w_T

n i.i.d. samples $\{A_i \in \mathbb{R}^{d \times d}\}_{i=1}^n$ satisfying Assumption 5 with parameters $(\Sigma, M, V, K, \kappa, a, \gamma^2)$, if

$$n \;=\; \tilde{O}\Big(e^{\kappa^2} + \frac{d^{1/2} (\log(1/\delta))^{3/2}}{\varepsilon} + \kappa M + \kappa^2 V + \frac{d\,\kappa\,\gamma\,(\log(1/\delta))^{1/2}}{\varepsilon} + \frac{d \log(1/\delta)}{\varepsilon} \Big) \;, \quad (4.9)$$

with a large enough constant and $\delta \le 1/n$, then there exists a positive universal constant c_1 and a choice of learning rate η_t that depends on $(t, M, V, K, a, \lambda_1, \lambda_1 - \lambda_2, n, d, \varepsilon, \delta)$ such that $T = \lfloor n/B \rfloor$ steps of DP-PCA in Algorithm 12 with choices of $\zeta = 0.01$ and $B = c_1 n / (\log n)^2$, outputs w_T such that with probability 0.99,

$$\sin\left(w_T, v_1\right) \;=\; \widetilde{O}\left(\kappa \Big(\sqrt{\frac{V}{n}} + \frac{\gamma d \sqrt{\log(1/\delta)}}{\varepsilon n} \Big) \right) \;, \quad (4.10)$$

where $\widetilde{O}(\cdot)$ hides poly-logarithmic factors in n, d, $1/\varepsilon$, and $\log(1/\delta)$ and polynomial factors in K.

We further interpret this analysis and show that (i) DP-PCA is nearly optimal when the data is from a Gaussian distribution by comparing against a lower bound (Theorem 4.5.3);

and (*ii*) DP-PCA significantly improves upon the private Oja's algorithm under Example 4.4.1. We discuss the necessity of some of the assumptions at the end of this section, including how to agnostically find the appropriate learning rate scheduling.

Near-optimality of DP-PCA under Gaussian distributions. Consider the case of i.i.d. samples $\{x_i\}_{i=1}^n$ from a Gaussian distribution from Remark 4.3.4.

Corollary 4.5.2 (Upper bound; Gaussian distribution)**.** *Under the hypotheses of Theorem 4.5.1 and* $\{A_i = x_i x_i^\top\}_{i=1}^n$ *with Gaussian random vectors* x_i*'s, after* $T = n/B$ *steps, DP-PCA outputs* w_T *that achieves, with probability* 0.99*,*

$$\sin(w_T, v_1) \;=\; \tilde{O}\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta)}}{\varepsilon n}\right)\right) \;. \tag{4.11}$$

We prove a nearly matching lower bound, up to factors of $\sqrt{\lambda_1/\lambda_2}$ and $\sqrt{\log(1/\delta)}$. One caveat is that the lower bound assumes *pure*-DP with $\delta = 0$. We do not yet have a lower bound technique for approximate DP that is tight, and all known approximate DP lower bounds have gaps to achievable upper bounds in its dependence in $\log(1/\delta)$, e.g., [25, 161]. We provide a proof in Appendix C.3.1.

Theorem 4.5.3 (Lower bound; Gaussian distribution)**.** *Let* $\mathcal{M}_\varepsilon$ *be a class of* $(\varepsilon, 0)$*-DP estimators that map* n *i.i.d. samples to an estimate* $\hat{v} \in \mathbb{R}^d$*. A set of Gaussian distributions with* (λ_1, λ_2) *as the first and second eigenvalues of the covariance matrix is denoted by* $\mathcal{P}_{(\lambda_1,\lambda_2)}$*. For* $d > c$ *where* $c > 0$ *is some absolute constant, there exists a universal constant* $C > 0$ *such that*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \;\geq\; C \min\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right) \;. \tag{4.12}$$

Comparisons with private Oja's algorithm. We demonstrate that DP-PCA can significantly improve upon Private Oja's Algorithm with Example 4.4.1, where DP-PCA achieves an error bound of $\sin(w_T, v_1) = \tilde{O}\big(\sigma\sqrt{d/n} + \sigma d\sqrt{\log(1/\delta)}/(\varepsilon n)\big)$. As the noise power σ^2 decreases DP-PCA achieves a vanishing error, whereas Private Oja's Algorithm has a non-vanishing error in Eq. (4.8). This follows from the fact that the second term in the error

bound in Eq. (4.10) scales as γ, which can be made arbitrarily smaller than the second term in Eq. (4.5) that scales as $(\gamma+1)$. Further, the error bound for DP-PCA holds for any $\varepsilon = O(1)$, whereas Private Oja's Algorithm requires significantly smaller $\varepsilon = O(\sqrt{\log(n/\delta)/n})$.

Remarks on the assumptions of Theorem 4.5.1. We have an exponential dependence of the sample complexity in the spectral gap, $n \geq \exp(\kappa^2)$. This ensures we have a large enough $T = \lfloor n/B \rfloor$ to reduce the non-dominant second term in Eq. (4.2), in balancing the learning rate η_t and T (which is explicitly shown in Eqs. C.54 and (C.55) in the Appendix). It is possible to get rid of this exponential dependence at the cost of an extra term of $\tilde{O}(\kappa^4\gamma^2 d^2 \log(1/\delta)/(\varepsilon n)^2)$ in the error rate in Eq. (4.10), by selecting a slightly larger $T = c\kappa^2 \log^2 n$. A Gaussian-like tail bound in Assumption A.4 is necessary to get the desired upper bound scaling as $\tilde{O}(d\sqrt{\log(1/\delta)}/(\varepsilon n))$ in Eq. 4.11, for example. The next lower bound shows that without such assumptions on the tail, the error due to privacy scales as $\Omega(\sqrt{d \wedge \log(1/\delta)/(\varepsilon n)})$. We believe that the dependence in δ is loose, and it might be possible to get a tighter lower bound using [132]. We provide a proof and other lower bounds in Appendix B.3.

Theorem 4.5.4 (Lower bound without Assumption A.4)**.** *Let $\mathcal{M}_\varepsilon$ be a class of (ε, δ)-DP estimators that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3 with $M = \tilde{O}(d + \sqrt{n\varepsilon/d})$, $V = O(d)$ and $\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\tilde{\mathcal{P}}} \mathbb{E}_{S\sim P^n} \left[\sin(\hat{v}(S), v_1)\right] \;\geq\; C\kappa \min\left(\sqrt{\frac{d \wedge \log\left((1-e^{-\varepsilon})/\delta\right)}{\varepsilon n}}, 1\right) . \tag{4.13}$$

Currently, DP-PCA requires choices of the learning rates, η_t, that depend on possibly unknown quantities. Since we can privately evaluate the quality of our solution, one can instead run multiple instances of DP-PCA with varying $\eta_t = c_1/(c_2 + t)$ and find the best choice of $c_1 > 0$ and $c_2 > 0$. Let $w_T(c_1, c_2)$ denote the resulting solution for one instance of $\{\eta_t = c_1/(c_2 + t)\}_{t=1}^T$. We first set a target error ζ. For each round $i = 1, \ldots$, we will run algorithm for $(c_1, c_2) = [2^{i-1}, 2^{-i+1}] \times [2^{-i+1}, 2^{-i+2} \ldots, 2^{i-1}]$ and $(c_1, c_2) = [2^{-i+1}, 2^{-i+2} \ldots, 2^{i-1}] \times [2^{i-1}, 2^{-i+1}]$, and compute each $\sin(w_T(c_1, c_2), v_1)$ privately, each with

privacy budget $\varepsilon_i = \frac{\varepsilon}{2^{i+1}(2i-1)}, \delta_i = \frac{\delta}{2^{i+1}(2i-1)}$. We terminate the algorithm once there there is a $w_T(c_1, c_2)$ satisfies $\sin(w_T(c_1, c_2), v_1) \leq \zeta$. It is clear that this search meta-algorithm terminate in logarithmic round, and the total sample complexity only blows up by a poly-log factor.

4.6 Private mean estimation for the minibatch stochastic gradients

DP-PCA critically relies on private mean estimation to reduce variance of the noise required to achieve (ε, δ)-DP. We follow a common recipe from [140, 130, 135, 32, 54]. First, we privately find an approximate range of the gradients in the minibatch (Alg. 21). Next, we apply the Gaussian mechanism to the truncated gradients where the truncation is tailored to the estimated range (Alg. 22).

Step 1: estimating the range. We need to find an approximate range of the minibatch of gradients in order to adaptively truncate the gradients and bound the sensitivity. Inspired by a private preconditioning mechanism designed for mean estimation with unknown covariance from [133], we propose to use privately estimated top eigenvalue of the covariance matrix of the gradients. For details on the version of the histogram learner we use in Alg. 21 in Appendix C.5.2, we refer to [160, Lemma D.1]. Unlike the private preconditioning of [133] that estimates all eigenvalues and requires $n = \widetilde{O}(d^{3/2}\log(1/\delta)/\varepsilon)$ samples, we only require the top eigenvalue and hence the next theorem shows that we only need $n = \widetilde{O}(d\log(1/\delta)/\varepsilon)$.

Theorem 4.6.1. *Algorithm 21 is (ε, δ)-DP. Let $g_i = A_i u$ for some fixed vector u, where A_i satisfies A.1 and A.4 in Assumption 5 such that the mean is $\mathbb{E}[g_i] = \Sigma u$ and the covariance is $\mathbb{E}[(g_i - \Sigma u)(g_i - \Sigma u)^\top] = \lambda_1^2 H_u$. With a large enough sample size scaling as*

$$B = O\left(\frac{K^2\, d\, \log(d\log(1/(\delta\zeta))/(\zeta\varepsilon))\log^{2a}(Bd/\zeta)\log(1/(\zeta\delta))}{\varepsilon}\right) = \tilde{O}\left(\frac{K^2\, d\, \log(1/\delta)}{\varepsilon}\right),$$

Algorithm 21 outputs $\hat{\Lambda}$ achieving $\hat{\Lambda} \in \left[(1/\sqrt{2})\lambda_1^2\|H_u\|_2, \sqrt{2}\lambda_1^2\|H_u\|_2\right]$ with probability $1-\zeta$, where the pair $(K > 0, a > 0)$ parametrizes the tail of the distribution in A.4 and $\tilde{O}(\cdot)$ hides logarithmic factors in $B, d, 1/\zeta, \log(1/\delta)$, and ε.

We provide a proof in Appendix C.5.2. There are other ways to privately estimate the range. Some approaches require known bounds such as $\sigma_{\min}^2 \leq \lambda_1^2(H_u)_{ii} \leq \sigma_{\max}^2$ for all $i \in [d]$ [140], and other agnostic approaches are more involved such as instance optimal universal estimators of [72].

Step 2: Gaussian mechanism for mean estimation. Once we have a good estimate of the top eigenvalue from the previous section, we use it to select the bin size of the private histogram and compute the truncated empirical mean. Since truncated empirical mean has a bounded sensitivity, we can use Gaussian mechanism to achieve DP. The algorithm is now standard in DP mean estimation, e.g., [140, 130]. However, the analysis is slightly different since our assumptions on g_i's are different. For completeness, we provide the Algorithm 22 in Appendix C.5.3.

The next lemma shows that the Private Mean Estimation is (ε, δ)-DP, and with high probability clipping does not apply to any of the gradients. The returned private mean, therefore, is distributed as a spherical Gaussian centered at the empirical mean of the gradients. This result requires that we have a good estimate of the top eigenvalue from Alg. 21 such that $\hat{\Lambda} \simeq \lambda_1^2 \|H_u\|_2$. This analysis implies that we get an unbiased estimate of the gradient mean (which is critical in the analysis) with noise scaling as $\tilde{O}(\lambda_1 \gamma \sqrt{d \log(1/\delta)}/(\varepsilon B))$, where $\gamma^2 = \max_{u:\|u\|=1} \|H_u\|_2$ (which is critical in getting the tight sample complexity in the second term of the final utility guarantee in Eq. (4.10)). We provide a proof in Appendix C.5.3.

Lemma 4.6.2. *For $\varepsilon \in (0, 0.9)$ and any $\delta \in (0, 1)$, Algorithm 22 is (ε, δ)-DP. Let $g_i = A_i u$ for some fixed vector u, where A_i satisfies A.1 and A.4 in Assumption 5 such that the mean is $\mathbb{E}[g_i] = \Sigma u$ and the covariance is $\mathbb{E}[(g_i - \Sigma u)(g_i - \Sigma u)^\top] = \lambda_1^2 H_u$. If $\hat{\Lambda} \in [\lambda_1^2 \|H_u\|_2/\sqrt{2}, \sqrt{2}\lambda_1^2 \|H_u\|_2]$, $\delta \leq 1/B$, and $B = \Omega((\sqrt{d \log(1/\delta)}/\varepsilon) \log(d/(\zeta\delta)))$ then, with probability $1 - \zeta$, $g_i \in \bar{g} + \left[-3K\sqrt{\hat{\Lambda}} \log^a(Bd/\zeta), 3K\sqrt{\hat{\Lambda}} \log^a(Bd/\zeta)\right]^d$ for all $i \in [B]$.*

4.6.1 Proof sketch of Theorem 4.5.1

We choose $B = \Theta(n/\log^2 n)$ such that we access the dataset only $T = \Theta(\log^2 n)$ times. Hence we do not need to rely on amplification by shuffling. To add Gaussian noise that scales as the standard deviation of the gradients in each minibatch (as opposed to potentially excessively large mean of the gradients), DP-PCA adopts techniques from recent advances in private mean estimation. Namely, we first get a private and accurate estimate of the range from Theorem 4.6.1. Using this estimate, $\hat{\Lambda}$, Private Mean Estimation returns an unbiased estimate of the empirical mean of the gradients, as long as no truncation has been applied as ensured by Lemma 4.6.2. This gives

$$w_t' \leftarrow w_{t-1} + \eta_t \left(\frac{1}{B} \sum_{i=1}^{B} A_{B(t-1)+i} w_{t-1} + \beta_t z_t \right) , \tag{4.14}$$

for $z_t \sim \mathcal{N}(0, \mathbf{I})$ and $\beta_t = \frac{8K\sqrt{2\hat{\Lambda}_t}\log^a(Bd/\zeta)\sqrt{2d\log(2.5/\delta)}}{\varepsilon B}$. Using rotation invariance of spherical Gaussian random vectors and the fact that $\|w_{t-1}\| = 1$, we can reformulate it as

$$w_t' \leftarrow w_{t-1} + \eta_t \underbrace{\left(\frac{1}{B} \sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t \right)}_{\tilde{A}_t} w_{t-1} . \tag{4.15}$$

This process can be analyzed with Theorem 4.2.2 with $\tilde{A}_t$ substituting A_t.

4.7 Discussion

Under the canonical task of computing the principal component from i.i.d. samples, we show the first result achieving an optimal error rate. This critically relies on two ideas: minibatch SGD and private mean estimation. In particular, private mean estimation plays a critical role in the case when the range of the gradients is significantly smaller than the norm; we achieve an optimal error rate that cannot be achieved with the standard recipe of gradient clipping.

Assumption A.4 can be relaxed to heavy-tail bounds with bounded k-th moment on A_i, in which case we expect the second term in Eq. (4.10) to scale as $O(d(\sqrt{\log(1/\delta)}/\varepsilon n)^{1-1/k})$, drawing analogy from a similar trend in a computationally inefficient DP-PCA without

spectral gap [161, Corollary 6.10]. When a fraction of data is corrupted, recent advances in [212, 145, 121] provide optimal algorithms for PCA. However, existing approach of [161] for robust and private PCA is computationally intractable. Borrowing ideas from robust and private mean estimation in [160], one can design an efficient algorithm, but at the cost of sub-optimal sample complexity. It is an interesting direction to design an optimal and robust version of DP-PCA. Our lower bounds are loose in its dependence in $\log(1/\delta)$. Recently, a promising lower bound technique has been introduced in [132] that might close this gap.

There are two ways to extend our framework to general rank-r PCA, whose analyses are promising research directions. First, applying Hotelling's deflation method [111], we can iteratively find the PCA components one by one, by alternating our DP-PCA and deflation. For example, in one step of the iteration, we only update the current iterate vector in the directions orthogonal to all the previously found PCA components. Repeating this steps gives the estimates of the top principal components. Secondly, we can directly apply Oja's algorithm. We keep track of a r-dimensional subspace in the Oja's update rule for PCA, and perform QR decomposition to keep the iterates on the Grassmannian manifold. It might be possible to extend the analysis of [113] to analyze the private version.

Chapter 5

LABEL-ROBUST DIFFERENTIALLY PRIVATE LINEAR REGRESSION

5.1 Introduction

Differential Privacy (DP) [78] is a standard notion of privacy widely adopted by both industry and government [189, 82, 84, 2]. With widespread usage of ML and statistical techniques, DP becomes even more critical to ensure private information of participating individuals is not revealed in any form via the learned model. An statistical estimator is said to be (ε, δ)-differentially private if presence/absence of an individual's data point in the dataset does not significantly change the estimated output. Smaller $\varepsilon > 0$ and $\delta \in [0, 1]$ imply stronger privacy guarantees.

While privacy preserving statistical estimators have been studied extensively in recent past, several critical questions remain open (see App. D.1 for a survey). Consider the canonical statistical task of linear regression with n i.i.d. samples, $\{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$, drawn from $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top w^* + z_i$, $z_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbb{E}[x_i z_i] = 0$ for some true parameter $w^* \in \mathbb{R}^d$. The error is measured in $(1/\sigma)\|\hat{w} - w^*\|_\Sigma := (1/\sigma)\|\Sigma^{1/2}(\hat{w} - w^*)\|$, which correctly accounts for the signal-to-noise ratio in each direction; in the direction of large eigenvalue of Σ, we have larger signal in x_i but the noise z_i remains the same. We expect smaller errors in those directions, which is accounted for in the error measure $(1/\sigma)\|\hat{w} - w^*\|_\Sigma$.

Minimax optimal sample complexity for estimating the optimal linear regression model with DP was recently established. For the lower bound, using recently introduced score attack technique, [41, Theorem 3.1] shows that $n = \Omega(d/\alpha^2 + d/(\varepsilon\alpha))$ samples are necessary to achieve an error of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \alpha$ (in expectation). For the matching upper bound, High-dimensional Propose-Test-Release (HPTR) in [161] and Robust-to-Private in

[17] show that $n = \tilde{O}(d/\alpha^2 + d/(\varepsilon\alpha))$ samples are also sufficient. The first term of d/α^2 is the fundamental sample complexity even if privacy is not required, and the second term of $d/(\varepsilon\alpha)$ is the cost of privacy.

This implies that, statistically, the problem appears to be solved. However, computationally, the problem is still open despite multiple studies of the problem. That is, the statistical optimal algorithms still take exponential time.

After a series of efforts in computationally efficient approaches as surveyed in App. D.1, [199] achieves the best known sample complexity of $n = \tilde{O}(d/\alpha^2 + \kappa d/(\varepsilon\alpha) + \kappa^2 d/\varepsilon)$, where κ is the condition number of the covariance Σ of the covariates. Compared to HPTR, the cost of computational efficiency is factor of κ in the second term and the third term that is unnecessary. As the condition number can be quite large, improving the dependence on κ is of utmost importance. Furthermore, the technique of [199] strictly requires sampling without replacement, whose analysis relies on having an explicit form of the end-to-end update. In particular, their analysis technique is not applicable to the case with corrupted samples.

In contrast, we propose a novel method (Alg. 13) that builds upon full-batch gradient descent and applies a carefully chosen adaptive clipping which is a general technique used in practice as well [1]. Together with an intuitive but intricate analysis technique, we improve the sample complexity to $n = \tilde{O}(d/\alpha^2 + \kappa^{1/2} d/(\varepsilon\alpha))$.

Corollary 5.1.1 (Corollary of Thm. 5.3.1 for sub-Gaussian data)**.** *Alg. 13 is (ε, δ)-DP. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of i.i.d. samples with $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top w^* + z_i$ and $z_i \sim \mathcal{N}(0, \sigma^2)$ for some unknown true parameter $w^* = \Sigma^{-1}\mathbb{E}[y_i x_i] \in \mathbb{R}^d$ and unknown Σ and σ^2. Then $n = \tilde{O}(d/\alpha^2 + \kappa^{1/2} d/(\varepsilon\alpha))$ samples are sufficient for Alg. 13 to achieve $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \tilde{O}(\alpha)$ with high probability, where $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$.*

Due to space constraints, we focus on sub-Gaussian distributions in the main text and provide comparisons to prior work in Tab. 5.1. Our analysis in App. D.8 applies to a more general family of *light-tailed* distributions, called sub-Weibull. Next, when the noise in the samples is *heavy-tailed*, a similar algorithm can be applied with carefully chosen clipping

Table 5.1: Suppose data is drawn from a linear model in d-dimensions from sub-Gaussian covariates with covariance Σ and sub-Gaussian noise with variance σ^2. To achieve an error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \alpha$ with (ε, δ)-DP, `DP-RobGD` requires the least number of samples among computationally efficient algorithms. This improves over [199] by a factor of $\kappa^{1/2}$ in the second term, where κ is the condition number of Σ. We hide polylogarithmic factors in d, κ and $1/\delta$. ♠DP-Theil-Sen is only analyzed when $\kappa = 1$ and its dependence κ^c is unknown.

Algorithm	Runtime	Sample Complexity
`TukeyEM` [11]	poly	no guarantee
`DP-Theil-Sen` [187] ♠	poly	$\frac{d^2}{\alpha^2} + \frac{d}{\varepsilon\alpha}\kappa^c$
`DP-AMBSSGD` [199]	poly	$\frac{d}{\alpha^2} + \frac{d}{\varepsilon\alpha}\kappa + \frac{\kappa^2 d}{\varepsilon}$
`DP-RobGD` [Theorem 5.3.7]	poly	$\frac{d}{\alpha^2} + \frac{d}{\varepsilon\alpha}\kappa^{1/2}$
`HPTR` [161], `Robust-to-private` [17]	exp	$\frac{d}{\alpha^2} + \frac{d}{\varepsilon\alpha}$
Lower Bound [41]		$\frac{d}{\alpha^2} + \frac{d}{\varepsilon\alpha}$

thresholds to account for the heavier tail. Concretely, for k-th moment bounded distributions, the tail of the distribution gets increasingly heavier with smaller k. This would require larger number of samples to achieve the same accuracy, which is captured in our sample complexity of $n = \tilde{O}(d/\alpha^{2k/(k-1)} + \kappa^{1/2}d/(\varepsilon\alpha^{k/(k-1)}))$. We explain the heavy-tailed setting, provide a detailed analysis and a proof, and discuss the results in App. D.12. This is the first efficient algorithm with provable guarantees achieving (ε, δ)-DP.

Corollary 5.1.2 (informal version of Coro. D.12.5 for heavy-tailed noise)**.** *Alg. 25 is (ε, δ)-DP. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of i.i.d. samples with $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top w^* + z_i$, and the zero-mean, independent, and heavy-tailed noise z_i satisfies $\mathbb{E}[|z/\sigma|^k] = O(1)$ for some unknown true parameter $w^* \in \mathbb{R}^d$ and unknown Σ and σ^2. Then $n = \tilde{O}(d/\alpha^{2k/(k-1)} + \kappa^{1/2}d/(\varepsilon\alpha^{k/(k-1)}))$ samples are sufficient for Alg. 25 in App. D.12 to achieve an error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \tilde{O}(\alpha)$ with high probability, where $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$.*

Perhaps surprisingly, we show that Alg. 13 is also robust against label-corruption, where an adversary selects an arbitrary α_{corrupt} fraction of the data points and changes their response variables arbitrarily. Ideally, we want a robust algorithm against a stronger adversary who can corrupt the covariates also. However, even for a simpler problem of private mean estimation, achieving robustness against such a strong adversary with $O(d)$ samples requires heavy machinery (convex relaxations of sum-of-squares optimization) with significantly more computations (although polynomial) [109].

Our lower bound in Prop. 5.3.8, together with the lower bound in [41] on the uncorrupted case, shows that $n = \Omega(d/\alpha^2 + d/(\varepsilon\alpha))$ samples are necessary to achieve an error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = O(\alpha + \alpha_{\text{corrupt}})$. In particular, it is impossible to achieve an error below α_{corrupt} even if we have infinite samples (Prop. 5.3.8), and hence there is no need to aim for $\alpha < \alpha_{\text{corrupt}}$. This lower bound is matched by exponential time approaches, HPTR in [161] and Robust-to-Private in [17], which also guarantee robustness. Currently, there is no efficient algorithm that can guarantee both privacy and robustness for linear regression. To this end, we provide the first efficient algorithm guaranteeing both, with a sample complexity that is optimal up to a $\kappa^{1/2}$ factor.

Corollary 5.1.3 (Corollary of Thm. D.8.2 for sub-Gaussian data with adversarial label corruption)**.** *Under the hypotheses of Coro. 5.1.1, suppose α_{corrupt}-fraction of the labels are corrupted arbitrarily. Then $n = \tilde{O}(d/\alpha^2 + \kappa^{1/2}d/(\varepsilon\alpha))$ samples are sufficient for Alg. 13 to achieve an error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \tilde{O}(\alpha + \alpha_{\text{corrupt}})$ with high probability, where $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$.*

When $\alpha_{\text{corrupt}} = 0$, this recovers the non-robust result from Coro. 5.1.1. A similar robustness guarantee also holds for heavy-tailed settings. We provide a formal statement in App. D.12

Contributions. For a canonical problem of private linear regression under sub-Gaussian distributions, the best known efficient algorithm [199] requires

$$n = \tilde{O}\left(\frac{d}{\alpha^2} + \frac{\kappa d}{\varepsilon\alpha} + \frac{\kappa^2 d}{\varepsilon}\right),$$

to achieve $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \alpha$. We provide the first efficient algorithm that improves this to

$$n = \tilde{O}\left(\frac{d}{\alpha^2} + \frac{\kappa^{1/2} d}{\varepsilon\alpha}\right) ,$$

which nearly matches the exponential-time algorithms [161, 17] and the lower bound [41] up to $\kappa^{1/2}$ in the second term. For the same problem, we show that the same algorithm is the first to achieve robustness against adversarial corruption of the labels.

Under a heavy-tailed distribution of the noise, we provide the first computationally efficient algorithm, to the best of our knowledge, that achieves a sample complexity close to that of an exponential-time algorithm of [161]. There is no matching lower bound in the heavy-tailed setting. This is also the first efficient algorithm to achieve robustness against adversarial corruption of the labels under heavy-tailed noise.

5.2 Problem formulation and background

When there is no adversary, we present our results under the standard linear model with sub-Gaussian covariates and noise. In App. D.8, we present a more general family of (K, a)-sub-Weibull distributions that recovers the standard sub-Gaussian family as a special case when $a = 0.5$. The necessity of such assumptions on the tail is explained in Sec. 5.3.4.

Assumption 6 (sub-Gaussian model)**.** *We have i.i.d. samples $S = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ from a distribution $\mathcal{P}_{\Sigma,w^*,\sigma^2}$ of a linear model $y_i = \langle x_i, w^*\rangle + z_i$, where the input vector x_i has zero mean $\mathbb{E}[x_i] = 0$ and a positive definite covariance $\Sigma := \mathbb{E}[x_i x_i^\top] \succ 0$, and the (input dependent) label noise z_i has zero mean $\mathbb{E}[z_i] = 0$ and variance $\sigma^2 := \mathbb{E}[z_i^2]$. We further assume $\mathbb{E}[x_i z_i] = 0$, which is equivalent to assuming that the true parameter $w^* = \Sigma^{-1}\mathbb{E}[y_i x_i]$. We assume the marginal distributions of x_i and z_i are K-sub-Gaussian with $K = O(1)$, as defined below.*

Definition 5.2.1. *$x \in \mathbb{R}^d$ is K-sub-Gaussian if for all $v \in \mathbb{R}^d$, $\mathbb{E}\left[\exp\left(\frac{\langle v,x\rangle^2}{K^2\mathbb{E}[\langle v,x\rangle^2]}\right)\right] \le 2$.*

Given a dataset S that is i.i.d. sampled from $\mathcal{P}_{\Sigma^2,w^*,\sigma^2}$ satisfying Asmp. 6, our goal is to estimate w^* that minimizes $(1/\sigma)\|\hat{w} - w^*\|_\Sigma$ which is also equivalent to minimize the excess population risk, i.e., $\mathcal{L}(w^*) - \mathcal{L}(\hat{w})$ where $\mathcal{L}(w) := \mathbb{E}_{(x,y)\sim\mathcal{P}_{\Sigma,w^*,\sigma^2}}[(y - \langle w, x\rangle)^2]$.

Notations. A vector $x \in \mathbb{R}^d$ has the Euclidean norm $\|x\|$. For a matrix M, we use $\|M\|_2$ to denote the spectral norm. The error is measured in $\|\hat{w} - w^*\|_\Sigma := \|\Sigma^{1/2}(\hat{w} - w^*)\|$ for some PSD matrix Σ. The identity matrix is denoted by $\mathbf{I}_d \in \mathbb{R}^{d\times d}$. Let $[n] = \{1, 2, \ldots, n\}$. $\tilde{O}(\cdot)$ hides some constants terms, $K = \Theta(1)$, and poly-logarithmic terms in n, d, $1/\varepsilon$, $\log(1/\delta)$, $1/\zeta$, and $1/\alpha_{\text{corrupt}}$. For a vector $x \in \mathbb{R}^d$, we define $\text{clip}_a(x) := x \cdot \min\{1, a/\|x\|\}$.

Background on DP. Differential Privacy is a standard measure of privacy leakage when data is accessed via queries, introduced by [78]. Two datasets S and S' are said to be neighbors if they differ at most by one entry, which is denoted by $S \sim S'$. A stochastic query q is said to be (ε, δ)-differentially private for some $\varepsilon > 0$ and $\delta \in [0, 1]$, if $\mathbb{P}(q(S) \in A) \le e^\varepsilon \mathbb{P}(q(S) \in A) + \delta$, for all neighboring datasets $S \sim S'$ and all subset A of the range of the query. We build upon two widely used DP primitives, the Gaussian mechanism and the private histogram. A central concept in DP mechanism design is the *sensitivity* of a query, defined as $\Delta_q := \sup_{S\sim S'} \|q(S) - q(S')\|$. We describe Gaussian mechanism and private histogram in App. D.2.

5.2.1 Comparisons with the prior work

The state-of-the-art approach introduced by [199] is based on DP-SGD [182], where privacy is ensured by gradient norm clipping and the Gaussian mechanism. Two additional technical components are adaptive clipping and streaming SGD. Adaptive clipping with an appropriate threshold θ_t ensures that no data point is clipped (under the sub-Gaussian assumption), while providing a bound on the sensitivity of the average mini-batch gradient (to ensure we do not add too much noise). The streaming approach, where each data point is only touched once and discarded, ensures independence between the past iterate w_t and the gradients at round $t + 1$, which the analysis critically relies on. For $T = \tilde{\Theta}(\kappa)$ iterations where κ is the condition number of the covariance Σ, the dataset $S = \{(x_i, y_i)\}_{i=1}^n$ is partitioned into $\{B_t\}_{t=0}^{T-1}$ subsets of equal size: $|B_t| = \tilde{\Theta}(n/\kappa)$. At each round t, the gradients are clipped and averaged with additive Gaussian noise chosen to satisfy (ε, δ)-DP:

$$w_{t+1} \leftarrow w_t - \eta\Big(\frac{1}{|B_t|}\sum_{i\in B_t} \text{clip}_{\theta_t}(x_i(w_t^\top x_i - y_i)) + \frac{\theta_t\sqrt{2\log(1.25/\delta)}}{\varepsilon|B_t|}\nu_t\Big)\,, \tag{5.1}$$

where $\nu_t \sim \mathcal{N}(0, \mathbf{I}_d)$. In [199], a slight variation of this streaming SGD is shown to achieve an error of $(1/\sigma)\|w_T - w^*\|_\Sigma = \alpha$ with $n = \tilde{O}(d/\alpha^2 + \kappa d/(\varepsilon\alpha) + \kappa^2 d/\varepsilon)$ samples (Row 3 in Tab. 5.1).

Our technical innovations. Our approach builds upon such gradient based methods but makes several important innovations. First, we use full-batch gradient descent, as opposed to the streaming SGD above. Using all n samples reduces the sensitivity of the per-round gradient average by a κ factor, and thus decreases the privacy noise added in each iteration. This improves the second term of sample complexity from $\kappa d/(\varepsilon\alpha)$ to $\kappa^{1/2} d/(\varepsilon\alpha)$ and removes the third term completely. However, full-batch GD loses the independence that the streaming SGD enjoyed between w_t and the samples used in the round $t+1$. This dependence makes the analysis more challenging. We instead propose using the *resilience* to precisely track the bias and variance of the (dependent) full-batch average gradient. Resilience is a central concept in robust statistics that links the tail-property of the distribution to the bias, which we explain in Sec. 5.5.

Next, one critical component in achieving this improved sample complexity is the new analysis technique we introduce for tracking the end-to-end gradient updates. Since our gradient descent algorithm is not guaranteed to make progress every step, we cannot use the vanilla one-step analysis. Taking the full end-to-end analysis by expanding the whole gradient trajectory will introduce too many correlated cross-terms which are very hard to control. Therefore, we leverage an every κ-step analysis and show that the objective function at least decreases geometrically every κ steps. To be more specific, our analysis technique in App. D.8 (steps 3 and 4) opens up the iterative updates from the beginning to the end, and exploits the fact that $\lambda_{\max}((\eta\Sigma)^{1/2}(1-\eta\Sigma)^i(\eta\Sigma)^{1/2})$ is upper bounded by $1/(i+1)$ when $\|\eta\Sigma\| \le 1$. This technique is critical in achieving the near-optimal dependence in κ. This might be of independent interest to other analysis of gradient-based algorithms. We refer to the beginning of step 3 in App. D.8 for a detailed explanation.

Finally, we propose a novel clipping that separately clips x_i and $(w_t^\top x_i - y_i)$ in the gradient, $(w_t^\top x_i - y_i)x_i$. This is critical in achieving robustness to label-corruption, as we explain in

Sec. 5.3.1.

5.3 Label-robust and private linear regression

We introduce a novel gradient descent approach. This achieves an improved sample complexity compared to the state-of-the-art algorithm and robustness against label corruption.

5.3.1 Algorithm

The skeleton of our approach in Alg. 13 is the general DP-SGD [1, 182] with adaptive clipping [12]. We partition the dataset into three equal-sized subsets: S_1, S_2, S_3. S_1 and S_2 are used in adaptively estimating the clipping thresholds, and S_3 is re-used every step to compute the average gradient.

The standard adaptive clipping, e.g., [12, 199], is not robust against label-corruption. Under sub-Gaussian distribution, a positive fraction of the covariates, x_i's, can be close to the origin. If the adversary chooses to corrupt those points with small norm, $\|x_i\|$, they can make large changes in the corrupted residual, $(y_i - w_t^\top x_i)$, while evading the standard clipping by the norm of the gradient; the norm of the gradient, $\|x_i(y_i - w_t^\top x_i)\| = \|x_i\| \, |y_i - w_t^\top x_i|$, can remain under the threshold. This is problematic, since the bias due to the corrupted samples in the gradient scales proportionally to the magnitude of the residual (after clipping). To this end, we propose clipping the norm and the residual separately: $\text{clip}_\Theta(x_i)\text{clip}_{\theta_t}\left(w_t^\top x_i - y_i\right)$. This keeps the sensitivity of gradient average bounded by $\Theta(\theta_t)$. The subsequent Gaussian mechanism in line 11 ensures $(\varepsilon_0, \delta_0)$-DP at each round. Applying advanced composition in Lemma 2.3.4 of T rounds, this ensures end-to-end (ε, δ)-DP.

Novel adaptive clipping. When clipping with $\text{clip}_\Theta(x_i)$, the only purpose of clipping the covariate by its norm, $\|x_i\|$, is to bound the sensitivity of the resulting clipped gradient. In particular, we do not need to make it robust as there is no corruption in the covariates. Ideally, we want to select the smallest threshold Θ that does not clip any of the covariates. Since the norm of a covariate is upper bounded by $\|x_i\|^2 \leq K^2 \text{Tr}(\Sigma) \log(1/\zeta)$ with probability $1 - \zeta$ (Lemma D.10.3), we estimate the unknown $\text{Tr}(\Sigma)$ using Private Norm Estimator in

Alg. 24 in App. D.6 and set the norm threshold $\Theta = K\sqrt{2\Gamma \log(n/\zeta)}$ (Alg. 13 line 4). The n in the logarithm ensures that the union bound holds.

When clipping with $\text{clip}_{\theta_t}(w_t^\top x_i - y_i)$, the purpose of clipping the residual by its magnitude, $|y_i - w_t^\top x_i| = |(w^* - w_t)^\top x_i + z_i|$, is to bound the sensitivity of the gradient and also to provide robustness against label-corruption. We want to choose a threshold that only clips corrupt data points and at most a few clean data points. In order to achieve an error $(1/\sigma)\|w_T - w^*\|_\Sigma = \alpha$, we know that any set of $(1-\alpha)$ fraction of the clean data points is sufficient to get a good estimate of the average gradient. By clipping at $|(w^* - w_t)^\top x_i + z_i|^2 \leq (\|w_t - w^*\|_\Sigma^2 + \sigma^2) C K^2 \log(1/(2\alpha))$, Lemma D.10.3 guarantees that the unclipped subset will be large enough, i.e., $(1-\alpha)n$. At the same time, this threshold on the residual is small enough to guarantee robustness against the label-corrupted samples. We introduce the robust and DP Distance Estimator in Alg. 23 to estimate the unknown (squared and shifted) distance, $\|w_t - w^*\|_\Sigma^2 + \sigma^2$, and set the distance threshold $\theta_t = 2\sqrt{2\gamma_t}\sqrt{9C_2K^2\log(1/(2\alpha))}$ (Alg. 13 line 7). Both norm and distance estimation rely on DP histogram (Lemma A.2.1), but over a set of statistics computed on partitioned datasets, which we explain in detail in App. D.3.

5.3.2 Analysis without adversarial corruption

We show that Alg. 13 achieves an improved sample complexity. We provide the proof for a more general class of distributions in App. D.8 and a sketch of the proof in Sec. 5.5. We address the necessity of the assumptions in Sec. 5.3.4, along with some lower bounds.

Theorem 5.3.1. *Alg. 13 is (ε, δ)-DP. Under sub-Gaussian model of Asmp. 6, for any failure probability $\zeta \in (0,1)$ and target error rate α, if the sample size is large enough such that*

$$n = \tilde{O}\left(K^2 d \log^2\left(\frac{1}{\zeta}\right) + \frac{d + \log(1/\zeta)}{\alpha^2} + \frac{K^2 d T^{1/2} \log(\frac{1}{\delta})\sqrt{\log(\frac{1}{\zeta})}}{\varepsilon\alpha} \right), \tag{5.2}$$

with a large enough constant, then the choices of a step size $\eta = 1/(C\lambda_{\max}(\Sigma))$ for any $C \geq 1.1$ and the number of iterations, $T = \tilde{\Theta}\left(\kappa \log\left(\|w^*\|\right)\right)$ for a condition number of the

Algorithm 13: Robust and Private Linear Regression

Input: $S = \{(x_i, y_i)\}_{i=1}^{3n}$, DP parameters (ε, δ), T, learning rate η, failure probability ζ, target error α, distribution parameter K

Partition dataset S into three equal sized disjoint subsets $S = S_1 \cup S_2 \cup S_3$.

$\delta_0 \leftarrow \frac{\delta}{2T}$, $\varepsilon_0 \leftarrow \frac{\varepsilon}{4\sqrt{T\log(1/\delta_0)}}$, $\zeta_0 \leftarrow \frac{\zeta}{3}$, $w_0 \leftarrow 0$

$\Gamma \leftarrow \text{PrivateNormEstimator}(S_1, \varepsilon_0, \delta_0, \zeta_0)$ `// using Alg. 24, App. D.6`

$\Theta \leftarrow K\sqrt{2\Gamma}\log^a(n/\zeta_0)$

for $t = 0, 1, 2, \ldots, T-1$ **do**

  $\gamma_t \leftarrow \text{PrivateDistanceEstimator}(S_2, w_t, \varepsilon_0, \delta_0, \alpha, \zeta_0)$ `// using Alg. 23, App. D.3`

  $\theta_t \leftarrow 2\sqrt{2\gamma_t} \cdot \sqrt{9C_2K^2\log(1/(2\alpha))}$.

  Sample $\nu_t \sim \mathcal{N}(0, \mathbf{I}_d)$

  $\tilde{g}_i^{(t)} \leftarrow \text{clip}_\Theta(x_i)\text{clip}_{\theta_t}(x_i^\top w_t - y_i)$

  $\phi_t = (\sqrt{2\log(1.25/\delta_0)}\Theta\theta_t)/(\varepsilon_0 n)$

  $w_{t+1} \leftarrow w_t - \eta\left(\frac{1}{n}\sum_{i\in S_3}\tilde{g}_i^{(t)} + \phi_t\nu_t\right)$

Return w_T

covariance $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$*, ensures that, with probability* $1-\zeta$*, Alg. 13 achieves*

$$\mathbb{E}_{\nu_1,\cdots,\nu_T\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\,\|w_T - w^*\|_\Sigma^2\,\right] = \tilde{O}\Big(\,K^4\sigma^2\alpha^2\log^2\left(\frac{1}{\alpha}\right)\Big)\,, \tag{5.3}$$

where the expectation is taken over the noise added for DP, and $\tilde{O}$ *and* $\tilde{\Theta}(\cdot)$ *hide logarithmic terms in* $K, \sigma, d, n, 1/\varepsilon, \log(1/\delta), 1/\alpha$*, and* κ*.*

Remark 5.3.2. *Omitting some constant and logarithmic terms, Alg. 13 requires*

$$n \;=\; \tilde{O}\Big(\,\frac{d}{\alpha^2} + \frac{\kappa^{1/2}d}{\varepsilon\alpha}\Big)\,, \tag{5.4}$$

samples to ensure an error rate of $(1/\sigma^2)\mathbb{E}[\|w_T - w^*\|_\Sigma^2] = \tilde{O}(\alpha^2)$*. From [41, Theorem 3.1], there exists an* $n = \Omega(d/\alpha^2 + d/(\varepsilon\alpha))$ *lower bound, and our upper bound matches this lower bound up to a factor of* $\kappa^{1/2}$ *in the second term and other logarithmic factors.* (5.4) *is the best known rate among all efficient private linear regression algorithms, strictly improving upon the state-of-the-art. The best existing efficient algorithm by [199] requires* $n = \tilde{O}(d/\alpha^2 + \kappa d/(\varepsilon\alpha) + \kappa^2 d/\varepsilon)$ *to achieve the same error rate. Compared to* (5.4)*, the second term is larger by a factor of* $\kappa^{1/2}$ *compared to the second term in* (5.4)*. Further, [199] requires* $\kappa^2 d/\varepsilon$*, which is not needed in* (5.4)*.*

Remark 5.3.3. *Consider the standard settings of linear regression with* $x_i \sim \mathcal{N}(0, \mathbf{I}_d)$ *and* $z_i \sim \mathcal{N}(0, \sigma^2)$ *such that the condition number is one, our bound given by Eq* (5.4) *nearly matches the lower bound ([41, Theorem 3.1]) up to logarithmic factors.*

Remark 5.3.4. *Note that the leading term in Eq* (5.4) *is the first term* d/α^2 *when target error* $\alpha \le \varepsilon/\kappa^{1/2}$*. Our first term is independent of* κ*, which matches the lower bound for non-private linear regression.*

Remark 5.3.5. *The third term* $\kappa^2 d/\varepsilon$ *in* [199] *is independent of error rate* α *but scales as* κ^2*. This term is required to ensure the privacy noise added in each iteration is small enough for their DP-SGD to make progress (Appendix. B.2.2 in [199]). Our algorithm is based on full-batch gradient descent, which uses all* n *samples and thus reduces the sensitivity of*

gradient average by a κ factor. As a result, we show in (D.51) that our algorithm only requires $n = \tilde{O}((1/\varepsilon)\sqrt{\kappa^{1/2}d/\alpha})$ to make progress for each iteration. This is strictly smaller than our dominant term $\kappa^{1/2}d/(\varepsilon\alpha)$ and does not show up in our final guarantee. We provide a formal proof in App. D.8.

Remark 5.3.6. *One of the key innovations in Alg. 13 is the adaptive distance estimator (Alg. 23 in App. D.3). The goal is to privately estimate the (shifted) distance of the current estimate, i.e., $\|w_t - w^*\|_\Sigma + \sigma^2$, without the knowledge of w^*. We show in Thm. D.3.1 that our novel distance estimator only requires an error-independent sample complexity $n = \tilde{O}(\kappa^{1/2}d/\varepsilon)$ to achieve a constant multiplicative error. Note that the DP-STAT (Algorithm 3 in [199]) can also be used to estimate the distance. But it requires the knowledge of domain size $\|w^*\|_\Sigma + \sigma$. We completely remove this requirement, improve the dependence on K and $\log(n)$, and show it is also robust, as introduced in the next section. We provide the algorithms and analysis in App. D.3 and the formal proof in App. D.4.*

5.3.3 Robustness against label corruption

We assume there exists a good dataset S_{good} that satisfies Asmp. 6. We only get access to a label-corrupted dataset under the standard definition of *label corruption*, e.g., [31]. There are variations in literature on the definition, which we survey in App. D.1.

Assumption 7 (α_{corrupt}-corruption)**.** *Given a dataset $S_{\text{good}} = \{(x_i, y_i)\}_{i=1}^n$, an adversary inspects all the data points, selects $\alpha_{\text{corrupt}}n$ data points denoted as S_r, and replaces the labels with arbitrary labels while keeping the covariates unchanged. We let S_{bad} denote this set of $\alpha_{\text{corrupt}}n$ newly labelled examples by the adversary. Let the resulting set be $S_{\text{corrupt}} := S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$.*

Our goal is to estimate the unknown parameter w^*, given corrupted dataset S_{corrupt}, distribution parameter K, and (an upper bound on) the corruption level α_{corrupt}.

Under the *non-private scenario*, i.e., $\varepsilon = \infty$, recent advances led to optimal algorithms for linear regression that are robust to label corruptions [31, 55]; if the corruption level is

smaller than the target error rate, i.e., $\alpha_{\text{corrupt}} \leq \alpha$, then $n = \tilde{O}(d/\alpha^2)$ samples are sufficient to achieve an error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = \alpha$. The sample complexity of d/α^2 is optimal as it matches the information theoretic lower bound. The condition $\alpha_{\text{corrupt}} \leq \alpha$ is necessary since it is information theoretically impossible to achieve error α less than α_{corrupt}, as we prove in Prop. 5.3.8. Setting the target error to the minimum possible value of $\alpha = \alpha_{\text{corrupt}}$, we say that these algorithms achieve optimal robustness since the minimum robust error rate of $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = O(\alpha_{\text{corrupt}})$ can be achieved with minimal sample complexity of $n = \tilde{O}(d/\alpha^2_{\text{corrupt}})$. We aim to achieve such optimal robustness simultaneously with differential privacy in a computationally efficient manner.

Theorem 5.3.7. *Under sub-Gaussian model of Asmp. 6 and α_{corrupt}-corruption of Asmp. 7, if the corruption level is below the target error rate, $\alpha \geq \alpha_{\text{corrupt}}$, then $n = \tilde{O}(d/\alpha^2 + \kappa^{1/2}d/(\varepsilon\alpha))$ samples are sufficient for Alg. 13 to achieve an error rate of $(1/\sigma^2)\mathbb{E}[\|\hat{w} - w^*\|^2_\Sigma] = \tilde{O}(\alpha^2)$.*

This is the first efficient approach to achieve robustness and (ε, δ)-DP simultaneously. The existing such algorithms take exponential time [161, Corollary C.2] and [17], but achieve optimal sample complexity of $n = O(d/\alpha^2 + d/(\varepsilon\alpha))$. Notice that there is no dependence on κ. It remains an open question if *computationally efficient* private linear regression algorithms can achieve such an optimal κ-independent sample complexity. We make the first advance towards this ambitious goal with the above theorem. Our sample complexity is sub-optimal only by a factor of $\sqrt{\kappa}$ in the second term. This is achieved by individually clipping the covariate, x_i, and the residual, $(w_t^\top x_i - y_i)$, in Alg. 13 and carefully tracking the bias of clipping with the use of resilience in the analysis in App. D.8.

5.3.4 Lower bounds

Necessity of our assumptions. A tail assumption on the covariate x_i such as Asmp. 6 is necessary to achieve $n = O(d)$ sample complexity in (5.4). Even when the covariance Σ is close to identity, without further assumptions on the tail of covariate x, the result in [27] implies that for $\delta < 1/n$, it is necessary for an (ε, δ)-DP estimator to have $n = \Omega(d^{3/2}/(\varepsilon\alpha))$

samples to achieve $\|\hat{w} - w^*\|_\Sigma = \tilde{O}(\alpha)$ (see Eq. (3) in [206]). Note that this lower bound is a factor $d^{1/2}$ larger than our upper bound that benefits from the additional tail assumption.

A tail assumption on the noise z_i such as Asmp. 6 is necessary to achieve $n = O(d/(\varepsilon\alpha))$ dependence on the sample complexity in (5.4). For heavy-tailed noise, such as k-th moment bounded noise, the dependence can be significantly larger. [161, Proposition C.5] implies that for $\delta = e^{-\Theta(d)}$ and 4-th moment bounded x_i and z_i, any (ε, δ)-DP estimator requires $n = \Omega(d/(\varepsilon\alpha^2))$, which is a factor of $1/\alpha$ larger, to achieve $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma = \tilde{O}(\alpha)$.

The assumption that only labels are corrupted is critical for Alg. 13. The average of the clipped gradients can be significantly more biased, if the adversary can place the covariates of the corrupted samples in the same direction. In particular, the bound on the bias of our gradient step in (D.36) in App. D.8 would no longer hold. Against such strong attacks, one requires additional steps to estimate the mean of the gradients robustly and privately, similar to those used in robust private mean estimation [160, 146, 108, 15]. There is no known linear-time algorithm to achieve this, and this is outside the scope of this work.

Lower bounds under label corruption. Under the α_{corrupt} label corruption setting (Asmp. 7), even with infinite data and without privacy constraints, no algorithm is able to learn w^* with ℓ_2 error better than α_{corrupt}. We provide a formal derivation for completeness.

Proposition 5.3.8. *Let $\mathcal{D}_{\Sigma,\sigma^2,w^*,K}$ be a class of distributions on (x_i, y_i) from sub-Gaussian model in Asmp. 6. Let $S_{n,\alpha}$ be an α-corrupted dataset of n i.i.d. samples from some distribution $\mathcal{D} \in \mathcal{D}_{\Sigma,\sigma^2,w^*,K}$ under Asmp. 7. Let $\mathcal{M}$ be a class of estimators that are functions over $S_{n,\alpha}$. Then there exists a constant c such that* $\min_{n,\hat{w}\in\mathcal{M}} \max_{S_{n,\alpha},\mathcal{D}\in\mathcal{D}_{\Sigma,\sigma^2,w^*,K},w^*,K} \mathbb{E}[\|\hat{w} - w^*\|_\Sigma^2] \geq c\,\alpha^2\,\sigma^2$.

A proof is provided in App. D.9.1. A similar lower bound can be found in [22, Theorem 6.1].

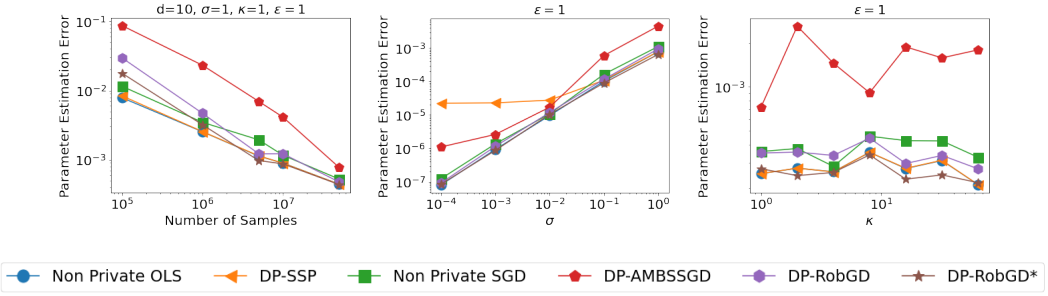

Figure 5.1: Performance of various techniques on DP linear regression. $d = 10$ in all the experiments. $n = 10^7, \kappa = 1$ in the 2^{nd} experiment. $n = 10^7, \sigma = 1$ in the 3^{rd} experiment, where κ is the condition number of Σ and σ^2 is the variance of the label noise z_i.

5.4 Experimental results

5.4.1 DP Linear Regression

We present experimental results comparing our proposed technique (DP-RobGD) with other baselines. We consider non-corrupted regression in this section and defer corrupted regression to the App. D.11. We begin by describing the problem setup and the baseline algorithms first.

Experiment Setup. We generate data for all the experiments using the following generative model. The parameter vector w^* is uniformly sampled from the surface of a unit sphere. The covariates $\{x_i\}_{i=1}^n$ are first sampled from $\mathcal{N}(0, \Sigma)$ and then projected to unit sphere. We consider diagonal covariances Σ of the following form: $\Sigma[0,0] = \kappa$, and $\Sigma[i,i] = 1$ for all $i \geq 1$. Here $\kappa \geq 1$ is the condition number of Σ. We generate noise z_i from uniform distribution over $[-\sigma, \sigma]$. Finally, the response variables are generated as follows $y_i = x_i^\top w^* + z_i$. All the experiments presented below are repeated 5 times and the averaged results are presented. We set the DP parameters (ϵ, δ) as $\epsilon = 1, \delta = \min(10^{-6}, n^{-2})$. Experiments for $\epsilon = 0.1$ can be found in Fig. D.1 in the App. D.11.

Baseline Algorithms. We compare our estimator with the following baseline algorithms:

- *Non private algorithms:* ordinary least squares (β_{ols}), one-pass stochastic gradient descent with tail-averaging (SGD). For SGD, step-size is $1/(2\lambda_{\max})$ and minibatch size is n/T,

where $T = 3\kappa \log n$.

- *Private algorithms:* sufficient statistics perturbation (DP-SSP) [88, 206], differentially private stochastic gradient descent (DP-AMBSSGD) [199]. DP-SSP had the best empirical performance among numerous techniques studied by [206], and DP-AMBSSGD has the best known theoretical guarantees. The DP-SSP algorithm involves releasing $X^T X$ and $X^T \mathbf{y}$ differentially privately and computing $\widehat{(X^T X)}^{-1} \widehat{X^T \mathbf{y}}$. DP-AMBSSGD is a private version of SGD where the DP noise is set adaptively according to the excess error in each iteration. For both algorithms, we use the hyper-parameters recommended in their respective papers. To improve the performance of DP-AMBSSGD, we reduce the theoretical clipping threshold by a constant factor.

DP-RobGD. We implement Alg. 13 with the following key changes. Instead of relying on PrivateNormEstimator to estimate Γ, we set it to its true value $\mathrm{Tr}(\Sigma)$. This is done for a fair comparison with DP-AMBSSGD which assumes the knowledge of $\mathrm{Tr}(\Sigma)$. Next, we use 20% of the samples to compute γ_t in line 5 (instead of the 50% stated in Alg. 13). In our experiments we also present results for a variant of our algorithm called DP-RobGD* which outputs the best iterate based on γ_t, instead of the last iterate. One could also perform tail-averaging instead of picking the best iterate. Both these modifications are primarily used to reduce the variance in the output of Alg. 13 and achieved similar performance in our experiments.

Results. Figure 5.1 presents the performance of various algorithms as we vary n, κ, σ. It can be seen that DP-RobGD outperforms DP-AMBSSGD in almost all the settings (and DP-RobGD* outperforms DP-RobGD in all cases). DP-SSP has poor performance when the noise σ is low, but performs slightly better than DP-RobGD in other settings. A major drawback of DP-SSP is its computational complexity which scales as $O(nd^2 + d^\omega)$. In contrast, the computational complexity of DP-RobGD has smaller dependence on d and scales as $\tilde{O}(nd\kappa)$. Thus the latter is more computationally efficient for high-dimensional problems. More experimental results on both robust and private linear regression can be found in the

App. D.11.

5.5 Sketch of the main ideas in the analysis

We provide the main ideas behind the proof of Thm. 5.3.1. The privacy proof is straightforward since no matter what clipping threshold we use the noise we add is always proportionally to the clipping threshold which guarantees privacy. In the remainder, we focus on the utility analysis.

The proof of the utility heavily relies on the *resilience* [185] (also known as *stability* [67]), which states that given a large enough sample set S, various statistics (for example, sample mean and sample variance) of any large enough subset of S will be close to each other. We define resilience as follows.

Definition 5.5.1 ([161, Definition 23])**.** *For some $\alpha \in (0,1)$, $\rho_1 \in \mathbb{R}_+$, $\rho_2 \in \mathbb{R}_+$, and $\rho_3 \in \mathbb{R}_+$, $\rho_4 \in \mathbb{R}_+$, we say dataset $S_{\text{good}} = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ is $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient with respect to (w^*, Σ, σ) for some $w^* \in \mathbb{R}^d$, positive definite $\Sigma \succ 0 \in \mathbb{R}^{d\times d}$, and $\sigma > 0$ if for any $T \subset S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$, the following holds for all $v \in \mathbb{R}^d$:*

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\langle v, x_i\rangle(y_i - x_i^\top w^*)\right| \leq \rho_1\sqrt{v^\top\Sigma v}\,\sigma \quad , \tag{5.5}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle^2 - v^\top\Sigma v\right| \leq \rho_2 v^\top\Sigma v \ , \tag{5.6}$$

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}(y_i - x_i^\top w^*)^2 - \sigma^2\right| \leq \rho_3\sigma^2 \ , \tag{5.7}$$

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\langle v, x_i\rangle\right| \leq \rho_4\sqrt{v^\top\Sigma v} \ . \tag{5.8}$$

We give an overview of the proof for non-robust case as follows. First, we introduce some notations. Let $g_i^{(t)} := (x_i^\top w_t - y_i)x_i$ be the raw gradient and $\tilde{g}_i^{(t)} := \text{clip}_\Theta(x_i)\text{clip}_{\theta_t}(x_i^\top w_t - y_i)$ be the clipped gradient. Note that when the data follows from our distributional assumption, with high probability, samples are not clipped by the norm: $\text{clip}_\Theta(x_i) = x_i$. We can write

down one step of gradient update (see Alg. 13) as follows:

$$w_{t+1} - w^* = \underbrace{\left(\mathbf{I} - \frac{\eta}{n}\sum_{i\in S} x_i x_i^\top\right)(w_t - w^*)}_{(i)} + \underbrace{\frac{\eta}{n}\sum_{i\in S} x_i z_i}_{(ii)} + \underbrace{\frac{\eta}{n}\sum_{i\in S}(g_i^{(t)} - \tilde{g}_i^{(t)})}_{(iii)} - \underbrace{\eta\phi_t\nu_t}_{(iv)} .$$

In the above equation, the first term is a contraction, meaning w_t is moving toward w^*. The second term captures the noise from the randomness in the samples. The third term captures the bias introduced by the clipping operation, and the fourth term captures the added noise for privacy. The second term is standard and relatively easy to control, and our main focus is on the last two terms.

The third term $(\eta/n)\sum_{i\in S}(g_i^{(t)} - \tilde{g}_i^{(t)})$ can be controlled using the resilience property. We prove that with our estimated threshold, the clipping will only affect a small amount of datapoints, whose contribution to the gradient is small collectively.

Now we have controlled the deterministic bias. Then, we upper bound the fourth term, which is the noise for the purpose of privacy, and show the expected prediction error decrease in every gradient step. The difficulty is that, since our clipping threshold is adaptive, the decrease of the estimation error depends on the estimation error of all the previous steps. This causes that in some iterations, the estimation error actually increases. In order to get around this, we split the iterations into length κ chunks, and argue that the maximum estimation error in a chunk must be a constant factor smaller than the previous chunk. This implies we will reach the desired error within $\tilde{O}(\kappa)$ steps.

5.6 Discussion

We provide a novel variant of DP-SGD algorithm for differentially private linear regression under label corruption. We show the first near-optimal rate that achieves privacy and robustness to label corruptions simultaneously. When there is no label corruption, our result also improves upon the state-of-the-art method [199] in terms of the condition number κ. Compared to [199], our algorithm has two innovations: 1) we introduce a novel adaptive clipping, which is critical in achieving robustness against label corruptions; and 2) we use

full batch gradient descent and a novel convergence analysis to get the near-optimal sample complexity.

BIBLIOGRAPHY

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018. KDD Invited Talk.

[3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

[4] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. *arXiv preprint arXiv:2010.09929*, 2020.

[5] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust, and optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 483–496, 2023.

[7] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *Proceedings on Privacy Enhancing Technologies*, 2022.

[8] Zeyuan Allen-Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and regret minimization beyond matrix multiplicative updates. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 237–245, 2015.

[9] Edoardo Amaldi and Viggo Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical computer science*, 147(1-2):181–210, 1995.

[10] Kareem Amin, Travis Dick, Alex Kulesza, Andrés Munoz Medina, and Sergei Vassilvitskii. Differentially private covariance estimation. In *NeurIPS*, pages 14190–14199, 2019.

[11] Kareem Amin, Matthew Joseph, Mónica Ribero, and Sergei Vassilvitskii. Easy differentially private linear regression. *arXiv preprint arXiv:2208.07353*, 2022.

[12] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*, 2021.

[14] Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.

[15] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pages 1075–1076. PMLR, 2022.

[16] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

[17] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*, 2023.

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

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

[20] Marco Avella-Medina and Victor-Emmanuel Brunel. Differentially private sub-gaussian location estimators. *arXiv preprint arXiv:1906.11923*, 2019.

[21] Ainesh Bakshi and Pravesh Kothari. List-decodable subspace recovery via sum-of-squares. *arXiv preprint arXiv:2002.05139*, 2020.

[22] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

[23] S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 169–212, 2017.

[24] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309. PMLR, 2016.

[25] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[26] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[28] Amos Beimel, Shay Moran, Kobbi Nissim, and Uri Stemmer. Private center points and learning of halfspaces. *arXiv preprint arXiv:1902.10731*, 2019.

[29] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 2107–2116, 2017.

[30] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

[31] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28:721–729, 2015.

[32] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*, 2020.

[33] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

[34] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34:7950–7964, 2021.

[35] Victor-Emmanuel Brunel and Marco Avella-Medina. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

[36] Mark Bun, Gautam Kamath, Thomas Steinke, and Steven Z Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems*, pages 156–167, 2019.

[37] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *ITCS*, pages 369–380, 2016.

[38] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *J. Mach. Learn. Res.*, 20:94–1, 2019.

[39] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *arXiv preprint arXiv:1906.02830*, 32, 2019.

[40] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[41] T Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

[42] Clément L Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. *arXiv preprint arXiv:1905.11947*, 2019.

[43] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[44] Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1327. NIH Public Access, 2012.

[45] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[46] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14(1):2905–2943, 2013.

[47] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[48] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[49] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms*, pages 2755–2771. SIAM, 2019.

[50] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pages 727–757. PMLR, 2019.

[51] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

[52] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. List decodable mean estimation in nearly linear time. *arXiv preprint arXiv:2005.09796*, 2020.

[53] Christopher A. Choquette-Choo, Krishnamurthy Dvijotham, Krishna Pillutla, Arun Ganesh, Thomas Steinke, and Abhradeep Thakurta. Correlated noise provably beats independent noise for differentially private learning, 2023.

[54] Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

[55] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, volume 32, pages 13188–13198, 2019.

[56] Jules Depersin. A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*, 2020.

[57] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[58] Jules Depersin and Guillaume Lecué. On the robustness to adversarial corruption and to heavy-tailed data of the stahel-donoho median of means. *arXiv preprint arXiv:2101.09117*, 2021.

[59] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[60] Aditya Dhar and Jason Huang. Designing differentially private estimators in high dimensions. *arXiv preprint arXiv:2006.01944*, 2020.

[61] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

[62] Ilias Diakonikolas, Gautam Kamath, Daniel 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.

[63] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and

Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019.

[64] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being Robust (in High Dimensions) Can Be Practical. *arXiv e-prints*, page arXiv:1703.00893, March 2017.

[65] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

[66] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

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

[68] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[69] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

[70] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[71] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

[72] Wei Dong and Ke Yi. Universal private estimators. *arXiv preprint arXiv:2111.02598*, 2021.

[73] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32:6067–6077, 2019.

[74] David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston., 1982.

[75] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191. PMLR, 2019.

[76] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[77] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[78] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[79] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[80] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[81] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[82] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[83] Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Tight and robust private mean estimation with few users. *arXiv preprint arXiv:2110.11876*, 2021.

[84] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 2016(3):41–61, 2016.

[85] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[86] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964. IEEE, 2022.

[87] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544. PMLR, 2018.

[88] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

[89] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: z-test and tight confidence intervals. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2545–2554. PMLR, 2019.

[90] Arun Ganesh, Mahdi Haghifam, Thomas Steinke, and Abhradeep Thakurta. Faster differentially private convex optimization via second-order methods. *arXiv preprint arXiv:2305.13209*, 2023.

[91] Arun Ganesh, Daogao Liu, Sewoong Oh, and Abhradeep Thakurta. Private (stochastic) non-convex optimization revisited: Second-order stationary points and excess risks. *arXiv preprint arXiv:2302.09699*, 2023.

[92] Arpita Gang, Bingqing Xiang, and Waheed U Bajwa. Distributed principal subspace analysis for partitioned big data: Algorithms, analysis, and implementation. *IEEE Transactions on Signal and Information Processing over Networks*, 7:699–715, 2021.

[93] Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

[94] Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

[95] Quan Geng and Pramod Viswanath. The optimal mechanism in differential privacy. In *2014 IEEE international symposium on information theory*, pages 2371–2375. IEEE, 2014.

[96] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. Robust and private learning of halfspaces. *arXiv preprint arXiv:2011.14580*, 2020.

[97] Antonious M Girgis, Deepesh Data, Suhas Diggavi, Ananda Theertha Suresh, and Peter Kairouz. On the renyi differential privacy of the shuffle model. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2321–2341, 2021.

[98] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

[99] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 1986.

[100] Moritz Hardt. Robust subspace iteration and privacy-preserving spectral analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1624–1626. IEEE, 2013.

[101] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

[102] Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 1255–1268, 2012.

[103] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340, 2013.

[104] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Defense against backdoor attacks via robust covariance estimation. In *International Conference on Machine Learning*, pages 4129–4139. PMLR, 2021.

[105] Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[107] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. *arXiv preprint arXiv:2111.12981*, 2021.

[108] 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 SIGACT Symposium on Theory of Computing*, pages 1406–1417, 2022.

[109] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023.

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

[111] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[112] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*, 2021.

[113] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pages 2463–2498. PMLR, 2021.

[114] Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. *arXiv preprint arXiv:2106.00463*, 2021.

[115] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[116] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[117] Hafiz Imtiaz and Anand D Sarwate. Differentially private distributed principal component analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2206–2210. IEEE, 2018.

[118] Hafiz Imtiaz and Anand D Sarwate. Distributed differentially private algorithms for matrix and tensor factorization. *IEEE journal of selected topics in signal processing*, 12(6):1449–1464, 2018.

[119] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pages 1147–1164. PMLR, 2016.

[120] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34, 2021.

[121] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33:15689–15701, 2020.

[122] He Jia and Santosh Vempala. Robustly clustering a mixture of gaussians. *arXiv preprint arXiv:1911.11838*, 2019.

[123] Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Steven Z Wu. Locally private gaussian estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

[124] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[125] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, pages 2879–2887, 2014.

[126] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014.

[127] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385, 2015.

[128] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[129] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902, 2019.

[130] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[131] 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.

[132] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, 35:24405–24418, 2022.

[133] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Conference on Learning Theory*, pages 544–572. Proceedings of Machine Learning Research, 2022.

[134] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated gaussians. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–62. IEEE, 2020.

[135] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. *arXiv preprint arXiv:2002.09464*, 2020.

[136] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1395–1414. SIAM, 2013.

[137] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 7423–7432, 2019.

[138] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. In *2nd Symposium on Simplicity in Algorithms*, 2019.

[139] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[140] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[141] Ashish Khetan, Zachary C Lipton, and Animashree Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[142] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

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

[144] Weihao Kong, Rajat Sen, Pranjal Awasthi, and Abhimanyu Das. Trimmed maximum likelihood estimation for robust learning in generalized linear models. *arXiv preprint arXiv:2206.04777*, 2022.

[145] Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *Advances in Neural Information Processing Systems*, 33, 2020.

[146] Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Conference on Learning Theory*, pages 723–777. Proceedings of Machine Learning Research, 2022.

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

[148] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

[149] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

[150] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[151] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

[152] Jing Lei. Differentially private m-estimators. *Advances in Neural Information Processing Systems*, 24:361–369, 2011.

[153] Jerry Li. CSE 599-M, Lecture Notes: Robustness in Machine Learning , 2019. URL: `https://jerryzli.github.io/robust-ml-fall19/lec7.pdf`.

[154] Jerry Li and Guanghao Ye. Robust gaussian covariance estimation in nearly-matrix multiplication time. *Advances in Neural Information Processing Systems*, 33, 2020.

[155] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

[156] Xiaohui Liu. Fast implementation of the tukey depth. *Computational Statistics*, 32(4):1395–1410, 2017.

[157] Xiaohui Liu, Karl Mosler, and Pavlo Mozharovskyi. Fast computation of tukey trimmed regions and median in dimension p$>$ 2. *Journal of Computational and Graphical Statistics*, 28(3):682–697, 2019.

[158] Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[159] Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. DP-PCA: Statistically optimal and differentially private pca. In *Advances in Neural Information Processing Systems*, 2022.

[160] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34:3887–3901, 2021.

[161] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. Proceedings of Machine Learning Research, 2022.

[162] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019.

[163] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[164] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[165] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[166] Jason Milionis, Alkis Kalavasis, Dimitris Fotakis, and Stratis Ioannidis. Differentially private regression with unbounded covariates. In *International Conference on Artificial Intelligence and Statistics*, pages 3242–3273. Proceedings of Machine Learning Research, 2022.

[167] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964, 2016.

[168] Darakhshan J Mir. *Differential privacy: an exploration of the privacy-utility landscape*. Rutgers The State University of New Jersey-New Brunswick, 2013.

[169] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[170] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 313–322, 2019.

[171] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[172] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[173] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[174] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

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

[176] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

[177] Peter J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(37):283–297, 1985.

[178] Holger Sambale. Some notes on concentration for α-subexponential random variables. *arXiv preprint arXiv:2002.10761*, 2020.

[179] Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *Journal of Machine Learning Research*, 16:3475–3486, 2015.

[180] Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.

[181] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

[182] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

[183] Shuang Song, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783*, 2020.

[184] Werner A Stahel. *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zurich, 1981.

[185] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

[186] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[187] Thomas Steinke and Alexander Knop. Differentially Private Linear Regression via Medians. `https://openreview.net/pdf?id=JSBgIaxAXk9`, 2022. [Online; accessed 28-April-2023].

[188] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

[189] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[190] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[191] Henri Theil. A rank-invariant method of linear and polynomial regression analysis. *Indagationes mathematicae*, 12(85):173, 1950.

[192] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. *Advances in neural information processing systems*, 31, 2018.

[193] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[194] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[195] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[196] John W Tukey and Donald H McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352, 1963.

[197] Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ilias Zadik. Optimal private median estimation under minimal distributional assumptions. *Advances in Neural Information Processing Systems*, 33:3301–3311, 2020.

[198] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

[199] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

[200] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[201] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

[202] Vincent Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Artificial intelligence and statistics*, pages 1278–1286. PMLR, 2012.

[203] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[204] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private

stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

[205] Sen Wang and J Morris Chang. Differentially private principal component analysis over horizontally partitioned data. In *2018 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 1–8. IEEE, 2018.

[206] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

[207] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

[208] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

[209] Lu Wei, Anand D Sarwate, Jukka Corander, Alfred Hero, and Vahid Tarokh. Analysis of a privacy-preserving pca algorithm using random matrix theory. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1335–1339. IEEE, 2016.

[210] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322, 2017.

[211] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, pages 1689–1698. PMLR, 2015.

[212] Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. *arXiv preprint arXiv:1002.4658*, 2010.

[213] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

[214] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning markov random fields. *arXiv preprint arXiv:2002.09463*, pages 11129–11140, 2020.

[215] Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *arXiv preprint arXiv:2206.00363*, 2022.

[216] Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. DPZero: dimension-independent and differentially private zeroth-order optimization. *arXiv preprint arXiv:2310.09639*, 2023.

[217] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

[218] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. When does the tukey median work? *arXiv preprint arXiv:2001.07805*, 2020.

[219] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.

Appendix A

APPENDICES FOR CHAPTER 2

A.1 Proof of Theorem 5 on the accuracy of the exponential mechanism for Tukey median

First, the $(\varepsilon, 0)$-differential privacy guarantee of private Tukey median follows as a corollary of Proposition 2.2.2, by noting that sensitivity of $n\, D_{\text{Tukey}}(\mathcal{D}_n, x)$ is one, where $\mathcal{D}_n$ is a dataset of size n. This follows from the fact that for any fixed x and v, $|\{z \in \mathcal{D}_n : (v^\top(x - z)) \geq 0\}|$ is the number of samples on one side of the hyperplane, which can change at most by one if we change one sample in $\mathcal{D}$.

Next, given n i.i.d samples $X_1, X_2, \ldots X_n$ from distribution p, denote $\hat{p}_n$ as the empirical distribution defined by the samples $X_1, X_2, \ldots X_n$. Denote $\tilde{p}_n$ as the distribution that is corrupted from $\hat{p}_n$. We slightly overload the definition of Tukey depth to denote $D_{\text{Tukey}}(p, x)$ as the Tukey depth of point $x \in \mathbb{R}^d$ under distribution p, which is defined as

$$D_{\text{Tukey}}(p, x) = \inf_{v \in \mathbb{R}^d} \mathbb{P}_{z \sim p}(v^\top (x - z) \geq 0).$$

Note that this is the standard definition of Tukey depth. First we show that for n large enough, the Tueky depth for the empirical distribution is close to that of the true distribution. We provide proofs of the following lemmas later in this section.

Lemma A.1.1. *With probability* $1 - \delta$*, for any* p *and* $x \in \mathbb{R}^d$*,*

$$|D_{Tukey}(p, x) - D_{Tukey}(\hat{p}_n, x)| \leq C \cdot \sqrt{\frac{d + 1 + \log(1/\delta)}{n}}.$$

The proof of Lemma A.1.1can be found in §A.1.1. This allows us to use the known Tukey depths of a Gaussian distribution to bound the Tukey depths of the corrupted empirical one. We use this to show that there is a strict separation between the Tueky depth of a point in

$S_1 = \{x : \|x - \mu\| \le \alpha\}$ and a point in $S_2 = \{x : \|x - \mu\| \ge 10\alpha\}$. The proof of Lemma A.1.2 can be found in §A.1.2.

Lemma A.1.2. *Define $p = \mathcal{N}(\mu, I)$, and assume $\alpha < 0.01$. Given that $n = \Omega(\alpha^{-2}(d + \log(1/\delta)))$, with probability $1 - \delta$,*

1. *For any point $x \in \mathbb{R}^d$, $\|x - \mu\| \le \alpha$, it holds that*

$$D_{Tukey}(\tilde{p}_n, x) \ge \frac{1}{2} - 2\alpha$$

2. *For any point $x \in \mathbb{R}^d$, $\|x - \mu\| \ge 10\alpha$, it holds that*

$$D_{Tukey}(\tilde{p}_n, x) \le \frac{1}{2} - 5\alpha.$$

This implies that most of the probability mass of the exponential mechanism is concentrated inside a ball of radius $O(\alpha)$ around the true mean μ. Hence, with high probability, the exponential mechanism outputs an approximate mean that is $O(\alpha)$ close to the true one. The following lemma finishes the proof the the desired claim, whose proof can be found in §A.1.3.

Lemma A.1.3 (Utility)**.** *Denote $\tilde{p}_n$ as the distribution that is corrupted from $\hat{p}_n$. Suppose x is sampled from $[-2R, 2R]^d$ with density $r(x) \propto \exp(-(1/2)\varepsilon n D_{Tukey}(\tilde{p}_n, x))$, then given $n = \Omega(\,(d/(\alpha\varepsilon))\log(dR/\eta\alpha) + (1/\alpha^2)(d + \log(1/\eta))\,)$ and $\mu \in [-R, R]^d$, and $R \ge \alpha$,*

$$\mathbb{P}(\|x - \mu\| \le 5\alpha) \ge 1 - \eta\,.$$

A.1.1 Proof of Lemma A.1.1

From the VC inequality ([59], Chap 2, Chapter 4.3) and the fact that the family of sets $\{\{z|v^\top z \ge t\}|\|v\| = 1, t \in \mathbb{R}, v \in \mathbb{R}^d\}$ has VC dimension $d + 1$, there exists some universal constant C such that with probability at least $1 - \delta$

$$\sup_{t\in\mathbb{R}, v\in\mathbb{R}^d, \|v\|=1} |\mathbb{P}_{z\sim p}(v^\top z \ge t) - \mathbb{P}_{z\sim\hat{p}_n}(v^\top z \ge t)| \le C \cdot \sqrt{\frac{d + 1 + \log(1/\delta)}{n}},$$

which implies, for any $x \in \mathbb{R}^d$,

$$\sup_{v \in \mathbb{R}^d} |\mathbb{P}_{z \sim p}(v^\top (x - z) \geq 0) - \mathbb{P}_{z \sim \hat{p}_n}(v^\top (x - z) \geq 0)| \leq C \cdot \sqrt{\frac{d + 1 + \log(1/\delta)}{n}},$$

by letting $t = v^\top x$. We conclude the proof since

$$\begin{aligned}
&|D_{\text{Tukey}}(p, x) - D_{\text{Tukey}}(\hat{p}_n, x)| \\
&= |\inf_{v \in \mathbb{R}^d} \mathbb{P}_{z \sim p}(v^\top (x - z) \geq 0) - \inf_{v \in \mathbb{R}^d} \mathbb{P}_{z \sim \hat{p}_n}(v^\top (x - z) \geq 0)| \\
&\leq \sup_{v \in \mathbb{R}^d} |\mathbb{P}_{z \sim p}(v^\top (x - z) \geq 0) - \mathbb{P}_{\hat{p}_n}(v^\top (x - z) \geq 0)| \\
&\leq C \cdot \sqrt{\frac{d + 1 + \log(1/\delta)}{n}}.
\end{aligned}$$

A.1.2 Proof of Lemma A.1.2

For the first claim, we first prove a lower bound on $D_{\text{Tukey}}(p, x)$. Since $p = \mathcal{N}(\mu, I)$, for any $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$,

$$\begin{aligned}
&\mathbb{P}_{z \sim p}(v^\top (z - x) \geq 0) \\
&= \mathbb{P}_{z \sim N(0,1)}(z \geq v^\top (x - \mu)) \\
&= \int_{v^\top (x - \mu)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz \\
&\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} v^\top (x - \mu) \\
&\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \|x - \mu\|_2 \\
&\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \alpha
\end{aligned}$$

Thus,

$$\begin{aligned}
&D_{\text{Tukey}}(p, x) \\
&= \inf_{v \in \mathbb{R}^d} \mathbb{P}_{z \sim p}(v^\top (x - z) \geq 0) \\
&\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \alpha
\end{aligned}$$

Then Lemma A.1.1 implies that with probability $1-\delta$

$$D_{\text{Tukey}}(\hat{p}_n, x) \geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\alpha - C \cdot \sqrt{\frac{d+1+\log(1/\delta)}{n}}.$$

Since the corruption can change at most α probability mass, it holds that $|D_{\text{Tukey}}(\tilde{p}_n, x) - D_{\text{Tukey}}(\hat{p}_n, x)| \leq \alpha$. Setting $n = \Omega(\alpha^{-2}(d+\log(1/\delta)))$ yields

$$\begin{aligned} D_{\text{Tukey}}(\tilde{p}_n, x) &\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\|x-\mu\|_2 - C \cdot \sqrt{\frac{d+1+\log(1/\delta)}{n}} - \alpha \\ &\geq \frac{1}{2} - 2\alpha. \end{aligned}$$

For the second claim, note that

$$\begin{aligned} & D_{\text{Tukey}}(p, x) \\ \leq\ & \int_{v^\top(x-\mu)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz \\ \overset{(a)}{\leq}\ & \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \exp(-(20\alpha)^2/2) \cdot 20\alpha \\ \overset{(b)}{\leq}\ & \frac{1}{2} - 7\alpha \end{aligned}$$

where (a) holds since $\|x-\mu\| \geq 20\alpha$, and it is easy to verify that (b) holds for $\alpha \leq 0.01$. The second claim holds since

$$\begin{aligned} & D_{\text{Tukey}}(\tilde{p}_n, x) \\ \leq\ & D_{\text{Tukey}}(\hat{p}_n, x) + \alpha \\ \leq\ & D_{\text{Tukey}}(p, x) + \alpha + C \cdot \sqrt{\frac{d+1+\log(1/\delta)}{n}} \\ \overset{(a)}{\leq}\ & D_{\text{Tukey}}(p, x) + 2\alpha \\ \leq\ & \frac{1}{2} - 5\alpha, \end{aligned}$$

where (a) holds by setting $n = \Omega(\alpha^{-2}(d+\log(1/\delta)))$.

A.1.3 Proof of Lemma A.1.3

Let $r(x) = \frac{1}{A}\exp(-\varepsilon n D_{\text{Tukey}}(\tilde{p}_n, x))$ where A is the normalization factor. Then

$$\mathbb{P}(\|x-\mu\| \le \alpha) \ge \frac{1}{A}\exp(\varepsilon n(\frac{1}{2}-2\alpha)) \cdot \frac{\pi^{d/2}}{\Gamma(d/2+1)}\alpha^d,$$

using the fact that $\mu \in [-R,R]^d$ and that $R \ge \alpha$, and

$$\mathbb{P}(\|x-\mu\| \ge 5\alpha) \le \frac{1}{A}\exp(\varepsilon n(\frac{1}{2}-10\alpha)) \cdot (4R)^d.$$

Hence

$$\log(\frac{\mathbb{P}(\|x-\mu\| \le \alpha)}{\mathbb{P}(\|x-\mu\| \ge 5\alpha)}) \ge \varepsilon n(3\alpha) - C \cdot d\log(dR/\alpha),$$

where C is an absolute constant. If we set $n = \Omega(\frac{d\log(dB/\delta\alpha)}{\alpha\varepsilon})$, we get that

$$\frac{\mathbb{P}(\|x-\mu\| \le \alpha)}{\mathbb{P}(\|x-\mu\| \ge 5\alpha)} \ge \frac{10}{\delta},$$

which implies that with probability at least $1-\delta$, $\|x-\mu\| \le 5\alpha$.

A.2 Estimating the range with DPRANGE

Algorithm 14: Differentially private range estimation (DPRANGE) [139, Algorithm 1]

Input: $\mathcal{D}_n = \{x_i\}_{i=1}^n$, R, ε, δ, $\sigma = 1$

for $j \leftarrow 1$ **to** d **do**

  Run the histogram learner of Lemma A.2.1 with privacy parameters $\left(\min\{\varepsilon, 0.9\}/2\sqrt{2d\log(2/\delta)}, \delta/(2d)\right)$ and bins $B_l = (2\sigma\ell, 2\sigma(\ell+1)]$ for all $\ell \in \{-\lceil R/2\sigma\rceil - 1, \ldots, \lceil R/2\sigma\rceil\}$ on input $\mathcal{D}_n$ to obtain noisy estimates $\{\tilde{h}_{j,l}\}_{l=-\lceil R/2\sigma\rceil-1}^{\lceil R/2\sigma\rceil}$

  $\bar{x}_j \leftarrow 2\sigma \cdot \arg\max_{\ell \in \{-\lceil R/2\sigma\rceil-1,\ldots,\lceil R/2\sigma\rceil\}} \tilde{h}_{j,\ell}$

Output: $(\bar{x}, B = 8\sigma\sqrt{\log(dn/\zeta)})$

A.2.1 Proof of Lemma 2.3.5

Assuming the distribution is σ^2 sub-Gaussian, we use $\mathcal{P}$ to denote the sub-Gaussian distribution. Denote $I_l = [2\sigma l, 2\sigma(l+1)]$ as the interval of the l'th bin. Denote the population probability in the l'th bin $h_{j,l} = \mathbb{P}_{x\sim\mathcal{P}}[x_j \in I_l]$, empirical probability in the l'th bin $\tilde{h}_{j,l} = \frac{1}{n}\sum_{x_i\in\mathcal{D}} \mathbf{1}\{x_{i,j} \in I_l\}$, and the noisy version $\hat{h}_{j,l}$ computed by the histogram learner of Lemma A.2.1. Notice that Lemma A.2.1 with d compositions (Lemma 2.3.4) immediately implies that our algorithm is (ε, δ)-differentially private.

For the utility of the algorithm, we will first show that for all dimension $j \in [d]$, the output $|\bar{x}_j - \mu_j| = O(\sigma)$. Note that by the definition of σ^2-subgaussian, it holds that for all $i \in [d]$, $\mathbb{P}[|x_i - \mu_i| \geq z] \leq 2\exp(-z^2/\sigma^2)$ where x is drawn from distribution $\mathcal{P}$. This implies that $\mathbb{P}[|x_i - \mu_i| \geq 2\sigma] \leq 2\exp(-4) \leq 0.04$. Suppose the k'th bin contains μ_j, namely $\mu_j \in I_k$. Then it is clear that $[\mu_j - 2\sigma, \mu_j + 2\sigma] \subset (I_{k-1} \cup I_k \cup I_{k+1})$. This implies $h_{j,k-1} + h_{j,k} + h_{j,k+1} \geq 1 - 0.04 = 0.96$, hence $\min(h_{j,k-1}, h_{j,k}, h_{j,k+1}) \geq 0.32$.

Recall that $\mathcal{G}$ is the set of clean data drawn from distribution P. By Dvoretzky-Kiefer-Wolfowitz inequality and an union bound over $j \in [d]$, we have that with probability $1-\zeta$, $\max_{j,l}(|h_{j,l} - \frac{1}{n}\sum_{x\in G} x_j|) \leq \sqrt{\frac{\log(d/\zeta)}{n}}$. The deviation due to corruption is at most α on each bin, hence we have $\max_{j,l}(|h_{j,l} - \hat{h}_{j,l}) \leq \sqrt{\frac{\log(d/\zeta)}{n}} + \alpha$. Lemma A.2.1 and a union bound over $j \in [d]$ implies that with probability $1-\zeta$, $\max_{j,l}(|\tilde{h}_{j,l} - \hat{h}_{j,l}|) \leq \beta$ when $n \geq \Omega(\min\left\{\frac{\sqrt{d\log(1/\delta)}}{\varepsilon\beta}\log(dR/\zeta), \frac{\sqrt{d\log(1/\delta)}}{\varepsilon\beta}\log(d/\zeta\delta)\right\})$.

Assuming that $n = \Omega\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\min\left\{\log(dR/\zeta), \log(d/\zeta\delta)\right\}\right)$, we have that with probability $1-\zeta$, $\max_{j,l}(|h_{j,l} - \hat{h}_{j,l}|) \leq 0.01 + \alpha$. Using the assumption that $\alpha \leq 0.1$, since $\min(h_{j,k-1}, h_{j,k}, h_{j,k+1}) - 0.11 \geq 0.31 \geq 0.04 + 0.11 \geq \max_{l\neq k-1,k,k+1} h_{j,l} + 0.11$. This implies that with probability $1-\zeta$, the algorithm choose the bin from $k-1, k, k+1$, which means the estimate $|\bar{x}_j - \mu| \leq 4\sigma$. By the tail bound of sub-Gaussian distribution and a union bound over n, d, we have that with probability $1-\zeta$, for all $x_i \in \mathcal{D}$ and $j \in [d]$, $x_{i,j} \in [\bar{x}_j - 8\sigma\sqrt{\log(nd/\zeta)}, \bar{x}_j + 8\sigma\sqrt{\log(nd/\zeta)}]$.

Lemma A.2.1 (Histogram Learner, Lemma 2.3 in [139]). *For every $K \in \mathbb{N} \cup \infty$, domain Ω,*

for every collection of disjoint bins $B_1, \ldots, B_K$ *defined on* Ω, $n \in \mathbb{N}$, $\varepsilon, \delta \in (0, 1/n)$, $\beta > 0$ *and* $\alpha \in (0,1)$ *there exists an* (ε, δ)*-differentially private algorithm* $M : \Omega^n \to \mathbb{R}^K$ *such that for any set of data* $X_1, \ldots, X_n \in \Omega^n$

1. $\hat{p}_k = \frac{1}{n} \sum_{X_i \in B_k} 1$

2. $(\tilde{p}_1, \ldots, \tilde{p}_K) \leftarrow M(X_1, \ldots, X_n)$, *and*

3. $$n \geq \min \left\{ \frac{8}{\varepsilon\beta} \log(2K/\alpha), \frac{8}{\varepsilon\beta} \log(4/\alpha\delta) \right\}$$

then,

$$\mathbb{P}(|\tilde{p}_k - \hat{p}_k| \leq \beta) \geq 1 - \alpha$$

Proof. This is an intermediate result in the proof of Lemma 2.3 in [139]. □

A.3 Differentially private robust filtering with DPFILTER

A.3.1 Proofs of the sensitivity of the filtering in Lemma 2.3.6 and Lemma A.5.1

Proof of Lemma 2.3.6. We only need to show that one step of the proposed filter is a contraction. To this end, we only need to show contraction for two datasets at distance 1, i.e., $d_\triangle(\mathcal{D}, \mathcal{D}') = 1$. For fixed (μ, v) and Z, we apply filter to set of scalars $(v^\top(\mathcal{D} - \mu))^2$ and $(v^\top(\mathcal{D}' - \mu))^2$, whose distance is also one. If the entries that are different (say $a \in \mathcal{D}$ and $a' \in \mathcal{D}'$) are both below the subset of the top $2n\alpha$ points (as in Definition 2.3.1), then the same set of points will be removed for both and the distance is preserved $d_\triangle(S(\mathcal{D}), S(\mathcal{D}')) = 1$. If they are both above the top $2n\alpha$ subset, then either both are removed, one of them is removed, or both remain. The rest of the points that are removed coincide in both sets. Hence, $d_\triangle(S(\mathcal{D}), S(\mathcal{D}')) \leq 1$. If a is below and a' is above the top $2n\alpha$ subset of respective datasets, then either a' is not removed (in which case $d_\triangle(S(\mathcal{D}), S(\mathcal{D}')) = 1$) or a' is removed (in which case $S(\mathcal{D}) = S(\mathcal{D}') \cup \{a\}$ and the distance remains one).

Note that when there are ties, it is critical to resolve them in a consistent manner in both datasets $\mathcal{D}$ and $\mathcal{D}'$. The tie breaking rule of Definition 2.3.1 is critical in sorting those samples with the same score τ_i's in a consistenet manner.

Proof of Lemma A.5.1. The analysis of contraction of the filtering step in DPMMWFILTER is analogous to that of DPFILTER in Lemma 2.3.6.

Algorithm 15: Interactive version of DPFILTER

Input: $\alpha \in (0,1)$, $T \in \mathbb{Z}_+$

$\varepsilon_1 \leftarrow \min\{\varepsilon, 0.9\}/(4\sqrt{2T\log(2/\delta)})$, $\delta_1 \leftarrow \delta/(8T)$

for $t = 1, \ldots, T$ **do**

  $n_t \leftarrow q_{\text{size}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \varepsilon_1)$

  if $n_t < 3n/4$ **then**

    terminate

  $\mu_t \leftarrow q_{\text{mean}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \varepsilon_1)$

  if $\lambda_t \le (C - 0.01)\alpha \log 1/\alpha$ **then**

    Output: μ_t

  $\lambda_t \leftarrow q_{\text{norm}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \mu_t, \varepsilon_1)$

  $v_t \leftarrow q_{\text{PCA}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \mu_t, \varepsilon_1, \delta_1))$

  $Z_t \leftarrow \text{Unif}([0,1])$

Output: μ_t

Filter $(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]})$**:**

  $S_0 \leftarrow [n]$

  for $\ell = 1, \ldots, t-1$ **do**

    $S_\ell \leftarrow S_{\ell-1} \setminus \{i \in S_{\ell-1} : i \in \mathcal{T}_{2\alpha} \text{ for } \{\tau_j = (v_\ell^\top (x_j - \mu_\ell))^2\}_{j\in S_{\ell-1}} \text{ and } \tau_i \ge d\, B^2\, Z_\ell\}$

$q_{\text{mean}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \varepsilon)$**:**

  Filter$(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]})$

  return $\mu_t \leftarrow (1/|S_{t-1}|)\big(\sum_{i\in S_{t-1}} x_i\big) + \text{Lap}(2B/(n\varepsilon))$

$q_{\text{PCA}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \mu_t, \varepsilon, \delta)$**:**

  Filter$(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]})$

  return $v_t \leftarrow$ top singular vector of $\Sigma_{t-1} =$

  $(1/n)\sum_{i\in S_{t-1}} (x_i - \mu_t)(x_i - \mu_t)^\top + \mathcal{N}(0, (B^2 d\sqrt{2\log(1.25/\delta)}/(n\varepsilon))^2 \mathbf{I}_{d^2\times d^2})$

$q_{\text{norm}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \mu_t, \varepsilon)$**:**

  Filter$(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]})$

  return $\lambda_t \leftarrow \|(1/n)\sum_{i\in S_{t-1}} (x_i - \mu_t)(x_i\mu_t)^\top\|_2 + \text{Lap}(2B^2 d/(n\varepsilon))$

$q_{\text{size}}(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]}, \varepsilon)$**:**

  Filter$(\{(\mu_\ell, v_\ell, Z_\ell)\}_{\ell\in[t-1]})$

  return $n_t \leftarrow |S_{t-1}| + \text{La}[(1/\varepsilon)$

A.3.2 Proof of part 1 of Lemma 2.3.7 on differential privacy of DPFILTER

We explicitly write out how many times we access the database and how much privacy is lost each time in an interactive version of DPFILTER in Algorithm 15, which performs the same operations as DPFILTER. In order to apply Lemma 2.3.4, we cap ε at 0.9 in initializing ε_1. We call q_{mean}, q_{PCA}, q_{norm} and q_{size} T times, each with $(\varepsilon_1, \delta_1)$ guarantee. In total this accounts for (ε, δ) privacy loss, using Lemma 2.3.4 and our choice of ε_1 and δ_1.

This proof is analogous to the proof of DP for DPMMWFILTER in §A.5.1, and we omit the details here. We will assume for now that $|S_r| \geq n/2$ for all $r \in [t]$ and prove privacy. This happens with probability larger than $1 - \delta_1$, hence ensuring the privacy guarantee. In all sub-routines, we run Filter$(\cdot)$ in Algorithm 15 to simulate the filtering process so far and get the current set of samples S_t. Lemma 2.3.6 allows us to prove privacy of all interactive mechanisms. This shows that the two data datasets S_t and S_t' are neighboring, if they are resulting from the identical filtering but starting from two neighboring datasets $\mathcal{D}_n$ and $\mathcal{D}_n'$. As all four sub-routines are output perturbation mechanisms with appropriately chosen sensitivities, they satisfy the desired $(\varepsilon_1, \delta_1)$-DP guarantees. Further, the probability that $n_t > 3/4n$ and $|S_t| \leq n/2$ is less than δ_1 for $n = \tilde{\Omega}((1/\varepsilon_1)\log(1/\delta_1))$.

A.3.3 Proof of part 2 of Lemma 2.3.7 on accuracy of DPFILTER

The following theorem analyzing DPFILTER implies the desired Lemma 2.3.7 when the good set is α-subgaussian good, which follows from A.6.3 and the assumption that $n = \widetilde{\Omega}(d/\alpha^2)$.

Theorem 25 (Anlaysis of DPFILTER)**.** *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1, where $\alpha \leq c$ for some universal constant $c \in (0, 1/2)$. Let S_{good} be α-subgaussian good with respect to $\mu \in \mathbb{R}^d$. Suppose $\mathcal{D} = \{x_i \in \bar{x} + [-B/2, B/2]^d\}_{i=1}^n$ be the projected dataset where all of the uncorrupted samples are contained in $\bar{x} + [-B/2, B/2]^d$. If $n = \widetilde{\Omega}\left(d^2 B^3 \log(1/\delta)/(\varepsilon\alpha)\right)$, then* DPFILTER *terminates after at most $O\left(dB^2\right)$ iterations and outputs S_t such that with probability 0.9, we have $|S_t \cap S_{\text{good}}| \geq (1 - 10\alpha)n$ and*

$$\|\mu(S_t) - \mu\|_2 \lesssim \alpha\sqrt{\log 1/\alpha}\,.$$

To prove this theorem, we use the following lemma to first show that we do not remove too many uncorrupted samples. The upper bound on the accuracy follows immediately from Lemma A.6.7 and the stopping criteria of the algorithm.

Lemma A.3.1. *If* $n \gtrsim \frac{B^2 d^{3/2}}{\varepsilon_1 \alpha \log 1/\alpha} \log(1/\delta)$, $\lambda_t \geq (C - 0.01) \cdot \alpha \log 1/\alpha$ *and* $|S_t \cap S_{\text{good}}| \geq (1 - 10\alpha)n$, *then there exists constant* $C > 0$ *such that for each iteration* t, *with probability* $1 - O(1/d)$, *we have Eq.* (A.3) *holds. If this condition holds, we have*

$$\mathbb{E}\,|(S_t \setminus S_{t+1}) \cap S_{\text{good}}| \leq \mathbb{E}\,|S_t \setminus S_{t+1} \cap S_{\text{bad}}| \ .$$

We measure the progress by by summing the number of clean samples removed up to iteration t and the number of remaining corrupted samples, defined as $d_t \triangleq |(S_{\text{good}} \cap S) \setminus S_t| + |S_t \setminus (S_{\text{good}} \cap S)|$. Note that $d_1 = \alpha n$, and $d_t \geq 0$. At each iteration, we have

$$\mathbb{E}[d_{t+1} - d_t | d_1, d_2, \cdots, d_t] \ = \ \mathbb{E}\left[|S_{\text{good}} \cap (S_t \setminus S_{t+1})| - |S_{\text{bad}} \cap (S_t \setminus S_{t+1})|\right] \ \leq \ 0,$$

from the Lemma A.3.1. Hence, d_t is a non-negative super-martingale. By optional stopping theorem, at stopping time, we have $\mathbb{E}[d_t] \leq d_1 = \alpha n$. By Markov inequality, d_t is less than $10\alpha n$ with probability 0.9, i.e. $|S_t \cap S_{\text{good}}| \geq (1 - 10\alpha)n$. The desired bound follows from induction and Lemma A.6.7.

Now we bound the number of iterations under the conditions of Lemma A.3.2. Let $W_t = |S_t \setminus S_{t-1}|/n$. Since Eq. (A.4), we have

$$\mathbb{E}[W_t] \geq \frac{1}{n} \sum_{i \in \mathcal{T}_{2\alpha}} \frac{\tau_i}{dB^2} \geq \frac{0.7\|M(S_{t-1}) - \mathbf{I}\|_2}{\alpha d B^2} \geq \frac{0.7C\alpha \log(1/\alpha)}{dB^2} \ .$$

Let T be the stopping time. We know $\sum_{t=1}^{T} W_t \leq 10\alpha$. By Wald's equation, we have

$$\mathbb{E}[\sum_{t=1}^{T} W_t] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}[W_t]] \geq \mathbb{E}[T] \frac{0.7C\alpha \log(1/\alpha)}{dB^2} \ .$$

This means $\mathbb{E}[T] \leq (15dB^2)/(C \log(1/\alpha))$. By Markov inequality we know with probability 0.9, we have $T = O(dB^2/\log(1/\alpha))$.

A.3.3.1 Proof of Lemma A.3.1

The expected number of removed good points and bad points are proportional to the $\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}}\tau_i$ and $\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}\tau_i$. It suffices to show

$$\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}}\tau_i \;\le\; \sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}\tau_i\;.$$

Assuming we have $\|M(S_{t-1})-I\|_2 \ge C\alpha\log 1/\alpha$ for some $C>0$ sufficiently large, it suffices to show

$$\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}\tau_i \ge \frac{1}{1000}\|M(S_{t-1})-\mathbf{I}\|_2\;.$$

First of all, we have

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_{t-1}}\tau_i - 1 &= v_t^\top M(S_{t-1})v_t - 1\\ &= v_t^\top\left(M(S_{t-1})-\mathbf{I}\right)v_t\end{aligned}$$

Lemma A.6.6 shows that the magnitude of the largest eigenvalue of $M(S_{t-1})-\mathbf{I}$ is positive since the magnitudes negative eigenvalues are all less than $c\alpha\log 1/\alpha$. So we have

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_{t-1}}\tau_i - 1 &\ge \|M(S_{t-1})-\mathbf{I}\|_2 - O(\alpha\log 1/\alpha) &\text{(A.1)}\\ &\ge 0.9\|M(S_{t-1})-\mathbf{I}\|_2\;, &\text{(A.2)}\end{aligned}$$

where the first inequality follows from Lemma A.3.3, and the second inequality follows from our choice of large constant C. The next lemma regularity conditions for τ_i's for each iteration is satisfied.

Lemma A.3.2. *If $n \gtrsim \frac{B^2d^{3/2}}{\varepsilon_1\alpha\log 1/\alpha}\log(1/\delta)$, then there exists a large constant $C>0$ such that, with probability $1-O(1/d)$, we have*

1.

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}}\tau_i \le \frac{1}{1000}\left\|M(S_{t-1})-\mathbf{I}\right\|_2\;. \qquad \text{(A.3)}$$

2. For all $i \notin \mathcal{T}_{2\alpha}$,

$$\alpha\tau_i \leq \frac{1}{1000}\|M(S_{t-1}) - \mathbf{I}\|_2 \,.$$

3.

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_{t-1}} (\tau_i - 1) \leq \frac{1}{1000}\|M(S_{t-1}) - \mathbf{I}\|_2 \,.$$

Thus, by combining with Lemma A.3.2, we have

$$\frac{1}{n}\sum_{i\in S_{t-1}\cap S_{\text{bad}}} \tau_i \;\geq\; 0.8\|M(S_{t-1}) - \mathbf{I}\|_2 \,.$$

We now have

$$\begin{aligned}
\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}} \tau_i &\geq 0.8\|M(S_{t-1}) - \mathbf{I}\|_2 - \sum_{i\in S_{\text{bad}}\cap S_{t-1}\setminus\mathcal{T}_{2\alpha}} \tau_i \\
&\geq 0.8\|M(S_{t-1}) - \mathbf{I}\|_2 - \max_{i\in S_{\text{bad}}\cap S_{t-1}\setminus\mathcal{T}_{2\alpha}} \alpha\tau_i \\
&\geq 0.8\|M(S_{t-1}) - \mathbf{I}\|_2 - \frac{1}{1000}\|M(S_{t-1}) - \mathbf{I}\|_2 \qquad \text{(A.4)} \\
&\geq \frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}} \tau_i \,,
\end{aligned}$$

which completes the proof.

A.3.3.2 Proof of Lemma A.3.2

By our choice of sample complexity n, with probability $1 - O(1/dB^2)$, we have $\|\mu(S_{t-1}) - \mu_t\|_2^2 \lesssim \alpha\log 1/\alpha$, $v_t^\top (M(S_{t-1}) - \mathbf{I})\, v_t \gtrsim \|M(S_{t-1}) - \mathbf{I}\|_2 - \alpha\log 1/\alpha$ (Lemma A.3.3), and $\|M(S_{t-1}) - \mathbf{I}\|_2 \geq C\alpha\log 1/\alpha$ simultaneously hold before stopping.

Lemma A.3.3. *If*

$$n \gtrsim \frac{d^{3/2}B^2}{\eta\varepsilon_1}\sqrt{2\ln\frac{1.25}{\delta}}\log\frac{1}{\zeta} \,,$$

then with probability $1-\zeta$, we have

$$v_t^\top (M(S_{t-1}) - \mathbf{I})\, v_t \geq \|M(S_{t-1}) - \mathbf{I}\|_2 - 2\eta - \frac{2|S_{t-1}|}{n}\|\mu_t - \mu(S_{t-1})\|_2^2$$

We first consider the upper bound of the good points.

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}}\tau_i = \frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}}\langle x_i-\mu_t, v_t\rangle^2\\
&\overset{(a)}{\le}\frac{2}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}}\langle x_i-\mu, v_t\rangle^2 + \frac{2}{n}|S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}|\langle \mu-\mu_t, v_t\rangle^2\\
&\le O(\alpha\log 1/\alpha) + \alpha\left(\|\mu-\mu(S_{t-1})\|_2 + \|\mu_t-\mu(S_{t-1})\|_2\right)^2\\
&\overset{(b)}{\le} O(\alpha\log 1/\alpha) + \alpha\left(O(\alpha\sqrt{\log 1/\alpha}) + \sqrt{\alpha\left(\|M(S_{t-1})-\mathbf{I}\|_2 + O(\alpha\log 1/\alpha)\right)} + O(\sqrt{\alpha\log 1/\alpha})\right)^2\\
&\le O(\alpha\log 1/\alpha) + \alpha^2\|M(S_{t-1})-\mathbf{I}\|_2\\
&\overset{(c)}{\le}\frac{1}{1000}\|M(S_{t-1})-I\|_2
\end{aligned}$$

where the (a) is implied by the fact that for any vector x, y, z, we have $(x-y)(x-y)^\top \preceq 2(x-z)(x-z)^\top + 2(y-z)(y-z)^\top$, (b) follows from Lemma A.6.7 and c follows from our choice of large constant C.

Since $|S_{\text{bad}}\cap\mathcal{T}_{2\alpha}|\le\alpha n$, we know $|S_{\text{good}}\cap\mathcal{T}_{2\alpha}|\ge\alpha n$, so we have for $i\notin\mathcal{T}_{2\alpha}$,

$$\alpha\tau_i \le \frac{\alpha}{|S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}|}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}\cap S_{t-1}}\tau_i \le \frac{1}{1000}\|M(S_{t-1})-\mathbf{I}\|_2\,.$$

Since $|S_{\text{good}}\cap S_{t-1}|\ge(1-10\alpha)n$, we have

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_{t-1}}\tau_i = \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_{t-1}}\langle x_i-\mu(S_{t-1}), v_t\rangle^2 \tag{A.5}$$

$$= \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_{t-1}}\langle x_i-\mu(S_{\text{good}}\cap S_{t-1}), v_t\rangle^2 + \frac{|S_{\text{good}}\cap S_{t-1}|}{n}\langle\mu(S_{\text{good}}\cap S_{t-1})-\mu(S_{t-1}), v_t\rangle^2 \tag{A.6}$$

$$\overset{(a)}{\le} c\alpha\log 1/\alpha + 1 + \|\mu(S_{\text{good}}\cap S_{t-1})-\mu(S_{t-1})\|_2^2 \tag{A.7}$$

$$\le c\alpha\log 1/\alpha + 1 + \left(\|\mu(S_{\text{good}}\cap S_{t-1})-\mu\|_2 + \|\mu-\mu(S_{t-1})\|_2\right)^2 \tag{A.8}$$

$$\overset{(b)}{\le} c\alpha\log 1/\alpha + 1 + \alpha\|M(S_{t-1})-\mathbf{I}\|_2 + O(\alpha\log 1/\alpha) \tag{A.9}$$

$$\overset{(c)}{\le}\frac{1}{1000}\|M(S_{t-1})-\mathbf{I}\|_2\,, \tag{A.10}$$

where (a) follows from Lemma A.6.6, and (b) follows from Lemma A.6.7, and (c) follows from our choice of large constant C.

A.3.3.3 Proof of Lemma A.3.3

Proof. We have following identity.

$$\begin{aligned} & \frac{1}{n}\sum_{i\in S_{t-1}}(x_i-\mu_t)(x_i-\mu_t)^\top \\ = \; & \frac{1}{n}\sum_{i\in S_{t-1}}(x_i-\mu(S_{t-1}))(x_i-\mu(S_{t-1}))^\top + \frac{|S_{t-1}|}{n}(\mu(S_{t-1})-\mu_t)(\mu(S_{t-1})-\mu_t)^\top . \end{aligned}$$

So we have,

$$\begin{aligned} & v_t^\top\left(M(S_{t-1})-\mathbf{I}\right)v_t \\ \geq \; & v_t^\top\left(\frac{1}{n}\sum_{i\in S_{t-1}}(x_i-\mu_t)(x_i-\mu_t)^\top-\mathbf{I}\right)v_t - \frac{|S_{t-1}|}{n}\|\mu_t-\mu(S_{t-1})\|_2^2 \\ \geq \; & \|M(S_{t-1})-\mathbf{I}\|_2 - 2\eta - \frac{2|S_{t-1}|}{n}\|\mu_t-\mu(S_{t-1})\|_2^2 \end{aligned}$$

where the last inequality follows from Lemma A.6.6, which shows that the magnitude of the largest eigenvalue of $M(S_{t-1})-\mathbf{I}$ must be positive. □

A.3.4 Proof of Theorem 6

Differential privacy guarantee. To achieve (ε_0,δ_0) end-to-end target privacy guarantee, Algorithm 3 separates the privacy budget into two. The $(0.01\varepsilon_0, 0.01\delta_0)$-DP guarantee of DPRANGE follows from Lemma 2.3.5. The $(0.99\varepsilon_0, 0.99\delta_0)$-DP guarantee of DPFILTER follows from Lemma 2.3.7.

Accuracy. From Lemma 2.3.5 DPRANGE is guaranteed to return a hypercube that includes all clean data in the dataset. It follows from Lemma 2.3.7 that when $n = \widetilde{\Omega}(d/\alpha^2 + d^2\log(1/\delta)/(\varepsilon\alpha))$, we have $\|\mu-\hat{\mu}\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$.

A.4 Differentially private 1D filter with DP-1DFILTER

A.4.1 Proof of Lemma 2.3.9

1. **Threshold ρ sufficiently reduces the total score.**

   Let ρ be the threshold picked by the algorithm. Let $\hat{\tau}_i$ denote the minimum value of the interval of the bin that τ_i belongs to. It holds that

   $$\begin{aligned}
   &\frac{1}{n}\sum_{\tau_i\geq\rho, i\in[n]}(\tau_i-\rho) \geq \frac{1}{n}\sum_{\hat{\tau}_i\geq\rho, i\in[n]}(\hat{\tau}_i-\rho)\\
   &= \sum_{\tilde{\tau}_j\geq\rho, j\in[2+\log(B^2d)]}(\tilde{\tau}_j-\rho)h_j\\
   &\overset{(a)}{\geq} \sum_{\tilde{\tau}_j\geq\rho, j\in[2+\log(B^2d)]}(\tilde{\tau}_j-\rho)\tilde{h}_j - O\left(\log(B^2d)\cdot B^2d\cdot\frac{\sqrt{\log(\log(B^2d)\log d)\log(1/\delta)}}{\varepsilon n}\right)\\
   &\overset{(b)}{\geq} 0.31\tilde{\psi} - \tilde{O}(\frac{B^2d}{\varepsilon n})\\
   &\overset{(c)}{\geq} 0.3\psi - \tilde{O}(\frac{B^2d}{\varepsilon n}),
   \end{aligned}$$

   where (a) holds due to the accuracy of the private histogram (Lemma A.6.12), (b) holds by the definition of ρ in our algorithm, and (c) holds due to the accuracy of $\tilde{\psi}$. This implies if $\rho<1$, then $\frac{1}{n}\sum_{\tau_i<\rho}(\tau_i-1)$ is negative and if $\rho\geq 1$, then

   $$\frac{1}{n}\sum_{\tau_i<\rho}(\tau_i-1) = \psi - \frac{1}{n}\sum_{\tau_i\geq\rho}(\tau_i-1) \leq \psi - \frac{1}{n}\sum_{\tau_i\geq\rho}(\tau_i-\rho) \leq 0.7\psi + \tilde{O}(B^2d/\varepsilon n).$$

   By Lemma A.4.1, it holds that

   $$\begin{aligned}
   \frac{1}{n}\sum_{i\in S\setminus\mathcal{T}_{2\alpha}}(\tau_i-1) &= \psi - \frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}}(\tau_i-1) - \frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}(\tau_i-1)\\
   &\leq \psi - \frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}(\tau_i-1)\\
   &\leq (2/1000)\psi
   \end{aligned}$$

   And we conclude that

   $$\frac{1}{n}\sum_{\tau_i<\rho \text{ or } i\notin\mathcal{T}_{2\alpha}}(\tau_i-1) \leq 0.71\psi + \tilde{O}(B^2d/\varepsilon n) \leq 0.75\psi$$

2. **Threshold ρ removes more bad data points than good data points.**

   Define C_2 to be the threshold such that $\frac{1}{n}\sum_{\tau_i > C_2}(\tau_i - C_2) = (2/3)\psi$. Suppose $2^b \leq C_2 \leq 2^{b+1}$, $\frac{1}{n}\sum_{\hat{\tau}_i \geq 2^{b-1}}(\hat{\tau}_i - 2^{b-1}) \geq (1/3)\psi$ because $\forall \tau_i \geq C_2$, $(\hat{\tau}_i - 2^{b-1}) \geq \frac{1}{2}(\tau_i - C_2)$. Trivially $C_2 \geq 1$ due to the fact that $\frac{1}{n}\sum_{\tau_i \geq 1}\tau_i - 1 \geq \psi$. Then we have the threshold picked by the algorithm $\rho \geq 2^{b-1}$, which implies $\rho \geq \frac{1}{4}C_2$. Suppose $\rho < C_2$, since $\rho \geq \frac{1}{4}C_2$, we have

$$\begin{aligned}
\Big(\sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}, \tau_i < \rho} \tau_i + \sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}, \tau_i \geq \rho} \rho\Big) &\geq \frac{1}{4}\Big(\sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}, \tau_i < C_2} \tau_i + \sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}, \tau_i \geq C_2} C_2\Big) \\
&\overset{(a)}{\geq} \frac{10}{4}\Big(\sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}, \tau_i < C_2} \tau_i + \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}, \tau_i \geq C_2} C_2\Big) \\
&\overset{(b)}{\geq} \frac{10}{4}\Big(\sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}, \tau_i < \rho} \tau_i + \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}, \tau_i >= \rho} \rho\Big),
\end{aligned}$$

   where (a) holds by Lemma A.4.2, and (b) holds since $\rho \leq C_2$. If $\rho \geq C_2$, the statement of the Lemma A.4.2 directly implies Equation (2.4).

Lemma A.4.1. *[Conditions for τ_i's]*

Suppose

$$\begin{aligned}
\frac{1}{n}\sum_{i \in S_{\text{good}} \cap S}(\tau_i - 1) &\leq \psi/1000 \\
\frac{1}{n}\sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}} \tau_i &\leq \psi/1000
\end{aligned}$$

then, we have

$$\begin{aligned}
\alpha\tau_{2\alpha n} &\leq \psi/1000 \\
\frac{1}{n}\sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}}(\tau_i - 1) &\geq (998/1000)\psi
\end{aligned}$$

Proof. Since $|S_{\text{good}} \cap \mathcal{T}_{2\alpha}| \geq \alpha n$, it holds

$$\alpha\tau_{2\alpha n} \leq \psi/1000.$$

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha}}(\tau_i-1) &= \frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S}(\tau_i-1)-\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S\setminus\mathcal{T}_{2\alpha}}(\tau_i-1)\\ &\geq (999/1000)\psi-\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S\setminus\mathcal{T}_{2\alpha}}(\tau_i-1)\\ &\geq (999/1000)\psi-(1/1000)\psi\\ &= (998/1000)\psi\end{aligned}$$

□

Lemma A.4.2. *Assuming that the conditions in Lemma A.4.1 holds, and for any C such that*

$$\frac{1}{n}\sum_{i\in S,\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S,\tau_i\geq C}(C-1)\geq(1/3)\psi\ ,$$

we have

$$\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i<C}\tau_i+\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i\geq C}C\geq 10(\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha},\tau_i<C}\tau_i+\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha},\tau_i\geq C}C)$$

Proof. First we show an upper bound on $S_{\text{good}}\cap\mathcal{T}_{2\alpha}$:

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha},\tau_i<C}\tau_i+\frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha},\tau_i\geq C}C\leq\frac{1}{n}\sum_{i\in S_{\text{good}}\cap\mathcal{T}_{2\alpha}}\tau_i\leq\psi/1000.$$

Then we show an lower bound on $S_{\text{bad}}\cap\mathcal{T}_{2\alpha}$:

$$\begin{aligned}&\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i>C}(C-1)\\ =\ &\frac{1}{n}\sum_{i\in S,\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S,\tau_i\geq C}(C-1)\\ &-(\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i\geq C}(C-1))\\ \geq\ &(1/3-1/1000)\psi\ .\end{aligned}$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i<C}\tau_i+\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i>C}C\geq\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap\mathcal{T}_{2\alpha},\tau_i>C}(C-1)\\
&=\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i<\rho}(\tau_i-1)+\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i>C}(C-1)\\
&-\left(\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S\setminus\mathcal{T}_{2\alpha},\tau_i<C}(\tau_i-1)+\frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S\setminus\mathcal{T}_{2\alpha},\tau_i>C}(C-1)\right)\\
&\geq(1/3-1/1000)\psi-\alpha\tau_{2\alpha n}\\
&\geq(1/3-2/1000)\psi
\end{aligned}$$

Combing the lower bound and the upper bound yields the desired statement □

A.5 Proof of the analysis of PRIME in Theorem 7

A.5.1 Proof of part 1 of Theorem 7 on differential privacy

Let (ε_0,δ_0) be the end-to-end target privacy guarantee. The $(0.01\varepsilon_0,0.01\delta_0)$-DP guarantee of DPRANGE follows from Lemma 2.3.5. We are left to show that DPMMWFILTER in Algorithm 6 satisfy $(0.99\varepsilon_0,0.99\delta_0)$-DP. To this end, we explicitly write out how many times we access the database and how much privacy is lost each time in an interactive version of DPMMWFILTER in Algorithm 17, which performs the same operations as DPMMWFILTER.

In order to apply Lemma 2.3.4, we cap ε at 0.9 in initializing ε_2. We call q_{spectral} and q_{size} T_1 times, each with (ε_1,δ_1) guarantee. In total this accounts for $(0.5\varepsilon,0.5\delta)$ privacy loss. The rest of the mechanisms are called $5T_1T_2$ times ($q_{\text{spectral}}(\cdot)$ and $q_{\text{MMW}}(\cdot)$ each call two DP mechanisms internally), each with (ε_2,δ_2) guarantee. In total this accounts for $(0.5\varepsilon,0.5\delta)$ privacy loss. Altogether, this is within the privacy budget of $(\varepsilon=0.99\varepsilon_0,\delta=0.99\delta_0)$.

We are left to show privacy of q_{spectral}, q_{MMW}, and q_{1Dfilter}, and q_{size} in Algorithm 16. We will assume for now that $|S_r^{(\ell)}|\geq n/2$ for all $\ell\in[T_1]$ and $r\in[T_2]$ and prove privacy. We show in the end that this happens with probability larger than $1-\delta_1$. In all sub-routines, we run Filter$(\cdot)$ in Algorithm 16 to simulate the filtering process so far and get the current set of

samples $S_{t_s}^{(s)}$. The following main technical lemma allows us to prove privacy of all interactive mechanisms. This is a counterpart of Lemma 2.3.6 used for DPFILTER. We provide a proof in §A.3.1.

Lemma A.5.1. *Let $S(\mathcal{D}_n) \subseteq \mathcal{D}_n$ denote the output of the simulated filtering process* Filter$(\cdot)$ *on $\mathcal{D}_n$ for a given set of parameters $(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]})$ in Algorithm 16. Then we have $d_\triangle(S(\mathcal{D}_n), S(\mathcal{D}')_n) \leq d_\triangle(\mathcal{D}_n, \mathcal{D}'_n)$, where $d_\triangle(\mathcal{D}, \mathcal{D}') \triangleq \max\{|\mathcal{D} \setminus \mathcal{D}'|, |\mathcal{D}' \setminus \mathcal{D}|\}$.*

This is a powerful tool for designing private mechanisms, as it guarantees that we can safely simulate the filtering process with privatized parameters and preserve the neighborhood of the dataset; if $\mathcal{D}_n \sim \mathcal{D}'_n$ are neighboring (i.e., $d_\Delta(\mathcal{D}_n, \mathcal{D}'_n) \leq 1$) then so are the filtered pair $S(\mathcal{D}_n)$ and $S(\mathcal{D}'_n)$ (i.e., $d_\Delta(S(\mathcal{D}_n), S(\mathcal{D}'_n)) \leq 1$). Note that in all the interactive mechanisms in Algorithm 16, the noise we need to add is proportional to the set sensitivity of Filter$(\cdot)$ defined as $\Delta_{\text{set}} \triangleq \max_{\mathcal{D}_n \sim \mathcal{D}'_n} d_\Delta(S(\mathcal{D}_n), S(\mathcal{D}'_n))$. If the repeated application of the Filter$(\cdot)$ is not a contraction in $d_\Delta(\cdot, \cdot)$, this results in a sensitivity blow-up. Fortunately, the above lemma ensures contraction of the filtering, proving that $\Delta_{\text{set}} = 1$. Hence, it is sufficient for us to prove privacy for two neighboring filtered sets $S \sim S'$ (as opposed to proving privacy for two neighboring original datasets before filtering $\mathcal{D}_n \sim \mathcal{D}'_n$).

In q_{spectral}, λ satisfy $(\varepsilon, 0)$-DP as the L_1 sensitivity is $\Delta_1 = (1/n)B^2 d$ (Definition 2.1.2) and we add $\text{Lap}(\Delta_1/\varepsilon)$. The release of μ also satisfy (ε, δ)-DP as the L_2 sensitivity is $\Delta_2 = 2B\sqrt{d}/n$, assuming $|S| \geq n/2$ as ensured by the stopping criteria, and we add $\mathcal{N}(0, \Delta_2(2\log(1.25/\delta))/\varepsilon)^2\mathbf{I})$. Note that in the outer loop call of q_{spectral}, we only release μ once in the end, and hence we count q_{spectral} as one access. On the other hand, in the inner loop, we use both μ and λ from q_{spectral} so we count it as two accesses.

In q_{size}, the returned set size $(\varepsilon, 0)$-DP as the L_1 sensitivity is $\Delta_1 = 1$ and we add $\text{Lap}(\Delta_1/\varepsilon)$. One caveat is that we need to ensure that the stopping criteria of checking $n^{(s)} > 3n/4$ ensures that $|S_t^{(s)}| > n/2$ with probability at least $1 - \delta_1$. This guarantees that the rest of the private mechanisms can assume $|S_t^{(s)}| > n/2$ in analyzing the sensitivity. Since Laplace distribution follows $f(z) = (\varepsilon/2)e^{-\varepsilon|z|}$, we have $\mathbb{P}(n^{(s)} > 3n/4 \text{ and } |S_t^{(s)}| < n/2) \leq (1/2)e^{-n\varepsilon/4}$. Hence,

the desired privacy is ensured for $(1/2)e^{-n\varepsilon/4} \le \delta_1$ (i.e., $n \ge (4/\varepsilon_1)\log(1/(2\delta_1))$).

In q_{MMW}, Σ is (ε,δ)-DP as the L_2 sensitivity is $\Delta_2 = B^2 d/n$, and we add $\mathcal{N}(0, \Delta_2(2\log(1.25/\delta))/\varepsilon)^2\mathbf{I})$. ψ is $(\varepsilon, 0)$-DP as the L_1 sensitivity is $\Delta_1 = 2B^2 d/n$ and we add $\mathrm{Lap}(\Delta_1/\varepsilon)$. This is made formal in the following theorem with a proof. in §A.5.1.1. This algorithm is identical to the MOD-SULQ algorithm introduced in [33] and analyzed in [46, Theorem 5], up to the choice of the noise variance. But a tighter analysis improvees over the MOD-SULQ analysis from [46] by a factor of d in the variance of added Gaussian noise as noted in [80].

Lemma A.5.2 (Differentially Private PCA)**.** *Consider a dataset $\{x_i \in \mathbb{R}^d\}_{i=1}^n$. If $\|x_i\|_2 \le 1$ for all $i \in [n]$, the following privatized (centered) second moment matrix satisfies (ε,δ)-differential privacy:*

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top + Z\ ,$$

with $Z_{i,j} \sim \mathcal{N}(0, (\,(1/(n\varepsilon))\sqrt{2\log(1.25/\delta)}\,)^2)$ for $i \ge j$ and $Z_{i,j} = Z_{j,i}$ for $i < j$.

In q_{1Dfilter}, the (ε,δ) differential privacy follows from that of DP-1DFILTER proved in Lemma 2.3.9.

A.5.1.1 Proof of Lemma A.5.2

Consider neighboring two databases $\mathcal{D} = \{x_i\}_{i=1}^n$ and $\tilde{\mathcal{D}} = \mathcal{D} \cup \{\tilde{x}_n\} \setminus \{x_n\}$, and let $A = (1/n)\sum_{x_i\in\mathcal{D}} x_i x_i^\top$ and $\tilde{A} = (1/n)\sum_{x_i\in\tilde{\mathcal{D}}} x_i x_i^\top$. Let B and $\tilde{B}$ be the noise matrix. Let $G = A + B$ and $\tilde{G} = \tilde{A} + \tilde{B}$. At point H, we have

$$\begin{aligned}
\ell_{D,\tilde{D}} = \log\frac{f_G(H)}{f_{\tilde{G}}(H)} &= \sum_{1\le i\le j\le d}\left(-\frac{1}{2\beta^2}(H_{ij}-A_{ij})^2 + \frac{1}{2\beta^2}\left(H_{ij}-\hat{A}_{ij}\right)^2\right)\\
&= \frac{1}{2\beta^2}\sum_{1\le i\le j\le d}\left(\frac{2}{n}(H_{ij}-A_{ij})(x_{n,i}x_{n,j}-\hat{x}_{n,i}\hat{x}_{n,j}) + \frac{1}{n^2}(\hat{x}_{n,i}\hat{x}_{n,j}-x_{n,i}x_{n,j})^2\right)\ .
\end{aligned}$$

Since $\|x_n\|_2 \le 1$ and $\|\tilde{x}_n\|_2 \le 1$, we have $\sum_{1\le i\le j\le d}(\hat{x}_{n,i}\hat{x}_{n,j} - x_{n,i}x_{n,j})^2 = 1/2\|\tilde{x}_n\tilde{x}_n^\top - x_n x_n^\top\|_F^2 \le 2$.

Now we bound the first term,

$$\begin{aligned} 2\sum_{1\le i\le j\le d}(H_{ij}-A_{ij})\left(x_{n,i}x_{n,j}-\hat{x}_{n,i}\hat{x}_{n,j}\right) &= \left\langle H-A, x_nx_n^\top-\tilde{x}_n\tilde{x}_n^\top\right\rangle \\ &= x_n^\top Bx_n-\tilde{x}_n^\top B\tilde{x}_n \\ &\le 2\|B\|_2\ . \end{aligned}$$

So we have $|\ell_{D,\tilde{D}}|\le\varepsilon$ whenever $\|B\|_2\le n\varepsilon\beta^2-1/n$.

For any fixed unit vector $\|v\|_2=1$, we have

$$v^\top Bv=2\sum_{1\le i\le j\le d}B_{ij}v_iv_j\sim\mathcal{N}(0,2\sum_{1\le i\le j\le d}v_i^2v_j^2) \;=\; \mathcal{N}(0,1)\ .$$

Then we have

$$\begin{aligned} \mathbb{P}\left(|\ell_{D,\tilde{D}}|\ge\varepsilon\right) &\le \mathbb{P}\left(\|B\|_2\ge n\varepsilon\beta^2-1/n\right) \\ &= \mathbb{P}\left(\mathcal{N}(0,1)\ge n\varepsilon\beta^2-\frac{1}{n}\right) \\ &= \Phi\left(\frac{1}{n}-n\varepsilon\beta^2\right)\ , \end{aligned}$$

where Φ is CDF of standard Gaussian. According to Gaussian mechanism, if $\beta=(1/(n\varepsilon))\sqrt{2\log(1.25/\delta)}$, we have $\Phi\left(\frac{1}{n}-n\varepsilon\beta^2\right)\le\delta$.

Algorithm 16: Interactive differentially private mechanisms for DPMMWFilter

$q_{\text{spectral}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$**:**

  $S \leftarrow \text{Filter}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$

  $\mu \leftarrow (1/|S|)\big(\sum_{i\in S} x_i\big) + \mathcal{N}(0, (2B\sqrt{2d\log(1.25/\delta)}/(n\varepsilon))^2\mathbf{I})$

  $\lambda \leftarrow \|M(S) - \mathbf{I}\|_2 + \text{Lap}(2B^2d/(n\varepsilon))$

  return (μ, λ)

$q_{\text{size}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$**:**

  $S \leftarrow \text{Filter}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$

  return $|S| + \text{Lap}(1/\varepsilon)$

$q_{\text{MMW}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \alpha^{(s)}, \mu_t^{(s)}, \varepsilon, \delta)$**:**

  $S \leftarrow \text{Filter}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$

  $\Sigma_{t_s+1}^{(s)} \leftarrow M(S) + \mathcal{N}(0, (2B^2d\sqrt{2\log(1.25/\delta)}/(n\varepsilon))^2\mathbf{I})$

  $U \leftarrow (1/\text{Tr}(\exp(\alpha^{(s)}\sum_{r=1}^{t_s+1}(\Sigma_r^{(s)} - \mathbf{I})))) \exp(\alpha^{(s)}\sum_{r=1}^{t_s+1}(\Sigma_r^{(s)} - \mathbf{I}))$

  $\psi \leftarrow \langle M(S) - \mathbf{I}, U\rangle + \text{Lap}(2B^2d/(n\varepsilon))$

  return $(\Sigma_{t_s+1}^{(s)}, U, \psi)$

$q_{\text{1Dfilter}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \mu, U, \alpha, \varepsilon, \delta)$**:**

  $S \leftarrow \text{Filter}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon, \delta)$

  return $\rho \leftarrow$ DP-1Dfilter$(\mu, U, \alpha, \varepsilon, \delta, S)$

$\text{Filter}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]})$**:**

  $S^{(1)} \leftarrow [n]$

  for epoch $\ell = 1, \ldots, s$ **do**

    $\alpha^{(\ell)} \leftarrow 1/(100(0.1/C + 1.01)\lambda^{(\ell)})$

    $S_1^{(\ell)} \leftarrow S^{(\ell)}$

    for $r = 1, \ldots, t_s$ **do**

      $S_{r+1}^{(\ell)} \leftarrow S_r^{(\ell)} \setminus \{i \mid i \in \mathcal{T}_{2\alpha}$ for $\{\tau_j = (x_j - \mu_r^{(\ell)})^\top U_r^{(\ell)}(x_j - \mu_r^{(\ell)})\}_{j\in S_r^{(\ell)}}$ and $\tau_i \geq \rho_r^{(\ell)} Z_r^{(\ell)}\}$, where $\mathcal{T}_{2\alpha}$ is defined in Definition 2.3.1.

  Output: $S_{t_s}^{(s)}$

Algorithm 17: Interactive version of DPMMWFILTER

Input: $\alpha \in (0,1)$, T_1, T_2, $\varepsilon_1 = \varepsilon/(4T_1)$, $\delta_1 = \delta/(4T_1)$,
$\varepsilon_2 = \min\{0.9, \varepsilon\}/(4\sqrt{10T_1T_2\log(4/\delta)})$, $\delta_2 = \delta/(20T_1T_2)$

for epoch $s = 1, 2, \ldots, T_1$ **do**
- $(\mu^{(s)}, \lambda^{(s)}) \leftarrow q_{\text{spectral}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s-1]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s-1]}, \varepsilon_1, \delta_1)$
- $n^{(s)} \leftarrow q_{\text{size}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s-1]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s-1]}, \varepsilon_1, \delta_1)$
- **if** $n^{(s)} \le 3n/4$ **then** terminate
- **if** $\lambda^{(s)} \le C\alpha\log(1/\alpha)$ **then**
  - **Output:** $\mu^{(s)}$
- $\alpha^{(s)} \leftarrow 1/(100(0.1/C + 1.01)\lambda^{(s)})$
- $t_s \leftarrow 0$
- **for** $t = 1, 2, \ldots, T_2$ **do**
  - $(\mu_t^{(s)}, \lambda_t^{(s)}) \leftarrow q_{\text{spectral}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \varepsilon_2, \delta_2)$
  - **if** $\lambda_t^{(s)} \le 0.5\lambda^{(s)}$ **then**
    - terminate epoch
  - **else**
    - $(\Sigma_t^{(s)}, U_t^{(s)}, \psi_t^{(s)}) \leftarrow$ $q_{\text{PMMW}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \alpha^{(s)}, \mu_t^{(s)}, \varepsilon_2, \delta_2)$
    - **if** $\psi_t^{(s)} \le (1/5.5)\lambda_t^{(s)}$ **then**
      - $\alpha_t^{(s)} \leftarrow 0$
    - **else**
      - $Z_t^{(s)} \leftarrow \text{Unif}([0,1])$
      - $\rho_t^{(s)} \leftarrow q_{\text{1Dfilter}}(\{\{\Psi_r^{(\ell)}\}_{r\in[t_\ell]}\}_{\ell\in[s]}, \{(\mu^{(\ell)}, \lambda^{(\ell)})\}_{\ell\in[s]}, \mu_t^{(s)}, U_t^{(s)}, \alpha, \varepsilon_2, \delta_2)$
      - $\alpha_t^{(s)} \leftarrow \alpha$
  - $\Psi_t^{(s)} \leftarrow (\mu_t^{(s)}, \lambda_t^{(s)}, \Sigma_t^{(s)}, U_t^{(s)}, \psi_t^{(s)}, Z_t^{(s)}, \rho_t^{(s)}, \alpha_t^{(s)})$
  - $t_s \leftarrow t$

Output: $\mu_{t_{T_1}}^{(T_1)}$

A.5.2 Proof of part 2 of Theorem 7 on accuracy

The accuracy of PRIME follows from the fact that DPRANGE returns a hypercube that contains all the clean data with high probability (Lemma 2.3.5) and that DPMMWFILTER achieves the desired accuracy (Theorem 26) if the original uncorrupted dataset S_{good} is α-subgaussian good. S_{good} is α-subgaussian good if we have $n = \widetilde{\Omega}(d/\alpha^2)$ as shown in Lemma A.6.3. We present the proof of Theorem 26 below. Then, we are left to show Lemma 2.3.8 in the following section.

Theorem 26 (Analysis of accuracy of DPMMWFILTER)**.** *Let S be an α-corrupted sub-Gaussian dataset, where $\alpha \le c$ for some universal constant $c \in (0, 1/2)$. Let S_{good} be α-subgaussian good with respect to $\mu \in \mathbb{R}^d$. Suppose $\mathcal{D} = \{x_i \in \bar{x} + [-B/2, B/2]^d\}_{i=1}^n$ be the projected dataset. If $n \ge \widetilde{\Omega}\left(\frac{d^{3/2}B^2 \log(2/\delta)}{\varepsilon\alpha \log 1/\alpha}\right)$, then* DPMMWFILTER *terminates after at most $O(\log dB^2)$ epochs and outputs $S^{(s)}$ such that with probability 0.9, we have $|S_t^{(s)} \cap S_{\text{good}}| \ge (1 - 10\alpha)n$ and*

$$\|\mu(S^{(s)}) - \mu\|_2 \lesssim \alpha\sqrt{\log 1/\alpha}\,.$$

Moreover, each epoch runs for at most $O(\log d)$ iterations.

Proof. In $s = O(\log_{0.98}((C\alpha\log(1/\alpha))/\|M(S^{(1)}) - \mathbf{I}\|_2))$ epochs, Lemma 2.3.8 guarantees that we find a candidate set $S^{(s)}$ of samples with $\|M(S^{(s)} - \mathbf{I}\|_2 \le C\alpha\log(1/\alpha)$. Lemma 2.3.2 ensures that we get the desired bound of $\|\mu(S^{(s)}) - \mu\|_2 = O(\alpha\sqrt{\log(1/\alpha)})$ as long as $S^{(s)}$ has enough clean data, i.e., $|S^{(s)} \cap S_{\text{good}}| \ge n(1-\alpha)$. Since Lemma 2.3.8 gets invoked at most $O((\log d)^2)$ times, we can take a union bound, and the following argument conditions on the good events in Lemma 2.3.8 holding, which happens with probability at least 0.99. To turn the average case guarantee of Lemma 2.3.8 into a constant probability guarantee, we apply the optional stopping theorem. Recall that the s-th epoch starts with a set $S^{(s)}$ and outputs a filtered set $S_t^{(s)}$ at the t-th inner iteration. We measure the progress by by summing the number of clean samples removed up to epoch s and iteration t and the number of remaining corrupted samples, defined as $d_t^{(s)} \triangleq |(S_{\text{good}} \cap S^{(1)}) \setminus S_t^{(s)}| + |S_t^{(s)} \setminus (S_{\text{good}} \cap S^{(1)})|$. Note that

$d_1^{(1)} = \alpha n$, and $d_t^{(s)} \geq 0$. At each epoch and iteration, we have

$$\mathbb{E}[d_{t+1}^{(s)} - d_t^{(s)} | d_1^{(1)}, d_2^{(1)}, \cdots, d_t^{(s)}] \;=\; \mathbb{E}\left[|S_{\text{good}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})| - |S_{\text{bad}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})|\right] \;\leq\; 0,$$

from part 1 of Lemma 2.3.8. Hence, $d_t^{(s)}$ is a non-negative super-martingale. By the optional stopping theorem, at stopping time, we have $\mathbb{E}[d_t^{(s)}] \leq d_1^{(1)} = \alpha n$. By the Markov inequality, $d_t^{(s)}$ is less than $10\alpha n$ with probability 0.9, i.e., $|S_t^{(s)} \cap S_{\text{good}}| \geq (1 - 10\alpha)n$. The desired bound in Theorem 26 follows from Lemma 2.3.2.

□

A.5.3 Proof of Lemma 2.3.8

In this section we state the formal version of this lemma and provide a proof.

Lemma A.5.3 (formal version of Lemma 2.3.8)**.** *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1. For an epoch $s \in [T_1]$ and an iteration $t \in [T_2]$, under the hypotheses of Lemma A.5.4, if S_{good} is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$ as in Definition A.6.2, $n = \widetilde{\Omega}(d^{3/2}\log(1/\delta)/(\varepsilon\alpha))$, and $|S_t^{(s)} \cap S_{\text{good}}| \geq (1 - 10\alpha)n$ then with probability $1 - O(1/\log^3 d)$ the conditions in Eqs. (A.11) and (A.12) hold. When these two conditions hold, more corrupted samples are removed in expectation than the uncorrupted samples, i.e., $\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| \leq \mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}|$. Further, for an epoch $s \in [T_1]$ there exists a constant $C > 0$ such that if $\|M(S^{(s)}) - \mathbf{I}\|_2 \geq C\,\alpha\log(1/\alpha)$, then with probability $1 - O(1/\log^2 d)$, the s-th epoch terminates after $O(\log d)$ iterations and outputs $S^{(s+1)}$ such that $\|M(S^{(s+1)}) - \mathbf{I}\|_2 \leq 0.98\|M(S^{(s)}) - \mathbf{I}\|_2$.*

Lemma A.5.3 is a combination of Lemma A.5.4 and Lemma A.5.5. We state the technical lemmas and subsequently provide the proofs.

Lemma A.5.4. *For an epoch s and an iteration t such that $\lambda^{(s)} > C\alpha\log(1/\alpha)$, $\lambda_t^{(s)} > 0.5\lambda_0^{(s)}$, and $n^{(s)} > 3n/4$, if $n \gtrsim \frac{B^2(\log B)d^{3/2}\log(1/\delta)}{\varepsilon\alpha}$ and $|S_t^{(s)} \cap S_{\text{good}}| \geq (1 - 10\alpha)n$ then with probability $1 - O(1/\log^3 d)$, the conditions in Eqs. (A.11) and (A.12) hold. When these two conditions hold we have $\mathbb{E}|S_t^{(s)} \setminus S_{t+1}^{(s)} \cap S_{\text{good}}| \leq \mathbb{E}|S_t^{(s)} \setminus S_{t+1}^{(s)} \cap S_{\text{bad}}|$. If $n \gtrsim \frac{B^2(\log B)d^{3/2}\log(1/\delta)}{\varepsilon\alpha}$, $\psi_t^{(s)} > \frac{1}{5.5}\lambda_t^{(s)}$,*

and $n^{(s)} > 3n/4$*, then we have with probability* $1 - O(1/\log^3 d)$*,* $\left\langle M(S_{t+1}^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle \le 0.76 \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle$.

Lemma A.5.5. *For an epoch* s *and for all* $t = 0, 1, \cdots, T_2 = O(\log d)$ *if Lemma A.5.4 holds,* $n^{(s)} > 3n/4$*, and* $n \gtrsim \frac{B^2(\log B)d^{3/2}\log(1/\delta)}{\varepsilon\alpha}$*, then we have* $\|M(S^{(s+1)}) - \mathbf{I}\|_2 \le 0.98\|M(S^{(s)}) - \mathbf{I}\|_2$ *with probability* $1 - O(1/\log^2 d)$.

A.5.3.1 Proof of Lemma A.5.4

Proof of Lemma A.5.4. To prove that we make progress for each iteration, we first show our dataset satisfies regularity conditions in Eqs. (A.11) and (A.12) that we need for DP-1DFILTER. Following Lemma A.5.6 implies with probability $1 - 1/(\log^3 d)$, our scores satisfies the regularity conditions needed in Lemma 2.3.9.

Lemma A.5.6. *For each epoch* s *and iteration* t*, under the hypotheses of Lemma A.5.4, with probability* $1 - O(1/\log^3 d)$*, we have*

$$\frac{1}{n} \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}} \tau_i \le \psi/1000 \tag{A.11}$$

$$\frac{1}{n} \sum_{i \in S_{\text{good}} \cap S_t^{(s)}} (\tau_i - 1) \le \psi/1000\,, \tag{A.12}$$

where $\psi \triangleq \frac{1}{n}\sum_{i \in S_t^{(s)}}(\tau_i - 1)$.

Then by Lemma 2.3.9 our DP-1DFILTER gives us a threshold ρ such that

$$\sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\} \le \sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\}\,.$$

Conditioned on the hypotheses and the claims of Lemma 2.3.9, according to our filter rule from Algorithm 6, we have

$$\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| = \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\}$$

and

$$\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}| = \sum_{i \in S_{\text{bad}} \cap \mathcal{T}_{2\alpha}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\} .$$

This implies $\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| \le \mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}|$. At the same time, Lemma 2.3.9 gives us a ρ such that with probability $1 - O(\log^3 d)$

$$\frac{1}{n}\sum_{i \in S_{t+1}^{(s)}} (\tau_i - 1) - 2\alpha \le \frac{1}{n}\sum_{\tau_i \le \rho} (\tau_i - 1) \le \frac{3}{4} \cdot \frac{1}{n}\sum_{i \in S_t^{(s)}} (\tau_i - 1) .$$

Hence, we have

$$\begin{aligned}
\left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle - \left\langle M(S_{t+1}^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle &= \frac{1}{n}\sum_{i \in S_t^{(s)} \setminus S_{t+1}^{(s)}} (\tau_i - 1) \\
&\ge \frac{1}{4n}\sum_{i \in S_t^{(s)}} (\tau_i - 1) - 2\alpha \\
&\overset{(a)}{\ge} \frac{1}{4} \cdot \frac{998}{1000} \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle ,
\end{aligned}$$

where (a) follows from our assumption on λ_t and stopping criteria. Rearranging the terms completes the proof. □

A.5.3.2 Proof of Lemma A.5.6

Proof of Lemma A.5.6. First of all, Lemma A.6.9, Lemma A.6.10 and Lemma A.6.11 gives us following Lemma A.5.7, which basically shows with enough samples, we can make sure the noises added for privacy guarantees are small enough with probability $1 - O(1/\log^3 d)$.

Lemma A.5.7. *For $\alpha \in (0, 0.5)$, if $n \gtrsim \frac{B^2(\log B)d^{3/2}\log(1/\delta)}{\varepsilon\alpha}$ and $n^{(s)} > 3n/4$ then we have with probability $1 - O(1/\log^3 d)$, following conditions simultaneously hold:*

1. $\|\mu_t^{(s)} - \mu(S_t^{(s)})\|_2^2 \le 0.001\alpha \log 1/\alpha$
2. $|\psi_t^{(s)} - \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle| \le 0.001\alpha \log 1/\alpha$

3. $\left|\lambda_t^{(s)} - \|M(S_t^{(s)}) - \mathbf{I}\|_2\right| \leq 0.001\alpha \log 1/\alpha$

4. $\left|\lambda^{(s)} - \|M(S^{(s)}) - \mathbf{I}\|_2\right| \leq 0.001\alpha \log 1/\alpha$

5. $\left\|M(S_{t+1}^{(s)}) - \Sigma_t^{(s)}\right\|_2 \leq 0.001\alpha \log 1/\alpha$

6. $\|\mu^{(s)} - \mu(S^{(s)})\|_2^2 \leq 0.001\alpha \log 1/\alpha$

Now under above conditions, since $\lambda_1^{(s)} > C\alpha \log 1/\alpha$, we have $\|M(S_t^{(s)}) - \mathbf{I}\|_2 > 0.5(C - 0.002)\alpha \log 1/\alpha$. Using the fact that $\mu(S_t^{(s)}) = (1/n)\sum_{i \in S_t^{(s)}} x_i$, we also have

$$\begin{aligned}
& \frac{1}{n}\sum_{i \in S_t^{(s)}} (\tau_i - 1) \\
= \; & \frac{1}{n}\sum_{i \in S_t^{(s)}} \left\langle \left(x_i - \mu_t^{(s)}\right)\left(x_i - \mu_t^{(s)}\right)^\top - \mathbf{I}, U_t^{(s)} \right\rangle \\
= \; & \frac{1}{n}\sum_{i \in S_t^{(s)}} \left\langle \left(x_i - \mu(S_t^{(s)})\right)\left(x_i - \mu(S_t^{(s)})\right)^\top - \mathbf{I}, U_t^{(s)} \right\rangle \\
& + \; \frac{|S_t^{(s)}|}{n}\left\langle \left(\mu(S_t^{(s)}) - \mu_t^{(s)}\right)\left(\mu(S_t^{(s)}) - \mu_t^{(s)}\right)^\top, U_t^{(s)} \right\rangle \\
= \; & \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle + \frac{|S_t^{(s)}|}{n}\left\langle \left(\mu(S_t^{(s)}) - \mu_t^{(s)}\right)\left(\mu(S_t^{(s)}) - \mu_t^{(s)}\right)^\top, U_t^{(s)} \right\rangle .
\end{aligned}$$

Thus, from the first and the second claims in Lemma A.5.7, we have

$$|\psi - \psi_t^{(s)}| \leq 0.002\ \alpha \log 1/\alpha\ . \tag{A.13}$$

For an epoch s and an iteration t, since $\alpha n \le S_{\text{good}} \cap \mathcal{T}_{2\alpha} \cap S_t^{(s)} \le 2\alpha n$, we have

$$
\begin{aligned}
\frac{1}{n} \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha} \cap S_t^{(s)}} \tau_i &= \frac{1}{n} \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha} \cap S_t^{(s)}} \left\langle (x_i - \mu_t^{(s)})(x_i - \mu_t^{(s)})^\top, U_t^{(s)} \right\rangle \\
&\overset{(a)}{\le} \frac{2}{n} \sum_{i \in S_{\text{good}} \cap \mathcal{T}_{2\alpha} \cap S_t^{(s)}} \left\langle (x_i - \mu)(x_i - \mu)^\top, U_t^{(s)} \right\rangle + \frac{2|S_{\text{good}} \cap \mathcal{T}_{2\alpha} \cap S_t^{(s)}|}{n} \left\langle (\mu - \mu_t^{(s)})(\mu - \mu_t^{(s)})^\top, U_t^{(s)} \right\rangle \\
&\overset{(b)}{\le} O(\alpha \log 1/\alpha) + 4\alpha \left\langle (\mu - \mu_t^{(s)})(\mu - \mu_t^{(s)})^\top, U_t^{(s)} \right\rangle \\
&\le O(\alpha \log 1/\alpha) + 4\alpha \|\mu_t^{(s)} - \mu\|_2^2 \\
&\le O(\alpha \log 1/\alpha) + 4\alpha \left(\|\mu - \mu(S_t^{(s)})\|_2 + \|\mu(S_t^{(s)}) - \mu_t^{(s)}\|_2 \right)^2 \\
&\overset{(c)}{\le} O\left(\alpha \log 1/\alpha\right) + 4\alpha \left(O\left(\alpha \sqrt{\log 1/\alpha}\right) + \sqrt{\alpha \left(O\left(\alpha \log 1/\alpha\right) + \|M(S_t^{(s)}) - \mathbf{I}\|_2 \right)} + \|\mu(S_t^{(s)}) - \mu_t^{(s)}\|_2 \right)^2 \\
&\le O(\alpha \log 1/\alpha) + 8\alpha^2 \left(\|M(S_t^{(s)}) - \mathbf{I}\|_2 + O\left(\alpha \log 1/\alpha\right) \right) + O(8\alpha^3 \log 1/\alpha) + 8\alpha^2 \log 1/\alpha \\
&\overset{(d)}{\le} \frac{1}{1000} \left(\frac{\|M(S_t^{(s)}) - \mathbf{I}\|_2 - 0.001\ \alpha \log 1/\alpha}{5.5} - 0.002\ \alpha \log 1/\alpha \right) \\
&\le \frac{\psi_t^{(s)} - 0.002\ \alpha \log 1/\alpha}{1000} \\
&\le \frac{\psi}{1000}\,,
\end{aligned}
$$

where (a) follows from the fact that for any vector x, y, z, we have $(x-y)(x-y)^\top \preceq 2(x-z)(x-z)^\top + 2(y-z)(y-z)^\top$, (b) follows from Lemma A.6.4, (c) follows from Lemma A.6.7, (d) follows from our choice of large constant C, and in the last inequality we used Eq. (A.13).

Similarly we have

$$
\begin{aligned}
& \frac{1}{n} \sum_{i \in S_{\text{good}} \cap S_t^{(s)}} (\tau_i - 1) \\
= \; & \frac{1}{n} \sum_{i \in S_{\text{good}} \cap S_t^{(s)}} \left\langle (x_i - \mu_t^{(s)})(x_i - \mu_t^{(s)})^\top - \mathbf{I}, U_t^{(s)} \right\rangle \\
= \; & \frac{1}{n} \sum_{i \in S_{\text{good}} \cap S_t^{(s)}} \left\langle \left(x_i - \mu(S_{\text{good}} \cap S_t^{(s)})\right) \left(x_i - \mu(S_{\text{good}} \cap S_t^{(s)})\right)^\top - \mathbf{I}, U_t^{(s)} \right\rangle \\
& + \frac{|S_{\text{good}} \cap S_t^{(s)}|}{n} \left\langle \left(\mu(S_{\text{good}} \cap S_t^{(s)}) - \mu_t^{(s)}\right) \left(\mu(S_{\text{good}} \cap S_t^{(s)}) - \mu_t^{(s)}\right)^\top, U_t^{(s)} \right\rangle \\
\overset{(a)}{\leq} \; & O\left(\alpha \log 1/\alpha\right) + \left\| \mu(S_{\text{good}} \cap S_t^{(s)}) - \mu_t^{(s)} \right\|_2^2 \\
\leq \; & O\left(\alpha \log 1/\alpha\right) + \left(\left\| \mu(S_{\text{good}} \cap S_t^{(s)}) - \mu \right\|_2 + \left\| \mu - \mu(S_t^{(s)}) \right\|_2 \right)^2 + 0.001\, \alpha \log 1/\alpha \\
\overset{(b)}{\leq} \; & O\left(\alpha \log 1/\alpha\right) + \left(O(\alpha \sqrt{\log 1/\alpha}) + \sqrt{\alpha(\|M(S_t^{(s)}) - \mathbf{I}\|_2 + O(\alpha \log 1/\alpha))} \right)^2 + 0.001\, \alpha \log 1/\alpha \\
\leq \; & O\left(\alpha \log 1/\alpha\right) + \alpha \left(\|M(S_t^{(s)}) - \mathbf{I}\|_2 + O\left(\alpha \log 1/\alpha\right) \right) + O(\alpha^2 \log 1/\alpha) + +0.001\, \alpha \log 1/\alpha \\
\overset{(c)}{\leq} \; & \frac{1}{1000} \left(\frac{\|M(S_t^{(s)}) - \mathbf{I}\|_2 - 0.001\, \alpha \log 1/\alpha}{5.5} - 0.002\, \alpha \log 1/\alpha \right) \\
\leq \; & \frac{\psi_t^{(s)} - 0.002\, \alpha \log 1/\alpha}{1000} \\
\leq \; & \frac{\psi}{1000}\,,
\end{aligned}
$$

where (a) follows from Lemma A.6.4, (b) follows from Lemma A.6.5 and Lemma A.6.7 and (c) follows from our choice of large constant C.

□

A.5.3.3 Proof of Lemma A.5.5

Proof of Lemma A.5.5. Under the conditions of Lemma A.5.7, we have picked n large enough such that with probability $1 - O(1/\log^3 d)$, we have

$$\|\Sigma_t^{(s)} - \mathbf{I}\|_2 \approx_{0.01} \|M(S_t^{(s)}) - \mathbf{I}\|_2\,.$$

By Lemma A.5.4, we now have

$$\begin{aligned}\left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle &\leq 0.76 \left\langle M(S_{t-1}^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle \\ &\leq 0.76 \left\langle M(S_1^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle \\ &\leq 0.76 \|M(S_1^{(s)}) - \mathbf{I}\|_2 \,. \end{aligned} \tag{A.14}$$

Since $\lambda_1^{(s)} > C\alpha \log 1/\alpha$, we have $\|M(S_t^{(s)}) - \mathbf{I}\|_2 > 0.5(C - 0.002)\alpha \log 1/\alpha$. Combining the above inequality and the fifth claim of Lemma A.5.7 together, we have

$$\left\langle \Sigma_t^{(s)} - \mathbf{I}, U_t^{(s)} \right\rangle \leq \left\langle M(S_t^{(s)}) - \mathbf{I}, U_t^{(s)} \right\rangle + \|\Sigma_t^{(s)} - M(S_t^{(s)})\|_2 \leq 0.77 \|M(S_1^{(s)}) - \mathbf{I}\|_2 \,.$$

By Lemma A.6.1, we have $M(S_t^{(s)}) - \mathbf{I} \preceq M(S_1^{(s)}) - \mathbf{I}$. by our choice of $\alpha^{(s)}$, we have $\alpha^{(s)} \left(M(S_{t+1}^{(s)}) - \mathbf{I} \right) \preceq \frac{1}{100}\mathbf{I}$ and $\alpha^{(s)} \left(\Sigma_t^{(s)} - \mathbf{I} \right) \preceq \frac{1}{100}\mathbf{I}$. Therefore, by Lemma A.6.13, we have

$$\begin{aligned} &\left\| \sum_{t=1}^{T_2} \Sigma_t^{(s)} - \mathbf{I} \right\|_2 \\ \leq\ & \sum_{t=1}^{T_2} \left\langle \Sigma_t^{(s)} - \mathbf{I}, U_t^{(s)} \right\rangle + \alpha^{(s)} \sum_{t=0}^{T_2} \left\langle U_t^{(s)}, \left| \Sigma_t^{(s)} - \mathbf{I} \right| \right\rangle \|\Sigma_t^{(s)} - \mathbf{I}\|_2 + \frac{\log(d)}{\alpha^{(s)}} \\ \overset{(a)}{\leq}\ & \sum_{t=1}^{T_2} \left\langle \Sigma_t^{(s)} - \mathbf{I}, U_t^{(s)} \right\rangle + \frac{1}{100} \sum_{t=1}^{T_2} \left\langle U_t^{(s)}, \left| \Sigma_t^{(s)} - \mathbf{I} \right| \right\rangle + 200 \log(d) \|M(S_1^{(s)}) - \mathbf{I}\|_2 \end{aligned}$$

where (a) follows from our choice of $\alpha^{(s)}$ and C. By Lemma A.6.6, $M(S_t^{(s)}) - \mathbf{I} \succeq -c_1\alpha \log 1/\alpha \cdot I$ for $t = 1, 2, \cdots, T_2$, we have

$$|M(S_t^{(s)}) - \mathbf{I}| \preceq M(S_t^{(s)}) - \mathbf{I} + 2c_1\alpha \log 1/\alpha \, \mathbf{I},$$

and hence

$$\left\langle U_t^{(s)}, \left| M(S_t^{(s)}) - \mathbf{I} \right| \right\rangle \leq \left\langle U_t^{(s)}, M(S_t^{(s)}) - \mathbf{I} \right\rangle + 2c_1\alpha \log 1/\alpha$$

Meanwhile, we have

$$M(S_{t+1}^{(s)}) - \mathbf{I} - \|\Sigma_t^{(s)} - M(S_t^{(s)})\|_2 \, \mathbf{I} \preceq \Sigma_t^{(s)} - \mathbf{I} \preceq M(S_{t+1}^{(s)}) - \mathbf{I} + \|\Sigma_t^{(s)} - M(S_t^{(s)})\|_2 \, \mathbf{I} \,.$$

Hence,

$$|\Sigma_t^{(s)} - \mathbf{I}| \preceq M(S_t^{(s)}) - \mathbf{I} + (3\|\Sigma_t^{(s)} - M(S_t^{(s)})\|_2 + 2c_1\alpha \log 1/\alpha)\, \mathbf{I}$$

Together with Eq. (A.14), we have

$$\begin{aligned} & \left\langle U_t^{(s)}, \left|\Sigma_t^{(s)} - \mathbf{I}\right| \right\rangle \\ \leq\ & \left\langle U_t^{(s)}, M(S_t^{(s)}) - \mathbf{I} \right\rangle + 3\|\Sigma_t^{(s)} - M(S_t^{(s)})\|_2 + 2c_1\alpha \log 1/\alpha \\ \leq\ & 0.79 \left\|M(S_1^{(s)}) - \mathbf{I}\right\|_2 + 2c_1\alpha \log 1/\alpha\ . \end{aligned}$$

By Lemma A.6.6, we have $M(S_t^{(s)}) - \mathbf{I} \succeq -c_1\alpha \log 1/\alpha\, \mathbf{I}$. Also, we know $M(S_t^{(s)}) - \mathbf{I} \preceq M(S_1^{(s)}) - \mathbf{I}$. Then we have

$$\begin{aligned} & \left\|M(S_{T_2}^{(s)}) - \mathbf{I}\right\|_2 \\ \leq\ & \frac{1}{T_2}\left\|\sum_{i=1}^{T_2} M(S_t^{(s)}) - \mathbf{I}\right\|_2 \\ \leq\ & \frac{1}{T_2}\left\|\sum_{i=1}^{T_2} \Sigma_t^{(s)} - \mathbf{I}\right\|_2 + 0.001\ \alpha \log 1/\alpha \\ \leq\ & \frac{1}{T_2}\left(\sum_{t=1}^{T_2}\left\langle \Sigma_t^{(s)} - \mathbf{I}, U_t^{(s)}\right\rangle + \frac{1}{100}\sum_{t=1}^{T_2}\left\langle U_t^{(s)}, \left|\Sigma_t^{(s)} - \mathbf{I}\right|\right\rangle + 200\log(d)\|M(S_1^{(s)}) - \mathbf{I}\|_2\right) + 0.001\ \alpha \log 1/\alpha \\ \leq\ & 0.79\|M(S_1^{(s)}) - \mathbf{I}\|_2 + 2c_1\alpha \log 1/\alpha + \frac{200\log(d)}{T_2}\|M(S_1^{(s)}) - \mathbf{I}\|_2 + 0.001\ \alpha \log 1/\alpha \\ \leq\ & 0.98\ \|M(S_1^{(s)}) - \mathbf{I}\|_2\ , \end{aligned}$$

where the last inequality follows from our assumption that $\lambda_0^{(s)} > C\alpha \log 1/\alpha$, and conditions of Lemma A.5.7 hold and we have $\|M(S_t^{(s)}) - \mathbf{I}\|_2 > 0.5(C - 0.002)\alpha \log 1/\alpha$. □

A.6 Technical lemmas

A.6.1 Lemmata for sub-Gaussian regularity from [73]

Lemma A.6.1 ([73, Lemma 3.4])**.** *If $S' \subset S$, then $M(S') \preceq M(S)$.*

Definition A.6.2 ([73, Definition 4.1]). *Let D be a distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\mathbf{I}$. For $0 < \alpha < 1/2$, we say a set of points $S = \{X_1, X_2, \cdots, X_n\}$ is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$ if following inequalities are satisfied:*

- $\|\mu(S) - \mu\|_2 \lesssim \alpha\sqrt{\log 1/\alpha}$ *and* $\left\|\frac{1}{|S|}\sum_{i\in S}(X_i - \mu(S))(X_i - \mu(S))^\top - \mathbf{I}\right\|_2 \lesssim \alpha \log 1/\alpha$.
- *for any subset $T \subset S$ so that $|T| = 2\alpha|S|$, we have*

$$\left\|\frac{1}{|T|}\sum_{i\in T} X_i - \mu\right\|_2 \lesssim \sqrt{\log 1/\alpha} \quad \text{and} \quad \left\|\frac{1}{|T|}\sum_{i\in T}(X_i - \mu(S))(X_i - \mu(S))^\top - \mathbf{I}\right\|_2 \lesssim \log 1/\alpha\ .$$

Lemma A.6.3 ([73, Lemma 4.1]). *A set of i.i.d. samples from an identity covariance sub-Gaussian distribution of size $n = \Omega\left(\frac{d+\log 1/\delta}{\alpha^2 \log 1/\alpha}\right)$ is α-subgaussian good with respect to μ with probability $1 - \delta$.*

Lemma A.6.4 ([73, Fact 4.2]). *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1. If S_{good} is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$, then for any $T \subset S$ such that $|T| \le 2\alpha|S|$, we have for any unit vector $v \in \mathbb{R}^d$*

$$\frac{1}{|S|}\sum_{X_i\in T}\langle (X_i - \mu), v\rangle^2 \lesssim \alpha \log 1/\alpha\ .$$

For any subset $T \subset S$ such that $|T| \ge (1 - 2\alpha)|S|$, we have

$$\left\|\frac{1}{|S|}\sum_{i\in T}(x_i - \mu)(x_i - \mu)^\top - \mathbf{I}\right\|_2 \lesssim \alpha \log 1/\alpha \quad \text{and} \quad ,$$
$$\left\|\frac{1}{|S|}\sum_{i\in T}(x_i - \mu(T))(x_i - \mu(T))^\top - \mathbf{I}\right\|_2 \lesssim \alpha \log 1/\alpha$$

Lemma A.6.5 ([73, Corollary 4.3]). *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1. If S_{good} is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$, then for any $T \subset S$ such that $|T| \le 2\alpha|S|$, we have*

$$\left\|\frac{1}{|S|}\sum_{X_i\in T}(X_i - \mu)\right\|_2 \lesssim \alpha\sqrt{\log 1/\alpha}\ .$$

For any subset $T \subset S$ such that $|T| \geq (1-2\alpha)|S|$, we have

$$\|\mu(T) - \mu\|_2 \lesssim \alpha\sqrt{\log 1/\alpha}\ .$$

Lemma A.6.6 ([73, Lemma 4.5])**.** *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1. If S_{good} is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$, then for any $T \subset S$ such that $|T \cap S_{\text{good}}| \geq (1-2\alpha)|S|$, then there is some universal constant c_1 such that*

$$\frac{1}{|S|}\sum_{i\in T}(x_i - \mu(T))(x_i - \mu(T))^\top \succeq (1 - c_1\alpha\log 1/\alpha)\mathbf{I}\ .$$

Lemma A.6.7 ([73] Lemma 4.6)**.** *Let S be an α-corrupted sub-Gaussian dataset under Assumption 1. If S_{good} is α-subgaussian good with respect to $\mu \in \mathbb{R}^d$, then for any $T \subset S$ such that $|T \cap S_{\text{good}}| \geq (1-2\alpha)|S|$, we have*

$$\|\mu(T) - \mu\|_2 \leq \frac{1}{1-\alpha}\cdot\left(\sqrt{\alpha\left(\|M(T) - \mathbf{I}\|_2 + O\left(\alpha\log 1/\alpha\right)\right)} + O\left(\alpha\sqrt{\log 1/\alpha}\right)\right)\ .$$

A.6.2 Auxiliary Lemmas on Laplace and Gaussian mechanism

Lemma A.6.8 (Theorem A.1 in [79])**.** *Let $\varepsilon \in (0,1)$ be arbitrary. For $c^2 \geq 2\ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma^2 \geq c^2\Delta_2 f/\varepsilon$ is (ε,δ)-differentially private.*

Lemma A.6.9. *Let $Y \sim \text{Lap}(b)$. Then for all $h > 0$, we have $\mathbb{P}(|Y| \geq hb) = e^{-h}$.*

Lemma A.6.10 (Tail bound of χ-square distribution [203])**.** *Let $x_i \sim \mathcal{N}(0,\sigma^2)$ for $i = 1, 2, \cdots, d$. Then for all $\zeta \in (0,1)$, we have $\mathbb{P}(\|X\|_2 \geq \sigma\sqrt{d\log(1/\zeta)}) \leq \zeta$.*

Lemma A.6.11 ([190, Corollary 2.3.6])**.** *Let $Z \in \mathbb{R}^{d\times d}$ be a matrix such that $Z_{i,j} \sim \mathcal{N}(0,\sigma^2)$ for $i \geq j$ and $Z_{i,j} = Z_{j,i}$ for $i < j$. For $\forall\zeta \in (0,1)$, then with probability $1-\zeta$ we have $\|Z\|_2 \leq \sigma\sqrt{d}\log(1/\zeta)$.*

Lemma A.6.12 (Accuracy of the histogram using Gaussian Mechanism)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^{\mathcal{S}}$ be a histogram over K bins. For any dataset $D \in \mathcal{X}^n$ and ε, Gaussian Mechanism is an (ε,δ)-differentially private algorithm $M(D)$ such that given*

with probability $1-\zeta$ we have

$$\|M(D) - f(D)\|_\infty \leq O(\frac{\sqrt{\log(K/\zeta)\log(1/\delta)}}{\varepsilon n})\ .$$

Proof. First notice that the ℓ_2 sensitivity of histogram function f is $\sqrt{2}/n$. Thus, by Lemma A.6.8, by adding noise $\mathcal{N}(0, (\frac{2\sqrt{2\log(1.25/\delta)}}{n\varepsilon})^2)$ to each entry of f, we have a (ε, δ) differentially private algorithm. Since Gaussian tail bound implies that $\mathbb{P}_{x\sim\mathcal{N}(0,\sigma^2)}[x \geq \Omega(\sqrt{\log(K/\eta)}\sigma)] \leq \eta/K$, we have that with probability $1-\eta$, the ℓ_∞ norm of the added noise is bounded by $O(\frac{\sqrt{\log(1/\delta)\log(K/\eta)}}{n\varepsilon})$. This concludes the proof. □

A.6.3 Analysis of $\|M(S_t^{(s)}) - \mathbf{I}\|_2$ shrinking

For any symmetric matrix $A = \sum_{i=1}^d \lambda_i v_i v_i^\top$, we let $|A|$ denote $|A| = \sum_{i=1}^d |\lambda_i| v_i v_i^\top$.

Lemma A.6.13 (Regret bound, Special case of [8, Theorem 3.1])**.** *Let*

$$U_t \;=\; \frac{\exp(\alpha \sum_{k=1}^{t-1}(\Sigma_k - \mathbf{I}))}{\mathrm{Tr}(\exp(\alpha \sum_{k=1}^{t-1}(\Sigma_k - \mathbf{I})))}\,,$$

and α satisfies $\alpha(\Sigma_t - \mathbf{I}) \preceq I$ for all $k \in [T]$, then for all $U \succeq 0$, $\mathrm{Tr}(U) = 1$, it holds that

$$\sum_{t=1}^T \langle (\Sigma_t - \mathbf{I}), U - U_t\rangle \leq \alpha \sum_{t=1}^T \langle |(\Sigma_t - \mathbf{I}), U_t|\rangle \cdot \|(\Sigma_t - \mathbf{I})\|_2 + \frac{\log d}{\alpha}.$$

Rearranging terms, and taking a supremum over U, we obtain that

$$\|\sum_{t=1}^T (\Sigma_t - \mathbf{I})\|_2 \leq \sum_{t=1}^T \langle U_t, (\Sigma_t - \mathbf{I})\rangle + \alpha \sum_{t=1}^T \langle |(\Sigma_t - \mathbf{I}), U_t|\rangle \cdot \|(\Sigma_t - \mathbf{I})\|_2 + \frac{\log d}{\alpha}.$$

A.7 Exponential time DP robust mean estimation of sub-Gaussian and heavy tailed distributions

In this section, we give a self-contained proof of the privacy and utility of our exponential time robust mean estimation algorithm for sub-Gaussian and heavy tailed distributions. The proof relies on the resilience property of the uncorrupted data as shown in the following lemmas.

Lemma A.7.1 (Lemma 10 in [186])**.** *If a set of points $\{x_i\}_{i\in S}$ lying in $\mathbb{R}^d$ is (σ, α)-resilient around a point μ, then*

$$\|\frac{1}{|T'|}\sum_{i\in T'}(x_i - \mu)\|_2 \leq \frac{2-\alpha}{\alpha}\sigma.$$

for all sets T' of size at least $\alpha|S|$.

Lemma A.7.2 (Finite sample resilience of sub-Gaussian distributions [217, Theorem G.1])**.** *Let S_{good} be a set of i.i.d. points from a sub-Gaussian distribution $\mathcal{D}$ with a parameter $\mathbf{I}_d$. Given that $|S_{\text{good}}| = \Omega((d+\log(1/\zeta))/(\alpha^2 \log 1/\alpha))$, S_{good} is $(\alpha\sqrt{\log(1/\alpha)}, \alpha)$-resilient around its mean μ with probability $1-\zeta$.*

Lemma A.7.3 (Finite sample resilience of heavy-tailed distributions [217, Theorem G.2])**.** *Let S_{good} be a set of i.i.d. samples drawn from distribution $\mathcal{D}$ whose mean and covariance are μ, Σ respectively, and that $\Sigma \preceq I$. Given that $|S| = \Omega(d/(\zeta\alpha))$, there exists a constant c_ζ that only depends on ζ such that S_{good} is $(c_\zeta\sqrt{\alpha}, \alpha)$-resilient around μ with probability $1-\zeta$.*

A.7.1 Case of heavy-tailed distributions and a proof of Theorem 9

Lemma A.8.1 ensures that DPRANGE-HT returns samples in a bounded support of Euclidean distance $\sqrt{d}B/2$ with $B = 50/\sqrt{\alpha}$ where $(1-2\alpha)n$ samples are uncorrupted (αn is corrupted by adversary and αn can be corrupted by the pre-processing step). For a $(c_\zeta\sqrt{3\alpha}, 3\alpha)$-resilient dataset, we first show that $R(S)$ is robust against corruption.

Lemma A.7.4 (α-corrupted data has small $R(S)$)**.** *Let S be the set of 2α-corrupted data. Given that $n = \Omega(d/(\zeta\alpha))$, with probability $1-\zeta$, $R(S) \le c_\zeta\sqrt{3\alpha}$.*

This follows immediately by selecting S' to be the uncorrupted $(1-2\alpha)$ fraction of the dataset and applying $(c_\zeta\sqrt{3\alpha}, 3\alpha)$-resilience. After pre-processing, we have that $\|x_i - \bar{x}\|_2 \le B\sqrt{d}/2$, and then clearly $R(\cdot)$ has sensitivity $\Delta_R \le B\sqrt{d}/n$.

Lemma A.7.5 (Sensitivity and Privacy of $\hat{R}(S)$)**.** *Given that $\hat{R}(S) = R(S) + \text{Lap}(\frac{3B\sqrt{d}}{n\varepsilon})$, $\hat{R}(S)$ is $(\varepsilon/3, 0)$-differentially private. Further, with probability $1-\delta/3$, $|\hat{R}(S) - R(S)| \le \frac{3B\sqrt{d}\log(3/\delta)}{n\varepsilon}$.*

In the algorithm, we first compute $\hat{R}(S)$. If $\hat{R}(S) \ge 2c_\zeta\sqrt{\alpha}$, we stop and output $\emptyset$. Otherwise, we use exponential mechanism with score function $d(\hat{\mu}, S)$ to find an estimate $\hat{\mu}$. We prove the privacy guarantee of our algorithm as follows.

Lemma A.7.6 (Privacy)**.** *Algorithm 8 is (ε, δ)-differentially private if $n \ge 6B\sqrt{d}\log(3/\delta)/(c_\zeta\varepsilon\sqrt{\alpha})$.*

Proof. We consider neighboring datasets S, S' under the following two scenario

1. $R(S) > 3c_\zeta\sqrt{\alpha}$

   In this case, given that $n \geq \frac{6B\sqrt{d}\log(3/\delta)}{c_\zeta\sqrt{\alpha}\varepsilon}$, we have $\widehat{R}(S) > 2c_\zeta\sqrt{\alpha}$ and the output of the algorithm $\mathcal{A}(S) = \emptyset$ with probability at least $1 - \delta/3$, and $\mathcal{A}(S') = \emptyset$ with probability at least $1 - \delta/3$. Thus, for any set Q, $\mathbb{P}[\mathcal{A}(S) \in Q] \leq \mathbb{P}[\mathcal{A}(S') \in Q] + \delta/3$.

2. $R(S) \leq 3c_\zeta\sqrt{\alpha}$

   Lemma A.7.7 (Sensitivity of $d(\hat{\mu}, S)$)**.** *Given that $R(S) \leq 3c_\zeta\sqrt{\alpha}$, for any neighboring dataset S', $|d(\hat{\mu}, S) - d(\hat{\mu}, S')| \leq 12c_\zeta/(n\sqrt{\alpha})$.*

   In this case, the privacy guarantee of $\hat{R}(S)$ yields that $\mathbb{P}[\hat{R}(S) \in Q] \leq \exp(\varepsilon/3) \cdot \mathbb{P}[\hat{R}(S') \in Q]$. Lemma A.7.7 yields that $\mathbb{P}[\hat{\mu}(S) \in Q] \leq \exp(\varepsilon) \cdot \mathbb{P}[\hat{\mu}(S') \in Q]$. A simple composition of the privacy guarantee with DPRANGE-HT$(\cdot)$ and the exponential mechanism gives that

   $$\mathbb{P}[(\hat{R}(S), \hat{\mu}(S)) \in Q] \leq \exp(\varepsilon) \cdot \mathbb{P}[(\hat{R}(S'), \hat{\mu}(S')) \in Q] + \delta/3$$

   This implies that $\mathbb{P}[\mathcal{A}(S) \in Q] \leq \exp(\varepsilon) \cdot \mathbb{P}[\mathcal{A}(S') \in Q] + \delta/3$.

□

Lemma A.7.8 (Utility of the algorithm)**.** *For an 2α-corrupted dataset S, Algorithm 8 achieves $\|\hat{\mu} - \mu^*\|_2 \leq c_\zeta\sqrt{\alpha}$ with probability $1 - \zeta$, if $n = \Omega(d/(\alpha\zeta) + (d\log(dR/\alpha) + \log(1/\zeta)/(\varepsilon\alpha))$.*

Proof of Lemma A.7.8. Following the proof strategy of §A.1, we use the following lemma showing that $d(\hat{\mu}, S)$ is a good approximation of $\|\hat{\mu} - \mu^*\|_2$.

Lemma A.7.9 ($d(\mu, S)$ approximates $\|\mu - \mu^*\|$)**.** *Let S be the set of 2α-corrupted data. Given that $n = \Omega(d/(\zeta\alpha))$, with probability $1 - \zeta$,*

$$\left| d(\hat{\mu}, S) - \|\hat{\mu} - \mu^*\|_2 \right| \leq 7c_\zeta\sqrt{\alpha}\,.$$

This implies that the exponential mechanism achieves the following bounds.

$$\begin{aligned}\mathbb{P}(\|\hat{\mu}-\mu^*\| \le c_\zeta\sqrt{\alpha}) &\ge \frac{1}{A}e^{-\frac{\varepsilon\alpha n}{3}}\,\mathrm{Vol}(c_\zeta\sqrt{\alpha},d), \text{ and}\\ \mathbb{P}(\|\hat{\mu}-\mu^*\| \ge 22c_\zeta\sqrt{\alpha}) &\le \frac{1}{A}e^{-\frac{5\varepsilon\alpha n}{8}}\,(4R)^d\,,\end{aligned}$$

where A denotes the normalizing factor for the exponential mechanism and $\mathrm{Vol}(r,d)$ is the volume of a ball of radius r in d dimensions. It follows that

$$\begin{aligned}\log\Big(\frac{\mathbb{P}(\|\hat{\mu}-\mu^*\|_2 \le c_\zeta\sqrt{\alpha})}{\mathbb{P}(\|\hat{\mu}-\mu^*\|_2 \ge 22c_\zeta\sqrt{\alpha})}\Big) &\ge \frac{7}{24}\varepsilon\alpha n - C\,d\log(dR/\alpha)\\ &\ge \log(1/\zeta)\,,\end{aligned}$$

for $n=\Omega((d\log(dR/\alpha)+\log(1/\zeta))/(\varepsilon\alpha))$.

□

A.7.1.1 Proof of Lemma A.7.7

Since $R(S)\le 3c_\zeta\sqrt{\alpha}$, define S_{good} as the minimizing subset in Definition 2.4.2 such that

$$R(S)=\max_{T\subset S_{\text{good}},|T|=(1-\alpha)|S_{\text{good}}|}\|\mu(T)-\mu(S_{\text{good}})\|_2\,.$$

By this definition of S_{good} and Lemma A.7.1,

$$|v^\top(\mu(S_{\text{good}}\cap\mathcal{T}^v)-\mu(S_{\text{good}}))|\le 6c_\zeta\sqrt{1/\alpha},\text{ and}$$

$$|v^\top(\mu(S_{\text{good}}\cap\mathcal{B}^v)-\mu(S_{\text{good}}))|\le 6c_\zeta\sqrt{1/\alpha}.$$

Therefore,

$$\min_{i\in S_{\text{good}}\cap\mathcal{T}^v}|v^\top(x_i-\mu(S_{\text{good}}))|\le|v^\top(\mu(S_{\text{good}}\cap\mathcal{T}^v)-\mu(S_{\text{good}}))|\le 6c_\zeta\sqrt{1/\alpha},$$

and similarly

$$\min_{i\in S_{\text{good}}\cap\mathcal{B}^v}|v^\top(x_i-\mu(S_{\text{good}}))|\le|v^\top(\mu(S_{\text{good}}\cap\mathcal{B}^v)-\mu(S_{\text{good}}))|\le 6c_\zeta\sqrt{1/\alpha}$$

This implies

$$\min_{i\in S_{\text{good}}\cap\mathcal{T}^v} v^\top x_i - \max_{i\in S_{\text{good}}\cap\mathcal{B}^v} v^\top x_i \;\leq\; 12c_\zeta\sqrt{1/\alpha}\,. \tag{A.15}$$

This implies that distribution of one-dimensional points $S_{(v)} = \{v^\top x_i\}$ is dense at the boundary of top and bottom α quantiles, and hence cannot be changed much by changing one entry. Formally, consider a neighboring dataset S' (and the corresponding $S'_{(v)}$) where one point x_i in $\mathcal{M}^{(v)}(S)$ is replaced by another point $\tilde{x}_i$. If $v^\top\tilde{x}_i \in [\max_{i\in S_{\text{good}}\cap\mathcal{B}^v} v^\top x_i\,,\, \min_{i\in S_{\text{good}}\cap\mathcal{T}^v} v^\top x_i\,]$, then Eq. (A.15) implies that this only changes the mean by $6c_\zeta/(\sqrt{\alpha}n)$. Otherwise, $\mathcal{M}^v(S')$ will have x_i replaced by either $\arg\min_{i\in S_{\text{good}}\cap\mathcal{T}^v} v^\top x_i$ or $\arg\max_{i\in S_{\text{good}}\cap\mathcal{B}^v} v^\top x_i$. In both cases, Eq. (A.15) implies that this only changes the mean by $12c_\zeta/(\sqrt{\alpha}n)$. The other case of when the replaced sample $x_i \in S$ is not in $\mathcal{M}^v(S)$ follows similarly. From this, we upper bounds the maximum difference between S and S' when projected on v, that is

$$\left|v^\top\left(\mu(\mathcal{M}^v(S)) - \mu(\mathcal{M}^v(S'))\right)\right| \leq \frac{12c_\zeta}{\sqrt{\alpha}n}\,.$$

This implies the sensitivity of $d(\mu, S)$ is bounded by $6c_\zeta/(\sqrt{\alpha}n)$:

$$\begin{aligned}
|d(\mu,S) - d(\mu,S')| \;&=\; \Big|\max_{v\in\mathbb{S}^{d-1}} v^\top\mu(M^v(S)) \;-\; \max_{\tilde{v}\in\mathbb{S}^{d-1}} \tilde{v}^\top\mu(M^v(S'))\Big| \\
&\leq\; \max_{v\in\mathbb{S}^{d-1}} \left|\,v^\top(\mu(M^v(S)) \;-\mu(M^v(S'))\,)\right| \;\leq\; \frac{12c_\zeta}{\sqrt{\alpha}\,n}
\end{aligned}$$

A.7.1.2 Proof of Lemma A.7.9

First we show $|v^\top(\mu(\mathcal{M}^v) - \mu^*)| \leq 7c_\zeta\sqrt{\alpha}$. Notice that $|S_{\text{good}}\cap\mathcal{T}^v| \leq 3\alpha|S|$, and $|S_{\text{good}}\cap\mathcal{B}^v| \leq 3\alpha|S|$. By the $(c_\zeta\sqrt{3\alpha}, 3\alpha)$-resilience property, we have $|v^\top(\mu(S_{\text{good}}\cap\mathcal{T}^v) - \mu^*)| \leq c_\zeta\sqrt{3/\alpha}$, and $|v^\top(\mu(S_{\text{good}}\cap\mathcal{B}^v) - \mu^*)| \leq c_\zeta\sqrt{3/\alpha}$. Since $|S_{\text{good}}\cap\mathcal{M}^v| \geq (1-8\alpha)|S_{\text{good}}|$, by the $(c_\zeta\sqrt{8\alpha}, 8\alpha)$-resilience property,

$$|v^\top(\mu(S_{\text{good}}\cap\mathcal{M}^v) - \mu^*)| \;\leq\; c_\zeta\sqrt{8\alpha}\,.$$

Since $\mathcal{T}^v$, $\mathcal{B}^v$ are the largest and smallest $3\alpha n$ points respectively and $|S_{\text{bad}}| \leq 2\alpha n$, we get

$$|v^\top(\mu(S_{bad}\cap\mathcal{M}^v) - \mu^*)| \;\leq\; 2c_\zeta\sqrt{3/\alpha}.$$

Combining $S_{\text{good}} \cap \mathcal{M}^v$ and $S_{\text{bad}} \cap \mathcal{M}^v$ we get

$$\begin{aligned}
&|v^\top(\mu(\mathcal{M}^v) - \mu^*)| \\
&\leq \frac{|S_{bad} \cap \mathcal{M}^v|}{|\mathcal{M}^v|}|v^\top(\mu(S_{bad} \cap \mathcal{M}^v) - \mu^*)| + \frac{|\mu(S_{\text{good}} \cap \mathcal{M}^v|}{|\mathcal{M}^v|}|v^\top(\mu(S_{\text{good}} \cap \mathcal{M}^v) - \mu^*)| \\
&\leq 7c_\zeta\sqrt{\alpha}.
\end{aligned}$$

Finally we get that

$$\begin{aligned}
\left| d(\hat{\mu}, S) - \|\hat{\mu} - \mu^*\|_2 \right| &\overset{(a)}{=} \left| \max_{v \in \mathbb{S}^{d-1}} \left|v^\top \left(\mu(\mathcal{M}^{(v)}) - \hat{\mu}\right)\right| - \max_{v \in \mathbb{S}^{d-1}} |v^\top(\hat{\mu} - \mu^*)| \right| \\
&\overset{(b)}{\leq} \max_{v \in \mathbb{S}^{d-1}} \left|v^\top \left(\mu(\mathcal{M}^{(v)}) - \mu^*\right)\right| \\
&\leq 7c_\zeta\sqrt{\alpha},
\end{aligned}$$

where (a) holds by the definition of the distance :

$$\|\mu - \mu^*\|_2 \;=\; \max_{v \in \mathbb{S}^{d-1}} |v^\top(\mu - \mu^*)|,$$

and (b) holds by triangle inequality.

A.7.2 Case of sub-Gaussian distributions and a proof of Theorem 8

Th proof is analogous to the previous section, we only state the lemmas that differ. DPRANGE returns a hypercube $\bar{x} + [-B/2, B/2]^d$ that includes all uncorrupted data points with a high probability.

Lemma A.7.10 (α-corrupted data has small $R(S)$)**.** *Let S be the set of α-corrupted data. Given that $n = \Omega(\frac{d+\log(1/\zeta)}{\alpha^2 \log 1/\alpha})$, with probability $1-\zeta$, $R(S) \leq 3\,\alpha\sqrt{\log(1/3\alpha)}$.*

Lemma A.7.11 (Privacy)**.** *Algorithm 8 is (ε, δ)-differentially private if $n \geq 3B\sqrt{d}\log(3/\delta)/(\varepsilon\alpha\sqrt{\log(1/\alpha)})$.*

This follows from the following lemma.

Lemma A.7.12 (Sensitivity of $d(\hat{\mu}, S)$)**.** *Given that $R(S) \leq 3\alpha\sqrt{\log(1/\alpha)}$, for any neighboring dataset S', $|d(\hat{\mu}, S) - d(\hat{\mu}, S')| \leq 12\sqrt{\log 1/\alpha}/n$.*

Lemma A.7.13 ($d(\hat{\mu}, S)$ approximates $\|\hat{\mu} - \mu^*\|$)**.** *Let S be the set of α-corrupted data. Given that $n = \Omega(\frac{d+\log(1/\zeta)}{\alpha^2 \log 1/\alpha})$, with probability $1-\zeta$,*

$$\left| d(\hat{\mu}, S) - \|\hat{\mu} - \mu^*\|_2 \right| \leq 14\,\alpha\,\sqrt{\log 1/\alpha}\,.$$

This implies the following utility bound.

Lemma A.7.14 (Utility of the algorithm)**.** *For an α-corrupted dataset S, Algorithm 8 achieves $\|\hat{\mu}-\mu^*\|_2 \leq \alpha\sqrt{\log 1/\alpha}$ with probability $1-\zeta$, if $n = \Omega((d+\log(1/\zeta))/(\alpha^2 \log(1/\alpha)) + (d\log(dR/\alpha) + \log(1/\zeta)/(\varepsilon\alpha))$.*

A.8 Algorithm and analysis for covariance bounded distributions

Algorithm 18: Differentially private range estimation for covariance bounded distributions (DPRANGE-HT) [135, Algorithm 2]

Input: $S = \{x_i\}_{i=1}^n$, R, ε, δ, ζ

1 Randomly partition the dataset $S = \cup_{\ell\in[m]} S^{(\ell)}$ with $m = 200\log(2/\zeta)$

2 $\bar{x}^{(\ell)} \leftarrow$DPRANGE$(S^{(\ell)}, R, \varepsilon/m, \delta/m, \sigma = 40)$ for all $\ell \in [m]$

3 $\hat{x}_j \leftarrow \text{median}(\{\bar{x}_j^{(\ell)}\}_{\ell\in[m]})$ for all $j \in [d]$

Output: $(\hat{x}, B = 50/\sqrt{\alpha})$

Lemma A.8.1 (Analysis of DPRANGE-HT)**.** DPRANGE-HT *is (ε,δ)-differentially private. Under Assumption 2 and for $\alpha \in (0, 0.01)$, if $n = \Omega((1/\alpha)\log(1/\zeta) + (\sqrt{d\log(1/\delta)}\log(1/\zeta)/\varepsilon)\min\{\log(dR), \log($* DPRANGE-HT *returns a ball $\mathcal{B}_{\sqrt{d}B/2}(\bar{x})$ of radius $\sqrt{d}B/2$ centered at $\bar{x}$ that includes $(1-2\alpha)n$ uncorrupted samples where $B = 50/\sqrt{\alpha}$ with probability $1-\zeta$.*

A.8.1 Range estimation with DPRANGE-HT *and a proof of Lemma A.8.1*

We first show that applying the private histogram to each coordinate provides a robust estimate of the range, but with a constant probability 0.9.

Lemma A.8.2 (Robustness of a single private histogram)**.** *Under the α-corruption model of Assumption 2, if $n = \Omega((\sqrt{d\log(1/\delta)}/\varepsilon)\min\{\log(dR), \log(d/\delta)\})$, for $\alpha \in (0, 0.01)$,*

DPRANGE *in Algorithm 14 with a choice of* $\sigma = 40$ *and* $B = 120$ *returns intervals* $\{I_j\}_{j=1}^d$ *of size* $|I_j| = 240$ *such that* $\mu_j \in I_j$ *with probability 0.9 for each* $j \in [d]$.

Proof of Lemma A.8.2. The proof is analogous to §A.2.1 and we only highlight the differences here. By Lemma A.2.1 we know that $|\tilde{p}_k - \hat{p}_k| \leq 0.01$ with the assumption on n. The corruption can change the normalized count in each bin by $\alpha \leq 0.01$ by assumption. It follows from Chebyshev inequality that $\mathbb{P}(|x_{i,j} - \mu_j|^2 > \sigma^2) \leq 1/\sigma^2$. It follows from (e.g. [135, Lemma A.3]) that $\mathbb{P}(|\{i : x_{i,j} \notin [\mu - \sigma, \mu + \sigma]\}| > (100/\sigma^2)n) < 0.05$. Hence the maximum bin has $\tilde{p}_k \geq 0.5(1 - 100/\sigma^2) - 0.02$ and the true mean is in the maximum bin or in an adjacent bin. The largest non-adjacent bucket is at most $100/\sigma^2 + 0.02$. Hence, the choice of $\sigma = 40$ ensures that we find the μ within $3\sigma = 120$.

□

Following [135, Algorithm 2], we partition the dataset into $m = 200 \log(2/\zeta)$ subsets of an equal size n/m and apply the median-of-means approach. Applying Lemma A.8.2, it is ensured (e.g., by [135, Lemma A.4]) that more than half of the partitions satisfy that the center of the interval is within 240 away from μ, with probability $1 - \zeta$. Therefore the median of those m centers is within 240 from the true mean in each coordinate. This requires the total sample size larger only by a factor of $\log(d/\zeta)$.

To choose a radius $\sqrt{d}B/2$ ball around this estimated mean that includes $1 - \alpha$ fraction of the points, we choose $B = 25/\sqrt{\alpha}$. Since $\|\hat{\mu} - \mu\|_2 \leq 120\sqrt{d} \ll \sqrt{d}B/2$ for $\alpha \leq 0.01$, this implies that we can choose $\sqrt{d}B/2$-ball around the estimated mean with $B = 50/\sqrt{\alpha}$.

Let $z_i = \mathbb{I}(\|x_i - \mu\|_2 > \sqrt{d}B/2)$. We know that $\mathbb{E}[z_i] = \mathbb{P}[(\|x_i - \mu\|_2 > \sqrt{d}B/2)] \leq \mathbb{E}[\|x_i - \mu\|_2^2 (2/dB^2)] = (1/1250)\alpha$. Applying multiplicative Chernoff bound (e.g., in [135, Lemma A.3]), we get $|\{i : \|x_i - \mu\|_2 \leq \sqrt{d}B/2\}| \geq 1 - (3/2500)\alpha$ with probability $1 - \zeta$, if $n = \Omega((1/\alpha)\log(1/\zeta))$. This ensures that with high probability, $(1 - \alpha)$ fraction of the original uncorrupted points are included in the ball. Since the adversary can corrupt αn samples, at least $(1 - 2\alpha)n$ of the remaining good points will be inside the ball.

A.8.2 Proof of Theorem 10

The proof of the privacy guarantee of Algorithm 19 follows analogously from the proof of the privacy of PRIME and is omitted here. The accuracy guarantee follows form the following theorem and Lemma A.8.1.

Theorem 27 (Analysis of accuracy of DPMMWFILTER-HT)**.** *Let S be an α-corrupted covariance bounded dataset under Assumption 2, where $\alpha \leq c$ for some universal constant $c \in (0, 1/2)$. Let S_{good} be α-good with respect to $\mu \in \mathbb{R}^d$. Suppose $\mathcal{D} = \{x_i \in \mathcal{B}_{\sqrt{d}B/2}(\bar{x})\}_{i=1}^n$ be the projected dataset. If $n \geq \widetilde{\Omega}\left(\frac{d^{3/2}B^2 \log(1/\delta)}{\varepsilon}\right)$, then* DPMMWFILTER-HT *terminates after at most $O(\log dB^2)$ epochs and outputs $S^{(s)}$ such that with probability 0.9, we have $|S_t^{(s)} \cap S_{\text{good}}| \geq (1 - 10\alpha)n$ and*

$$\|\mu(S^{(s)}) - \mu\|_2 \lesssim \sqrt{\alpha}\,.$$

Moreover, each epoch runs for at most $O(\log d)$ iterations.

Algorithm 19: Differentially private filtering with matrix multiplicative weights (DPMMWFILTER-HT) for distributions with bounded covariance

Input: $S = \{x_i \in \mathcal{B}_{\sqrt{d}B/2}(\bar{x})\}_{i=1}^n$, $\alpha \in (0,1)$, $T_1 = O(\log B), T_2 = O(\log d)$, $B \in \mathbb{R}_+$, /2 (ε, δ)

Initialize $S^{(1)} \leftarrow [n]$, $\varepsilon_1 \leftarrow \varepsilon/(4T_1)$, $\delta_1 \leftarrow \delta/(4T_1)$, $\varepsilon_2 \leftarrow \min\{0.9, \varepsilon\}/(4\sqrt{10T_1T_2\log(4/\delta)})$, $\delta_2 \leftarrow \delta/(20T_1T_2)$, a large enough constant $C > 0$

for epoch $s = 1, 2, \ldots, T_1$ **do**

  $\lambda^{(s)} \leftarrow \|M(S^{(s)})\|_2 + \mathrm{Lap}(2B^2d/(n\varepsilon_1))$

  $n^{(s)} \leftarrow |S^{(s)}| + \mathrm{Lap}(1/\varepsilon_1)$

  if $n^{(s)} \le 3n/4$ **then** terminate

  if $\lambda^{(s)} \le C$ **then**

    Output: $\mu^{(s)} \leftarrow (1/|S^{(s)}|)\big(\sum_{i\in S^{(s)}} x_i\big) + \mathcal{N}(0, (2B\sqrt{2d\log(1.25/\delta_1)}/(n\,\varepsilon_1))^2\mathbf{I}_{d\times d})$

  $\alpha^{(s)} \leftarrow 1/(100(0.1/C + 1.05)\lambda^{(s)})$

  $S_1^{(s)} \leftarrow S^{(s)}$

  for $t = 1, 2, \ldots, T_2$ **do**

    $\lambda_t^{(s)} \leftarrow \|M(S_t^{(s)})\|_2 + \mathrm{Lap}(2B^2d/(n\varepsilon_2))$

    if $\lambda_t^{(s)} \le 2/3\lambda_0^{(s)}$ **then**

      terminate epoch

    else

      $\Sigma_t^{(s)} \leftarrow M(S_t^{(s)}) + \mathcal{N}(0, (2B^2d\sqrt{2\log(1.25/\delta_2)}/(n\varepsilon_2))^2\mathbf{I}_{d^2\times d^2})$

      $U_t^{(s)} \leftarrow (1/\operatorname{Tr}(\exp(\alpha^{(s)}\sum_{r=1}^t(\Sigma_r^{(s)})))) \exp(\alpha^{(s)}\sum_{r=1}^t(\Sigma_r^{(s)}))$

      $\psi_t^{(s)} \leftarrow \left\langle M(S_t^{(s)}), U_t{}^{(s)} \right\rangle + \mathrm{Lap}(2B^2d/(n\varepsilon_2))$

      if $\psi_t^{(s)} \le (1/5.5)\lambda_t^{(s)}$ **then**

        $S_{t+1}^{(s)} \leftarrow S_t^{(s)}$

      else

        $Z_t^{(s)} \leftarrow \mathrm{Unif}([0,1])$

        $\mu_t^{(s)} \leftarrow (1/|S_t^{(s)}|)\big(\sum_{i\in S_t} x_i\big) + \mathcal{N}(0, (2B\sqrt{2d\log(1.25/\delta_2)}/(n\,\varepsilon_2)\mathbf{I}_{d\times d})^2)$

        $\rho_t^{(s)} \leftarrow$ DP-1DFILTER-HT$(\mu_t^{(s)}, U_t^{(s)}, \alpha, \varepsilon_2, \delta_2, S_t^{(s)})$ [Algorithm 20]

        $S_{t+1}^{(s)} \leftarrow S_t^{(s)} \setminus \{i \mid \{\tau_j = (x_j - \mu_t^{(s)})^\top U_t^{(s)}(x_j - \mu_t^{(s)})\}_{j\in S_t^{(s)}}$ and $\tau_i \ge \rho_t^{(s)}\, Z_t^{(s)}\}$.

  $S^{(s+1)} \leftarrow S_t^{(s)}$

Output: $\mu^{(T_1)}$

Algorithm 20: Differentially private 1D-filter DP-1DFILTER-HT

Input: μ, U, $\alpha \in (0,1)$, target privacy (ε, δ), $S = \{x_i \in \mathcal{B}_{B\sqrt{d}/2}(\bar{x})\}$

Set $\tau_i \leftarrow (x_i - \mu)^\top U (x_i - \mu)$ for all $i \in S$

Set $\tilde{\psi} \leftarrow (1/n)\sum_{i\in S} \tau_i + \mathrm{Lap}(B^2 d/n\varepsilon))$

Compute a histogram over geometrically sized bins

$$I_1 = [1/4, 1/2), I_2 = [1/2, 1), \ldots, I_{2+\log(B^2 d)} = [2^{\log(B^2 d)-1}, 2^{\log(B^2 d)}]$$

$$h_j \leftarrow \frac{1}{n} \cdot |\{i \in S \,|\, \tau_i \in [2^{-3+j}, 2^{-2+j})\}| \,, \qquad \text{for all } j = 1, \cdots, 2 + \log(B^2 d)$$

Compute a privatized histogram $\tilde{h}_j \leftarrow h_j + \mathcal{N}(0, (2\sqrt{2d\log(1.25/\delta)}/(|S|\varepsilon))^2)$, for all $j \in [2 + \log(B^2 d)]$

Set $\tilde{\tau}_j \leftarrow 2^{-3+j}$, for all $j \in [2 + \log(B^2 d)]$

Find the largest $\ell \in [2 + \log(B^2 d)]$ satisfying $\sum_{j \geq \ell} (\tilde{\tau}_j - \tilde{\tau}_\ell)\, \tilde{h}_j \geq 0.31\tilde{\psi}$

Output: $\rho = \tilde{\tau}_\ell$

A.8.2.1 Analysis of DPMMWFILTER-HT *and a proof of Theorem 27*

Algorithm 19 is a similar matrix multiplicative weights based filter algorithm for distributions with bounded covariance. Similarly, we first state following Lemma A.8.3 and prove Theorem 27 given Lemma A.8.3

Lemma A.8.3. *Let S be an α-corrupted bounded covariance dataset under Assumption 2. For an epoch s and an iteration t such that $\lambda^{(s)} > C$, $\lambda_t^{(s)} > 2/3\lambda_0^{(s)}$, and $n^{(s)} > 3n/4$, if $n \gtrsim \frac{B^2(\log B)d^{3/2}\log(1/\delta)}{\varepsilon}$ and $|S_t^{(s)} \cap S_{\mathrm{good}}| \geq (1 - 10\alpha)n$, then with probability $1 - O(1/\log(d)^3)$, we have the condition in Eq. (A.16) holds. When this condition holds, we have more corrupted samples are removed in expectation than the uncorrupted samples, i.e., $\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\mathrm{good}}| \leq \mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\mathrm{bad}}|$. Further, for an epoch $s \in [T_1]$ there exists a constant $C > 0$ such that if $\|M(S^{(s)})\|_2 \geq C$, then with probability $1 - O(1/\log^2 d)$, the s-th epoch terminates after $O(\log d)$ iterations and outputs $S^{(s+1)}$ such that $\|M(S^{(s+1)})\|_2 \leq 0.98\|M(S^{(s)})\|_2$.*

Now we define $d_t^{(s)} \triangleq |(S_{\text{good}} \cap S^{(1)}) \setminus S_t^{(s)}| + |S_t^{(s)} \setminus (S_{\text{good}} \cap S^{(1)})|$. Note that $d_1^{(1)} = \alpha n$, and $d_t^{(s)} \geq 0$. At each epoch and iteration, we have

$$\mathbb{E}[d_{t+1}^{(s)} - d_t^{(s)} | d_1^{(1)}, d_2^{(1)}, \cdots, d_t^{(s)}] \;=\; \mathbb{E}\left[|S_{\text{good}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})| - |S_{\text{bad}} \cap (S_t^{(s)} \setminus S_{t+1}^{(s)})|\right] \;\leq\; 0,$$

from the part 1 of Lemma A.8.3. Hence, $d_t^{(s)}$ is a non-negative super-martingale. By optional stopping theorem, at stopping time, we have $\mathbb{E}[d_t^{(s)}] \leq d_1^{(1)} = \alpha n$. By Markov inequality, $d_t^{(s)}$ is less than $10\alpha n$ with probability 0.9, i.e. $|S_t^{(s)} \cap S_{\text{good}}| \geq (1 - 10\alpha)n$. The desired bound in Theorem 27 follows from Lemma A.8.11.

A.8.2.2 Proof of Lemma A.8.3

Lemma A.8.3 is a combination of Lemma A.8.4, Lemma A.8.5 and Lemma A.8.6. We state the technical lemmas and subsequently provide the proofs.

Lemma A.8.4. *For each epoch s and iteration t, under the hypotheses of Lemma A.8.3 then with probability $1 - O(1/\log^3 d)$, we have*

$$\frac{1}{n} \sum_{i \in S_{\text{good}} \cap S_t^{(s)}} \tau_i \;\leq\; \psi/1000\,, \tag{A.16}$$

where $\psi \triangleq \frac{1}{n}\sum_{i \in S_t^{(s)}} \tau_i$.

Lemma A.8.5. *For each epoch s and iteration t, under the hypotheses of Lemma A.8.3, if condition Eq. (A.16) holds, then we have $\mathbb{E}|S_t^{(s)} \setminus S_{t+1}^{(s)} \cap S_{\text{good}}| \leq \mathbb{E}|S_t^{(s)} \setminus S_{t+1}^{(s)} \cap S_{\text{bad}}|$ and with probability $1 - O(1/\log^3 d)$, and $\left\langle M(S_{t+1}^{(s)}), U_t^{(s)} \right\rangle \leq 0.76 \left\langle M(S_t^{(s)}), U_t^{(s)} \right\rangle$.*

Lemma A.8.6. *For epoch s, suppose for $t = 0, 1, \cdots, T_2$ where $T_2 = O(\log d)$, if Lemma A.8.5 holds, $n \gtrsim \frac{B^2(\log B) d^{3/2} \log(1/\delta)}{\varepsilon\alpha}$, and $n^{(s)} > 3n/4$, then we have $\|M(S^{(s+1)})\|_2 \leq 0.98 \|M(S^{(s)})\|_2$ with probability $1 - O(1/\log^2 d)$.*

A.8.2.3 Proof of Lemma A.8.4

Proof. By Lemma A.6.9, Lemma A.6.10 and Lemma A.6.11, we can pick $n = \widetilde{\Omega}\left(\frac{B^2 d^{3/2} \log}{\varepsilon}\right)$ such that with probability $1 - O(1/\log^3 d)$, following conditions simultaneously hold:

1. $\|\mu_t^{(s)} - \mu(S_t^{(s)})\|_2^2 \le 0.001$
2. $|\psi_t^{(s)} - \left\langle M(S_t^{(s)}), U_t^{(s)} \right\rangle| \le 0.001$
3. $\left|\lambda_t^{(s)} - \|M(S_t^{(s)})\|_2\right| \le 0.001$
4. $\left|\lambda^{(s)} - \|M(S^{(s)})\|_2\right| \le 0.001$
5. $\left\|M(S_{t+1}^{(s)}) - \Sigma_t^{(s)}\right\|_2 \le 0.001$
6. $\|\mu^{(s)} - \mu(S^{(s)})\|_2^2 \le 0.001$.

Then we have

$$\begin{aligned}
\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_t^{(s)}} \tau_i &= \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S_t^{(s)}} \left\langle (x_i-\mu_t^{(s)})(x_i-\mu_t^{(s)})^\top, U_t^{(s)} \right\rangle \\
&\overset{(a)}{\le} \frac{2}{n}\sum_{i\in S_{\text{good}}\cap S_t^{(s)}} \left\langle (x_i-\mu(S_{\text{good}}\cap S_t^{(s)}))(x_i-\mu(S_{\text{good}}\cap S_t^{(s)}))^\top, U_t^{(s)} \right\rangle \\
&\quad + \frac{2|S_{\text{good}}\cap S_t^{(s)}|}{n}\left\langle (\mu(S_{\text{good}}\cap S_t^{(s)})-\mu_t^{(s)})(\mu(S_{\text{good}}\cap S_t^{(s)})-\mu_t^{(s)})^\top, U_t^{(s)} \right\rangle \\
&\le 2\left\langle M((S_{\text{good}}\cap S_t^{(s)}), U_t^{(s)} \right\rangle + 2\|\mu_t^{(s)} - \mu(S_{\text{good}}\cap S_t^{(s)})\|_2^2 \\
&\overset{(b)}{\le} 2 + 2\left(\|\mu_t^{(s)}-\mu\|_2 + \|\mu(S_{\text{good}}\cap S_t^{(s)})-\mu\|_2\right)^2 \\
&\overset{(c)}{\le} 2 + 2\left(0.01 + 2\sqrt{\alpha\|M(S_t^{(s)})\|_2} + 3\sqrt{\alpha}\right)^2 \\
&\le 3 + 8\alpha\|M(S_t^{(s)})\|_2 + 32\alpha \\
&\overset{(d)}{\le} \frac{\psi_t^{(s)} - 0.002}{1000} \\
&\le \frac{\psi}{1000}\,,
\end{aligned}$$

where (a) follows from the fact that for any vector x, y, z, we have $(x-y)(x-y)^\top \preceq 2(x-z)(x-z)^\top + 2(y-z)(y-z)^\top$, (b) follows from α-goodness of S_{good}, (c) follows from Lemma A.8.11 and (d) follows from our choice of large constant C and sample complexity n.

□

A.8.2.4 Proof of Lemma A.8.5

Proof. Lemma A.8.4 implies with probability $1-O(1/\log^3 d)$, our scores satisfies the condition in Eq. (A.16). Then by Lemma A.8.7 our DP-1DFILTER-HT gives us a threshold ρ such that

$$\sum_{i\in S_{\text{good}}\cap S_t^{(s)}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\} \le \sum_{i\in S_{\text{bad}}\cap S_t^{(s)}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\} \ .$$

According to our filter rule from Algorithm 20, we have

$$\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| = \sum_{i\in S_{\text{good}}\cap S_t^{(s)}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\}$$

and

$$\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}| = \sum_{i\in S_{\text{bad}}\cap S_t^{(s)}} \mathbf{1}\{\tau_i \le \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\} \ .$$

This implies $\mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{good}}| \le \mathbb{E}|(S_t^{(s)} \setminus S_{t+1}^{(s)}) \cap S_{\text{bad}}|$.

At the same time, Lemma A.8.7 gives us a ρ such that with probability $1 - O(\log^3 d)$, we have

$$\frac{1}{n}\sum_{i\in S_{t+1}^{(s)}} \tau_i \le \frac{1}{n}\sum_{\tau_i\le\rho, i\in S_t^{(s)}} \tau_i \le \frac{3}{4}\cdot\frac{1}{n}\sum_{i\in S_t^{(s)}} \tau_i \ .$$

Hence, we have

$$\begin{aligned}
\left\langle M(S_{t+1}^{(s)}), U_t^{(s)} \right\rangle &= \left\langle \frac{1}{n}\sum_{i\in S_{t+1}^{(s)}} (x_i - \mu(S_{t+1}^{(s)}))(x_i - \mu(S_{t+1}^{(s)}))^\top, U_t^{(s)} \right\rangle \\
&\le \left\langle \frac{1}{n}\sum_{i\in S_{t+1}^{(s)}} (x_i - \mu(S_t^{(s)}))(x_i - \mu(S_t^{(s)}))^\top, U_t^{(s)} \right\rangle \\
&\le \frac{1}{n}\sum_{i\in S_{t+1}^{(s)}} \tau_i + \|\mu_t^{(s)} - \mu(S_t^{(s)})\|_2^2 \\
&\le \frac{3}{4n}\sum_{i\in S_t^{(s)}} \tau_i + 0.01 \\
&\overset{(a)}{\le} 0.76 \left\langle M(S_t^{(s)}), U_t^{(s)} \right\rangle \ ,
\end{aligned}$$

where (a) follows from our assumption that $\psi_t^{(s)} > \frac{1}{5.5}\lambda_t^{(s)} > \frac{2}{16.5}C$.

□

A.8.2.5 Proof of Lemma A.8.6

Proof. If Lemma A.8.5 holds, we have

$$\begin{aligned}\left\langle M(S_t^{(s)}), U_t^{(s)} \right\rangle &\leq 0.76 \left\langle M(S_{t-1}^{(s)}), U_t^{(s)} \right\rangle \\ &\leq 0.76 \left\langle M(S_1^{(s)}), U_t^{(s)} \right\rangle \\ &\leq 0.76 \|M(S_1^{(s)})\|_2\end{aligned}$$

We pick n large enough such that with probability $1 - O(\log^3 d)$,

$$\|\Sigma_t^{(s)}\|_2 \approx_{0.05} \|M(S_t^{(s)})\|_2 .$$

Thus, we have

$$\left\langle \Sigma_t^{(s)}, U_t^{(s)} \right\rangle \leq 0.81 \|M(S_1^{(s)})\|_2 .$$

By Lemma A.6.1, we have $M(S_t^{(s)}) \preceq M(S_1^{(s)})$. by our choice of $\alpha^{(s)}$, we have $\alpha^{(s)} M(S_{t+1}^{(s)}) \preceq \frac{1}{100}\mathbf{I}$ and $\alpha^{(s)}\Sigma_t^{(s)} \preceq \frac{1}{100}\mathbf{I}$. Therefore, by Lemma A.6.13 we have

$$\begin{aligned}&\left\| \sum_{i=1}^{T_2} \Sigma_t^{(s)} \right\|_2 \\ \leq\ & \sum_{t=1}^{T_2} \left\langle \Sigma_t^{(s)}, U_t^{(s)} \right\rangle + \alpha^{(s)} \sum_{t=0}^{T_2} \left\langle U_t^{(s)}, \left| \Sigma_t^{(s)} \right| \right\rangle \|\Sigma_t^{(s)}\|_2 + \frac{\log(d)}{\alpha^{(s)}} \\ \overset{(a)}{\leq}\ & \sum_{t=1}^{T_2} \left\langle \Sigma_t^{(s)}, U_t^{(s)} \right\rangle + \frac{1}{100} \sum_{t=1}^{T_2} \left\langle U_t^{(s)}, \left| \Sigma_t^{(s)} \right| \right\rangle + 200 \log(d) \|M(S_1^{(s)})\|_2\end{aligned}$$

where (a) follows from our choice of $\alpha^{(s)}$, C, and n.

Meanwhile, we have

$$|\Sigma_t^{(s)}| \preceq M(S_t^{(s)}) + 0.15\, \mathbf{I} .$$

Thus we have

$$\left\langle U_t^{(s)}, \left|\Sigma_t^{(s)}\right|\right\rangle \ \leq 0.91 \ \left\|M(S_1^{(s)})\right\|_2$$

Then we have

$$\begin{aligned}
\left\|M(S_{T_2}^{(s)})\right\|_2 &\leq \frac{1}{T_2}\left\|\sum_{i=1}^{T_2} M(S_t^{(s)})\right\|_2 \\
&\leq \frac{1}{T_2}\left\|\sum_{i=1}^{T_2}\Sigma_t^{(s)}\right\|_2 + 0.05\ \|M(S_1^{(s)})\|_2 \\
&\leq \frac{1}{T_2}\left(\sum_{t=1}^{T_2}\left\langle \Sigma_t^{(s)}, U_t^{(s)}\right\rangle + \frac{1}{100}\sum_{t=1}^{T_2}\left\langle U_t^{(s)}, \left|\Sigma_t^{(s)}\right|\right\rangle + 200\log(d)\|M(S_1^{(s)})\|_2\right) + 0.05\ \|M(S_1^{(s)})\|_2 \\
&\leq 0.91\|M(S_1^{(s)})\|_2 + \frac{200\log(d)}{T_2}\|M(S_1^{(s)})\|_2 + 0.05\ \|M(S_1^{(s)})\|_2 \\
&\leq 0.98\ \|M(S_1^{(s)})\|_2
\end{aligned}$$

□

A.8.2.6 Proof of Private 1-D filter for distributions with bounded covariance

Lemma A.8.7 (Private 1-D filter: picking threshold privately for distributions with bounded covariance)**.** *Algorithm* DP-1DFILTER-HT*$(\mu, U, \alpha, \varepsilon, \delta, S)$ running on a dataset $\{\tau_i = (x_i - \mu)^\top U(x_i - \mu)\}_{i\in S}$ is (ε, δ)-DP. Define $\psi \triangleq \frac{1}{n}\sum_{i\in S}\tau_i$. If τ_i's satisfy*

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S}\tau_i \ \leq \ \psi/1000\ ,$$

and $n \geq \tilde{\Omega}\left(\frac{B^2 d}{\varepsilon}\right)$ then DP-1DFILTER-HT *outputs a threshold ρ such that*

$$2\Big(\sum_{i\in S_{\text{good}}\cap S}\mathbf{1}\{\tau_i \leq \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\}\Big) \leq \sum_{i\in S_{\text{bad}}\cap S}\mathbf{1}\{\tau_i \leq \rho\}\frac{\tau_i}{\rho} + \mathbf{1}\{\tau_i > \rho\}\ , \tag{A.17}$$

and with probability $1 - O(1/\log^3 d)$,

$$\frac{1}{n}\sum_{\tau_i<\rho}\tau_i \ \leq \ 0.75\psi\ .$$

Proof. 1. ρ **cuts enough**

Let ρ be the threshold picked by the algorithm. Let $\hat{\tau}_i$ denote the minimum value of the interval of the bin that τ_i belongs to. It holds that

$$\begin{aligned}
\frac{1}{n}\sum_{\tau_i\geq\rho, i\in[n]}(\tau_i-\rho) &\geq \frac{1}{n}\sum_{\hat{\tau}_i\geq\rho, i\in[n]}(\hat{\tau}_i-\rho)\\
&= \sum_{\tilde{\tau}_j\geq\rho, j\in[2+\log(B^2d)]}(\tilde{\tau}_j-\rho)h_j\\
&\overset{(a)}{\geq} \sum_{\tilde{\tau}_j\geq\rho, j\in[2+\log(B^2d)]}(\tilde{\tau}_j-\rho)\tilde{h}_j - O\left(\log(B^2d)\cdot B^2d\cdot\frac{\sqrt{\log(\log(B^2d)\log d)\log(1/\delta)}}{\varepsilon n}\right)\\
&\overset{(b)}{\geq} 0.31\tilde{\psi} - \tilde{O}(\frac{B^2d}{\varepsilon n})\\
&\overset{(c)}{\geq} 0.3\psi - \tilde{O}(\frac{B^2d}{\varepsilon n}),
\end{aligned}$$

where (a) holds due to the accuracy of the private histogram (Lemma A.6.12), (b) holds by the definition of ρ in our algorithm, and (c) holds due to the accuracy of $\tilde{\psi}$. This implies

$$\frac{1}{n}\sum_{\tau_i<\rho}\tau_i \leq \psi - \frac{1}{n}\sum_{\tau_i\geq\rho}(\tau_i-\rho) \leq 0.7\psi + \tilde{O}(B^2d/\varepsilon n).$$

2. ρ **doesn't cut too much**

Define C_2 to be the threshold such that $\frac{1}{n}\sum_{\tau_i>C_2}(\tau_i - C_2) = (2/3)\psi$. Suppose $2^b \leq C_2 \leq 2^{b+1}$, we have $\sum_{\hat{\tau}_i\geq 2^{b-1}}(\hat{\tau}_i - 2^{b-1}) \geq (1/3)\psi$ because $\forall\tau_i \geq C_2$, $(\hat{\tau}_i - 2^{b-1}) \geq \frac{1}{2}(\tau_i - C_2)$. Then the threshold picked by the algorithm $\rho \geq 2^{b-1}$, which implies $\rho \geq \frac{1}{4}C_2$. Suppose $\rho < C_2$, since $\rho \geq \frac{1}{4}C_2$

$$\begin{aligned}
\sum_{i\in S_{\text{bad}}\cap S,\tau_i<\rho}\tau_i + \sum_{i\in S_{\text{bad}}\cap S,\tau_i\geq\rho}\rho &\geq \frac{1}{4}(\sum_{i\in S_{\text{bad}}\cap S,\tau_i<C_2}\tau_i + \sum_{i\in S_{\text{bad}}\cap S,\tau_i\geq C_2}C_2)\\
&\overset{(a)}{\geq} \frac{10}{4}(\sum_{i\in S_{\text{good}}\cap S,\tau_i<C_2}\tau_i + \sum_{i\in S_{\text{good}}\cap S,\tau_i\geq C_2}C_2)\\
&\overset{(b)}{\geq} \frac{10}{4}(\sum_{i\in S_{\text{good}}\cap S,\tau_i<\rho}\tau_i + \sum_{i\in S_{\text{good}}\cap S,\tau_i\geq\rho}\rho),
\end{aligned}$$

where (a) holds by Lemma A.8.8, and (b) holds since $\rho \leq C_2$. If $\rho \geq C_2$, the statement of the Lemma A.8.8 directly implies Equation (A.17).

Lemma A.8.8. *Assuming that the condition in Eq.(A.16) holds, then for any C such that*

$$\frac{1}{n}\sum_{i\in S,\tau_i<C}\tau_i + \frac{1}{n}\sum_{i\in S,\tau_i\geq C} C \geq (1/3)\psi\ ,$$

we have

$$\sum_{i\in S_{\text{bad}}\cap S,\tau_i<C}\tau_i + \sum_{i\in S_{\text{bad}}\cap S,\tau_i\geq C} C \geq 10(\sum_{i\in S_{\text{good}}\cap S,\tau_i<C}\tau_i + \sum_{i\in S_{\text{good}}\cap S,\tau_i\geq C} C)$$

Proof. First we show an upper bound on S_{good}:

$$\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i<C}\tau_i + \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i\geq C} C \leq \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S}\tau_i \leq \psi/1000.$$

Then we show an lower bound on S_{bad}:

$$\begin{aligned}
& \frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i<C}\tau_i + \frac{1}{n}\sum_{i\in S_{\text{bad}}\cap S,\tau_i>C} C \\
= \quad & \frac{1}{n}\sum_{i\in S,\tau_i<C}\tau_i + \frac{1}{n}\sum_{i\in S,\tau_i\geq C} C \\
& -(\frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i<C}\tau_i + \frac{1}{n}\sum_{i\in S_{\text{good}}\cap S,\tau_i\geq C} C) \\
\geq \quad & (1/3 - 1/1000)\psi\ .
\end{aligned}$$

Combing the lower bound and the upper bound yields the desired statement □

□

A.8.2.7 Regularity lemmas for distributions with bounded covariance

Definition A.8.9 ([73, Definition 3.1])**.** *Let D be a distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \preceq \mathbf{I}$. For $0 < \alpha < 1/2$, we say a set of points $S = \{X_1, X_2, \cdots, X_n\}$ is α-good with respect to $\mu \in \mathbb{R}^d$ if following inequalities are satisfied:*

- $\|\mu(S) - \mu\|_2 \leq \sqrt{\alpha}$
- $\left\| \frac{1}{|S|} \sum_{i \in S} (X_i - \mu(S)) (X_i - \mu(S))^\top \right\|_2 \leq 1.$

Lemma A.8.10 ([73, Lemma 3.1])**.** *Let D be a distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \preceq \mathbf{I}$. Let $S = \{X_1, X_2, \cdots, X_n\}$ be a set of i.i.d. samples of D. If $n = \Omega(d \log(d)/\alpha)$, then with probability $1 - O(1)$, there exists a set $S_{\text{good}} \subseteq S$ such that S_{good} is α-good with respect to μ and $|S_{\text{good}}| \geq (1 - \alpha)n$.*

Lemma A.8.11 ([73, Lemma 3.2])**.** *Let S be an α-corrupted bounded covariance dataset under Assumption 2. If S_{good} is α-good with respect to μ, then for any $T \subset S$ such that $|T \cap S_{\text{good}}| \geq (1 - \alpha)|S|$, we have*

$$\|\mu(T) - \mu\|_2 \leq \frac{1}{1 - 2\alpha} \cdot \left(2\sqrt{\alpha \|M(T)\|_2} + 3\sqrt{\alpha} \right) .$$

A.9 Experiments

We evaluate our PRIME and compare with DP mean estimators [129] on synthetic dataset in Figure 2.1, which consists of 10^6 samples from $(1 - \alpha)\mathcal{N}(0, \mathbf{I}) + \alpha\mathcal{N}(\mu_{\text{bad}}, \mathbf{I})$. The main focus of this evaluation is to compare the estimation error and demonstrate the robustness of PRIME under differential privacy guarantees. Our choice of experiment settings and hyper parameters are following: $d \approx 100$, $\mu_{\text{bad}} = (1.5, 1.5, \cdots, 1.5)_d$, $(\varepsilon, \delta) = (10, 0.01)$, $\alpha = 0.1$, $R = 10$, $C = 2$.

Our implementation is based on Python with basic Numpy library. We run on a 2018 Macbook Pro machine. For each choice of d in our settings, PRIME takes less than 2 minutes and stops after roughly 3 epochs. The source code for reproducing Figure 2.1 is available at `https://github.com/xiyangl3/robust_dp`.

Appendix B

APPENDICES FOR CHAPTER 3

B.1 General case: utility analysis of HPTR

We prove the following theorem that provides a utility guarantee for HPTR output $\hat{\theta}$ measured in $D_\phi(\hat{\theta}, \theta)$.

Theorem 28. *For a given dataset S, a target error function $D_\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$, probability $\zeta \in (0,1)$, and privacy (ε, δ),* HPTR *achieves $D_\phi(\hat{\theta}, \theta) = c_0\rho$ for some $\rho > 0$ and any constant $c_0 > 3c_1$ with probability $1-\zeta$ if there exist constants $c_1, c_2 > 0$ and $(\Delta \in \mathbb{R}^+, \rho \in \mathbb{R}^+)$ such that with the choice of $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, $\tau = (c_0 + c_1)\rho$, the following assumptions are satisfied:*

(a) (Bounded volume) $(7/8)\tau - (k^*+1)\Delta > 0$,

$$\begin{aligned} \frac{\mathrm{Vol}(B_{\tau+(k^*+1)\Delta+c_1\rho,S})}{\mathrm{Vol}(B_{(7/8)\tau-(k^*+1)\Delta-c_1\rho,S})} &\leq e^{c_2 p}\ , \text{ and} \\ \frac{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta},\theta) \leq (c_0+2c_1)\rho\})}{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta},\theta) \leq c_1\rho\})} &\leq e^{c_2 p}\ , \end{aligned}$$

(b) (Local sensitivity) For all S' within Hamming distance k^* from S, $\max_{S''\sim S'} \|D_{S''}(\hat{\mu}) - D_{S'}(\hat{\mu})\| \leq \Delta$ *for all $\hat{\mu} \in B_{\tau+(k^*+3)\Delta,S}$,*

(c) (Bounded sensitivity) $\Delta \leq \frac{(c_0-3c_1)\rho\varepsilon}{32(c_2 p+(\varepsilon/2)+\log(16/\delta\zeta))}$, *and*

(d) (Robustness) $|D_\phi(\hat{\theta},\theta) - D_S(\hat{\theta})| \leq c_1\rho$ *for all $\hat{\theta} \in B_{\tau,S}$.*

The parameter $\rho \in \mathbb{R}_+$ represents the target error up to a constant factor and depends on the resilience of the underlying distribution $P_{\theta,\phi}$ that the samples are drawn from. We

explicitly prescribe how to choose the parameter ρ for each problem instance in Sections 3.3, 3.4, 3.5, and 3.6. Following the standard analysis techniques for exponential mechanisms, we show that the output concentrates around an inner set $\{\hat{\theta} : D_\phi(\hat{\theta}, \theta) \leq c_0\rho\}$, by comparing its probability mass with an outer set $\{\hat{\theta} : D_\phi(\hat{\theta}, \theta) \geq c_1\rho\}$. This uses the ratio of the volumes in the assumption (a) and the closeness of the error metric and $D(\hat{\theta})$ in the assumption (d). When there is a strict gap between the two, which happens if $\varepsilon\rho/\Delta \gg p + \log(1/\zeta)$ as in the assumption (c), this implies $D_\phi(\hat{\theta}, \theta) \leq c_0\rho$ with probability $1 - \zeta$. We provide a proof in Section B.1.2.

A major challenge in analyzing HPTR is in showing that the safety test threshold $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$ is not only large enough to ensure that datasets with safety violation is screened with probability $1 - \delta/2$ but also small enough such that good datasets satisfying the assumptions (a), (b), and (c) pass the test with probability $1 - \zeta/2$. We establish this first in Section B.1.1.

B.1.1 Large safety margin

In this section, we show in Lemma B.1.3 that under the assumptions of Theorem 28, we get a large enough margin for safety such that we pass the safety test with high probability. We follow the proof strategy introduced in [34] adapted to our more general framework. A major challenge is the lack of a uniform bound on the sensitivity, which the analysis of [34] relies on. We generalize the analysis by showing that while the data does not satisfy uniform sensitivity bound, we can still exploit its *local* sensitivity bound in the assumption (b).

The following main technical lemma is a counter part of [34, Lemma 3.7], where we have an extra challenge that the sensitivity bound is only local; there exists $\hat{\theta}$ far from θ where the sensitivity bound fails. We rely on the assumption (b) to resolve it. Let $w_S(B) \triangleq \int_B \exp\{-(\varepsilon/4\Delta)D_S(\hat{\mu})\}d\hat{\mu}$ be the weight of a subset $B \subset \mathbb{R}^p$. The following lemma will be used to show that the denominator of the exponential distribution in RELEASE step does not change too fast between two neighboring datasets.

Lemma B.1.1. *Under the assumption (b) and $\delta \in (0, 1/2)$, for a dataset S' at Hamming distance at most k^* from S, if $w_{S'}(B_{\tau-\Delta,S'}) \geq (1-\delta) w_{S'}(B_{\tau+\Delta,S'})$ then $S' \in \mathsf{SAFE}_{\varepsilon,4e^{2\varepsilon}\delta,\tau}$.*

Proof. We follow the proof strategy of [34, Lemma 3.7] but there are key differences due to the fact that we do not have a universal sensitivity bound, but only local bound. In particular, we first establish that under the local sensitivity assumption, $B_{\tau,S''} \subseteq B_{\tau+\Delta,S'}$ for all $S'' \sim S'$, which will be used heavily throughout the proof. Since $D_{S''}(\hat{\theta}) \leq D_{S'}(\hat{\theta}) + \Delta$ for all $\hat{\theta} \in B_{\tau+(k^*+3)\Delta,S}$, we have $B_{\tau,S''} \cap B_{\tau+(k^*+3)\Delta,S} \subseteq B_{\tau+\Delta,S'}$. We are left to show that $B_{\tau,S''} \setminus B_{\tau+(k^*+3)\Delta,S} = \emptyset$, which follows from the fact that $(B_{\tau,S''} \setminus B_{\tau+(k^*+1.5)\Delta,S}) \cap B_{\tau+(k^*+3)\Delta,S}) = \emptyset$ and $D_{S''}(\hat{\theta})$ is a Lipschitz continuous function. Similarly, it follows that $B_{\tau-\Delta,S'} \subseteq B_{\tau,S''}$. In particular, this implies that $B_{\tau,S'} \subseteq B_{\tau+(k^*+3)\Delta,S}$ for any S' with $d_H(S', S) \leq k^*$.

We first show that for any $E \subset B_{\tau,S'}$ one side of the $(\varepsilon/2, 4e^{\varepsilon/2}\delta)$-DP condition is met: $\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) \leq e^{\varepsilon/2}\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E) + 4e^{\varepsilon/2}\delta$ for all $S'' \sim S'$ where $r_{(\varepsilon,\Delta,\tau,S')}$ and $r_{(\varepsilon,\Delta,\tau,S'')}$ are the distributions used in the exponential mechanism as defined in (3.3) respectively. For $B = B_{\tau,S'} \cap B_{\tau,S''}$, we have

$$\begin{aligned}
\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) &= \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \cap B) + \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \setminus B) \\
&= \frac{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \cap B)}{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E \cap B)} \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E \cap B) + \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \setminus B) \\
&\leq \frac{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \cap B)}{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E \cap B)} \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E) + \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \notin B_{\tau,S''}) \,.
\end{aligned}$$

The ratio is bounded due to the local sensitivity bound at S' as

$$\begin{aligned}
\frac{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \cap B)}{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E \cap B)} &\leq e^{\varepsilon/4}\frac{w_{S''}(B_{\tau,S''})}{w_{S'}(B_{\tau,S'})} \\
&\leq e^{\varepsilon/2}\frac{w_{S'}(B_{\tau,S''})}{w_{S'}(B_{\tau,S'})} \\
&\leq e^{\varepsilon/2}\frac{w_{S'}(B_{\tau+\Delta,S})}{w_{S'}(B_{\tau,S'})} \leq e^{\varepsilon/2}(1+2\delta) \,,
\end{aligned}$$

where the second inequality follows from the fact that $w_{S''}(A) \leq e^{\varepsilon/6} w_{S'}(A)$ for any set $A \subset B_{\tau,S'} \cup B_{\tau,S''} \subseteq B_{\tau+(k^*+3)\Delta,S}$ and the third inequality follows from the fact that $B_{\tau,S''} \subseteq$

$B_{\tau+\Delta,S'}$. From the assumption on the weights, it follows that $w_{S'}(B_{\tau+\Delta,S'})/w_{S'}(B_{\tau,S'}) \leq w_{S'}(B_{\tau+\Delta,S'})/w_{S'}(B_{\tau-\Delta,S'}) \leq 1/(1-\delta) \leq 1+2\delta$ for $\delta < 1/2$. Similarly,

$$\begin{aligned}
\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \notin B_{\tau,S''}) &\leq \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \notin B_{\tau-\Delta,S'}) \\
&\leq 1 - \frac{w_{S'}(B_{\tau-\Delta,S'})}{w_{S'}(B_{\tau,S'})} \leq 1 - \frac{w_{S'}(B_{\tau-\Delta,S'})}{w_{S'}(B_{\tau+\Delta,S'})} \leq \delta \ .
\end{aligned}$$

Putting these together, we get $\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) \leq e^{\varepsilon/2}\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in E) + 4e^{\varepsilon/2}\delta$.

Next, we show the other side of the $(\varepsilon/2, 4e^{\varepsilon/2}\delta)$-DP condition: $\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E) \leq e^{\varepsilon/2}\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S)}}(\hat{\theta} \in E) + 4e^{2\varepsilon}\delta$ for all $S' \sim S$. We need to show an upper bound on the ratio:

$$\begin{aligned}
\frac{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S')}}(\hat{\theta} \in E \cap B)}{\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S)}}(\hat{\theta} \in E \cap B)} &\leq e^{\varepsilon/4}\frac{w_S(B_{\tau,S})}{w_{S'}(B_{\tau,S'})} \\
&\leq e^{\varepsilon/2}\frac{w_S(B_{\tau,S})}{w_S(B_{\tau,S'})} \\
&\leq e^{\varepsilon/2}\frac{w_S(B_{\tau,S})}{w_S(B_{\tau-\Delta,S})} \ \leq \ (1+2\delta)e^{\varepsilon/2} \ ,
\end{aligned}$$

For the probability outside $B_{\tau,S'}$,

$$\begin{aligned}
\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \notin B_{\tau,S'}) &\leq \mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S'')}}(\hat{\theta} \in B_{\tau+\Delta,S'} \setminus B_{\tau,S'}) \\
&\leq \frac{w_{S''}(B_{\tau+\Delta,S'} \setminus B_{\tau,S'})}{w_{S''}(B_{\tau,S''})} \\
&\leq e^{\varepsilon/2}\frac{w_{S'}(B_{\tau+\Delta,S'} \setminus B_{\tau,S'})}{w_{S'}(B_{\tau,S''})} \\
&\leq e^{\varepsilon/2}\frac{w_{S'}(B_{\tau+\Delta,S'}) - w_{S'}(B_{\tau,S'})}{w_{S'}(B_{\tau-\Delta,S'})} \\
&\leq e^{\varepsilon/2}(1+2\delta-1) \ = \ 2e^{\varepsilon/2}\delta \ .
\end{aligned}$$

where the first inequality follows from $B_{\tau,S''} \subseteq B_{\tau+\Delta,S'}$, the second inequality follows from $(B_{\tau+\Delta,S'} \setminus B_{\tau,S'}) \cap B_{\tau,S''} \subseteq B_{\tau+\Delta,S'} \setminus B_{\tau,S'}$, the third inequality follows from the fact that $B_{\tau,S''} \subseteq B_{\tau+\Delta,S'}$ and the local sensitivity assumption, and the last inequality follows from the weight assumption and $B_{\tau-\Delta,S'} \subseteq B_{\tau,S'}$.

□

The next lemma identifies the range of the threshold $k^* = O(\tau/\Delta)$ that ensures safety.

Lemma B.1.2. *Under the assumption (b), if there exists a $g > 0$ such that $\tau - \Delta(k^*+g+1) > 0$ and*

$$\frac{\mathrm{Vol}(B_{\tau+\Delta(k^*+1),S})}{\mathrm{Vol}(B_{\tau-\Delta(k^*+g+1),S})} e^{\frac{-\varepsilon g}{4}} \;\leq\; \frac{1}{8} e^{-\varepsilon/2}\delta \;, \tag{B.1}$$

then $S' \in \mathsf{SAFE}_{(\varepsilon/2,\delta/2,\tau)}$ for all S' within Hamming distance k^* from S.

Proof. Consider S' at Hamming distance k away from S. From Lemma B.1.1 it suffices to show that $w_{S'}(B_{\tau-\Delta,S'})/w_{S'}(B_{\tau+\Delta,S'}) \geq 1-\delta'$ for $\delta' = (1/8)e^{-\varepsilon/2}\delta$, which is equivalent to

$$w_{S'}(B_{\tau+\Delta,S'} \setminus B_{\tau-\Delta,S'})/w_{S'}(B_{\tau+\Delta,S'}) \leq \delta' \;.$$

The denominator is lower bounded by

$$\begin{aligned} w_{S'}(B_{\tau+\Delta,S'}) \;\geq\; w_{S'}(B_{\tau-\Delta(1+g),S'}) \;&\geq\; \mathrm{Vol}(B_{\tau-\Delta(1+g),S'}) e^{-\varepsilon(\tau-\Delta(1+g))/(4\Delta)} \\ &\geq\; \mathrm{Vol}(B_{\tau-\Delta(1+g+k),S}) e^{-\varepsilon(\tau-\Delta(1+g))/(4\Delta)} \;, \end{aligned}$$

where the last inequality uses the local sensitivity (the assumption (b)). The numerator is upper bounded by

$$w_{S'}(B_{\tau+\Delta,S'} \setminus B_{\tau-\Delta,S'}) \;\leq\; w_{S'}(B_{\tau+(k+1)\Delta,S} \setminus B_{\tau-\Delta,S'}) \;\leq\; \mathrm{Vol}(B_{\tau+(k+1)\Delta,S}) e^{-\varepsilon(\tau-\Delta)/(4\Delta)} \;,$$

where the first inequality uses the local sensitivity. Together, it follows that

$$\frac{w_{S'}(B_{\tau+\Delta,S'} \setminus B_{\tau-\Delta,S'})}{w_{S'}(B_{\tau+\Delta,S'})} \;\leq\; \frac{\mathrm{Vol}(B_{\tau+(k+1)\Delta,S}) e^{-\varepsilon(\tau-\Delta)/(4\Delta)}}{\mathrm{Vol}(B_{\tau-\Delta(1+g+k),S}) e^{-\varepsilon(\tau-\Delta(1+g))/(4\Delta)}} \;\leq\; \delta' \;=\; \frac{1}{8} e^{\varepsilon/2}\delta \;,$$

as $e^{-\varepsilon(\tau-\Delta)/(4\Delta)} / e^{-\varepsilon(\tau-\Delta(1+g))/(4\Delta)} = e^{-\varepsilon g/4}$, which implies safety.

□

We next show that $k^* = O((1/\varepsilon)\log(1/(\delta\zeta)))$ is sufficient to ensure a large enough safety margin of $m_\tau - k^* = \Omega((1/\varepsilon)\log(1/\zeta))$.

Lemma B.1.3. *Under the assumptions (a), (b), and (c) of Theorem 28, for $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$, if $d_H(S', S) \leq (2/\varepsilon)\log(4/(\zeta\delta))$ then $S' \in \mathsf{SAFE}_{(\varepsilon/2,\delta/2,\tau)}$.*

Proof. Applying Lemma B.1.2 with $k^* = (2/\varepsilon)\log(4/(\delta\zeta))$ and $g = (1/(8\Delta))\tau$, we require

$$\frac{\mathrm{Vol}(B_{\tau+\Delta(k^*+1),S})}{\mathrm{Vol}(B_{(7/8)\tau-\Delta(k^*+1),S})} e^{\frac{-\varepsilon\tau}{32\Delta}} \;\leq\; \frac{1}{8}e^{-\varepsilon/2}\delta\,.$$

From the assumption (a), it is sufficient to have

$$\exp\left\{c_2 p - \frac{\tau\varepsilon}{32\Delta}\right\} \;\leq\; \frac{1}{8}e^{-\varepsilon/2}\delta\,.$$

For $\Delta \leq (\tau\varepsilon)/(32(c_2 p + (\varepsilon/2) + \log(8/\delta)))$, which follows from the assumption (c), this is satisfied. □

B.1.2 Proof of Theorem 28

We first show that we pass the safety test with high probability. Define the error event E as the event that we output $\perp$ in the TEST step. From Lemma B.1.3, we have $m_\tau > (2/\varepsilon)\log(4/(\delta\zeta))$ under the assumptions (a), (b), and (c). This implies that

$$\mathbb{P}(E) \;=\; \mathbb{P}\big(\, m_\tau + \mathrm{Lap}(2/\varepsilon) < (2/\varepsilon)\log(2/\delta)\,\big) \;\leq\; \frac{\zeta}{2}\,.$$

We next show that resilience implies good utility (once safety test has passed). We want the exponential mechanism to output an accurate $\hat{\theta}$ near θ with high probability, i.e., $\mathbb{P}_{\hat{\theta}\sim r_{(\varepsilon,\Delta,\tau,S)}}(D_\phi(\hat{\theta},\theta) \geq c_0\rho) \leq \zeta/2$. We omit the subscript in the probability for brevity, and it is assumed that randomness is in the sampling of the exponential mechanism. We want to bound by $\zeta/2$ the failure probability:

$$\begin{aligned}
\mathbb{P}\big(\, D_\phi(\hat{\theta},\theta) \geq c_0\rho\,\big) \;&\leq\; \frac{\mathbb{P}\big(\, D_\phi(\hat{\theta},\theta) \geq c_0\rho\,\big)}{\mathbb{P}\big(\, D_\phi(\hat{\theta},\theta) \leq c_1\rho_1\,\big)} \\
&\leq\; \frac{\mathrm{Vol}(B_{\tau,S})}{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta},\theta) \leq c_1\rho\})} \frac{\max_{\hat{\theta}:D_\phi(\hat{\theta},\theta)\geq c_0\rho}\mathbb{P}(\hat{\theta})}{\min_{\hat{\theta}:D_\phi(\hat{\theta},\theta)\leq c_1\rho_1}\mathbb{P}(\hat{\theta})}\,,
\end{aligned}$$

as long as $\{\hat{\theta} : D_\phi(\hat{\theta},\theta) \leq c_0\rho\} \subseteq B_{\tau,S}$ (otherwise we are under-estimating the volume), which follows from the assumption (d); $D_S(\hat{\theta}) \leq (D_\phi(\hat{\theta},\theta) + c_1\rho) \leq (c_0 + c_1)\rho = \tau$.

Similarly, since $\hat{\theta} \in B_{\tau,S}$ implies $D_\phi(\hat{\theta}, \theta) \le \tau + c_1\rho = (c_0 + 2c_1)\rho$, the volume ratio is bounded by

$$\frac{\mathrm{Vol}(B_{\tau,S})}{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta}, \theta) \le c_1\rho)} \;\le\; \frac{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta}, \theta) \le (c_0 + 2c_1)\rho\})}{\mathrm{Vol}(\{\hat{\theta} : D_\phi(\hat{\theta}, \theta) \le c_1\rho\})} \;\le\; e^{c_2 p} \,,$$

under the assumption (a). The probability ratio can be bounded similarly. From the assumption (d), we have

$$\frac{\max_{\hat{\theta}: D_\phi(\hat{\theta},\theta) \ge c_0\rho} \mathbb{P}(\hat{\theta})}{\min_{\hat{\theta}: D_\phi(\hat{\theta},\theta) \le c_1\rho} \mathbb{P}(\hat{\theta})} \;\le\; \exp\Big\{ -\frac{\varepsilon}{4\Delta}\big((c_0 - c_1) - (2c_1)\big)\rho \Big\} \;\le\; \exp\Big\{ -\frac{\varepsilon(c_0 - 3c_1)\rho}{4\Delta} \Big\} \,.$$

When $e^{c_2 p - (\varepsilon(c_0 - 3c_1)\rho/(4\Delta))}) \le \zeta/2$, we have the desired bound. This is guaranteed with our assumption (c).

B.2 Auxiliary lemmas

Lemma B.2.1. *For any symmetric $\Sigma \succ 0$ and vector $u \in \mathbb{R}^d$,*

$$\max_{v: \|v\|=1} \frac{\langle v, u\rangle}{v^\top \Sigma v} = \big\|\Sigma^{-1/2} u\big\| \;. \tag{B.2}$$

Proof. This follows analogously from the proof of Lemma 3.3.1. □

Lemma B.2.2. *Let $\Sigma, A \in \mathbb{R}^{d\times d}$ be a symmetric matrix. If $-c\mathbf{I}_{d\times d} \preceq \Sigma^{-1/2} A \Sigma^{-1/2} - \mathbf{I}_{d\times d} \preceq c\mathbf{I}_{d\times d}$ for some $c > 0$, then we have for any $u \in \mathbb{R}^d$,*

$$\|\Sigma^{-1/2}(A - \Sigma)u\| \le c\|\Sigma^{1/2} u\| \;. \tag{B.3}$$

Proof. Using the fact that $-\mathbf{I}_{d\times d} \preceq M \preceq \mathbf{I}_{d\times d}$ implies $-\mathbf{I}_{d\times d} \preceq M^2 \preceq \mathbf{I}_{d\times d}$, for any symmetric matrix M, we know

$$-c^2 \mathbf{I}_{d\times d} \preceq \Sigma^{-1/2}(A - \Sigma)\Sigma^{-1}(A - \Sigma)\Sigma^{-1/2} \preceq c^2 \mathbf{I}_{d\times d} \,, \tag{B.4}$$

which implies that

$$-c^2 \Sigma \preceq (A - \Sigma)\Sigma^{-1}(A - \Sigma) \preceq c^2 \Sigma \,. \tag{B.5}$$

Thus, we know

$$\|\Sigma^{-1/2}(A-\Sigma)u\|^2 = u^\top(A-\Sigma)\Sigma^{-1}(A-\Sigma)u \le c^2 u^\top \Sigma u = c^2\|\Sigma^{1/2}u\|^2 \,. \tag{B.6}$$

□

B.3 Existing lower bounds

Theorem B.3.1 (Lower bound for DP Gaussian mean estimation with known covariance [129, Lemma 6.7])**.** *Let $\hat{\mu} : \mathbb{R}^{n\times d} \to [-R\sigma, R\sigma]^d$ be an (ε,δ)-differentially private estimator (with $\delta \le \sqrt{d}/(48\sqrt{2}Rn\sqrt{\log(48\sqrt{2}Rn/\sqrt{d})})$) such that for every Gaussian distribution $P = \mathcal{N}(\mu, \sigma^2\mathbf{I}_{d\times d})$ (for $-R\sigma \le \mu_j \le R\sigma$ where $j \in [d]$) and*

$$\mathbb{E}_{S\sim P^n}\left[\|\hat{\mu}(S)-\mu\|^2\right] \le \alpha^2 \le \frac{d\sigma^2R^2}{6} \,, \tag{B.7}$$

then $n \ge \frac{d\sigma}{24\alpha\varepsilon}$.

Theorem B.3.2 (Lower bound for DP covariance bounded mean estimation [135, Theorem 6.1])**.** *Suppose $\hat{\mu}$ is an $(\varepsilon, 0)$-DP estimator such that, for every product distribution $P \in \mathbb{R}^d$ such that $\mathbb{E}[P] = \mu$, $\sup_{v:\|v\|=1}\mathbb{E}_{x\sim P}[\langle v, x-\mu\rangle^2] \le 1$ and*

$$\mathbb{E}_{S\sim P^n}\left[\|\hat{\mu}(S)-\mu\|^2\right] \le \alpha^2 \,. \tag{B.8}$$

Then $n = \Omega\left(d/(\varepsilon\alpha^2)\right)$

Theorem B.3.3 (Lower bound on the error rate for hypercontractive linear regression with independent noise[22, Theorem 6.1])**.** *Consider linear model $y = \langle\beta, x\rangle + \eta$, where optimal hyperplane β is used to generate data, and the noise η is independent of the samples x. Then there exists two distribution D_1 and D_2 over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is (κ_k, k)-hypercontractive yet $\|\Sigma^{1/2}(\beta_1-\beta_2)\| = \Omega(\sqrt{\kappa_k}\gamma\alpha^{1-1/k})$, where β_1 and β_2 are the optimal hyperplanes for D_1 and D_2 respectively, $\gamma < 1/\alpha^{1/k}$ and the noise η is uniform over $[-\gamma,\gamma]$.*

Theorem B.3.4 (Lower bound on the error rate for hypercontractive linear regression with dependent noise[22, Theorem 6.2]). *There exists two distributions D_1, D_2 over $\mathbb{R}^2 \times \mathbb{R}$ such that the marginal distribution over $\mathbb{R}^2$ has covariance Σ and is κ_k, k-hypercontractive yet $\|\Sigma^{1/2}(\beta_1 - \beta_2)\| = \Omega(\sqrt{\kappa_k}\gamma\alpha^{1-2/k})$, where β_1 and β_2 are least square solutions for D_1 and D_2, respectively, $\gamma < 1/\alpha^{1/k}$ and the noise is a function of the marginal distribution of $\mathbb{R}^2$,*

Theorem B.3.5 (Lower bound for DP sub-Gaussian linear regression [40, Theorem 4.1]). *Given i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^n$ drawn from model $y_i = \langle \beta, x_i \rangle + \eta_i$, where $\eta_i \sim \mathcal{N}(0, \gamma^2)$, $\beta \in \Theta = \{\beta \in \mathbb{R}^d : \|\beta\| \leq 1\}$, $\mathbb{P}(\|x\| \leq 1) = 1$, $\Sigma = \mathbb{E}[xx^\top]$ is diagonal and satisfies $0 < 1/L < d\lambda_{\min}(\Sigma) \leq d\lambda_{\max}(\Sigma) < L$ for some constant $L = O(1)$. Denote this class of distribution as $\mathcal{P}_{\gamma,\Theta,\Sigma}$. Denote $\mathcal{M}_{\varepsilon,\delta}$ as a class of (ε, δ)-DP algorithms. Then suppose $\varepsilon \in (0, 1)$, $\delta \in (0, n^{-(1+w)})$ for some fixed $w > 0$, then there exists a constant such that*

$$\inf_{\hat{\beta} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\Sigma \succ 0, P \in \mathcal{P}_{\gamma,\Theta,\Sigma}} \mathbb{E}_{P^n}\left[\|\Sigma^{1/2}(\hat{\beta}(S) - \beta)\|^2\right] \geq c\gamma^2\left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right). \tag{B.9}$$

Theorem B.3.6 (Lower bound of linear regression [179, Theorem 1]). *A multiset of i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^n$ is drawn from distribution $P \in \mathbb{R}^d \times \mathbb{R}$ in a class $\mathcal{P}_{B,Y}$, where $|y| \leq Y$, $\|x\| \leq 1$ and $\beta \in \Theta_B = \{\beta \in \mathbb{R}^d : \|\beta\| \leq B\}$. Then there exists a constant c such that*

$$\inf_{\hat{\beta} \in \Theta_B} \sup_{P \in \mathcal{P}_{B,Y}} \mathbb{E}_{P^n}\left[\left(y - \left\langle \hat{\beta}(S), x \right\rangle\right)^2 - \min_{\beta \in \Theta_B}(y - \langle \beta, x \rangle)^2\right] \geq c \min\left\{Y^2, B^2, \frac{dY^2}{n}, \frac{BY}{\sqrt{n}}\right\} \tag{B.10}$$

Theorem B.3.7 (Lower bound of Gaussian DP covariance estimation [129, Lemma 6.11]). *Let $\widehat{\Sigma} : \mathbb{R}^{n \times d} \to \Theta$ be an $(\varepsilon, 0)$-DP estimator (where Θ is the space of all $d \times d$ PSD matriaces), and for every $\mathcal{N}(0, \Sigma)$ over $\mathbb{R}^d$ such that $1/2\mathbf{I}_{d\times d} \preceq \Sigma \preceq 3/2\mathbf{I}_{d\times d}$,*

$$\mathbb{E}_{S \sim \mathcal{N}(0,\Sigma)^n}\left[\|\widehat{\Sigma}(S) - \Sigma\|_F^2\right] \leq \frac{\alpha^2}{64}, \tag{B.11}$$

then $n \geq \Omega(d^2/(\varepsilon\alpha))$.

Appendix C

APPENDICES FOR CHAPTER 4

C.1 Related work

Our work builds upon a series of advances in private SGD [131, 27, 26, 85, 149, 204, 112] to make advance in understanding the tradeoff of privacy and sample complexity for PCA. Such tradeoffs have been studied extensively in canonical statistical estimation problems of mean (and covariance) estimation and linear regression.

Private mean estimation. As one of the most fundamental problem in private data analysis, mean estimation is initially studied under the bounded support assumptions, and the optimal error rate is now well understood. More recently, [25] considered the private mean estimation problem for k-th moment bounded distributions where the support of the data is *unbounded* and provided minimax error bound in various settings. [140] studied private mean estimation from Gaussian sample, and obtained an optimal error rate. There has been a lot of recent interests on private mean estimation under various assumptions, including mean and covariance joint estimation [130, 32], heavy-tailed mean estimation [135], mean estimation for general distributions [87, 197], distribution adaptive mean estimation [39], estimation for unbounded distribution parameters [133], mean estimation under pure differential privacy [107], local differential privacy [75, 76, 89, 123], user-level differential privacy [83], Mahalanobis distance[34] and robust and differentially private mean estimation [160, 146, 161].

Private linear regression The goal of private linear regression is to learn a linear predictor of response variable y from a set of examples $\{x_i, y_i\}_{i=1}^n$ while guarantee the privacy of the examples. Again, the work on private linear regression can be divided into two categories: deterministic and randomized. In the deterministic setting where the data is deterministically given without any probabilistic assumptions, significant advances in DP

linear regression has been made [201, 142, 168, 71, 27, 208, 88, 167, 206, 180]. In the randomized settings where each example $\{\mathbf{x}_i, y_i\}$ is drawn i.i.d. from a distribution [166], [77] proposes an exponential time algorithm that achieves asymptotic consistency. [40] provides an efficient and minimax optimal algorithm under sub-Gaussian design and nearly identity covariance assumptions. Very recently, [161] for the first time gives an exponential time algorithm that achieves minimax risk for general covariance matrix under sub-Gaussian and hypercontractive assumptions. [199] gives the first computationally efficient algorithm to achieve nearly optimal risk using DP-SGD with adaptive clipping.

Private PCA without spectral gap. There is a long line of work in Private PCA [102, 103, 101, 33, 46, 136, 80, 24]. We explain the closely related ones in Section 4.2.3, with analysis when the covariance matrix has a spectral gap.

When there is no spectral gap, one can still learn a principal component. However, since the principal component is not unique, the error is typically measured in how much of the variance is captured in the estimated direction: $1 - \hat{v}^\top \Sigma \hat{v}/\|\Sigma\|$. [46] introduces an exponential mechanism (from [164]) which samples an estimate from a distribution $f_{\widehat{\Sigma}}(\hat{v}) = (1/C)\exp\{((\varepsilon n)/c^2)\hat{v}^\top \widehat{\Sigma}\hat{v}\}$, where C is a normalization constant to ensure that the pdf integrates to one. This achieves a stronger pure DP, i.e., $(\varepsilon, 0)$-DP, but is computationally expensive; [46] does not provide a tractable implementation and [136] provides a polynomial time implementation with time complexity at least cubic in d. This achieves error rate $1 - \hat{v}^\top \Sigma \hat{v}/\|\Sigma\| = \tilde{O}(d^2/(\varepsilon n))$ in [46, Theorem 7], which, when there is a spectral gap, translates into

$$\sin(\hat{v}, v_1)^2 \;=\; \tilde{O}\Big(\frac{\kappa d^2}{\varepsilon n}\Big)\,, \tag{C.1}$$

with high probability. Closest to our setting is the analyses in [161, Corollary 6.5] that proposed an exponential mechanism that achieves $1-\hat{v}^\top \Sigma \hat{v}/\|\Sigma\| = \tilde{O}(\sqrt{d/n}+(d+\log(1/\delta))/(\varepsilon n))$ with high probability under (ε, δ)-DP and Gaussian samples, but this algorithm is computationally intractable. This is shown to be tight when there is no spectral gap. When there is a spectral

gap, this translates into

$$\sin(\hat{v}, v_1)^2 \;=\; \tilde{O}\Big(\kappa\Big(\sqrt{\frac{d}{n}} + \frac{d + \log(1/\delta)}{\varepsilon n}\Big)\Big)\,. \tag{C.2}$$

Distributed PCA. In distributed PCA, the dataset is stored across different local servers [117, 118, 205, 92]. [117, 118, 205] consider differentially private distributed PCA under the assumption that the examples are deterministic and have norms bounded by a fixed and known constant. The algorithms appeared in [117, 118, 205] are based on the Gaussian mechanism [80] on local server and an aggregator in the central server. The resulting utility guarantees are the same as those from [80], which are discussed in Section 4.2.3.

C.2 Preliminaries

Since we focus on one-pass algorithms where a data point is only accessed once, we need a basic parallel composition of DP.

Lemma C.2.1 (Parallel composition [165])**.** *Consider a sequence of interactive queries $\{q_k\}_{k=1}^K$ each operating on a subset S_k of the database and each satisfying (ε, δ)-DP. If S_k's are disjoint then the composition $(q_1(S_1), q_2(S_2), \ldots, q_K(S_K))$ is (ε, δ)-DP.*

We also utilize the following serial composition theorem.

Lemma C.2.2 (Serial composition [79])**.** *If a database is accessed with an $(\varepsilon_1, \delta_1)$-DP mechanism and then with an $(\varepsilon_2, \delta_2)$-DP mechanism, then the end-to-end privacy guarantee is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.*

When we apply private histogram learner to each coordinate, we require more advanced composition theorem from [128].

Lemma C.2.3 (Advanced composition [128])**.** *For $\varepsilon \leq 0.9$, an end-to-end guarantee of (ε, δ)-differential privacy is satisfied if a database is accessed k times, each with a $(\varepsilon/(2\sqrt{2k\log(2/\delta)}), \delta/(2k))$-differential private mechanism.*

C.3 Converse results

When privacy is not required, we know from Theorem 4.2.2 that under Assumptions A.1-A.3, we can achieve an error rate of $\tilde{O}(\kappa\sqrt{V/n})$. In the regime of $V = O(d)$ and $\kappa = O(1)$, $n = O(d)$ samples are enough to achieve an arbitrarily small error. The next lower bounds shows that we need $n = O(d^2)$ samples when $(\varepsilon = O(1), 0)$-DP is required, showing that private PCA is significantly more challenging than a non-private PCA, when assuming only the support and moment bounds of assumptions A.1-A.3. We provide a proof in Appendix C.3.3.

Theorem C.3.1 (Lower bound without Assumption A.4)**.** *Let $\mathcal{M}_\varepsilon$ be a class of $(\varepsilon, 0)$-DP estimators that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3 with $M = O(d \log n)$ and $V = O(d)$ is denoted by $\tilde{\mathcal{P}}_{(\lambda_1,\lambda_2)}$. There exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\tilde{\mathcal{P}}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \geq C \min\left(\frac{\kappa d^2}{\varepsilon n}\sqrt{\frac{\lambda_2}{\lambda_1}}, \sqrt{\frac{\lambda_2}{\lambda_1}}\right). \tag{C.3}$$

We next provide the proofs of all the lower bounds.

C.3.1 Proof of Theorem 4.5.3 on the lower bound for Gaussian case

Our proof is based on following differentially private Fano's method [3, Corollary 4].

Theorem C.3.2 (DP Fano's method [3, Corollary 4])**.** *Let $\mathcal{P}$ denote family of distributions of interest and $\theta : \mathcal{P} \to \Theta$ denote the population parameter. Our goal is to estimate θ from i.i.d. samples $x_1, x_2, \ldots, x_n \sim P \in \mathcal{P}$. Let $\hat{\theta}_\varepsilon$ be an $(\varepsilon, 0)$-DP estimator. Let $\rho : \Theta \times \Theta \to \mathbb{R}^+$ be a pseudo-metric on parameter space Θ. Let $\mathcal{V}$ be an index set with finite cardinality. Define $\mathcal{P}_\mathcal{V} = \{P_v, v \in \mathcal{V}\} \subset \mathcal{P}$ be an indexed family of probability measures on measurable set $(\mathcal{X}, \mathcal{A})$. If for any $v \neq v' \in \mathcal{V}$,*

1. $\rho(\theta(P_v), \theta(P_{v'})) \geq \tau$,

2. $D_{\mathrm{KL}}(P_v, P_{v'}) \leq \beta$,

3. $D_{\mathrm{TV}}(P_v, P_{v'}) \le \phi$,

then

$$\inf_{\hat{\theta}_\varepsilon} \max_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}\left[\rho(\hat{\theta}_\varepsilon(S), \theta(P))\right] \ge \max\left(\frac{\tau}{2}\left(1 - \frac{n\beta + \log(2)}{\log(|\mathcal{V}|)}\right), 0.4\tau \min\left(1, \frac{|\mathcal{V}|}{e^{10n\phi\varepsilon}}\right)\right) . \tag{C.4}$$

For our problem, we are interested in Gaussian $\mathcal{P}_\Sigma$ and metric $\rho(u, v) = \sin(u, v)$. Using Theorem C.3.2, it suffices to construct such indexed set $\mathcal{V}$ and the indexed distribution family $\mathcal{P}_\mathcal{V}$. We use the same construction as in [202, Theorem 2.1] introduced to prove a lower bound for the (non-private) sparse PCA problem. The construction is given by the following lemma.

Lemma C.3.3 ([202, Lemma 3.1.2])**.** *Let* $d > 10$. *For* $\alpha \in (0, 1]$, *there exists* $\mathcal{V}_\alpha \subset \mathbb{S}_2^{d-1}$ *and an absolute constant* $c_1 > 0.0233$ *such that for every* $v \ne v' \in \mathcal{V}_\alpha$, $\alpha/\sqrt{2} \le \|v - v'\|_2 \le \sqrt{2}\alpha$ *and* $\log(|\mathcal{V}_\alpha|) \ge c_1 d$.

Fix $\alpha \in (0, 1]$. For each $v \in \mathcal{V}_\alpha$, we define $\Sigma_v = (\lambda_1 - \lambda_2)vv^\top + \lambda_2 \mathbf{I}_d$ and $P_v = \mathcal{N}(0, \Sigma_v)$. It is easy to see that Σ_v has eigenvalues $\lambda_1 > \lambda_2 = \cdots = \lambda_n$. The top eigenvector of Σ_v is v. Using Lemma C.6.4, we know for any $v \ne v' \in \mathcal{V}$,

$$\frac{\alpha}{2} \le \frac{1}{\sqrt{2}}\|v - v'\| \le \rho(v, v') = \sqrt{1 - \langle v, v'\rangle^2} \le \|v - v'\| \le \sqrt{2}\alpha . \tag{C.5}$$

Using [202, Lemma 3.1.3], we know

$$D_{\mathrm{KL}}(P_v, P_{v'}) = \frac{(\lambda_1 - \lambda_2)^2}{\lambda_1\lambda_2}(1 - \langle v, v'\rangle^2) \le \frac{(\lambda_1 - \lambda_2)^2\alpha^2}{\lambda_1\lambda_2} . \tag{C.6}$$

Using Pinsker's inequality, we have

$$D_{\mathrm{TV}}(P_v, P_{v'}) \le \sqrt{\frac{D_{\mathrm{KL}}(P_v, P_{v'})}{2}} \le \alpha\sqrt{\frac{(\lambda_1 - \lambda_2)^2}{2\lambda_1\lambda_2}} . \tag{C.7}$$

Now we set

$$\alpha := \min\left(1, \max\left(\sqrt{\frac{dc_1\lambda_1\lambda_2}{2n(\lambda_1 - \lambda_2)^2}}, \frac{c_1 d}{10n\varepsilon}\sqrt{\frac{2\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)^2}}\right)\right) \tag{C.8}$$

Combining all cases, it follows from Theorem C.3.2 and $d > 10$ that there exists a constant C such that

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_\Sigma} \mathbb{E}_{S \sim P^n} \left[\sin(\hat{v}(S), v_1(\Sigma)) \right] \geq C \min \left(\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n} \right) \sqrt{\frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2}}, 1 \right) . \tag{C.9}$$

C.3.1.1 Proof of Lemma C.3.3

We first point out that Lemma C.3.3 is a special case of [202, Lemma 3.1.2]. Here is the original statement from [202].

Lemma C.3.4 ([202, Lemma 3.1.2]). *Define* $\mathbb{B}_q^p(R_q) = \left\{ \theta \in \mathbb{R}^p : \sum_{j=1}^p |\theta_j|^q \leq R_q \right\}$. *Let* $\bar{R}_q = R_q - 1 \geq 1$ *and* $p \geq 5$. *There exists a finite subset* $\Theta_\epsilon \subset \mathbb{S}_2^{p-1} \cap \mathbb{B}_q^p(R_q)$ *and an absolute constant* $c > 0$ *such that every distinct pair* $\theta_1, \theta_2 \in \Theta_\epsilon$ *satisfies*

$$\epsilon / \sqrt{2} < \|\theta_1 - \theta_2\|_2 \leq \sqrt{2}\epsilon$$

and

$$\log |\Theta_\epsilon| \geq c \left(\frac{\bar{R}_q}{\epsilon^q} \right)^{\frac{2}{2-q}} \left[\log(p-1) - \log \left(\frac{\bar{R}_q}{\epsilon^q} \right)^{\frac{2}{2-q}} \right]$$

for all $q \in [0, 1]$ *and* $\epsilon \in (0, 1]$.

Assume $d \geq 10$ and set $q = 0$ and $R_q = \frac{d}{8} + 1$. Lemma C.3.4 implies that there exists a finite subset $\mathcal{V}_\alpha \subset \mathbb{S}_2^{d-1} \cap \mathbb{B}_q^d \left(\frac{d}{8} + 1 \right)$ and an absolute constant c such that for $v \neq v' \in \mathcal{V}_\alpha$ satisfies

$$\frac{\alpha}{\sqrt{2}} \leq \|v - v'\| \leq \sqrt{2}\alpha \tag{C.10}$$

and

$$\log(|\mathcal{V}_\alpha|) \geq c \frac{d}{8} \left(\log(d-1) - \log(\frac{d}{8}) \right) = \frac{cd}{8} \log \left(8 \left(1 - \frac{1}{d} \right) \right) \geq \frac{cd}{8} \log(6.3) . \tag{C.11}$$

For completeness, we also provide a direct proof of Lemma C.3.3, following the proof strategy of Lemma C.3.4. The following lemma is a variant of classic Varshamov-Gilbert bounds that appeared in [163, Lemma 4.10]. A similar lemma can be also found in [3, Lemma 6].

Lemma C.3.5 ([163, Lemma 4.10]). *Let l be a positive integer that is at most $k/4$. Then there exists a subset $\Theta \subset \{0,1\}^k$ and absolute constant $c' > 0.233$ such that*

1. *For any $w \in \Theta$, $\|w\|_0 = l$,*
2. *For any $w \neq w' \in \Theta$, $\|w - w'\|_0 \geq l/2$,*
3. *$\log(|\Theta|) \geq c' l \log(k/l)$.*

For $d \geq 10$, let $k = d - 1$ and l be an integer between 1 and $(d-1)/4$. We will choose l later. Let Θ be such a set that satisfies the conditions in Lemma C.3.5. Now for $\alpha \in (0, 1]$, we construct $\mathcal{V}_\alpha$. Define $f : \{0,1\}^{d-1} \to \mathbb{R}^d$ as follows.

$$f(w) = \left(\sqrt{1-\alpha^2}, \frac{w\alpha}{\sqrt{l}}\right) \in \mathbb{R}^d. \tag{C.12}$$

Let

$$\mathcal{V}_\alpha := \{f(w) : w \in \Theta\} . \tag{C.13}$$

It is easy to see that

$$\|f(w)\| = \sqrt{1 - \alpha^2 + \|w\|^2\alpha^2/l} = 1 . \tag{C.14}$$

For any $v \neq v' \in \mathcal{V}_\alpha$, if $v = f(w)$ and $v' = f(w')$, we know

$$\frac{\alpha}{\sqrt{2}} \leq \|v - v'\| = \sqrt{\frac{\|w - w'\|^2\alpha^2}{l}} \leq \sqrt{2}\alpha \tag{C.15}$$

where the last inequality follows from the fact that $\|w - w'\|_0 \leq 2l$.

Note that above inequalities hold for any l between 1 and $(d-1)/4$. Let $l = (d-1)/8$. Then we have

$$\log(|\mathcal{V}_\alpha|) = \log(|\Theta|) \geq c'((d-1)/8) \log\left(\frac{d-1}{(d-1)/8}\right) \geq \frac{c'd}{10} \tag{C.16}$$

for any $d \geq 2$.

C.3.2 *Proof of Theorem 4.5.4*

We first construct an indexed set $\mathcal{V}$ and indexed distribution family $\mathcal{P}_{\mathcal{V}}$ such that $x_i x_i^\top$ satisfies A.1, A.2 and A.3 in Assumption 5. Our construction is defined as follows.

By [3, Lemma 6] , there exists a finite set $\mathcal{V} \subset \mathbb{S}_2^{d-1}$, with cardinality $|\mathcal{V}| \geq 2^d$, such that for any $v \neq v' \in \mathcal{V}$, $\|v - v'\| \geq 1/2$.

Let $f_{(0,\mathbf{I}_d)}$ denotes the density function of $\mathcal{N}(0, \mathbf{I}_d)$. Let Q_v be a uniform distribution on two point masses $\{\pm\alpha^{-\frac{1}{4}}v\}$. Let Q_0 be Gaussian distribution $\mathcal{N}(0, \mathbf{I}_d)$. For $\alpha \in (0, 1]$, we construct $P_v := (1-\alpha)Q_0 + \alpha Q_v$. It is easy to see that P_v is a distribution over $\mathbb{R}^d$ with the following density function.

$$P_v(x) = \begin{cases} \frac{\alpha}{2}, & \text{if } x = -\alpha^{-\frac{1}{4}}v \ , \\ \frac{\alpha}{2}, & \text{if } x = \alpha^{-\frac{1}{4}}v \ , \\ (1-\alpha)f_{(0,\mathbf{I}_d)}(x) & \text{otherwise} \end{cases} . \tag{C.17}$$

The mean of P_v is 0. The covariance of P_v is $\Sigma_v = (1-\alpha)\mathbf{I}_d + \sqrt{\alpha}vv^\top$. The top eigenvalue is $\lambda_1 = 1 - \alpha + \sqrt{\alpha}$, the top eigenvector is v, and the second eigenvalue is $\lambda_2 = 1 - \alpha$. And $\kappa = O(\alpha^{-1/2})$.

If $x = \alpha^{-1/4}v$, then $\|xx^\top - \Sigma_v\|_2 = O(\alpha^{-1/2})$. If $x \sim \mathcal{N}(0, \mathbf{I}_d)$, we know $\|xx^\top - \Sigma_v\|_2 = O(d)$. This implies P_v satisfies A.2 in Assumption 5 with $M = O((d + \alpha^{-1/2})\log(n))$ for n i.i.d. samples.

It is easy to see that $\|\mathbb{E}[(xx^\top - \Sigma_v)(xx^\top - \Sigma_v)^\top]\|_2 = O(d)$. This means P_v satisfies A.3 in Assumption 5 with $V = O(d)$.

By the fact that $\mathbb{E}[\langle x, u\rangle^2] = O(1)$ and $\mathbb{E}[\langle x, u\rangle^4] = O(1)$ for any unit vector u, we have $\gamma^2 = \|\mathbb{E}[(xx^\top - \Sigma_v)uu^\top(xx^\top - \Sigma_v)^\top]\|_2 = O(1)$ for any unit vector u.

Our proof technique is based on following lemma.

Lemma C.3.6 ([25, Theorem 3])**.** *Fix $\alpha \in (0, 1]$. Define $P_v = (1-\alpha)Q_0 + \alpha Q_v$ for $v \in \mathcal{V}$ such that such that $\rho(\theta(P_v), \theta(P_{v'})) \geq 2t$. Let $\hat{\theta}$ be a (ε, δ) differentially private estimator.*

Then,

$$\frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} P_v\left(\rho\left(\widehat{\theta},\theta(P_v)\right)\geq t\right)\geq \frac{(|\mathcal{V}|-1)\cdot\left(\frac{1}{2}e^{-\varepsilon\lceil n\alpha\rceil}-\delta\frac{1-e^{-\varepsilon\lceil n\alpha\rceil}}{1-e^{-\varepsilon}}\right)}{1+(|\mathcal{V}|-1)\cdot e^{-\varepsilon\lceil n\alpha\rceil}}\,. \tag{C.18}$$

Set $\rho(\theta(P_v),\theta(P_{v'}))=\sin(v,v')/\kappa$. By Lemma C.6.4, $\rho(\theta(P_v),\theta(P_{v'}))\geq\|v-v'\|/\kappa=\Omega(\sqrt{\alpha})$.

Lemma C.3.6 implies

$$\sup_{P\in\tilde{\mathcal{P}}}\mathbb{E}_{S\sim P^n}[\sin(\hat{v}(S),v_1(\Sigma))]\geq\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathbb{E}_{S\sim P_v^n}[\sin(\hat{v}(S),v_1(\Sigma_v))] \tag{C.19}$$

$$=\kappa t\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}P_v\left(\frac{\sin(\hat{v}(S),v_1(\Sigma_v))}{\kappa}\geq t\right) \tag{C.20}$$

$$\gtrsim\kappa t\frac{(2^d-1)\cdot\left(\frac{1}{2}e^{-\varepsilon\lceil n\alpha\rceil}-\frac{\delta}{1-e^{-\varepsilon}}\right)}{1+(2^d-1)e^{-\varepsilon\lceil n\alpha\rceil}}\,, \tag{C.21}$$

For $d\geq 2$, we know $2^d-1\geq e^{d/2}$. We choose

$$\alpha=\min\left\{\frac{1}{n\varepsilon}\left(\frac{d}{2}-\varepsilon\right),\frac{1}{n\varepsilon}\log\left(\frac{1-e^{-\varepsilon}}{4\delta e^{\varepsilon}}\right),1\right\}\,. \tag{C.22}$$

This implies

$$\frac{1}{2}e^{-\varepsilon\lceil n\alpha\rceil}-\frac{\delta}{1-e^{-\varepsilon}}\geq\frac{1}{4}e^{-\varepsilon(n\alpha+1)}>0\,. \tag{C.23}$$

So we have there exists a constant C such that

$$\inf_{\hat{v}}\sup_{P\in\tilde{\mathcal{P}}}\mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S),v_1(\Sigma))\right]\geq C\kappa\sqrt{\alpha}\frac{\frac{1}{4}e^{d/2}e^{-\varepsilon(n\alpha+1)}}{1+e^{d/2}e^{-\varepsilon(n\alpha+1)}} \tag{C.24}$$

$$\gtrsim\kappa\min\left(1,\sqrt{\frac{d\wedge\log\left((1-e^{-\varepsilon})/\delta\right)}{n\varepsilon}}\right)\,. \tag{C.25}$$

C.3.3 Proof of Theorem C.3.1

Similar to the proof of Theorem 4.5.3, we use DP Fano's method in Theorem C.3.2. It suffices to construct an indexed set $\mathcal{V}$ and indexed distribution family $\mathcal{P}_{\mathcal{V}}$ such that $x_ix_i^\top$ satisfies A.1, A.2 and A.3 in Assumption 5. Our construction is defined as follows.

Let $\lambda_1 > \lambda_2 > 0$. By Lemma C.3.3, there exists a finite set $\mathcal{V}_\alpha \subset \mathbb{S}_2^{d-1}$, with cardinality $|\mathcal{V}_\alpha| = 2^{\Omega(d)}$, such that for any $v \neq v' \in \mathcal{V}_\alpha$, $\alpha/\sqrt{2} \leq \|v - v'\| \leq \sqrt{2}$, where $\alpha := \sqrt{\lambda_2/\lambda_1}$.

Let $f_{(0,S)}$ denotes the density function of $\mathcal{N}(0, S)$. We construct P_v over $\mathbb{R}^d$ for $v \in \mathcal{V}_\alpha$ with the following density function.

$$P_v(x) = \begin{cases} \frac{1-\lambda_2/\lambda_1}{2d}, & \text{if } x = -\sqrt{d\lambda_1}v\ , \\ \frac{1-\lambda_2/\lambda_1}{2d}, & \text{if } x = \sqrt{d\lambda_1}v\ , \\ 1 - \frac{1-\lambda_2/\lambda_1}{d}\ f_{(0, \frac{\lambda_2}{1-\frac{1-\lambda_2/\lambda_1}{d}}\mathbf{I}_d)}(x) & \text{otherwise} \end{cases} . \tag{C.26}$$

The mean of P_v is 0. The covariance of P_v is $\Sigma_v := (\lambda_1 - \lambda_2)vv^\top + \lambda_2\mathbf{I}_d$. It is easy to see that the top eigenvalue is λ_1, the top eigenvector is v, and the second eigenvalue is λ_2.

If $x = \sqrt{d\lambda_1}v$, then $\|xx^\top - \Sigma_v\|_2 = \|(d\lambda_1 - \lambda_1 + \lambda_2) - \lambda_2\mathbf{I}_d\|_2 = O(d\lambda_1)$. If $x \sim \mathcal{N}(0, \frac{\lambda_2}{1-\frac{1-\lambda_2/\lambda_1}{d}}\mathbf{I}_d)$, by the fact that $\frac{\lambda_2}{1-\frac{1-\lambda_2/\lambda_1}{d}} \leq \lambda_1$, we know $\|xx^\top - \Sigma_v\|_2 \leq O(d\lambda_1)$. This implies P_v satisfies A.2 in Assumption 5 with $M = O(d\log(n))$ for n i.i.d. samples.

Similarly, $\|\mathbb{E}[(xx^\top - \Sigma_v)(xx^\top - \Sigma_v)^\top]\|_2 \leq \|d(\lambda_1^2 - \lambda_1\lambda_2)vv^\top + d\lambda_2\lambda_1 + 3\Sigma_v\Sigma_v^\top\|_2 = O(d\lambda_1^2)$. This means P_v satisfies A.3 in Assumption 5 with $V = O(d)$.

For $v \neq v' \in \mathcal{V}_\alpha$, we have $D_{\mathrm{TV}}(P_v, P_{v'}) = (1 - \lambda_2/\lambda_1)/d$. By Lemma C.6.4, $\sin(v, v') \geq \|v - v'\|/\sqrt{2} \geq (\sqrt{\lambda_2/\lambda_1})/2$.

By Theorem C.3.2, there exists a constant C such that

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_\Sigma} \mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S), v_1(\Sigma))\right] \geq C \min\left(\sqrt{\frac{\lambda_2}{\lambda_1}}, \frac{d^2}{n\varepsilon}\sqrt{\frac{\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)^2}}\right) . \tag{C.27}$$

C.4 The analysis of Private Oja's Algorithm

We analyze Private Oja's Algorithm in Algorithm 11.

C.4.1 Proof of privacy in Lemma 4.3.1

We use following Theorem C.4.1 to prove our privacy guarantees.

Theorem C.4.1 (Privacy amplification by shuffling [86, Theorem 3.8])**.** *For any domain $\mathcal{D}$, let $\mathcal{R}^{(i)} : \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(i-1)} \times \mathcal{D} \to \mathcal{S}^{(i)}$ for $i \in [n]$ (where $\mathcal{S}^{(i)}$ is the range space of $\mathcal{R}^{(i)}$) be a sequence of algorithms such that $\mathcal{R}^{(i)}(z_{1:i-1}, \cdot)$ is an $(\varepsilon_0, \delta_0)$-DP local randomizer for all values of auxiliary inputs $z_{1:i-1} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(i-1)}$. Let $\mathcal{A}_S : \mathcal{D}^n \to \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(n)}$ be the algorithm that given a dataset $x_{1:n} \in \mathcal{D}^n$, samples a uniform random permutation π over $[n]$, then sequentially computes $z_i = \mathcal{R}^{(i)}(z_{1:i-1}, x_{\pi(i)})$ for $i \in [n]$ and outputs $z_{1:n}$. Then for any $\delta \in [0, 1]$ such that $\varepsilon_0 \le \log\left(\frac{n}{16\log(2/\delta)}\right)$, $\mathcal{A}_s$ is $(\varepsilon, \delta + O(e^{\varepsilon}\delta_0 n))$-DP, where*

$$\varepsilon = O\left((1 - e^{-\varepsilon_0})\left(\frac{\sqrt{e^{\varepsilon_0}\log(1/\delta)}}{\sqrt{n}} + \frac{e^{\varepsilon_0}}{n}\right)\right) . \tag{C.28}$$

Let $\mathcal{R}^{(t)}(w_{t-1}, A_{\pi(t)}) := w_t$. Let $\varepsilon_0 = \frac{\sqrt{2\log(1.25/\delta_0)}}{\alpha}$. We show $\mathcal{R}^{(t)}(w_{t-1}, \cdot)$ is an $(\varepsilon_0, \delta_0)$-DP local randomizer.

If there is no noise in each update step, the update rule is

$$w_t' \leftarrow w_{t-1} + \eta_t \mathrm{clip}_\beta\left(A_t w_{t-1}\right) , \tag{C.29}$$

$$w_t \leftarrow w_{t-1}/\|w_{t-1}\| \tag{C.30}$$

The sensitivity of w_t' is $2\beta\eta_t$ with respect to a difference in A_t. By Gaussian mechanism in Lemma D.2.1 and post processing property of local differential privacy, we know w_t is $(\varepsilon_0, \delta_0)$-DP local randomizer.

Assume that $\varepsilon_0 = \frac{\sqrt{2\log(1.25/\delta_0)}}{\alpha} \le \frac{1}{2}$. By Theorem C.4.1, for $\hat{\delta} \in [0, 1]$ such that $\varepsilon_0 \le$

$\log\left(\frac{n}{16\log(2/\hat{\delta})}\right)$, Algorithm 11 is $(\hat{\varepsilon}, \hat{\delta} + O(e^{\hat{\varepsilon}}\delta_0 n))$-DP and for some constant $c_1 > 0$,

$$\hat{\varepsilon} \le c_1 \left((1 - e^{-\varepsilon_0}) \left(\frac{\sqrt{e^{\varepsilon_0}\log(1/\hat{\delta})}}{\sqrt{n}} + \frac{e^{\varepsilon_0}}{n} \right) \right) \tag{C.31}$$

$$\le c_1 \left((e^{0.5} - e^{-0.5\varepsilon_0}) \frac{\sqrt{\log(1/\hat{\delta})}}{\sqrt{n}} + \frac{e^{\varepsilon_0} - 1}{n} \right) \tag{C.32}$$

$$\le c_1 \left(((1 + \varepsilon_0) - (1 - \varepsilon_0/2)) \frac{\sqrt{\log(1/\hat{\delta})}}{\sqrt{n}} + \frac{1 + 2\varepsilon_0 - 1}{n} \right) \tag{C.33}$$

$$= c_1\varepsilon_0 \left(\frac{1}{2}\sqrt{\frac{\log(1/\hat{\delta})}{n}} + \frac{2}{n} \right) \tag{C.34}$$

$$\le c_2 \frac{\sqrt{\log(1/\delta_0)}}{\alpha} \sqrt{\frac{\log(1/\hat{\delta})}{n}}\,, \tag{C.35}$$

for some absolute constant $c_2 > 0$.

Set $\hat{\delta} = \delta/2$, $\delta_0 = c_3\delta/(e^{\hat{\varepsilon}}n)$ for some $c_3 > 0$ and $\alpha = C'\log(n/\delta)/(\varepsilon\sqrt{n})$. We have

$$\hat{\varepsilon} \le c_2 \frac{\sqrt{\log(e^{\hat{\varepsilon}}n/(c_3\delta))}}{\alpha} \sqrt{\frac{\log(2/\delta)}{n}} \tag{C.36}$$

$$= \frac{\sqrt{\log(e^{\hat{\varepsilon}}n/(c_3\delta))\log(2/\delta)}}{C'\log(n/\delta)} \cdot \varepsilon. \tag{C.37}$$

For any $\varepsilon \le 1$, by Eq. (C.37), there exists some sufficiently large $C' > 0$ such that $\hat{\varepsilon} \le \varepsilon$. Recall that we assume $\varepsilon_0 = \frac{\sqrt{2\log(1.25/\delta_0)}}{\alpha} \le \frac{1}{2}$. This means $\varepsilon = O(\sqrt{\frac{\log(n/\delta)}{n}})$.

C.4.2 Proof of clipping in Lemma 4.3.2

Let $z_t = A_t w_{t-1}$. Let $\mu_t := \mathbb{E}[z_t] = \Sigma w_{t-1}$. By Lemma 4.2.1, we know for any $\|v\| = 1$, with probability $1 - \zeta$,

$$|v^\top(z_t - \mu_t)| \le K\gamma\lambda_1 \log^a(2/\zeta)\,. \tag{C.38}$$

Applying union bound over all basis vectors $v \in \{e_1, \ldots, e_d\}$ and all samples, we know

with probability $1-\zeta$, for all $j \in [d]$ and $t \in [n]$

$$|z_{t,j}| \le K\gamma\lambda_1 \log^a(2nd/\zeta) + \lambda_1 \,. \tag{C.39}$$

This implies that with probability $1-\zeta$, for all $t \in [n]$, we have

$$\|z_t\| \le (K\gamma \log^a(2nd/\zeta) + 1)\lambda_1\sqrt{d} \,. \tag{C.40}$$

C.4.3 Proof of utility in Theorem 4.3.3

Lemma 4.3.2 implies that with probability $1-O(\zeta)$, Algorithm 11 does not have any clipping. Under this event, the update rule becomes

$$w_t' \leftarrow w_{t-1} + \eta_t \left(A_t + 2\alpha\beta G_t\right) w_{t-1} \tag{C.41}$$

$$w_t \leftarrow w_t'/\|w_t'\| \,, \tag{C.42}$$

where $\beta = (K\gamma \log^a(nd/\zeta) + 1)\lambda_1\sqrt{d}$ and each entry in $G_t \in \mathbb{R}^{d\times d}$ is i.i.d. sampled from standard Gaussian $\mathcal{N}(0,1)$. This follows form the fact that $\|w_{t-1}\| = 1$ and $G_t w_{t-1} \sim \mathcal{N}(0, \mathbf{I}_d)$.

Let $B_t = A_t + 2\alpha\beta G_t$. We show B_t satisfies the three conditions in Theorem 4.2.2 ([119, Theorem 4.12]). It is easy to see that $\mathbb{E}[B_t] = \Sigma$ from Assumption A.1. Next, we show upper bound of $\max\left\{\left\|\mathbb{E}\left[(B_t-\Sigma)(B_t-\Sigma)^\top\right]\right\|_2, \left\|\mathbb{E}\left[(B_t-\Sigma)^\top(B_t-\Sigma)\right]\right\|_2\right\}$. We have

$$\begin{aligned}
&\left\|\mathbb{E}\left[(B_t-\Sigma)(B_t-\Sigma)^\top\right]\right\|_2 \\
=\ &\left\|\mathbb{E}[(A_t+2\alpha\beta G_t-\Sigma)(A_t+2\alpha\beta G_t-\Sigma)^\top]\right\|_2 \\
\le\ &\left\|\mathbb{E}[(A_t-\Sigma)(A_t-\Sigma)^\top]\right\|_2 + 4\alpha^2\beta^2\|\mathbb{E}[G_tG_t^\top]\|_2 \\
\le\ &V\lambda_1^2 + 4\alpha^2\beta^2 C_2 d \,,
\end{aligned} \tag{C.43}$$

where the last inequality follows from Lemma C.6.3 and $C_2 > 0$ is an absolute constant. Let $\widetilde{V} := V\lambda_1^2 + 4\alpha^2\beta^2 C_2 d$. Similarly, we can show that $\left\|\mathbb{E}\left[(B_t-\Sigma)^\top(B_t-\Sigma)\right]\right\|_2 \le \widetilde{V}$.

By Lemma C.6.2, we know with probability $1-\zeta$, for all $t \in [T]$,

$$\begin{aligned}
&\|B_t - \Sigma\|_2 \\
=&\|A_t + 2\alpha\beta G_t - \Sigma\|_2 \\
\leq&\|A_t - \Sigma\|_2 + 2\alpha\beta\|G_t\|_2 \\
\leq& M\lambda_1 + 2C_3\alpha\beta\left(\sqrt{d} + \sqrt{\log(n/\zeta)}\right) .
\end{aligned}$$

Let $\widetilde{M} := M\lambda_1 + 2C_3\alpha\beta\left(\sqrt{d} + \sqrt{\log(n/\zeta)}\right)$.

Under the event that $\|B_t - \Sigma\|_2 \leq \widetilde{M}$ for all $t \in [n]$, we apply Theorem 4.2.2 with a learning rate $\eta_t = \frac{h}{(\lambda_1-\lambda_2)(\xi+t)}$ where

$$\xi = 20\max\left(\frac{\widetilde{M}h}{(\lambda_1-\lambda_2)}, \frac{\left(\widetilde{V} + \lambda_1^2\right)h^2}{(\lambda_1-\lambda_2)^2\log(1+\frac{\zeta}{100})}\right) . \tag{C.44}$$

Then Theorem 4.2.2 implies that with probability $1-\zeta$,

$$\sin^2(w_n, v_1) \leq \frac{C\log(1/\zeta)}{\zeta^2}\left(d\left(\frac{\xi}{n}\right)^{2h} + \frac{h^2\widetilde{V}}{(2h-1)(\lambda_1-\lambda_2)^2 n}\right) , \tag{C.45}$$

for some positive constant C.

Set $\alpha = \frac{C'\log(n/\delta)}{\varepsilon\sqrt{n}}$, the above bound implies

$$\sin^2(w_n, v_1) \leq \frac{C\log(1/\zeta)}{\zeta^2}\left(\frac{h^2V\lambda_1^2}{(2h-1)(\lambda_1-\lambda_2)^2 n} + \frac{(K\gamma\log^a(nd/\zeta)+1)^2\lambda_1^2\log^2(n/\delta)d^2h^2}{(2h-1)(\lambda_1-\lambda_2)^2\varepsilon^2n^2} + d\left(\tilde{\xi}\right)^h\right) , \tag{C.46}$$

where $\tilde{\xi} = (\xi/n)^2$, and

$$\begin{aligned}
\tilde{\xi} := \max\Bigg(&\frac{M^2\lambda_1^2h^2}{(\lambda_1-\lambda_2)^2n^2} + \frac{(K\gamma\log^a(nd/\zeta)+1)^2\lambda_1^2\log^3(n/\delta)h^2d^2}{(\lambda_1-\lambda_2)^2\varepsilon^2n^3}, \\
&\frac{V^2\lambda_1^4h^4}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})n^2} + \frac{(K\gamma\log^a(nd/\zeta)+1)^4\lambda_1^4\log^4(n/\delta)h^4d^4}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})\varepsilon^4n^4} \\
&+\frac{\lambda_1^4h^4}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})n^2}\Bigg) .
\end{aligned} \tag{C.47}$$

For $\zeta = O(1)$ and $K = O(1)$, selecting $h = c \log n$, and assuming

$$n \geq C \left(\frac{M\lambda_1 \log(n)}{\lambda_1 - \lambda_2} + \frac{(K\gamma \log^a(nd/\zeta) + 1)^{2/3} \lambda_1^{2/3} \log(n/\delta) \log^{2/3}(n) d^{2/3}}{(\lambda_1 - \lambda_2)^{2/3} \varepsilon^{2/3}} + \frac{V\lambda_1^2 (\log(n))^2}{(\lambda_1 - \lambda_2)^2} + \frac{(K\gamma \log^a(nd/\zeta) + 1)\lambda_1 \log(n/\delta) \log(n) d}{(\lambda_1 - \lambda_2)\varepsilon} + \frac{\lambda_1^2 \log^2(n)}{(\lambda_1 - \lambda_2)^2} \right) , \tag{C.48}$$

with large enough positive constants c, and C, we have $\tilde{\xi} \leq 1$ and $d\tilde{\xi}^\alpha \leq 1/n^2$. Hence it is sufficient to have

$$n = \tilde{O}\Big(\frac{\lambda_1^2}{(\lambda_1 - \lambda_2)^2} + \frac{M\lambda_1}{\lambda_1 - \lambda_2} + \frac{V\lambda_1^2}{(\lambda_1 - \lambda_2)^2} + \frac{d\,(\gamma + 1)\lambda_1\, \log(1/\delta)}{(\lambda_1 - \lambda_2)\varepsilon} \Big) ,$$

with a large enough constant.

C.5 The analysis of DP-PCA

We provides the proofs for Theorem 4.5.1, Theorem 4.6.1, and Lemma 4.6.2 that guarantees the privacy and utility of DP-PCA.

C.5.1 Proof of Theorem 4.5.1 on the privacy and utility of DP-PCA

From Theorem 4.6.1 we know that Alg. 21 returns $\hat{\Lambda}$ satisfying $2\hat{\Lambda} \geq \lambda_1^2 \|H_u\|_2$ with high probability. Then, from Lemma 4.6.2, we know that with high probability Alg 22 returns an unbiased estimate of the gradient mean with added Gaussian noise. Under this case, the update rule becomes

$$w_t' \leftarrow w_{t-1} + \eta_t \left(\frac{1}{B} \sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t \right) w_{t-1} \tag{C.49}$$

$$w_t \leftarrow w_t' / \|w_t'\| , \tag{C.50}$$

where $\beta_t = \frac{8K\sqrt{2\hat{\Lambda}_t} \log^a(Bd/\zeta)\sqrt{2d \log(2.5/\delta)}}{\varepsilon B}$, $\hat{\Lambda}_t$ denote the estimated eigenvalue of covariance of the gradients at t-th iteration, and each entry in $G_t \in \mathbb{R}^{d \times d}$ is i.i.d. sampled from standard Gaussian $\mathcal{N}(0, 1)$. This follows form the fact that $\|w_{t-1}\| = 1$ and $G_t w_{t-1} \sim \mathcal{N}(0, \mathbf{I}_d)$.

Let $\beta := \frac{16K\gamma\lambda_1 \log^a(Bd/\zeta)\sqrt{2d\log(2.5/\delta)}}{\varepsilon B}$ such that $\beta \geq \beta_t$, which follows from the fact that $\hat{\Lambda} \leq \sqrt{2}\lambda_1^2\|H_u\|_2 \leq \sqrt{2}\lambda_1^2\gamma^2$ (Theorem 4.6.1 and Assumption A.4). Let $B_t = (1/B)\sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t$. We show B_t satisfies the three conditions in Theorem 4.2.2 ([119, Theorem 4.12]). It is easy to see that $\mathbb{E}[B_t] = \Sigma$ from Assumption A.1. Next, we show upper bound of $\max\left\{\left\|\mathbb{E}\left[(B_t-\Sigma)(B_t-\Sigma)^\top\right]\right\|_2, \left\|\mathbb{E}\left[(B_t-\Sigma)^\top(B_t-\Sigma)\right]\right\|_2\right\}$. We have

$$
\begin{aligned}
&\left\|\mathbb{E}\left[(B_t-\Sigma)(B_t-\Sigma)^\top\right]\right\|_2 \\
=& \left\|\mathbb{E}[(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t - \Sigma)(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t - \Sigma)^\top]\right\|_2 \\
\leq& \left\|\mathbb{E}[(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} - \Sigma)(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} - \Sigma)^\top]\right\|_2 + \beta^2\|\mathbb{E}[G_t G_t^\top]\|_2 \\
=& V\lambda_1^2/B + \beta^2\|\mathbb{E}[G_t G_t^\top]\|_2 \\
\leq& V\lambda_1^2/B + \beta^2 C_2 d\,, \qquad \text{(C.51)}
\end{aligned}
$$

where the last inequality follows from Lemma C.6.3 and $C_2 > 0$ is an absolute constant. Let $\widetilde{V} := V\lambda_1^2/B + \beta^2 C_2 d$. Similarly, we can show that $\left\|\mathbb{E}\left[(B_t-\Sigma)^\top(B_t-\Sigma)\right]\right\|_2 \leq \widetilde{V}$. By Lemma C.6.5 and Lemma C.6.2, we know with probability $1-\zeta$, for all $t \in [T]$,

$$
\begin{aligned}
&\|B_t - \Sigma\|_2 \\
=& \left\|\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t - \Sigma\right\|_2 \\
\leq& C_3\left(\frac{M\lambda_1\log(dT/\zeta)}{B} + \sqrt{\frac{V\lambda_1^2\log(dT/\zeta)}{B}} + \beta\left(\sqrt{d} + \sqrt{\log(T/\zeta)}\right)\right)\,.
\end{aligned}
$$

Let $\widetilde{M} := C_3\left(\frac{M\lambda_1\log(dT/\zeta)}{B} + \sqrt{\frac{V\lambda_1^2\log(dT/\zeta)}{B}} + \beta\left(\sqrt{d} + \sqrt{\log(T/\zeta)}\right)\right)$. Under the event that $\|B_t - \Sigma\|_2 \leq \widetilde{M}$ for all $t \in [T]$, we apply Theorem 4.2.2 with a learning rate $\eta_t = \frac{\alpha}{(\lambda_1-\lambda_2)(\xi+t)}$ where

$$
\xi = 20\max\left(\frac{\widetilde{M}\alpha}{(\lambda_1-\lambda_2)}, \frac{\left(\widetilde{V}+\lambda_1^2\right)\alpha^2}{(\lambda_1-\lambda_2)^2\log(1+\frac{\zeta}{100})}\right)\,. \qquad \text{(C.52)}
$$

Then Theorem 4.2.2 implies that with probability $1-\zeta$,

$$\sin^2\left(w_T, v_1\right) \leq \frac{C\log(1/\zeta)}{\zeta^2}\left(d\left(\frac{\xi}{T}\right)^{2\alpha} + \frac{\alpha^2\widetilde{V}}{(2\alpha-1)\left(\lambda_1-\lambda_2\right)^2 T}\right), \tag{C.53}$$

for some positive constant C. Using $n = BT$ and Eq. (C.51), the above bound implies

$$\sin^2\left(w_T, v_1\right) \leq \frac{C\log(1/\zeta)}{\zeta^2}\left(\frac{\alpha^2 V\lambda_1^2}{(2\alpha-1)\left(\lambda_1-\lambda_2\right)^2 n} + \frac{K^2\gamma^2\lambda_1^2\log^{2a}(nd/(T\zeta))\log(1/\delta)d^2\alpha^2 T}{(2\alpha-1)(\lambda_1-\lambda_2)^2\varepsilon^2 n^2} + d\left(\tilde{\xi}\right)^{\alpha}\right). \tag{C.54}$$

where $\tilde{\xi} = (\xi/T)^2$, and

$$\begin{aligned}\tilde{\xi} := \max\Bigg(&\frac{M^2\lambda_1^2\alpha^2\log^2(dT/\zeta)}{(\lambda_1-\lambda_2)^2 n^2} + \frac{V\lambda_1^2\log(dT/\zeta)\alpha^2}{(\lambda_1-\lambda_2)^2 nT} + \frac{K^2\gamma^2\lambda_1^2\log^{2a}(nd/(T\zeta))\log(1/\delta)\log(T/\zeta)\alpha^2 d^2}{(\lambda_1-\lambda_2)^2\varepsilon^2 n^2},\\ &\frac{V^2\lambda_1^4\alpha^4}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})n^2} + \frac{K^4\gamma^4\lambda_1^4\log^{4a}(nd/(T\zeta))\log^2(1/\delta)\alpha^4 d^4 T^2}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})\varepsilon^4 n^4}\\ &+\frac{\lambda_1^4\alpha^4}{(\lambda_1-\lambda_2)^4\log^2(1+\frac{\zeta}{100})T^2}\Bigg).\end{aligned} \tag{C.55}$$

For $\zeta = O(1)$ and $K = O(1)$, selecting $\alpha = c\log n$, $T = c'(\log n)^2$, and assuming $\log n \geq \lambda_1^2/(\lambda_1-\lambda_2)^2$ and

$$\begin{aligned}n \geq C\Bigg(&\frac{M\lambda_1\log(n)\log(d\log(n))}{\lambda_1-\lambda_2} + \frac{\sqrt{V\lambda_1^2\log(dT)}}{(\lambda_1-\lambda_2)} + \frac{\gamma\lambda_1\log^a(nd/\log(n))\sqrt{\log(1/\delta)\log(\log(n))}\log(n)d}{(\lambda_1-\lambda_2)\varepsilon}\\ &+\frac{V\lambda_1^2(\log(n))^2}{(\lambda_1-\lambda_2)^2} + \frac{\gamma\lambda_1\log^a(nd/\log(n))\sqrt{\log(1/\delta)}(\log(n))^2 d}{(\lambda_1-\lambda_2)\varepsilon}\Bigg),\end{aligned} \tag{C.56}$$

with large enough positive constants c, c', and C, we have $\tilde{\xi} \leq 1$ and $d\tilde{\xi}^{\alpha} \leq 1/n^2$. Hence it is sufficient to have

$$n = \tilde{O}\Big(\exp(\lambda_1^2/(\lambda_1-\lambda_2)^2) + \frac{M\lambda_1}{\lambda_1-\lambda_2} + \frac{V\lambda_1^2}{(\lambda_1-\lambda_2)^2} + \frac{d\,\gamma\,\lambda_1\sqrt{\log(1/\delta)}}{(\lambda_1-\lambda_2)\varepsilon}\Big),$$

with a large enough constant.

C.5.2 Algorithm and proof of Theorem 4.6.1 on top eigenvalue estimation

Algorithm 21: Private Top Eigenvalue Estimation

Input: $S = \{g_i\}_{i=1}^{B}$, (ε, δ)-DP, failure probability ζ

Let $\tilde{g}_i \leftarrow g_{2i} - g_{2i-1}$ for $i \in 1, 2, \ldots, \lfloor B/2 \rfloor$. Let $\tilde{S} = \{\tilde{g}_i\}_{i=1}^{\lfloor B/2 \rfloor}$

Partition $\tilde{S}$ into $k = C_1 \log(1/(\delta\zeta))/\varepsilon$ subsets and denote each dataset as $G_j \in \mathbb{R}^{d \times b}$, where each dataset is of size $b = \lfloor B/2k \rfloor$

Let $\lambda_1^{(j)}$ be the top eigenvalue of $(1/b)G_j G_j^\top$ for $\forall j \in [k]$

Partition $[0, \infty)$ into

$\Omega \leftarrow \left\{\ldots, \left[2^{-2/4}, 2^{-1/4}\right) \left[2^{-1/4}, 1\right) \left[1, 2^{1/4}\right), \left[2^{1/4}, 2^{2/4}\right), \ldots\right\} \cup \{[0, 0]\}$

Run (ε, δ)-DP histogram learner of Lemma A.2.1 on $\{\lambda_1^{(j)}\}_{j=1}^{k}$ over Ω

if all the bins are empty **then** Return $\perp$

Let $[l, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram

Return $\hat{\Lambda} = l$

Taking the difference ensures that $\tilde{g}_i$ is zero mean, such that we can directly use the top eigenvalue of $(1/b)G_j G_j^\top$ for $j \in [k]$. We compute a histogram over those k top eigenvalues. This histogram is privatized by adding noise only to the occupied bins and thresholding small entries of the histogram to be zero. The choice $k = \Omega(\log(1/\zeta)/\varepsilon)$ ensures that the most occupied bin does not change after adding the DP noise to the histograms, and $k = \Omega(\log(1/\delta)/\varepsilon)$ is necessary for handling unbounded number of bins. We emphasize that we do not require any upper and lower bounds on the eigenvalue, thanks to the private histogram learner from [38, 140] that gracefully handles unbounded number of bins.

The privacy guarantee follows from the privacy guarantee of the histogram learner provided in Lemma A.2.1.

For utility analysis, we follow the analysis of [133, Theorem 3.1]. The main difference is that we prove a smaller sample complexity sine we only need the top eigenvalue, and we analyze a more general distribution family. The random vector $\tilde{g}_i$ is zero mean with covariance $2\lambda_1^2 H_u \in \mathbb{R}^{d \times d}$, where $H_u = \mathbb{E}[(A_i - \Sigma)uu^\top (A_i - \Sigma)^\top]/\lambda_1^2$, and $\tilde{g}_i$ satisfies with probability

$1-\zeta$,

$$|\langle \tilde{g}_i, v\rangle| \;\le\; 2K\lambda_1\sqrt{\|H_u\|_2}\log^a(1/\zeta)\;, \tag{C.57}$$

which follows from Lemma 4.2.1. Applying union bound over all basis vectors $v \in \{e_1,\ldots,e_d\}$, we know with probability $1-\zeta$,

$$\|\tilde{g}_i\| \;\le\; 2K\lambda_1\sqrt{d\|H_u\|_2}\log^a(d/\zeta)\;.$$

We next show that conditioned on event $\mathcal{E} = \{\|\tilde{g}_i\| \le 2K\lambda_1\sqrt{d\|H_u\|_2}\log^a(d/\zeta)\}$, the covariance $\mathbb{E}[\tilde{g}_i\tilde{g}_i^\top|\mathcal{E}]$ is close to the true covariance $\mathbb{E}[\tilde{g}_i\tilde{g}_i^\top] = 2\lambda_1^2 H_u$. Note that

$$\begin{aligned}\mathbb{E}[\tilde{g}_i\tilde{g}_i^\top|\mathcal{E}] \;&=\; \frac{\mathbb{E}[\tilde{g}_i\tilde{g}_i^\top\mathbb{I}\{\|\tilde{g}_i\| \le 2K\lambda_1\sqrt{d\|H_u\|_2}\log^a(d/\zeta)\}]}{\mathbb{P}(\mathcal{E})}\\ &\preceq\; \frac{\mathbb{E}[\tilde{g}_i\tilde{g}_i^\top]}{\mathbb{P}(\mathcal{E})} \;\preceq\; \frac{2\lambda_1^2 H_u}{1-\zeta}\;.\end{aligned} \tag{C.58}$$

We next show the empirical covariance $(1/b)\sum_{i=1}^b \tilde{g}_i\tilde{g}_i^\top$ concentrates around $2\lambda_1^2 H_u$. First of all, using union bound on Eq. (C.57), we have with probability $1-\zeta$, for all $i\in[b]$ and $j\in[d]$,

$$|\tilde{g}_{ij}| \le 2K\lambda_1\sqrt{\|H_u\|_2}\log^a(bd/\zeta)\;.$$

Under the event that $|\tilde{g}_{ij}| \le 2K\lambda_1\sqrt{\|H_u\|_2}\log^a(nd/\zeta)$ for all $i\in[b]$, $j\in[d]$, [203, Corrollary 6.20] together with Eq. (C.58) implies

$$\mathbb{P}\left(\left\|\frac{1}{b}\sum_{i=1}^b \tilde{g}_i\tilde{g}_i^\top - 2\lambda_1^2 H_u\right\|_2 \ge \alpha\right) \le 2d\exp\left(-\frac{b\alpha^2}{8K^2\lambda_1^2\|H_u\|_2\log^{2a}(\frac{bd}{\zeta})d(2\lambda_1^2\|H_u\|_2/(1-\zeta)+\alpha)}\right)\;.$$

The above bound implies that with probability $1-\zeta$,

$$\left\|\frac{1}{b}\sum_{i=1}^b \tilde{g}_i\tilde{g}_i^\top - \lambda_1^2 2H_u\right\|_2 \;=\; O\Big(K\lambda_1^2\|H_u\|_2\log^a(bd/\zeta)\sqrt{\frac{d\log(d/\zeta)}{b}} + K^2\lambda_1^2\|H_u\|_2\log^{2a}(bd/\zeta)\frac{d\log(d/\zeta)}{b}\Big)\;.$$

This means if $b = \Omega(K^2 d\log(dk/\zeta)\log^{2a}(bdk/\zeta))$, then with probability $1-\zeta$, for all $j\in[k]$, $(1-2^{1/8})\lambda_1^2\|H_u\|_2 \le \lambda_1^{(j)} \le (1+2^{1/8})\lambda_1^2\|H_u\|_2$. This means all of $\lambda_1^{(j)}$ must be within $2^{1/4}\lambda_1^2\|H_u\|_2$ interval. Thus, at most two consecutive buckets are filled with $\lambda_1^{(j)}$. By private histogram from Lemma A.2.1, if $k \ge \log(1/(\delta\zeta))/\varepsilon$, one of those two bins are released. The resulting total multiplicative error is bounded by $2^{1/2}$.

C.5.3 Algorithm and proof of Lemma 4.6.2 on DP mean estimation

Algorithm 22: Private Mean Estimation [140, 130]

Input: $S = \{g_i\}_{i=1}^B$, (ε, δ), target error α, failure probability ζ, approximate top eigenvalue $\hat{\Lambda}$

Let $\tau = 2^{1/4} K \sqrt{\hat{\Lambda}} \log^a(25)$.

for *j=1, 2, . . . , d* **do**

  Run $(\frac{\varepsilon}{4\sqrt{2d\log(4/\delta)}}, \frac{\delta}{4d})$-DP histogram learner of Lemma A.2.1 on $\{g_{ij}\}_{i\in[B]}$ over $\Omega = \{\cdots, (-2\tau, -\tau], (-\tau, 0], (0, \tau], (\tau, 2\tau], (2\tau, 3\tau] \cdots\}$.

  Let $[l, h]$ be the bucket that contains maximum number of points in the private histogram

  $\bar{g}_j \leftarrow l$

  Truncate the j-th coordinate of gradient $\{g_i\}_{i\in[B]}$ by $[\bar{g}_j - 3K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta), \bar{g}_j + 3K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta)]$.

  Let $\tilde{g}_i$ be the truncated version of g_i.

Compute empirical mean of truncated gradients $\tilde{\mu} = (1/B)\sum_{i=1}^B \tilde{g}_i$ and add Gaussian noise: $\hat{\mu} = \tilde{\mu} + \mathcal{N}\left(0, \left(\frac{12K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta)\sqrt{2d\log(2.5/\delta)}}{\varepsilon B}\right)^2 \mathbf{I}_d\right)$

Return $\hat{\mu}$

The histogram learner is called d times, each with $(\varepsilon/(4\sqrt{2d\log(4/\delta)}), \delta/(4d))$-DP guarantee, and the end-to-end privacy guarantee is $(\varepsilon/2, \delta/2)$ from Lemma 2.3.4 for $\varepsilon \in (0, 0.9)$. The sensitivity of the clipped mean estimate is $\Delta = \sqrt{d}6K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta)$. Gaussian mechanism with covariance $(2\Delta\sqrt{2\log(2.5/\delta)}/\varepsilon)^2\mathbf{I}_d$ satisfy $(\varepsilon/2, \delta/2)$-DP from Lemma D.2.1 for $\varepsilon \in (0, 1)$. Putting these two together, with serial composition of Lemma C.2.2, we get the desired privacy guarantee.

The proof of utility follows similarly as [160, Lemma D.2]. Let $I_l = (l\sqrt{\hat{\Lambda}}, (l+1)\sqrt{\hat{\Lambda}}]$. Denote the population probability of j-th coordinate at I_l as $h_{j,l} = \mathbb{P}(g_{ij} \in I_l)$. Denote the empirical probability as $\hat{h}_{j,l} = \frac{1}{B}\sum_{i=1}^B \mathbb{I}(g_{ij} \in I_l)$. Denote the private empirical probability being released as $\tilde{h}_{j,l}$.

Fix $j \in [d]$. Let I_k be the bin that contains the μ_j. Then we know $[\mu_j - K\lambda_1\sqrt{\|H_u\|_2}\log^a(25), \mu_j + K\lambda_1\sqrt{\|H_u\|_2}\log^a(25)] \subseteq [\mu_j - \tau, \mu_j + \tau] \subset (I_{k-1} \cup I_k \cup I_{k+1})$. By Lemma 4.2.1, we know $\mathbb{P}(|g_{ij} - \mu_j| \geq \tau) \leq \mathbb{P}(|g_{ij} - \mu_j| \geq K\lambda_1\sqrt{\|H_u\|_2}\log^a(25)) \leq 0.04$. This means $h_{(k-1),j} + h_{k,j} + h_{(k+1),j} \geq 0.96$ and $\min(h_{(k-1),j}, h_{k,j}, h_{(k+1),j}) \geq 0.32$.

By Dvoretzky-Kiefer-Wolfowitz inequality and an union bound over $j \in [d]$, we have that with probability $1-\zeta$, $\max_{j,l}|h_{j,l} - \hat{h}_{j,l}| \leq \sqrt{\log(d/\zeta)/B}$. Using Lemma A.2.1, if $B = \Omega((\sqrt{d\log(1/\delta)}/\varepsilon)\log(d/(\zeta\delta)))$, with probability $1-\zeta$, we have $\max_{j,l}|\tilde{h}_{j,l} - \hat{h}_{j,l}| \leq 0.005$. Thus, with our assumption on B, we can make sure with probability $1-\zeta$, $\max_{j,l}|\tilde{h}_{j,l} - h_{j,l}| \leq 0.01$. Then we have $\min(h_{(k-1),j}, h_{k,j}, h_{(k+1),j}) - 0.01 \geq 0.31 \geq 0.04 + 0.01 \geq \max_{l \neq k-1,k,k_1} h_{j,l} + 0.01$. This implies with probability $1-\zeta$, the algorithm must pick one of the bins from I_{k-1}, I_k, I_{k+1}. This means $|\bar{g}_j - \mu_j| \leq 2\tau \leq 2^{1.5}K\lambda_1\sqrt{\|H_u\|_2}\log^a(25)$. By tail bound of Lemma 4.2.1, we know for all $j \in [d]$ and $i \in [B]$, $|g_{ij} - \bar{g}_j| \leq |g_{ij} - \mu_j| + |\bar{g}_j - \mu_j| \leq 3K\lambda_1\sqrt{\|H_u\|_2}\log^a(Bd/\zeta)$. This completes our proof.

C.6 Technical lemmas

Lemma C.6.1. *Let $x \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$. Then there exists universal constant C such that with probability $1-\zeta$,*

$$\|x\|^2 \leq C\,\mathrm{Tr}(\Sigma)\log(1/\zeta)\,. \tag{C.59}$$

Proof. Let $\tilde{x} := \Sigma^{-1/2}x$. Then $\tilde{x}$ is also a Gaussian with $\tilde{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. By Hanson-Wright inequality ([200, Theorem 6.2.1]), there exists universal constant $c > 0$ such that with probability $1-\zeta$,

$$\|x\|^2 = \tilde{x}^\top\Sigma\tilde{x} \leq \mathrm{Tr}(\Sigma) + c(\|\Sigma\|_{\mathbf{F}} + \|\Sigma\|_2)\log(2/\zeta) \leq C\,\mathrm{Tr}(\Sigma)\log(1/\zeta)\,. \tag{C.60}$$

□

Lemma C.6.2 ([200, Theorem 4.4.5])**.** *Let $G \in \mathbb{R}^{d\times d}$ be a random matrix where each entry G_{ij} is i.i.d. sampled from standard Gaussian $\mathcal{N}(0,1)$. Then there exists universal constant $C > 0$ such that with probability $1 - 2e^{-t^2}$, $\|G\|_2 \leq C(\sqrt{d} + t)$ for $t > 0$.*

Lemma C.6.3. *Let $G \in \mathbb{R}^{d\times d}$ be a random matrix where each entry G_{ij} is i.i.d. sampled from standard Gaussian $\mathcal{N}(0,1)$. Then we have $\|\mathbb{E}[GG^\top]\|_2 \le C_2 d$ and $\|\mathbb{E}[G^\top G]\|_2 \le C_2 d$.*

Proof. By Lemma C.6.2, there exists universal constant $C_3 > 0$ such that

$$\mathbb{P}\left(\|G\| \ge C_1(\sqrt{d}+s)\right) \le e^{-s^2}, \quad \forall s > 0\,. \tag{C.61}$$

Then

$$\begin{aligned}
\|\mathbb{E}[GG^\top]\|_2 &\le \mathbb{E}[\|GG^\top\|_2] && \text{(C.62)}\\
&\le \mathbb{E}[\|G\|_2^2] && \text{(C.63)}\\
&= \int_0^\infty 2r\mathbb{P}(\|G\|_2 \ge r)dr \le C_1 d + C_3\int_{\sqrt{d}}^\infty 2re^{-\frac{(r-\sqrt{d})^2}{2}}d && \text{(C.64)}\\
&= C_1(d+\sqrt{2\pi d}+2) \le C_2 d\,, && \text{(C.65)}
\end{aligned}$$

where C_2 is an absolute constant. The proof for the second claim follows similarly. □

Lemma C.6.4. *Let $x, y \in \mathbb{S}_2^{d-1}$. Then*

$$1 - \langle x, y\rangle^2 \le \|x-y\|^2\,. \tag{C.66}$$

If $\|x-y\|^2 \le 2$, then

$$1 - \langle x, y\rangle^2 \ge \frac{1}{2}\|x-y\|^2\,. \tag{C.67}$$

The following lemma follows from matrix Bernstein inequality [193].

Lemma C.6.5. *Under A.1, A.2, and A.3, in Assumption 5, with probability $1-\zeta$,*

$$\Big\|\frac{1}{B}\sum_{i\in[B]} A_i - \Sigma\Big\|_2 \;=\; O\Big(\sqrt{\frac{\lambda_1^2 V\log(d/\zeta)}{B}} + \frac{\lambda_1 M\log(d/\zeta)}{B}\Big)\,. \tag{C.68}$$

Appendix D

APPENDICES FOR CHAPTER 5

D.1 Related work

Differentially private optimization. There is a long line of work at the intersection of differentially privacy and optimization [45, 142, 27, 182, 26, 210, 12, 85, 183, 16, 149, 131, 215, 91, 90, 216]. As one of the most well-studied problem in differentially privacy, DP Empirical Risk Minimization (DP-ERM) aims to minimize the empirical risk $(1/n)\sum_{i\in S}\ell(x_i; w)$ privately. The optimal excess empirical risk for approximate DP (i.e., $\delta > 0$) is known to be $GD\cdot\sqrt{d}/(\varepsilon n)$, where the loss ℓ is convex and G-Lipschitz with respect to the data, and D is the diameter of the convex parameter space [27]. This bound can be achieved by several DP-SGD methods, e.g., [182, 27], with different computational complexities. Differentially private stochastic convex optimization considers minimizing the population risk $\mathbb{E}_{x\sim\mathcal{D}}[\ell(x, w)]$, where data is drawn i.i.d. from some unknown distribution $\mathcal{D}$. Using some variations of DP-SGD, [26] and [85] achieves a population risk of $GD(1/\sqrt{n} + \sqrt{d}/(\varepsilon n))$.

DP linear regression. Applying above results for the linear model, by observing that $G = O(d)$ if $D = O(1)$, the sample complexity required for achieving generalization error is $n = d^2$. Existing works for DP linear regression, for example [201, 142, 168, 71, 208, 88, 167, 206, 180, 5] typically consider deterministic data. Under the i.i.d. Gaussian data setting, this translates into a sample complexity of $n = d^{3/2}/(\varepsilon\alpha)$, where the extra $d^{1/2}$ due to the fact that no statistical assumptions are made. For i.i.d. sub-Weibull data, recent work [199] achieved nearly optimal excess population risk $d/n + d^2/(\varepsilon^2 n^2)$ using DP-SGD with adaptive clipping, up to extra factors on the condition number. This is closest to our work and we provide detailed comparisons in Sections 5.2.1 and 5.3.2. Under Gaussian assumptions, [166] analyze linear regression algorithm with sub-optimal guarantees. [77, 11, 7, 161] also consider using

robust statistics like Tukey median [195] or Theil–Sen estimator [191] for differentially private regression. However, [77] and [11] lack utility guarantees and [7] is restricted to one-dimensional data. [161] achieves optimal sample complexity but takes exponential time. More recently, [53] uses private linear regression scenario to show that correlated noise provably improves upon vanilla DP-SGD.

Recent work [41] considers DP generalized linear model and provides a DP-SGD type algorithm that achieves nearly optimal error $d/n + d^2/(\varepsilon^2 n^2)$. Their result is not comparable to ours because they assume the norm of the gradient is bounded by a constant, while for linear regression, the norm of the gradient is $O(\sqrt{d})$.

Robust linear regression. Robust mean estimation and linear regression have been studied for a long time in the statistics community [196, 116, 195]. However, for high dimensional data, these estimators generalizing the notion of median to higher dimensions are typically computationally intractable. Recent advances in the filter-based algorithms, e.g., [65, 61, 62, 69, 49, 73], achieve nearly optimal guarantees for mean estimation in time linear in the dimension of the dataset. Motivated by the filter algorithms, [70, 63, 174, 173, 51, 121] achieved nearly optimal rate with d samples for robust linear regression, where both data x_i and label y_i are corrupted. Another type of efficient methods that achieve similar rates and sample complexity in polynomial time is based on sum-of-square proofs [143, 22], which can be computationally expensive in practice. [217] and [161] achieves nearly optimal rates using d samples but require exponential time complexities. An important special case of adversarial corruption is when the adversary only corrupts the response variable in supervised learning [141] and also in unsupervised learning [192]. For linear regression, when there is only label corruptions, [31, 55, 144] achieve nearly optimal rates with $O(d)$ samples. Under the oblivious label corruption model, i.e., the adversary only corrupts a fraction of labels in complete ignorance of the data, [30, 188] provide consistent estimator $\hat{w}_n$ such that $\lim_{n\to\infty} \mathbb{E}\left[\widehat{w}_n - w^*\right]_2 = 0$ with $O(d)$ samples.

Of these, [31, 55] are relevant to our work as they consider the same adversary model as Asmp. 7. When x_i's and z_i's are sampled from $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \sigma^2)$, [55] proposed a Huber

loss based estimator that achieves error rate of $\sigma^2\alpha^2\log^2(n/\delta)$ when $n = \tilde{O}\left(\kappa^2 d/\alpha^2\right)$. Under the same setting, [31] propoased a hard thresholding based estimator that achieves $\sigma^2\alpha^2$ error rate with $\tilde{O}\left(d/\alpha^2\right)$ sample complexity. Our results in Thm. 5.3.1 match these rates, except for the sub-optimal dependence on $\log^2(1/\alpha)$. Another line of work considered both label and covariate corruptions and developed optimal algorithms for parameter recovery [70, 63, 174, 173, 51, 121, 143, 22, 217, 56]. The best existing efficient algorithm , e.g. [173], achieves error rate of $\sigma^2\alpha^2\log(1/\alpha)$ when $n = \tilde{O}\left(d/\alpha^2\right)$, and the x_i and z_i are sampled from $\mathcal{N}(0, I)$ and $\mathcal{N}(0, \sigma^2)$.

Robust and private linear regression. Under the settings of both DP and data corruptions, the only algorithm by [161] achieves nearly optimal rates $\alpha\log(1/\alpha)\sigma$ with optimal sample complexities of $d/\alpha^2 + d/(\varepsilon\alpha)$. However, their algorithm requires exponential time complexities.

Robust and private mean estimation Based on sum-of-square proofs, recent works [109, 6] are able to achieve nearly optimal rates $\alpha\log(1/\alpha)$ with $\tilde{O}(d)$ samples for sub-Gaussian data with known covariance.

D.2 Preliminary on differential privacy

Our algorithm builds upon two DP primitive: Gaussian mechanism and private histogram. The Gaussian mechanism is one examples of a larger family of mechanisms known as output perturbation mechanisms. In practice, it is possible to get better utility trade-off for a output perturbation mechanism by carefully designing the noise, such as the stair-case mechanism which are shown to achieve optimal utility in the variance [94] and also in hypothesis testing [126]. However, the gain is only by constant factors, which we do not try to optimize in this work. We provide a reference for the Gaussian mechanism and private histogram below.

Lemma D.2.1 (Gaussian mechanism [79])**.** *For a query q with sensitivity Δ_q, the Gaussian mechanism outputs $q(S) + \mathcal{N}(0, (\Delta_q\sqrt{2\log(1.25/\delta)}/\varepsilon)^2\mathbf{I}_d)$ and achieves (ε, δ)-DP.*

When the database is accessed multiple times, we use the following composition theorems

to account for the end-to-end privacy leakage.

Lemma D.2.2 (Parallel composition [165]). *Consider a sequence of interactive queries $\{q_k\}_{k=1}^K$ each operating on a subset S_k of the database and each satisfying (ε, δ)-DP. If S_k's are disjoint then the composition $(q_1(S_1), q_2(S_2), \ldots, q_K(S_K))$ is (ε, δ)-DP.*

Lemma D.2.3 (Serial composition [79]). *If a database is accessed with an $(\varepsilon_1, \delta_1)$-DP mechanism and then with an $(\varepsilon_2, \delta_2)$-DP mechanism, then the end-to-end privacy guarantee is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.*

In most modern privacy analysis of iterative processes, advanced composition theorem from [128] gives tight accountant for the end-to-end privacy budget. It can be improved for specific mechanisms using tighter accountants, e.g., in [169, 97, 207, 219, 98].

Lemma D.2.4 (Advanced composition [128]). *For $\varepsilon \leq 0.9$, an end-to-end guarantee of (ε, δ)-differential privacy is satisfied if a database is accessed k times, each with a $(\varepsilon/(2\sqrt{2k\log(2/\delta)}), \delta/(2k))$-differential private mechanism.*

D.3 Adaptive clipping for the gradient norm

In the ideal clipping thresholds for the norm and the residual, there are unknown terms which we need to estimate adaptively, $(\|w_t - w^*\|_\Sigma^2 + \sigma^2)$ and $\mathrm{Tr}(\Sigma)$, up to constant multiplicative errors. We privately estimate the (squared and shifted) distance to optimum, $(\|w_t - w^*\|_\Sigma^2 + \sigma^2)$, with Alg. 23 and privately estimate the average input norm, $\mathbb{E}[\|x_i\|^2] = \mathrm{Tr}(\Sigma)$, with Alg. 24 in App. D.6. These are used to get the clipping thresholds in Alg. 13. We propose a trimmed mean approach below for distance estimation. The norm estimator is similar and is provided in App. D.6.

Private distance estimation using private trimmed mean. The goal is to estimate the (shifted) distance to optimum, $\|w_t - w^*\|_\Sigma^2 + \sigma^2$, up to some constant multiplicative error. Note that this is precisely the task of estimating the variance of the residual $b_i = y_i - w_t^\top x_i$. When there is no adversarial corruption and no privacy constraint, we can simply use the

empirical variance estimator $(1/n)\sum_{i\in[n]}(y_i - w_t^\top x_i)^2$ to obtain a good estimate. However, the empirical variance estimator is not robust against adversarial corruptions since one outlier can make the estimate arbitrarily large. A classical idea is using the *trimmed estimator* from [196], which throws away the 2α fraction of residuals b_i with the largest magnitude. For datasets with resilience property as assumed in this work, this will guarantee an accurate estimate of the distance to optimum in the presence of α fraction of corruptions.

To make the estimator private, it is tempting to simply add a Laplacian noise to the estimate. However, the sensitivity of the trimmed estimator is unknown and depends on the distance to the optimum that we aim to estimate; this makes it challenging to determine the variance of the Laplacian noise we add. Instead, we propose to partition the dataset into k batches, compute an estimate for each batch, and form a histogram with over those k estimates. Using a private histogram mechanism with geometrically increasing bin sizes, we propose using the bin with the most estimates to guarantee a constant factor approximation of the distance to the optimum. We describe the algorithm as follows.

Algorithm 23: Robust and Private Distance Estimator

Input: $S_2 = \{(x_i, y_i)\}_{i=1}^n$, current w_t, $(\varepsilon_0, \delta_0)$, failure probability ζ,

Let $b_i \leftarrow (y_i - w_t^\top x_i)^2$, $\forall i \in [n]$ and $\tilde{S} \leftarrow \{b_i\}_{i=1}^n$.

Partition $\tilde{S}$ into $k = \lceil C_1 \log(1/(\delta_0\zeta))/\varepsilon_0 \rceil$ subsets of equal size and let G_j be the j-th partition.

For $j \in [k]$, denote ψ_j as the 0.9-quantile of G_j and $\phi_j \leftarrow \frac{1}{|G_j|}\sum_{i\in G_j} b_i \mathbf{1}\{b_i \le \psi_j\}$.

Partition $[0, \infty)$ into geometrically increasing intervals

$\Omega := \{\ldots, [2^{-1}, 1), [1, 2), [2, 2^2), \ldots\} \cup \{[0, 0]\}$

Run $(\varepsilon_0, \delta_0)$-DP histogram of Lemma A.2.1 on $\{\phi_j\}_{j=1}^k$ over Ω

if all the bins are empty **then** Return $\perp$

Let $[\ell, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram

return ℓ

This algorithm gives an estimate of the distance up to a constant multiplicative error as we show in the following theorem. We provide a proof in App. D.4.

Theorem D.3.1. *Alg. 23 is $(\varepsilon_0, \delta_0)$-DP. For an α_{corrupt}-corrupted dataset S_2 that satisfy Asmp. 6 and Asmp. 7 and any $\zeta \in (0,1)$, if*

$$n = O\left(\frac{(d + \log((\log(1/(\delta_0\zeta)))/\varepsilon_0\zeta))(\log(1/(\delta_0\zeta)))}{\varepsilon_0}\right), \tag{D.1}$$

with a large enough constant, then with probability $1-\zeta$, Alg. 23 returns ℓ such that $\frac{1}{4}(\|w_t - w^*\|_\Sigma^2 + \sigma^2) \leq \ell \leq 4(\|w_t - w^*\|_\Sigma^2 + \sigma^2)$.

Note that in Thm. D.3.1, we only need to estimate distance up to a constant multiplicative error, as opposed to an error that depends on our final end-to-end desired level α. Consequently, we require smaller sample complexity (that doesn't depend on α) than other parts of our approach.

Remark D.3.2. *While DP-STAT (Algorithm 3 in [199]) can also be used to estimate $\|w_t - w^*\|_\Sigma + \sigma$ (and it would not change the ultimate sample complexity in its dependence on κ, d, ε, and n), there are three important improvements we make: (i) DP-STAT requires the knowledge of $\|w^*\|_\Sigma + \sigma$; (ii) our utility guarantee has improved dependence in K and $\log(n)$; and (iii) Alg. 23 is robust against label corruption.*

Upper bound on clipped good data points. Using the above estimated distance to the optimum in selecting a threshold θ_t, we also need to ensure that we do not clip too many clean data points. The tolerance in our algorithm to reach the desired level of accuracy is clipping $O(\alpha)$ fraction of clean data points. This is ensured by the following lemma, and we provide a proof in App. D.5.

Lemma D.3.3. *Under Asmp. 6 and for all $t \in [T]$, if $\theta_t \geq \sqrt{9C_2K^2\log(1/(2\alpha))}\cdot(\|w^* - w_t\|_\Sigma + \sigma)$ then $\left|\left\{i \in S_3 \cap S_{\text{good}} : \left|w_t^\top x_i - y_i\right| \geq \theta_t\right\}\right| \leq \alpha n$.*

D.4 Proof of Thm. D.3.1 on the private distance estimation

We present our formal theorem for the general sub-Weibull distribution as follows.

Theorem D.4.1. *Alg. 23 is $(\varepsilon_0, \delta_0)$-DP. For an α_{corrupt}-corrupted dataset S_2 satisfying Asmp. 8 and Asmp. 7 and an upper bound $\bar{\alpha}$ on α_{corrupt} that satisfy $37C_2K^2 \cdot \bar{\alpha}\log^{2a}(1/(6\bar{\alpha})) \leq 1/4$ and any $\zeta \in (0,1)$, if*

$$n = O\left(\frac{(d + \log((\log(1/(\delta_0\zeta)))/\varepsilon_0\zeta))(\log(1/(\delta_0\zeta)))}{\bar{\alpha}^2\varepsilon_0}\right), \tag{D.2}$$

with a large enough constant then, with probability $1-\zeta$, Alg. 23 returns ℓ such that $\frac{1}{4}(\|w_t - w^*\|_\Sigma^2 + \sigma^2) \leq \ell \leq 4(\|w_t - w^*\|_\Sigma^2 + \sigma^2)$.

We first analyze the privacy. Changing a data point (x_i, y_i) can affect at most one partition in $\{G_j\}_{j=1}^k$. This would affect at most two histogram bins, increasing the count of one bin by one and decreasing the count in another bin by one. Under such a bounded ℓ_1 sensitivity, the privacy guarantees follows from Lemma A.2.1.

Next, we analyze the utility. In the (private) histogram step, we claim that at most only two consecutive bins can be occupied by any ϕ_j's. This is also true for the private histogram, because the private histogram of Lemma A.2.1 adds noise to non-empty bins only. By Lemma A.2.1, if $k \geq c\log(1/(\delta_0\zeta_0))/\varepsilon_0$, one of these two intervals (the union of which contains the true distance $\|w_t - w^*\|_\Sigma^2 + \sigma^2$) is released. This results in a multiplicative error bound of four, as the bin size increments by a factor of two.

To show that only two bins are occupied, we show that all ϕ_j's are close to the true distance. We first show that each partition contains at most $2\alpha_{\text{corrupt}}$ fraction of corrupted samples and thus all partitions are $(2\bar{\alpha}, 6\bar{\alpha}, 6\hat{\rho}, 6\hat{\rho}, 6\hat{\rho}, 6\hat{\rho}')$-corrupt good, where $\hat{\rho}(C_2, K, a, \bar{\alpha}) = C_2K^2\bar{\alpha}\log^{2a}(1/6\bar{\alpha})$ and $\hat{\rho}'(C_2, K, a, \bar{\alpha}) = C_2K\bar{\alpha}\log^a(1/6\bar{\alpha})$, as defined in Definition D.10.6.

Let $B = \lfloor n/k \rfloor$ be the sample size in each partition. Let $\zeta_0 = \zeta/2$. Since the partition is drawn uniformly at random, for each partition G_j, the number of corrupted samples $\alpha' n$ satisfies $\alpha' n \sim \text{Hypergeometric}(n, \alpha_{\text{corrupt}}n, n/k)$. The tail bound gives that with probability

$1-\zeta_0$,

$$\alpha' \le \alpha_{\text{corrupt}} + (k/n)\log(2/\zeta_0) \le 2\bar{\alpha} \ ,$$

where the last inequality follows from the fact that the corruption level is bounded by $\alpha_{\text{corruption}} \le \bar{\alpha}$ and the assumption on the sample size in Eq. (D.2) which implies $n \gtrsim \log(1/(\delta_0\zeta_0))\log(1/\zeta_0)/(\bar{\alpha}\varepsilon_0)$.

For a particular subset G_j, Lemma D.10.7 implies that if $B = O((d+\log(1/\zeta_0))/\bar{\alpha}^2)$, then G_j is $(\alpha', 6\bar{\alpha}, 6\hat{\rho}, 6\hat{\rho}, 6\hat{\rho}, 6\hat{\rho}')$-corrupt good set with respect to (w^*, Σ, σ) from Asmp. 8. This means that there exists a constant $C_2 > 0$ such that for any $T_1 \subset S_{\text{good}}$ with $|T_1| \ge (1-6\bar{\alpha})B$, we have

$$\left| \frac{1}{|T_1|} \sum_{i \in T_1} \langle x_i, w^* - w_t \rangle^2 - \|w^* - w_t\|_\Sigma^2 \right| \le 6C_2K^2\bar{\alpha}\log^{2a}(1/(6\bar{\alpha}))\|w^* - w_t\|_\Sigma^2 \ ,$$

$$\left| \frac{1}{|T_1|} \sum_{i \in T_1} z_i^2 - \sigma^2 \right| \le 6C_2K^2\bar{\alpha}\log^{2a}(1/(6\bar{\alpha}))\sigma^2 \ ,$$

and

$$\left| \frac{1}{|T_1|} \sum_{i \in T_1} z_i \langle x_i, w^* - w_t \rangle \right| \le 6C_2K^2\bar{\alpha}\log^{2a}(1/(6\bar{\alpha}))\|w^* - w_t\|_\Sigma\sigma \ .$$

Note that for $i \in S_{\text{good}}$, $b_i = z_i^2 + 2z_i(w^* - w_t)^\top x_i + (w^* - w_t)^\top x_i x_i^\top (w^* - w_t)$. By the triangular inequality, we know, under above conditions,

$$\left| \frac{1}{|T_1|} \sum_{i \in T_1} b_i - \|w^* - w_t\|_\Sigma^2 - \sigma^2 \right| \le 12C_2K^2\bar{\alpha}\log^{2a}(1/(6\bar{\alpha}))(\|w^* - w_t\|_\Sigma^2 + \sigma^2) \ . \tag{D.3}$$

Which also implies that any subset $T_2 \subset S_{\text{good}}$ and $|T_2| \le 6\bar{\alpha}|S_{\text{good}}|$, we have

$$\left| \frac{1}{|T_2|} \sum_{i \in T_2} b_i - \|w^* - w_t\|_\Sigma^2 - \sigma^2 \right| \le 12C_2K^2\log^{2a}(1/(6\bar{\alpha}))(\|w^* - w_t\|_\Sigma^2 + \sigma^2) \ . \tag{D.4}$$

Recall that ψ_j is the $(1-3\bar{\alpha})$-quantile of the dataset G_j. Let $T := \{i \in S_{\text{good}} : b_i \le \psi_j\}$, where with a slight abuse of notations, we use S_{good} to denote the set of uncorrupted samples

corresponding to G_j and S_{bad} to denote the set of corrupted samples corresponding to G_j. Since the corruption is less than α', we know $(1-3\bar{\alpha}-\alpha')B \le |T| \le (1-3\bar{\alpha}+\alpha')B$. By our assumption that $\alpha' \le 2\bar{\alpha}$, we have $|\bar{E}| \ge (3\bar{\alpha}-\alpha')B \ge \bar{\alpha}B$ where $\bar{E} := S_{\text{good}} \setminus E$. Using Eq (D.4) with a choice of $T_2 = \bar{E}$, we get that

$$\min_{i\in\bar{E}} b_i - \|w^* - w_t\|_\Sigma^2 - \sigma^2 \le 12C_2K^2\log^{2a}(1/(6\bar{\alpha}))(\|w^* - w_t\|_\Sigma^2 + \sigma^2)\,. \tag{D.5}$$

This implies that

$$\psi_j \le 12C_2K^2\log^{2a}(1/(6\bar{\alpha}))(\|w^* - w_t\|_\Sigma^2 + \sigma^2). \tag{D.6}$$

Hence

$$\begin{aligned}
\left|\phi_j - \|w^* - w_t\|_\Sigma^2 - \sigma^2\right| &= \left|\frac{1}{B}\sum_{i\in G_j} b_i \cdot \mathbf{1}\{b_i \le \psi_j\} - \|w^* - w_t\|_\Sigma^2 - \sigma^2\right| \\
&= \left|\frac{1}{B}\sum_{i\in T} b_i - \|w^* - w_t\|_\Sigma^2 - \sigma^2\right| + \left|\frac{1}{B}\sum_{i\in S_{\text{bad}}} b_i \cdot \mathbf{1}\{b_i \le \psi_j\}\right| \\
&\le 37C_2K^2\cdot\bar{\alpha}\log^{2a}(1/(6\bar{\alpha}))(\|w^* - w_t\|_\Sigma^2 + \sigma^2),
\end{aligned} \tag{D.7}$$

where we applied Eq (D.6) and Eq (D.3) in the last inequality.

On a fixed partition G_j, we showed that if $B = O((d+\log(1/\zeta_0))/\bar{\alpha}^2)$ then, with probability $1-\zeta_0$, $|\phi_j - \|w^* - w_t\|_\Sigma^2 - \sigma^2| \le \frac{1}{4}(\|w^* - w_t\|_\Sigma^2 + \sigma^2)$, which follows from our assumption that $37C_2K^2\cdot\bar{\alpha}\log^{2a}(1/(6\bar{\alpha})) \le 1/4$. Using an union bound for all subsets, we know if $B = O((d+\log(k/\zeta_0))/\bar{\alpha}^2)$, then $1-\zeta_0$, $|\phi_j - \|w^* - w_t\|_\Sigma^2 - \sigma^2| \le \frac{1}{4}(\|w^* - w_t\|_\Sigma^2 + \sigma^2)$ holds for all $j \in [k]$. Since the upper bound lower bound ratio is $5/3$ which is less than 2. All the ϕ_j must lie in two bins, which will result in a factor of 4 multiplicative error.

D.5 Proof of Lemma D.3.3 on the upper bound on clipped good points

Let $\hat{\rho}(C_2,K,a,\alpha) = 2C_2K^2\alpha\log^{2a}(1/(2\alpha))$ and $\hat{\rho}'(C_2,K,a,\alpha) = 2C_2K\alpha\log^{a}(1/(2\alpha))$. Lemma D.10.7 implies that if $n = O((d+\log(1/\zeta))/(\alpha^2))$ with a large enough constant, then there exists

a universal constant C_2 such that S_3 is, with respect to (w^*, Σ, σ), $(\alpha_{\text{corrupt}}, 2\alpha, \hat{\rho}, \hat{\rho}, \hat{\rho}, \hat{\rho}')$-corrupt good. The rest of the proof is under this (deterministic) resilience condition. By the resilience property in Eq (5.6), we know for any $T \subset S_{\text{good}}$ with $|T| \geq (1-2\alpha)n$,

$$\left| \frac{1}{|T|} \sum_{i \in T} (w^* - w_t)^\top x_i x_i^\top (w^* - w_t) - \|w^* - w_t\|_\Sigma^2 \right| \leq 2C_2 K^2 \alpha \log^{2a}(1/(2\alpha)) \|w^* - w_t\|_\Sigma^2 \,. \tag{D.8}$$

Let $E := \left\{ i \in S_{\text{good}} : (w^* - w_t)^\top x_i x_i^\top (w^* - w_t) > \|w^* - w_t\|_\Sigma^2 (8C_2 K^2 \log^{2a}(1/(2\alpha)) + 1) \right\}$. Denote $\tilde{\alpha} := |E|/n$. We want to show that $\tilde{\alpha} \leq \alpha/2$. Let T be the set of points that contain the smallest $1 - \alpha/2$ fraction in $\{(w^* - w_t)^\top x_i x_i^\top (w^* - w_t)\}_{i \in S_{\text{good}}}$. We know $|T| = (1-\alpha/2)n \geq (1-2\alpha)n$. To prove by contradiction, suppose $\tilde{\alpha} > \alpha/2$, which means all data points in $S_{\text{good}} \setminus T$ are larger than $\|w^* - w_t\|_\Sigma^2 (8C_2 K^2 \log^{2a}(1/(2\alpha)) + 1)$. From resilience property in Eq (D.8), we know

$$\begin{aligned}
&\frac{1}{n} \sum_{i \in S_{\text{good}}} (w^* - w_t)^\top x_i x_i^\top (w^* - w_t) \\
&= \frac{1}{n} \sum_{i \in T} (w^* - w_t)^\top x_i x_i^\top (w^* - w_t) + \frac{1}{n} \sum_{i \in S_{\text{good}} \setminus T} (w^* - w_t)^\top x_i x_i^\top (w^* - w_t) \\
&\geq \left(1 - \frac{\alpha}{2}\right) \left(1 - 2C_2 K^2 \alpha \log^{2a}(\frac{1}{2\alpha})\right) \|w^* - w_t\|_\Sigma^2 + \frac{\alpha}{2} (8C_2 K^2 \log^{2a}(\frac{1}{2\alpha}) + 1) \|w^* - w_t\|_\Sigma^2 \\
&> (1 + 2C_2 K^2 \alpha \log^{2a}(1/2\alpha)) \|w^* - w_t\|_\Sigma^2 \,,
\end{aligned}$$

which contradicts Eq (D.8) for S_{good}. This shows $\tilde{\alpha} \leq \alpha/2$.

Similarly, we can show that $\left| \left\{ i \in S_{\text{good}} : z_i^2 > \sigma^2 (8C_2 K^2 \log^{2a}(1/(2\alpha)) + 1) \right\} \right| \leq \alpha/2$. This means the rest $(1-\alpha)n$ points in S_{good} satisfies $\sqrt{(w^* - w_t)^\top x_i x_i^\top (w^* - w_t)} + |z_i| \leq (\|w_t - w^*\| + \sigma) \sqrt{(8C_2 K^2 \log^{2a}(1/(2\alpha)) + 1)}$. Note that for all $i \in S_{\text{good}}$, we have

$$\begin{aligned}
|x_i^\top w_t - y_i| &= \left| x_i^\top (w_t - w^*) - z_i \right| \\
&\leq |x_i^\top (w_t - w^*)| + |z_i| \\
&\leq \left(\sqrt{(w^* - w_t)^\top x_i x_i^\top (w^* - w_t)} + |z_i| \right) \,.
\end{aligned}$$

By our assumption that $C_2K^2\log^{2a}(1/(2\bar{\alpha})) \geq 1$ which follows from Asmp. 7, we have

$$\left|\left\{i \in S_{\text{good}} : |x_i^\top w_t - y_i| \leq (\|w_t - w^*\|_\Sigma + \sigma)\sqrt{9C_2K^2\log^{2a}(1/(2\alpha))}\right\}\right| \geq (1-\alpha)n\ . \quad \text{(D.9)}$$

D.6 Private norm estimation: algorithm and analysis

Algorithm 24: Private Norm Estimator

Input: $S_1 = \{(x_i, y_i)\}_{i=1}^n$, target privacy $(\varepsilon_0, \delta_0)$, failure probability ζ.

Let $a_i \leftarrow \|x_i\|^2$. Let $\tilde{S} = \{a_i\}_{i=1}^n$.

Partition $\tilde{S}$ into $k = \lfloor C_1 \log(1/(\delta_0\zeta))/\varepsilon \rfloor$ subsets of equal size and let G_j be the j-th partition.

For each $j \in [k]$, denote $\psi_j = (1/|G_j|)\sum_{i\in G_j} a_i$.

Partition $[0,\infty)$ into bins of geometrically increasing intervals

$$\Omega := \left\{\ldots, \left[2^{-2/4}, 2^{-1/4}\right), \left[2^{-1/4}, 1\right), \left[1, 2^{1/4}\right), \left[2^{1/4}, 2^{2/4}\right), \ldots\right\} \cup \{[0,0]\}$$

Run $(\varepsilon_0, \delta_0)$-DP histogram learner of Lemma A.2.1 on $\{\psi_j\}_{j=1}^k$ over Ω

if all the bins are empty **then** Return $\perp$

Let $[\ell, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram

Return ℓ

Lemma D.6.1. *Alg. 24 is $(\varepsilon_0, \delta_0)$-DP. If $\{x_i\}_{i=1}^n$ are i.i.d. samples from (K, a)-sub-Weibull distributions with zero mean and covariance Σ and*

$$n = \tilde{O}\left(\frac{\log^{2a}(1/(\delta_0\zeta))}{\varepsilon_0}\right)\ ,$$

with a large enough constant then Alg. 24 returns Γ such that, with probability $1-\zeta$,

$$\frac{1}{\sqrt{2}}\operatorname{Tr}(\Sigma) \leq \Gamma \leq \sqrt{2}\operatorname{Tr}(\Sigma)\ .$$

We provide a proof in App. D.6.1.

D.6.1 Proof of Lemma D.6.1 on the private norm estimation

By Hanson-Wright inequality in Lemma D.10.1 and union bound, there exists constant $c > 0$ such that with probability $1 - \zeta$,

$$|\frac{1}{b}\sum_{i=1}^{b}\|x_i\|^2 - \mathrm{Tr}(\Sigma)| \le cK^2\,\mathrm{Tr}(\Sigma)\left(\sqrt{\frac{\log(1/\zeta)}{b}} + \frac{\log^{2a}(1/\zeta)}{b}\right), \tag{D.10}$$

This means there exists a constant $c' > 0$ such that if $b \ge c'K^2\log^{2a}(k/\zeta)$, then for all $j \in [k]$.

$$|\psi_j - \mathrm{Tr}(\Sigma)| \le 2^{1/8}\,\mathrm{Tr}(\Sigma) \tag{D.11}$$

With probability $1-\zeta$, $\{\psi_j\}_{j=1}^k$ lie in interval of size $2^{1/4}\,\mathrm{Tr}(\Sigma)$. Thus, at most two consecutive bins are filled with $\{\psi_j\}_{j=1}^k$. Denote them as $I = I_1 \cup I_2$. Our analysis indicates that $\mathbb{P}(\psi_i \in I) \ge 0.99$. By private histogram in Lemma A.2.1, if $k \ge \log(1/(\delta\zeta))/\varepsilon$, $|\hat{p}_I - \tilde{p}_I| \le 0.01$ where $\hat{p}_I$ is the empirical count on I and $\tilde{p}_I$ is the noisy count on I. Under this condition, one of these two intervals are released. This results in multiplicative error of $\sqrt{2}$.

D.7 Proof of the resilience in Lemma D.10.7

We apply following resilience property for general distribution characterized by Orlicz function from [217].

Lemma D.7.1 ([217, Theorem 3.4]). *Dataset $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ consists i.i.d. samples from a distribution $\mathcal{D}$. Suppose $\mathcal{D}$ is zero mean and satisfies $\mathbb{E}_{x\sim\mathcal{D}}\left[\psi\left(\frac{(v^\top x)^2}{\kappa^2\mathbb{E}_{x\sim\mathcal{D}}[(v^\top x)^2]}\right)\right] \le 1$ for all $v \in \mathbb{R}^d$, where $\psi(\cdot)$ is Orlicz function. Let $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}[xx^\top]$. Suppose $\alpha \le \bar{\alpha}$, where $\bar{\alpha}$ satisfies $(1+\bar{\alpha}/2)\cdot 2\kappa^2\bar{\alpha}\psi^{-1}(2/\bar{\alpha}) < 1/3$, $\bar{\alpha} \le 1/4$. Then there exists constant c_1, C_2 such that if $n \ge c_1((d+\log(1/\zeta))/(\alpha^2))$, with probability $1-\zeta$, for any $T \subset S$ of size $|T| \ge (1-\alpha)n$, the following holds:*

$$\left\|\Sigma^{-1/2}\left(\frac{1}{|T|}\sum_{i\in T}x_i\right)\right\| \le C_2\kappa\alpha\sqrt{\psi^{-1}(1/\alpha)} \tag{D.12}$$

and

$$\left\|\mathbf{I}_d - \Sigma^{-1/2}\left(\frac{1}{|T|}\sum_{i\in T} x_i x_i^\top\right)\Sigma^{-1/2}\right\|_2 \leq C_2\kappa^2\alpha\psi^{-1}(1/\alpha)\,. \tag{D.13}$$

Let $\psi(t) = e^{t^{1/(2a)}}$. It is easy to see that $\psi(t)$ is a valid Orlicz function. Then if x_i is (K,a)-sub-Weibull, then we know

$$\left\|\Sigma^{-1/2}\left(\frac{1}{|T|}\sum_{i\in T} x_i\right)\right\| \leq C_2 K\alpha\sqrt{\log^{2a}(1/\alpha)}\,, \tag{D.14}$$

and

$$\left\|\mathbf{I}_d - \Sigma^{-1/2}\left(\frac{1}{|T|}\sum_{i\in T} x_i x_i^\top\right)\Sigma^{-1/2}\right\|_2 \leq C_2 K^2\alpha\log^{2a}(1/\alpha)\,. \tag{D.15}$$

This implies

$$(1 - C_2K^2\alpha\log^{2a}(1/\alpha))\mathbf{I}_d \preceq \Sigma^{-1/2}\left(\frac{1}{|T|}\sum_{i\in T} x_i x_i^\top\right)\Sigma^{-1/2} \preceq (1 + C_2K^2\alpha\log^{2a}(1/\alpha))\mathbf{I}_d\,. \tag{D.16}$$

Using the fact that $C^\top AC \preceq C^\top BC$ if $A \preceq B$, we know

$$(1 - C_2K^2\alpha\log^{2a}(1/\alpha))\Sigma \preceq \frac{1}{|T|}\sum_{i\in T} x_i x_i^\top \preceq (1 + C_2K^2\alpha\log^{2a}(1/\alpha))\Sigma\,. \tag{D.17}$$

This implies resilience properties of x_i and z_i in (5.6) and (5.7) in Definition 5.5.1 respectively. Next, we show the resilience property of $x_i z_i$.

By $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$, for any fixed $v \in \mathbb{R}^d$,

$$\mathbb{E}[\exp\left(\left(\frac{|\langle x_i z_i, v\rangle|^2}{K^4\sigma^2 v^\top\Sigma v}\right)^{1/(4a)}\right)] \leq \mathbb{E}\left[\exp\left(\left(\frac{|\langle x_i, v\rangle|^2}{K^2 v^\top\Sigma v}\right)^{1/(2a)}/2\right)\exp\left(\left(\frac{z_i^2}{K^2\sigma^2}\right)^{1/(2a)}/2\right)\right] \tag{D.18}$$

$$\leq \frac{1}{2}\left(\mathbb{E}\left[\exp\left(\left(\frac{|\langle x_i, v\rangle|^2}{K^2 v^\top\Sigma v}\right)^{1/(2a)}\right)\right] + \mathbb{E}\left[\exp\left(\left(\frac{z_i^2}{K^2\sigma^2}\right)^{1/(2a)}\right)\right]\right) \tag{D.19}$$

$$\leq 2\,. \tag{D.20}$$

Since $\mathbb{E}[x_i z_i] = 0$, [217, Lemma E.3] implies that there exists constant $c_1, C_2 > 0$ such that if $n \geq c_1(d + \log(1/\zeta))/(\alpha^2)$, with probability $1 - \zeta$, for any $T \subset S_{\text{good}}$ of size $|T| \geq (1-\alpha)n$,

$$\left\| \Sigma^{-1} \left(\frac{1}{|T|} \sum_{i \in T} x_i z_i \right) \right\| \leq C_2 K^2 \sigma \alpha \log^{2a}(1/\alpha) \; . \tag{D.21}$$

D.8 Proof of Thm. 26 on the analysis of Alg. 13

We provide our main theorem under the following sub-Weibull assumptions.

Assumption 8 ($(\Sigma, \sigma^2, w^*, K, a)$-model)**.** *A multiset $S_{\text{good}} = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ of n i.i.d. samples is from a linear model $y_i = \langle x_i, w^* \rangle + z_i$, where the input vector x_i is zero mean, $\mathbb{E}[x_i] = 0$, with a positive definite covariance $\Sigma := \mathbb{E}[x_i x_i^\top] \succ 0$, and the (input dependent) label noise z_i is zero mean, $\mathbb{E}[z_i] = 0$, with variance $\sigma^2 := \mathbb{E}[z_i^2]$. We further assume $\mathbb{E}[x_i z_i] = 0$, which is equivalent to assuming that the true parameter $w^* = \Sigma^{-1}\mathbb{E}[y_i x_i]$. We assume that the marginal distribution of x_i is (K, a)-sub-Weibull and that of z_i is also (K, a)-sub-Weibull, as defined below.*

Sub-Weibull distributions provide Gaussian-like tail bounds determining the resilience of the dataset in Lemma D.10.7, which our analysis critically relies on and whose necessity is justified in Sec. 5.3.4.

Definition D.8.1 (sub-Weibull distribution [148])**.** *For some $K, a > 0$, we say a random vector $x \in \mathbb{R}^d$ is from a (K, a)-sub-Weibull distribution if for all $v \in \mathbb{R}^d$, $\mathbb{E}\left[\exp\left(\left(\frac{\langle v, x\rangle^2}{K^2 \mathbb{E}[\langle v, x\rangle^2]}\right)^{1/(2a)}\right)\right] \leq 2$.*

Theorem D.8.2. *Alg. 13 is (ε, δ)-DP. Under $(\Sigma, \sigma^2, w^*, K, a)$-model of Asmp. 8 and α_{corrupt}-corruption of Assumption 7 and for any failure probability $\zeta \in (0, 1)$ and target error rate $\alpha \geq \alpha_{\text{corrupt}}$. We further assume that the corruption level is bounded by $\alpha_{\text{corrupt}} \leq \bar{\alpha}$, where $\bar{\alpha}$ is a known positive constant satisfying $\bar{\alpha} \leq 1/10$, $72 C_2 K^2 \bar{\alpha} \log^{2a}(1/(6\bar{\alpha})) \log(\kappa) \leq 1/2$, and $2C_2 K^2 \log^{2a}(1/(2\bar{\alpha})) \geq 1$ for the (K, a)-sub-Weibull distribution of interest and a positive*

constant C_2 defined in Lemma D.10.7 that only depends on (K, a). If the sample size is large enough such that

$$n = \tilde{O}\left(K^2 d \log^{2a+1}\left(\frac{1}{\zeta}\right) + \frac{d + \log(1/\zeta)}{\alpha^2} + \frac{K^2 d T^{1/2} \log(\frac{1}{\delta}) \log^a(\frac{1}{\zeta})}{\varepsilon\alpha}\right), \tag{D.22}$$

with a large enough constant where $\tilde{O}$ hides poly-logarithmic terms in d, n, and κ, then the choices of a step size $\eta = 1/(C\lambda_{\max}(\Sigma))$ for any $C \geq 1.1$ and the number of iterations, $T = \tilde{\Theta}\left(\kappa \log\left(\|w^*\|\right)\right)$ for a condition number of the covariance $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, ensures that, with probability $1 - \zeta$, Alg. 13 achieves

$$\mathbb{E}_{\nu_1, \cdots, \nu_t \sim \mathcal{N}(0, \mathbf{I}_d)}\left[\, \|w_T - w^*\|_\Sigma^2 \,\right] = \tilde{O}\Big(K^4 \sigma^2 \alpha^2 \log^{4a}\left(\frac{1}{\alpha}\right)\Big), \tag{D.23}$$

where the expectation is taken over the noise added for DP, and $\tilde{\Theta}(\cdot)$ hides logarithmic terms in $K, \sigma, d, n, 1/\varepsilon, \log(1/\delta), 1/\alpha$, and κ.

The main theorem builds upon the following lemma that analyzes a (stochastic) gradient descent method, where the randomness is from the DP noise we add and the analysis only relies on certain deterministic conditions on the dataset including resilienece and concentration. Thm. D.8.2 follows in a straightforward manner by collecting Thm. D.3.1, Lemma D.6.1, Lemma D.3.3, and Lemma D.8.3.

Lemma D.8.3. *Alg. 13 is (ε, δ)-DP. Under Assumptions 8 and 7 for any $\zeta \in (0, 1)$ and $\alpha \geq \alpha_{\text{corrupt}}$ satisfying $K^2 \alpha \log^{2a}(1/\alpha) \log(\kappa) \leq c$ for some universal constant $c > 0$, if distance threshold is small enough such that*

$$\theta_t \;\leq\; 3C_2^{1/2} K \log^a(1/(2\alpha)) \cdot \left(\|w^* - w_t\|_\Sigma + \sigma\right), \tag{D.24}$$

and large enough such that the number of clipped clean data points is no larger than αn, at every round, the norm threshold is large enough such that

$$\Theta \;\geq\; K\sqrt{\mathrm{Tr}(\Sigma)} \log^a(n/\zeta), \tag{D.25}$$

and sample size is large enough such that

$$n = O\left(K^2 d \log(d/\zeta) \log^{2a}(n/\zeta) + \frac{d + \log(1/\zeta)}{\alpha^2} + \frac{K^2 T^{1/2} d \log(T/\delta) \log^a(n/(\alpha\zeta))}{\varepsilon\alpha}\right) , \tag{D.26}$$

with a large enough constant, then the choices of a step size, $\eta = 1/(C\lambda_{\max}(\Sigma))$ for some $C \geq 1.1$, and the number of iterations, $T = \tilde{\Theta}\left(\kappa \log\left(\|w^*\|\right)\right)$, ensures that Alg. 13 outputs w_T satisfying the following with probability $1 - \zeta$:

$$\mathbb{E}_{\nu_1, \cdots, \nu_t \sim \mathcal{N}(0, \mathbf{I}_d)}[\|w_T - w^*\|_\Sigma^2] \lesssim K^4 \sigma^2 \log^2(\kappa) \alpha^2 \log^{4a}(1/\alpha) , \tag{D.27}$$

where the expectation is taken over the noise added for DP and $\tilde{\Theta}(\cdot)$ hides logarithmic terms in $K, \sigma, d, n, 1/\varepsilon, \log(1/\delta), 1/\alpha$.

Proof of Lemma D.8.3. We first prove a set of deterministic conditions on the clean dataset, which is sufficient for the analysis of the gradient descent.

Step 1: Sufficient deterministic conditions on the clean dataset. Let S_{good} be the uncorrupted dataset for S_3 and S_{bad} be the corrupted datapoints in S_3. Let $G := S_{\text{good}} \cap S_3 = S_3 \setminus S_{\text{bad}}$ denote the clean data that remains in the input dataset. Let $\lambda_{\max} = \|\Sigma\|_2$. Define $\hat{\Sigma} := (1/n) \sum_{i \in G} x_i x_i^\top$, $\hat{B} := \mathbf{I}_d - \eta\hat{\Sigma}$. Lemma D.10.4 implies that if $n = O(K^2 d \log(d/\zeta) \log^{2a}(n/\zeta))$, then

$$0.9\Sigma \preceq \hat{\Sigma} \preceq 1.1\Sigma . \tag{D.28}$$

We pick step size η such that $\eta \leq 1/(1.1\lambda_{\max})$ to ensure that $\eta \leq 1/\|\hat{\Sigma}\|_2$. Since the covariates $\{x_i\}_{i \in S}$ are not corrupted, from Lemma D.10.3, we know with probability $1 - \zeta$, for all $i \in S_3$,

$$\|x_i\|^2 \leq K^2 \operatorname{Tr}(\Sigma) \log^{2a}(n/\zeta) . \tag{D.29}$$

Lemma D.10.7 implies that if $n = O((d + \log(1/\zeta))/(\alpha^2))$, then there exists a universal constant C_2 such that S_3 is, following Definition D.10.6, with respect to (w^*, Σ, σ), $(\alpha_{\text{corrupt}}, \alpha, C_2 K^2 \alpha \log^{2a}(1/\alpha), C_2 K^2 \alpha \log^{2a}(1/\alpha), C_2 K^2 \alpha \log^{2a}(1/\alpha), C_2 K \alpha \log^a(1/\alpha))$-corrupt

good. Such corrupt good sets have a sufficiently large, $1-\alpha_{\text{corrupt}}$, fraction of points that satisfy a good property that we need: resilience. The rest of the proof is under Eq (D.28), Eq (D.29), and that S_{good} is resilient.

Step 2: Upper bounding the deterministic noise in the gradient. In this step, we bound the deviation of the gradient from its mean. There are several sources of deviation: (i) clipping, (ii) adversarial corruptions, and (iii) randomness of the data noise and privacy noise. We will show that deviations from all these sources can be controlled deterministically under the corrupt-goodness (i.e., resilience).

Let $\phi_t = (\sqrt{2\log(1.25/\delta_0)}\Theta\theta_t)/(\varepsilon_0 n)$, which ensures that we add enough noise to guarantee $(\varepsilon_0, \delta_0)$-DP for each step of gradient descent. This follows from the standard Gaussian mechanism in Lemma D.2.1 and the fact that each gradient is clipped to the norm of $\Theta\theta_t$, resulting in a DP sensitivity of $\Theta\theta_t/n$. The fact that this sensitivity scales as $1/n$ is one of the main reasons for the performance gain we get over [199] that uses a minimatch of size n/κ with sensitivity scaling as κ/n. Define $g_i^{(t)} := x_i(x_i^\top w_t - y_i)$. For $i \in S_{\text{good}}$, we know $y_i = x_i^\top w^* + z_i$. Let $\tilde{g}_i^{(t)} = \text{clip}_\Theta(x_i)\text{clip}_{\theta_t}(x_i^\top w_t - y_i)$. Note that under Eq (D.29), $\text{clip}_\Theta(x_i) = x_i$ for all $i \in S_3$.

From Alg. 13, we can write one-step update rule as follows:

$$
\begin{aligned}
&w_{t+1} - w^* \\
=&w_t - \eta\left(\frac{1}{n}\sum_{i\in S}\tilde{g}_i^{(t)} + \phi_t\nu_t\right) - w^* \\
=&\left(\mathbf{I} - \frac{\eta}{n}\sum_{i\in G}x_i x_i^\top\right)(w_t - w^*) + \frac{\eta}{n}\sum_{i\in G}x_i z_i + \frac{\eta}{n}\sum_{i\in G}(g_i^{(t)} - \tilde{g}_i^{(t)}) - \eta\phi_t\nu_t - \frac{\eta}{n}\sum_{i\in S_{\text{bad}}}\tilde{g}_i^{(t)}
\end{aligned}
\tag{D.30}
$$

Let $E_t := \{i \in G : \theta_t \le |x_i^\top w_t - y_i|\}$ be the set of clipped clean data points such that $\sum_{i\in G}(g_i^{(t)} - \tilde{g}_i^{(t)}) = \sum_{i\in E_t}(g_i^{(t)} - \tilde{g}_i^{(t)})$. We define $\hat{v} := (1/n)\sum_{i\in G} x_i z_i$, $u_t^{(1)} := (1/n)\sum_{i\in E_t} x_i x_i^\top (w_t - w^*)$, $u_t^{(2)} := (1/n)\sum_{i\in E_t} -x_i z_i$, and $u_t^{(3)} := (1/n)\sum_{i\in S_{\text{bad}}\cup E_t}\tilde{g}_i^{(t)}$.

We can further write the update rule as:

$$w_{t+1} - w^* = \hat{B}(w_t - w^*) + \eta\hat{v} + \eta u_{t-1}^{(1)} + \eta u_{t-1}^{(2)} - \eta\phi_t\nu_t - \eta u_{t-1}^{(3)} \,. \tag{D.31}$$

We bound each term one-by-one. Since $G \subset S_{\text{good}}$ and $|G| = (1-\alpha_{\text{corrupt}})n$, using the resilience property in Eq (5.5), we know

$$\begin{aligned}
\|\Sigma^{-1/2}\hat{v}\| &= (1-\alpha_{\text{corrupt}}) \max_{\|v\|=1} \Sigma^{-1/2} \left\langle v, \frac{1}{(1-\alpha_{\text{corrupt}})n} \sum_{i\in G} x_i z_i \right\rangle \\
&\le (1-\alpha_{\text{corrupt}}) C_2 K^2 \alpha \log^{2a}(1/\alpha)\sigma \qquad \text{(D.32)} \\
&\le C_2 K^2 \alpha \log^{2a}(1/\alpha)\sigma \,. \qquad \text{(D.33)}
\end{aligned}$$

Let $\tilde{\alpha} = |E_t|/n$. By assumption, we know $\tilde{\alpha} \le \alpha$ (which holds for the given dataset due to Lemma D.3.3), and

$$\|\Sigma^{-1/2} u_t^{(1)}\| = \|\Sigma^{-1/2} \frac{1}{n} \sum_{i\in E_t} x_i x_i^\top (w_t - w^*)\| \,.$$

From Corollary D.10.8, we know

$$\begin{aligned}
&\left| \|\Sigma^{-1/2} \frac{1}{|E_t|} \sum_{i\in E_t} x_i x_i^\top (w_t - w^*)\| - \|w_t - w^*\|_\Sigma \right| \\
=&\left| \max_{u:\|u\|=1} \frac{1}{|E_t|} \sum_{i\in E_t} u^\top \Sigma^{-1/2} x_i x_i^\top (w_t - w^*)\| - \max_{v:\|v\|=1} v^\top \Sigma^{1/2}(w_t - w^*) \right| \\
\le& \max_{u:\|u\|=1} \left| \frac{1}{|E_t|} \sum_{i\in E_t} u^\top \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} \Sigma^{1/2}(w_t - w^*)\| - u^\top \Sigma^{1/2}(w_t - w^*) \right| \\
\le& \max_{u:\|u\|=1} \left| \frac{1}{|E_t|} \sum_{i\in E_t} u^\top \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right) \Sigma^{1/2}(w_t - w^*)\| \right| \\
=& \left\| \frac{1}{|E_t|} \sum_{i\in E_t} \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right) \Sigma^{1/2}(w_t - w^*) \right\| \\
\le& \left\| \frac{1}{|E_t|} \sum_{i\in E_t} \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right) \right\| \cdot \left\| \Sigma^{1/2}(w_t - w^*) \right\| \\
\le& \frac{2-\tilde{\alpha}}{\tilde{\alpha}} C_2 K^2 \alpha \log^{2a}(1/\alpha) \left\| w_t - w^* \right\|_\Sigma \,.
\end{aligned}$$

This implies that

$$\begin{aligned}\|\Sigma^{-1/2}u_t^{(1)}\| \;&\leq\; \|\Sigma^{-1/2}\frac{1}{n}\sum_{i\in E}x_i x_i^\top(w_t-w^*)\| \\ &\leq \left(\tilde{\alpha}+2C_2K^2\alpha\log^{2a}(1/\alpha)\right)\|w_t-w^*\|_\Sigma \\ &\leq 3C_2K^2\alpha\log^{2a}(1/\alpha)\,\|w_t-w^*\|_\Sigma\;, \end{aligned} \tag{D.34}$$

where the last inequality follows from the fact that $\tilde{\alpha} \leq \alpha$ and our assumption that $C_2K^2\log^{2a}(1/\bar{\alpha}) \geq 1$ from Asmp. 7. Similarly, we use resilience property in Eq (5.5) instead of Eq (5.6), we can show that

$$\|\Sigma^{-1/2}u_t^{(2)}\| \leq 3C_2K^2\alpha\log^{2a}(1/\alpha)\sigma\;. \tag{D.35}$$

Next, we consider $u_t^{(3)}$. Since $|S_{\text{bad}}| \leq \alpha_{\text{corrupt}}n$ and $|E_t| \leq \alpha n$, using Eq (5.8) and Corollary D.10.8, we have

$$\begin{aligned}\|\Sigma^{-1/2}u_t^{(3)}\| &= \max_{v:\|v\|=1}\frac{1}{n}\sum_{i\in S_{\text{bad}}\cup E_t}v^\top\Sigma^{-1/2}x_i\text{clip}_{\theta_t}(x_i^\top w_t-y_i) \\ &\leq 2C_2K\alpha\log^a(1/\alpha)\theta_t \\ &\leq 6C_2^{1.5}K^2\alpha\log^{2a}(1/\alpha)(\|w_t-w^*\|_\Sigma+\sigma)\;. \end{aligned} \tag{D.36}$$

Now we use Eq (D.33), Eq (D.34), Eq (D.35) and Eq (D.36) to bound the final error from update rule in Eq (D.31).

Step 3: Analysis of the t-steps recurrence relation. We have controlled the deterministic noise in the last step. In this step, we will upper bound the noise introduced by the Gaussian noise for the purpose of privacy, and show the expected distance to optimum decrease every step.

We want to emphasize that most of our technical contribution is in the convergence analysis (Step 3 and Step 4). More precisely, naive linear regression analysis can only show a suboptimal error rate of $\|\hat{w}-w^\star\|_\Sigma = \tilde{O}(\kappa\alpha\sigma)$ with sample size $n = \tilde{O}(d/\alpha^2+\kappa^{1/2}d/(\varepsilon\alpha))$.

Define $u_t = (\hat{v} + u_t^{(1)} + u_t^{(2)} - u_t^{(3)})$. This follows from Eq (D.31):

$$w_{t+1} - w^* = \hat{B}(w_t - w^*) + \eta u_t - \eta \phi_t \nu_t \tag{D.37}$$

$$= (\mathbf{I}_d - \eta\hat{\Sigma})(w_t - w^*) + \eta u_t - \eta \phi_t \nu_t \ . \tag{D.38}$$

From Eq (D.34), Eq (D.35) and Eq (D.36), it follows that

$$\|w_{t+1} - w^*\|_\Sigma \leq (1 - \frac{1}{\kappa})\|w_t - w^*\|_\Sigma + \alpha(\sigma + \|w_t - w^*\|_\Sigma)$$

where we omitted constants for simplicity, which after $T = \tilde{O}(\kappa)$ iterations achieves a *sub-optimal* error rate $\|w_T - w^*\|_\Sigma = \tilde{O}(\kappa\alpha\sigma)$.

One attempt to get around it is to take the Euclidean norm instead, which gives, after some calculations,

$$\mathbb{E}[\|w_{t+1} - w^*\|^2] \leq \mathbb{E}[\|w_t - w^*\|^2] - \eta\Big(\|w_t - w^*\|_\Sigma^2 - \alpha^2\sigma^2\Big) \ .$$

This implies that $\mathbb{E}[\|w_{t+1} - w^*\|^2]$ strictly decreases as long as $\|w_t - w^*\|_\Sigma^2 > C\alpha^2\sigma^2$, which is the desired statistical error level we are targeting. With this analysis, we can show that in $T = \tilde{O}(\kappa)$ iterations, there exists at least one model w_t that achieves $\mathbb{E}[\|w_t - w^*\|_\Sigma^2] = \tilde{O}(\alpha^2\sigma^2)$ among all the intermediate models we have seen.

However, the problem is that under differential privacy, there is no way we could select this good model w_t among T models that we have, as privacy-preserving techniques for model selection are not accurate enough to achieve the desired level of accuracy. Hence, we came up with the following novel analysis that does not suffer from such issues.

We can rewrite Eq (D.31) or Eq (D.37) as

$$w_{t+1} - w^* = \hat{B}(w_t - w^*) + \eta u_t - \eta \phi_t \nu_t \tag{D.39}$$

$$= \hat{B}^{t+1}(w_0 - w^*) + \eta \sum_{i=0}^{t} \hat{B}^i u_{t-i} - \eta \sum_{i=0}^{t} \phi_{t-i} \hat{B}^i \nu_{t-i} \ . \tag{D.40}$$

Taking expectations of $\hat{\Sigma}$-norm square with respect to $\nu_1, \cdots, \nu_t$, we have

$$\mathbb{E}_{\nu_1,\ldots,\nu_t\sim\mathcal{N}(0,\mathbf{I}_d)}\|w_{t+1}-w^*\|_{\hat{\Sigma}}^2 \tag{D.41}$$

$$\leq\ 2\|\hat{B}^{t+1}(w_0-w^*)\|_{\hat{\Sigma}}^2+2\mathbb{E}[\|\eta\sum_{i=0}^{t}\hat{B}^i u_{t-i}\|_{\hat{\Sigma}}^2]+\eta^2\sum_{i=0}^{t}\mathrm{Tr}(\hat{B}^{2i}\hat{\Sigma})\mathbb{E}[\phi_{t-i}^2] \tag{D.42}$$

$$\leq\ 2\|\hat{B}^{t+1}(w_0-w^*)\|_{\hat{\Sigma}}^2+2\eta^2\mathbb{E}[\sum_{i=0}^{t}\sum_{j=0}^{t}\|\hat{B}^i u_{t-i}\|_{\hat{\Sigma}}\|\hat{B}^j u_{t-j}\|_{\hat{\Sigma}}] \tag{D.43}$$

$$+\eta^2\sum_{i=0}^{t}\mathrm{Tr}(\hat{B}^{2i}\hat{\Sigma})\mathbb{E}[\phi_{t-i}^2]\ , \tag{D.44}$$

where at the second step we used the fact that $\nu_1, \nu_2, \cdots, \nu_t$ are independent isotropic Gaussian.

Note that

$$\begin{aligned}
\eta\|\hat{B}^i u_{t-i}\|_{\hat{\Sigma}} &= \eta\|\hat{\Sigma}^{1/2}\hat{B}^i\hat{\Sigma}^{1/2}\hat{\Sigma}^{-1/2}u_{t-i}\| \\
&\leq \eta\|\hat{\Sigma}^{1/2}\hat{B}^i\hat{\Sigma}^{1/2}\|_2\cdot\|\hat{\Sigma}^{-1/2}u_{t-i}\| \\
&\leq \eta\|\hat{\Sigma}^{1/2}\hat{B}^i\hat{\Sigma}^{1/2}\|_2\,\hat{\rho}(\alpha)\,(\|w_{t-i}-w^*\|_{\hat{\Sigma}}+\sigma) \\
&\leq \frac{1}{i+1}\hat{\rho}(\alpha)\,(\|w_{t-i}-w^*\|_{\hat{\Sigma}}+\sigma)\ ,
\end{aligned}$$

where $\hat{\rho}(\alpha) = 1.1(6C_2+6C_2^{1.5})K^2\alpha\log^{2a}(1/\alpha)$, and the second inequality follows from Eq (D.34), Eq (D.35), Eq (D.36) and the deterministic condition in Eq (D.28). Note that the last inequality is true because $\eta\leq 1/(1.1\lambda_{\max})$ and $\|\hat{\Sigma}^{1/2}\hat{B}^i\hat{\Sigma}^{1/2}\|_2\leq\|\mathbf{I}_d-\eta\hat{\Sigma}\|_2^i\|\hat{\Sigma}\|_2\leq\lambda_{\max}/(i+1)$.

This implies

$$\mathbb{E}[\eta^2\sum_{i=0}^{t}\sum_{j=0}^{t}\|\hat{B}^i u_{t-i}\|_{\hat{\Sigma}}\|\hat{B}^j u_{t-j}\|_{\hat{\Sigma}}] \tag{D.45}$$

$$\leq\ 4\,\mathbb{E}[\sum_{i=0}^{t}\sum_{j=0}^{t}\frac{\hat{\rho}(\alpha)^2}{(i+1)(j+1)}(\mathbb{E}[\|w_{t-i}-w^*\|_{\hat{\Sigma}}^2]+\mathbb{E}[\|w_{t-j}-w^*\|_{\hat{\Sigma}}^2]+\sigma^2) \tag{D.46}$$

$$\leq\ 8(\sum_{i=0}^{t}\frac{1}{i+1})^2\hat{\rho}(\alpha)^2(\max_i\mathbb{E}[\|w_{t-i}-w^*\|_{\hat{\Sigma}}^2]+\sigma^2) \tag{D.47}$$

$$\leq\ 8(\log t)^2\hat{\rho}(\alpha)^2(\max_i\mathbb{E}[\|w_{t-i}-w^*\|_{\hat{\Sigma}}^2]+\sigma^2)\ , \tag{D.48}$$

Then,

$$\|\hat{B}^{t+1}(w_0 - w^*)\|_{\hat{\Sigma}}^2 = \|\hat{\Sigma}^{1/2}\hat{B}^{t+1}\hat{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}(w_0 - w^*)\|^2$$
$$\leq (1 - \frac{1}{\kappa})^{2(t+1)}\|w_0 - w^*\|_{\hat{\Sigma}}^2 \leq e^{-2(t+1)/\kappa}\|w_0 - w^*\|_{\hat{\Sigma}}^2 ,$$

and for $n \gtrsim (1/\varepsilon)\sqrt{\kappa d \log(1/\delta)/\alpha}$,

$$\eta^2 \sum_{i=0}^{t} \mathrm{Tr}(\hat{B}^{2i}\hat{\Sigma})\mathbb{E}[\phi_{t-i}^2] \tag{D.49}$$

$$\leq \eta^2 \sum_{i=0}^{t} \|\mathbf{I}_d - \eta\hat{\Sigma}\|_2^{2i}\|\hat{\Sigma}\|_2 \cdot \frac{2\log(1.25/\delta_0)K^2\,\mathrm{Tr}(\Sigma)\log^{2a}(n/\zeta_0)C_2K^2\log^{2a}(1/(2\alpha))(\mathbb{E}[\|w_{t-i} - w^*\|_{\Sigma}^2] + \sigma^2)}{\varepsilon_0^2 n^2} \tag{D.50}$$

$$\leq 4\sum_{i=0}^{t}(\frac{1}{i+1})^2\hat{\rho}(\alpha)^2(\mathbb{E}[\|w_{t-i} - w^*\|_{\hat{\Sigma}}^2] + \sigma^2) . \tag{D.51}$$

We have

$$\mathbb{E}_{\nu_1,\ldots,\nu_t\sim\mathcal{N}(0,\mathbf{I}_d)}[\|w_{t+1} - w^*\|_{\hat{\Sigma}}^2] \leq 2e^{-2(t+1)/\kappa}\|w_0 - w^*\|_{\hat{\Sigma}}^2 + 20(\log t)^2\hat{\rho}(\alpha)^2(\max_{i\in[t]}\mathbb{E}[\|w_{t-i} - w^*\|_{\hat{\Sigma}}^2] + \sigma^2) .$$

Note that this also implies that

$$\mathbb{E}[\|(w_{t'+t} - w^*)\|_{\hat{\Sigma}}^2|w_{t'}] \leq 2e^{-2t/\kappa}\|w_{t'} - w^*\|_{\hat{\Sigma}}^2 + 20\hat{\rho}(\alpha)^2\sum_{i=0}^{t-1}(\frac{1}{i+1})^2(\mathbb{E}[\|w_{t'+t-i} - w^*\|_{\hat{\Sigma}}^2|w_{t'}] + \sigma^2) , \tag{D.52}$$

which implies

$$\mathbb{E}[\|(w_{t'+t} - w^*)\|_{\hat{\Sigma}}^2] \leq 2e^{-2t/\kappa}\mathbb{E}[\|w_{t'} - w^*\|_{\hat{\Sigma}}^2] + 20\hat{\rho}(\alpha)^2\sum_{i=0}^{t-1}(\frac{1}{i+1})^2(\mathbb{E}[\|w_{t'+t-i} - w^*\|_{\hat{\Sigma}}^2] + \sigma^2) \tag{D.53}$$

$$\leq 2e^{-2t/\kappa}\mathbb{E}[\|w_{t'} - w^*\|_{\hat{\Sigma}}^2] + 20(\log t)^2\hat{\rho}(\alpha)^2(\max_{i\in[t]}\mathbb{E}[\|w_{t'+t-i} - w^*\|_{\hat{\Sigma}}^2] + \sigma^2) \tag{D.54}$$

Step 4: End-to-end analysis of the convergence. In the last step, we shown that the amount of estimation error decrease depends on the estimation error of the previous t

steps. In order for the estimation error to decrease by a constant factor, we will take $t = \kappa$. Roughly speaking, we will prove that for every κ steps, the estimation error will decrease by a constant factor, if it is much larger than $O((\log\kappa)^2\hat{\rho}(\alpha)^2\sigma^2)$. This implies we will reach $O((\log\kappa)^2\hat{\rho}(\alpha)^2\sigma^2)$ error with in $\tilde{O}(\kappa)$ steps.

For any integer $s \geq 0$, as long as $\max_{i\in[(s-1)\kappa+1,s\kappa]} \mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2] \geq 2(\log\kappa)^2\hat{\rho}(\alpha)^2\sigma^2$,

$$\max_{i\in[s\kappa+1,(s+1)\kappa]} \mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2] \leq (\frac{1}{e^2} + (\log\kappa)^2\hat{\rho}(\alpha)^2) \max_{i\in[(s-1)\kappa+1,s\kappa]} \mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2] + (\log 2\kappa)^2\hat{\rho}(\alpha)^2\sigma^2 \quad \text{(D.55)}$$

Assuming $\hat{\rho}(\alpha)^2(\log\kappa)^2 \leq 1/2 - 1/e^2$, the maximum expected error in a length κ sequence decrease by a factor of $1/2$ every time.

Now we bound the maximum expected error in the first length κ sequence: $\max_{i\in[0,\kappa-1]} \mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2]$. Since

$$\mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2] \leq e^{-2i/\kappa}\|w_0 - w^*\|_{\hat{\Sigma}}^2 + (\log i)^2\hat{\rho}(\alpha)^2 \max_{j\in[0,i-1]} \mathbb{E}[\|w_j - w^*\|_{\hat{\Sigma}}^2] + (\log i)^2\hat{\rho}(\alpha)^2\sigma^2 \ .$$

As a function of i, $\max_{j\in[0,i-1]} \mathbb{E}[\|w_j - w^*\|_{\hat{\Sigma}}^2]$ only increase when it is smaller than

$$\frac{1}{1 - (\log i)^2\hat{\rho}(\alpha)^2}(\|w_0 - w^*\|_{\hat{\Sigma}}^2 + (\log i)^2\hat{\rho}(\alpha)^2\sigma^2) \ .$$

Thus we conclude

$$\max_{i\in[0,\kappa-1]} \mathbb{E}[\|w_i - w^*\|_{\hat{\Sigma}}^2] \leq \frac{1}{1 - (\log\kappa)^2\hat{\rho}(\alpha^2)}(\|w_0 - w^*\|_{\hat{\Sigma}}^2 + (\log\kappa)^2\hat{\rho}(\alpha^2)\sigma^2)$$

$s = \log(\|w^*\|/(\hat{\rho}(\alpha)\sigma))$ will give us

$$\mathbb{E}[\|w_{s\kappa+1} - w^*\|_{\hat{\Sigma}}^2] \leq (\log\kappa)^2\hat{\rho}(\alpha)^2\sigma^2 \ .$$

□

D.9 Lower bounds

D.9.1 Proof of Proposition 5.3.8 for label corruption lower bounds

We first prove the following lemma.

Lemma D.9.1. *Consider an α label-corrupted dataset $S = \{(x_i, y_i)\}_{i=1}^n$ with $\alpha < 1/2$, that is generated from either $x_i \sim \mathcal{N}(0,1), y_i \sim \mathcal{N}(0,1)$ or $x_i \sim \mathcal{N}(0,1), z_i \sim \mathcal{N}(0, 1-\alpha^2), y_i = \alpha x_i + z_i$. It is impossible to distinguish the two hypotheses with probability larger than $1/2$.*

In the first case,

$$(x_i, y_i) \sim \mathcal{P}_1 = \mathcal{N}(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}).$$

In the second case,

$$(x_i, y_i) \sim \mathcal{P}_2 = \mathcal{N}(0, \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}).$$

By simple calculation, it holds that $D_{KL}(\mathcal{P}_1 || \mathcal{P}_2) = -\frac{1}{2}\log(1-\alpha^2) \leq \alpha^2/2$ for all $\alpha < 1/2$. Then, Pinsker's inequality implies that $D_{TV}(\mathcal{P}_1 || \mathcal{P}_2) \leq \alpha/2$. Since the covariate x_i follows from the same distribution in the two cases, and the total variation distance between the two cases is less than $\alpha/2$. This means there is an label corruption adversary that change $\alpha/2$ fraction of y_i's in $\mathcal{P}_1$ to make it identical to $\mathcal{P}_2$. Therefore, no algorithm can distinguish the two cases with probability better than $1/2$ under α fraction of label corruption.

Since $\Sigma = 1$, $\sigma^2 \in [3/4, 1]$, the first case above has $w^* = 0$, and the second case has $w^* = \alpha$, this implies that no algorithm is able to achieve $\mathbb{E}[\|\hat{w} - w^*\|_\Sigma] < \sigma\alpha$ for all instances with $\|w^*\| \leq 1$ under α fraction of label corruption.

D.10 Technical Lemmas

Lemma D.10.1 (Hanson-Wright inequality for subWeibull distributions [178])**.** *Let $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ be a dataset consist of i.i.d. samples from (K,a)-subWeibull distributions, then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - \mathrm{Tr}(\Sigma)\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{nt^2}{K^4(\mathrm{Tr}(\Sigma))^2}, \left(\frac{nt}{K^2\,\mathrm{Tr}(\Sigma)}\right)^{\frac{1}{2a}}\right\}\right) . \quad \text{(D.56)}$$

Lemma D.10.2. *Let $Y \sim \mathrm{Lap}(b)$. Then for all $h > 0$, we have $\mathbb{P}(|Y| \geq hb) = e^{-h}$.*

Lemma D.10.3. *If $x \in \mathbb{R}^d$ is (K,a)-subWeibull for some $a \in [1/2, \infty)$. Then*

- *for any fixed* $v \in \mathbb{R}^d$*, with probability* $1-\zeta$*,*

$$\langle x, v\rangle^2 \le K^2 v^\top \Sigma v \log^{2a}(1/\zeta)\ . \tag{D.57}$$

- *with probability* $1-\zeta$*,*

$$\|x\|^2 \le K^2 \operatorname{Tr}(\Sigma) \log^{2a}(1/\zeta)\ . \tag{D.58}$$

We provide a proof in App. D.10.1.1.

Lemma D.10.4. *Dataset* $S = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ *consists i.i.d. samples from a zero mean distribution* $\mathcal{D}$*. Suppose* $\mathcal{D}$ *is* (K,a)*-subWeibull. Define* $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}[xx^\top]$*. Then there exists a constant* $c_1 > 0$ *such that with probability* $1-\zeta$*,*

$$\left\| \frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \Sigma \right\| \le c_1 \left(\frac{K^2 d \log(d/\zeta)\log^{2a}(n/\zeta)}{n} + \sqrt{\frac{K^2 d \log(d/\delta)\log^{2a}(n/\zeta)}{n}} \right) \|\Sigma\|_2\ . \tag{D.59}$$

Lemma D.10.5 (Lemma F.1 from [159])**.** *Let* $x \in \mathbb{R}^d \sim \mathcal{N}(0,\Sigma)$*. Then there exists universal constant* C_6 *such that with probability* $1-\zeta$*,*

$$\|x\|^2 \le C \operatorname{Tr}(\Sigma) \log(1/\zeta)\ . \tag{D.60}$$

Definition D.10.6 (Corrupt good set)**.** *We say a dataset* S *is* $(\alpha_{\text{corrupt}}, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$*-corrupt good with respect to* (w^*, Σ, σ) *if it is* α_{corrupt}*-corruption of an* $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$*-resilient dataset* S_{good}*.*

Lemma D.10.7. *Under Assumptions 8 and 7, there exists positive constants* c_1 *and* C_2 *such that if* $n \ge c_1((d+\log(1/\zeta))/\alpha^2$*, then with probability* $1-\zeta$*,* S_{good} *is, with respect to* (w^*, Σ, σ)*,* $(\alpha, C_2K^2\alpha\log^{2a}(1/\alpha), C_2K^2\alpha\log^{2a}(1/\alpha), C_2K^2\alpha\log^{2a}(1/\alpha), C_2K\alpha\log^{a}(1/\alpha))$*-resilient.*

We provide a proof in App. D.7.

Corollary D.10.8 (Lemma 10 from [185] and Lemma 25 from [161]). *For a $(\alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-resilient set S with respect to (w^*, Σ, γ) and any $0 \le \tilde{\alpha} \le \alpha$, the following holds for any subset $T \subset S$ of size at least $\tilde{\alpha}n$ and for any unit vector $v \in \mathbb{R}^d$:*

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\langle v, x_i\rangle(y_i - x_i^\top w^*)\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_1\sqrt{v^\top\Sigma v}\,\sigma\,, \tag{D.61}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle^2 - v^\top\Sigma v\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_2 v^\top\Sigma v\,, \tag{D.62}$$

$$\left|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}(y_i - x_i^\top w^*)^2 - \sigma^2\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_3\,\sigma^2\,, \quad \textit{and} \tag{D.63}$$

$$\left|\frac{1}{|T|}\sum_{x_i\in T}\langle v, x_i\rangle\right| \le \frac{2-\tilde{\alpha}}{\tilde{\alpha}}\rho_4\sqrt{v^\top\Sigma v}\,. \tag{D.64}$$

D.10.1 Proof of technical lemmas

D.10.1.1 Proof of Lemma D.10.3

Using Markov inequality,

$$\mathbb{P}\left(\langle v, x\rangle^2 \ge t^2\right) = \mathbb{P}\left(e^{\langle v,x\rangle^{1/a}} \ge e^{t^{1/a}}\right) \tag{D.65}$$

$$\le e^{-t^{1/a}}\mathbb{E}[e^{\langle v,x\rangle^{1/a}}] \tag{D.66}$$

$$\le e^{-t^{1/a}} e^{K(\mathbb{E}[\langle v,x\rangle^2])^{1/(2a)}} \tag{D.67}$$

$$= 2\exp\Big(-\Big(\frac{t^2}{K^2\mathbb{E}[\langle v,x\rangle^2]}\Big)^{1/(2a)}\Big)\,. \tag{D.68}$$

This implies for any fixed v, with probability $1-\zeta$,

$$\langle x, v\rangle^2 \le K^2 v^\top\mathbb{E}[xx^\top]v\log^{2a}(1/\zeta)\,. \tag{D.69}$$

For j-th coordinate, let $v = e_j$ where $j \in [d]$. Definition D.8.1 implies

$$\mathbb{E}\left[\exp\left(\left(\frac{x_j^2}{K^2\,\mathrm{Tr}(\Sigma)}\right)^{1/(2a)}\right)\right] \le \mathbb{E}\left[\exp\left(\left(\frac{x_j^2}{K^2\Sigma_{jj}}\right)^{1/(2a)}\right)\right] \le 2\,. \tag{D.70}$$

Note that $f(x) = x^\alpha$ is concave function for $\alpha \leq 1$ and $x > 0$. Then $(a_1 + \cdots a_k)^\alpha \leq a_1^\alpha + \cdots a_k^\alpha$ holds for any positive numbers $a_1, \cdots, a_k > 0$. By our assumption that $1/(2a) \leq 1$. , we have

$$\mathbb{E}[\exp\left(\left(\frac{\|x\|^2}{K^2 \operatorname{Tr}(\Sigma)}\right)^{1/(2a)}\right)] = \mathbb{E}[\exp\left(\left(\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{K^2 \operatorname{Tr}(\Sigma)}\right)^{1/(2a)}\right)] \tag{D.71}$$

$$\leq \mathbb{E}[\prod_{j=1}^{d} \exp\left(\left(\frac{x_j^2}{K^2 \operatorname{Tr}(\Sigma)}\right)^{1/(2a)}\right)] \tag{D.72}$$

$$\leq \left(\frac{\sum_{j=1}^{d} \mathbb{E}[\exp\left(\left(\frac{x_j^2}{K^2 \operatorname{Tr}(\Sigma)}\right)^{1/(2a)}\right)]}{d}\right)^d \tag{D.73}$$

$$\leq 2 \, . \tag{D.74}$$

By Markov inequality,

$$\mathbb{P}\left(\|x\| \geq t\right) = \mathbb{P}\left(e^{\|x\|^{1/a}} \geq e^{t^{1/a}}\right) \tag{D.75}$$

$$\leq e^{-t^{1/a}} \mathbb{E}[e^{\|x\|^{1/a}}] \tag{D.76}$$

$$\leq \exp\left(-\left(\frac{t^2}{K^2 \operatorname{Tr}(\Sigma)}\right)^{1/(2a)}\right) \, . \tag{D.77}$$

This implies with probability $1 - \zeta$,

$$\|x\|^2 \leq K^2 \operatorname{Tr}(\Sigma) \log^{2a}(1/\zeta) \, . \tag{D.78}$$

D.11 Experiments

D.11.1 DP Linear Regression

Experimental results for $\epsilon = 0.1$ can be found in Figure D.1. The observations are similar to the $\epsilon = 1$ case. In particular, DP-SSP has poor performance when σ is small. In other settings, DP-SSP has better performance than DP-RobGD.

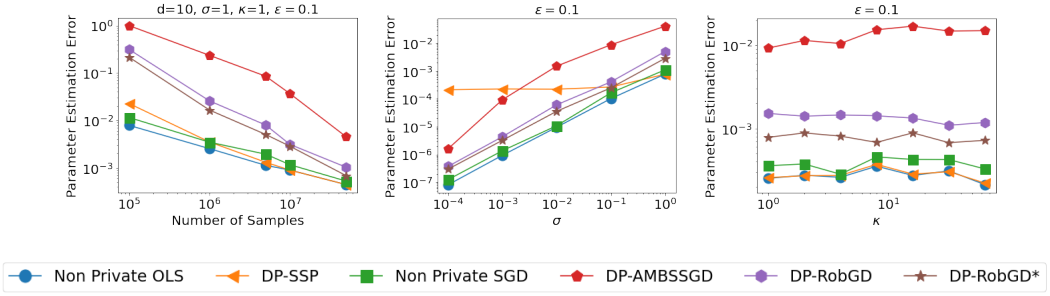

Figure D.1: Performance of various techniques on DP linear regression. $d = 10$ in all the experiments. $n = 10^7, \kappa = 1$ in the 2^{nd} experiment. $n = 10^7, \sigma = 1$ in the 3^{rd} experiment.

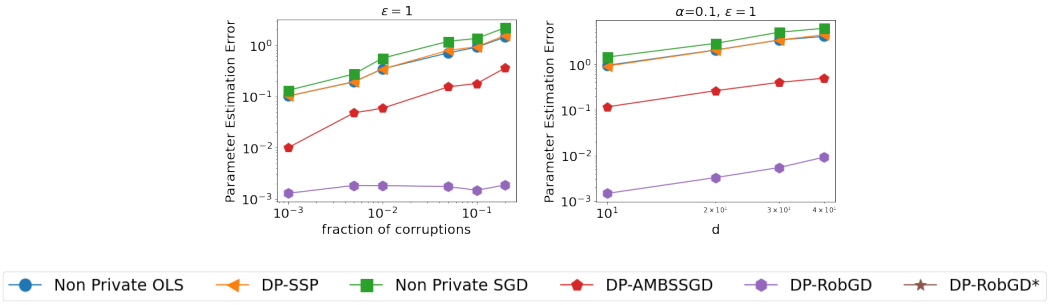

Figure D.2: Non-robustness of existing techniques to adversarial corruptions. $n = 10^7$, $\sigma = 1$ in both experiments.

D.11.2 DP Robust Linear Regression

We now illustrate the robustness of our algorithm. We consider the same experimental setup as in Sec. 5.4 and randomly corrupt α fraction of the response variables by setting them to 1000. Figure D.2 presents the results from this experiment. It can be seen that none of the baselines are robust to adversarial corruptions. They can be made arbitrarily bad by increasing the magnitude of corruptions. In contrast, DP-RobGD is able to handle the corruptions well.

D.11.3 Stronger adversary for DP Robust Linear Regression

In this section, we consider a stronger adversary for DP-RobGD than the one considered in Sec. 5.4. Recall, for the adversary model considered in Sec. 5.4, DP-RobGD was able to consistently estimate the parameter w^* (*i.e.,* the parameter recovery error goes down to 0 as $n \to \infty$). This is because the algorithm was able to easily identify the corruptions and ignore the corresponding points while performing gradient descent. We now construct a different instance where the corruptions are hard to identify. Consequently, DP-RobGD can no longer be consistent against the adversary. This hard instance is inspired by the lower bound in [22] (see Theorem 6.1 of [22]). This is a 2 dimensional problem where the first covariate is sampled uniformly from $[-1, 1]$. The second covariate, which is uncorrelated from the first, is sampled from a distribution with the following pdf

$$p(x^{(2)}) = \begin{cases} \frac{\alpha}{2} & \text{if } x^{(2)} \in \{-1, 1\} \\ \frac{1-\alpha}{2\alpha\sigma} & \text{if } x^{(2)} \in [-\sigma, \sigma] \\ 0 & \text{otherwise} \end{cases} .$$

We set $\sigma = 0.1$ in our experiments. The noise z_i is sampled uniformly from $[-\sigma, \sigma]$. We consider two possible parameter vectors $w^* = (1, 1)$ and $w^* = (1, -1)$. It can be shown that the total variation (TV) distance between these problem instances (each parameter vector corresponds to one problem instance) is $\Theta(\alpha)$ [22]. What this implies is that, one can corrupt at most α fraction of the response variables and convert one problem instance into another. Since the distance (in Σ norm) between the two parameter vectors is $\Omega(\alpha\sigma)$, any algorithm will suffer an error of $\Omega(\alpha\sigma)$.

We generate 10^7 samples from this problem instance and add corruptions that convert one problem instance to the other. Figure D.3 presents the results from this experiment. It can be seen that our algorithm works as expected. In particular, it is not consistent in this setting. Moreover, the parameter recovery error increases with the fraction of corruptions.

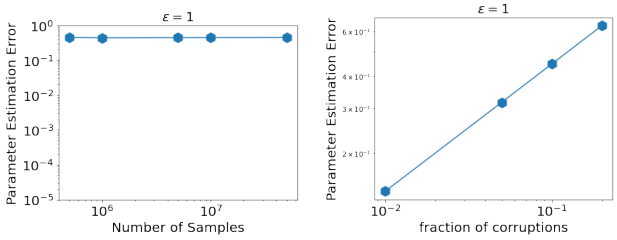

Figure D.3: Performance against the stronger adversary

D.12 Heavy-tailed noise

We study the heavy-tailed regression settings where the label noise z_i is hypercontractive, which is common in robust linear regression literature [143, 161]. We define (κ_2, k)-hypercontractivity as follows. This is a heavy-tailed distribution we have bound only up to the k-th moment.

Definition D.12.1. *For integer $k \geq 4$, a distribution $P_{\mu,\Sigma}$ is (κ_2, k)-hypercontractive if for all $v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim P_X}[|\langle v, (x-\mu)\rangle|^k] \leq \kappa_2^k (v^\top \Sigma v)^{k/2}$, where Σ is the covariance.*

We give a formal description of our setting in Asmp. 9. Note that we consider the input vector x_i to be sub-Weibull and label noise z_i to be hypercontractive. If both x_i and z_i are hypercontractive, the uncorrupted set S_{good} is known to be not resilient [217, 161]. However, by [217, Lemma G.10], we can clip x_i by $O(\sqrt{d}\|\Sigma\|_2)$, and obtain a $(\alpha, O(\kappa\alpha^{1-1/k}), O(\kappa\alpha^{1-2/k}), O(\kappa\alpha^{1-2/k}), O(\kappa\alpha^{1-1/k}))$-resilient set [161, Lemma 4.19]. This would result in sub-optimal error rate $\tilde{O}(\kappa\alpha^{1-2/k})$, which depends on condition number κ. For convenience, in this section, we further assume that x_i and z_i are independent. In the dependent case, the only thing we need to change is the ρ_1 resilience from $O(\alpha^{1-1/k})$ to $O(\alpha^{1-2/k})$ in Lemma D.12.2. This would result in $O(\alpha^{1-3/k})$ error rate if we plug this new resilience in Thm. D.12.3.

Assumption 9 ($(\Sigma, \sigma^2, w^*, K, a, \kappa_2, k)$-model)**.** *A multiset $S_{\text{good}} = \{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}_{i=1}^n$ of n i.i.d. samples is from a linear model $y_i = \langle x_i, w^*\rangle + z_i$, where the input vector x_i is zero mean, $\mathbb{E}[x_i] = 0$, with a positive definite covariance $\Sigma := \mathbb{E}[x_i x_i^\top] \succ 0$, and the independent label noise z_i is zero mean, $\mathbb{E}[z_i] = 0$, with variance $\sigma^2 := \mathbb{E}[z_i^2]$. We assume that the marginal*

distribution of x_i is (K, a)-sub-Weibull and that of z_i is (κ_2, k)-hypercontractive, as defined above.

This is similar to the light-tailed case in Asmp. D.8.1. The main difference is that the noise z_i is heavy-tailed and independent of the input x_i.

Assumption 10 (α_{corrupt}-corruption)**.** *Given a dataset $S_{\text{good}} = \{(x_i, y_i)\}_{i=1}^n$, an adversary inspects all the data points, selects $\alpha_{\text{corrupt}} n$ data points denoted as S_r, and replaces the labels with arbitrary labels while keeping the covariates unchanged. We let S_{bad} denote this set of $\alpha_{\text{corrupt}} n$ newly labelled examples by the adversary. Let the resulting set be $S := S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$. We further assume that the corruption rate is bounded by $\alpha_{\text{corrupt}} \leq \bar{\alpha}$, where $\bar{\alpha}$ is a positive constant that depends on κ_2, k, K, $\log(\kappa)$, a and ζ.*

Compared to Asmp. 7, this only difference is in the conditions on $\bar{\alpha}$. Similar as Lemma D.10.7, we have the following lemma showing that under Asmp. 9, the uncorrupted dataset can S_{good} is corrupt-good, which means that it can be seen as being corrupted from a resilient set. We provide the proof in App. D.12.2.

Lemma D.12.2. *A multiset of i.i.d. labeled samples $S_{\text{good}} = \{(x_i, y_i)\}_{i=1}^n$ is generated from a linear model: $y_i = \langle x_i, w^* \rangle + z_i$, where feature vector x_i has zero mean and covariance $\mathbb{E}[x_i x_i^\top] = \Sigma \succ 0$, independent label noise z_i has zero mean and covariance $\mathbb{E}[z_i^2] = \sigma^2 > 0$. Suppose x_i is (K, a)-sub-Weibull, z_i is (κ_2, k)-hypercontractive, then there exist constants $c_1, C_2 > 0$ such that, for any $0 < \alpha \leq \tilde{\alpha} \leq c$ where $c \in (0, 1/2)$ is some absolute constant if*

$$n \geq c_1 \left(\frac{d}{\zeta^{2(1-1/k)} \alpha^{2(1-1/k)}} + \frac{k^2 \alpha^{2-2/k} d \log d}{\zeta^{2-4/k} \kappa_2^2} + \frac{\kappa_2^2 d \log d}{\alpha^{2/k}} + \frac{d + \log(1/\zeta)}{\tilde{\alpha}^2} \right) , \tag{D.79}$$

then with probability $1 - \zeta$, S_{good} is $(0.2\alpha, \alpha, C_2 k (ka)^a K \kappa_2 \alpha^{1-1/k} \zeta^{-1/k}, C_2 K^2 \tilde{\alpha} \log^{2a}(1/\tilde{\alpha}), C_2 k^2 \kappa_2^2 \alpha^{1-2/k} \zeta^{-2/k}, C_2 K \tilde{\alpha} \log^a(1/\tilde{\alpha}))$-corrupt good with respect to (w^*, Σ, σ).

In the rest of this section, we assume we have a $(O(\alpha), \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set under Asmp. 9 and present following algorithm and our main theorem under this setting in Thm. D.12.3. We also provide the proof in App. D.12.1.

Algorithm 25: Robust and Private Linear Regression for heavy-tailed noise

Input: dataset $S = \{(x_i, y_i)\}_{i=1}^{3n}$, (ε, δ), T, learning rate η, failure probability ζ, target error rate α, distribution parameter (K, a)

Partition dataset S into three equal sized disjoint subsets $S = S_1 \cup S_2 \cup S_3$.

$\delta_0 \leftarrow \delta/(2T)$, $\varepsilon_0 \leftarrow \varepsilon/(4\sqrt{T\log(1/\delta_0)})$, $\zeta_0 \leftarrow \zeta/3$, $w_0 \leftarrow 0$

$\Gamma \leftarrow \text{PrivateNormEstimator}(S_1, \varepsilon_0, \delta_0, \zeta_0)$, $\Theta \leftarrow K\sqrt{2\Gamma}\log^a(n/\zeta_0)$

for $t = 1, 2, \ldots, T-1$ **do**

  $\gamma_t \leftarrow \text{RobustPrivateDistanceEstimator}(S_2, w_t, \varepsilon_0, \delta_0, \alpha, \zeta_0)$

  $\theta_t \leftarrow 2\sqrt{2\gamma_t} \cdot \sqrt{\max\{8\rho_2/\alpha, 8\rho_3/\alpha\} + 1}$.

  Sample $\nu_t \sim \mathcal{N}(0, \mathbf{I}_d)$

  $w_{t+1} \leftarrow w_t - \eta\left(\frac{1}{n}\sum_{i\in S_3}\left(\text{clip}_\Theta(x_i)\text{clip}_{\theta_t}\left(w_t^\top x_i - y_i\right)\right) + \frac{\sqrt{2\log(1.25/\delta_0)}\Theta\theta_t}{\varepsilon_0 n}\cdot\nu_t\right)$

Return w_T

Theorem D.12.3. *Alg. 25 is (ε, δ)-DP. Under $(\Sigma, \sigma^2, w^*, K, a, \kappa_2, k)$-model of Asmp. 9 and α_{corrupt}-corruption of Assumption 10 and for any failure probability $\zeta \in (0,1)$ and target error rate $\alpha \geq 1.2\alpha_{\text{corrupt}}$, if the dataset S is $(0.2\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to (w^*, Σ, σ) and sample size is large enough such that*

$$n = O\left(K^2 d\log(d/\zeta)\log^{2a}(n/\zeta) + \frac{K^2 dT^{1/2}\log(T/\delta)\log^a(n/(\alpha\zeta))\sqrt{8\max\{\rho_2/\alpha, \rho_3/\alpha\}+1}}{\varepsilon\hat{\rho}(\alpha)}\right), \tag{D.80}$$

where $\hat{\rho}(\alpha) = \max\{\rho_1, 3\rho_2, 2\rho_4\sqrt{8\max\{\rho_2/\alpha, \rho_3/\alpha\}+1}\}$, then the choices of a small enough step size, $\eta \leq 1/(1.1\lambda_{\max}(\Sigma))$, and the number of iterations, $T = \tilde{\Theta}\left(\kappa\log\left(\|w^*\|\right)\right)$ for a condition number of the covariance $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, ensures that, with probability $1-\zeta$, Alg. 13 achieves

$$\mathbb{E}_{\nu_1,\cdots,\nu_t\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\,\|w_T - w^*\|_\Sigma^2\,\right] \;=\; \tilde{O}\left(\hat{\rho}^2(\alpha)\sigma^2\right), \tag{D.81}$$

where the expectation is taken over the noise added for DP, and $\tilde{\Theta}(\cdot)$ hides logarithmic terms in $K, \kappa_2, \sigma, d, n, 1/\varepsilon, \log(1/\delta), 1/\alpha$, and κ.

By Lemma D.12.2, if we set $\tilde{\alpha} = \alpha^{1-1/k}$, $\rho_1 = C_2 k (ka)^a K \kappa_2 \alpha^{1-1/k} \zeta^{-1/k}$, $\rho_2 = C_2 K^2 \alpha^{1-1/k} \log^{2a}(1/\alpha^{1-1/k})$ $C_2 k^2 \kappa_2^2 \alpha^{1-2/k} \zeta^{-2/k}$, and $\rho_4 = C_2 K \alpha^{1-1/k} \log^a(1/\alpha^{1-1/k})$, we have following corollary.

Corollary D.12.4. *Under the same hypotheses of Thm. D.12.3 and under α_{corrupt}-corruption model of Asmp. 10, if $1.2\alpha_{\text{corrupt}} \le \alpha$ and $K, a, \kappa_2, k = O(1)$, then $n = \tilde{O}(d/(\zeta^{2-2/k}\alpha^{2-2/k}) + \kappa^{1/2} d \log(1/\delta)/(\varepsilon \alpha^{1-1/k}))$ samples are sufficient for Alg. 25 to achieve an error rate of $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma^2 = \tilde{O}(\zeta^{-2/k}\alpha^{2-4/k})$ with probability $1-\zeta$, where $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, $\tilde{O}(\cdot)$ hides logarithmic terms in $\sigma, d, n, 1/\varepsilon, \log(1/\delta), \log(1/\zeta)$ and κ.*

Simiarly, if we set $\tilde{\alpha} = \alpha$, $\rho_1 = C_2 k (ka)^a K \kappa_2 \alpha^{1-1/k} \zeta^{-1/k}$, $\rho_2 = C_2 K^2 \alpha \log^{2a}(1/\alpha)$,$\rho_3 = C_2 k^2 \kappa_2^2 \alpha^{1-2/k} \zeta^{-2/k}$, and $\rho_4 = C_2 K \alpha \log^a(1/\alpha)$, we have following corollary.

Corollary D.12.5. *Under the same hypotheses of Thm. D.12.3 and under α_{corrupt}-corruption model of Asmp. 10, if $1.2\alpha_{\text{corrupt}} \le \alpha$ and $K, a, \kappa_2, k = O(1)$, then $n = \tilde{O}(d/(\zeta^{2-2/k}\alpha^{2-2/k}) + \kappa^{1/2} d \log(1/\delta)/(\varepsilon\alpha) + (d + \log(1/\zeta)/\alpha^2))$ samples are sufficient for Alg. 25 to achieve an error rate of $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma^2 = \tilde{O}(\zeta^{-2/k}\alpha^{2-2/k})$ with probability $1-\zeta$, where $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, $\tilde{O}(\cdot)$ hides logarithmic terms in $\sigma, d, n, 1/\varepsilon, \log(1/\delta), \log(1/\zeta)$ and κ.*

As a comparison, we also apply the exponential-time robust linear regression algorithm HPTR by [161] under our setting.

Theorem D.12.6 ([161, Theorem 12])**.** *There exist positive constants c and C such that for any $((2/11)\alpha, \alpha, \rho_1, \rho_2, \rho_3, \rho_4)$-corrupt good set S with respect to $(w^*, \Sigma \succ 0, \sigma > 0)$ satisfying $\alpha < c$, $\rho_1 < c$, $\rho_2 < c$, $\rho_3 < c$,and $\rho_4^2 \le c\alpha$,* HPTR *achieves $(1/\sigma)\|(\hat{\beta} - \beta)\|_\Sigma \le 32\rho_1$ with probability $1-\zeta$, if*

$$n \;\ge\; C\,\frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha}\,. \tag{D.82}$$

We set $\tilde{\alpha} = \alpha^{1-1/k}$, $\rho_1 = C_2 k (ka)^a K \kappa_2 \alpha^{1-1/k} \zeta^{-1/k}$, $\rho_2 = C_2 K^2 \alpha^{1-1/k} \log^{2a}(1/\alpha^{1-1/k})$,$\rho_3 = C_2 k^2 \kappa_2^2 \alpha^{1-2/k} \zeta^{-2/k}$, and $\rho_4 = C_2 K \alpha^{1-1/k} \log^a(1/\alpha^{1-1/k})$, we have the following utility gaurentees.

Corollary D.12.7. *Under the hypothesis of Asmp. 9, there exists a constant $c > 0$ such that for any $\alpha \le c$, $(ka)^a K\kappa_2\alpha^{1-1/k}\zeta^{-1/k} \le c$, $k^2\kappa_2^2\alpha^{1-2/k}\zeta^{-2/k} \le c$ and $K^2\alpha^{1-2/k}\log^{2a}(1/\alpha^{1-1/k}) \le c$, it is sufficient to have a dataset of size*

$$n = O\Big(\frac{d}{\zeta^{2(1-1/k)}\alpha^{2(1-1/k)}} + \frac{k^2\alpha^{2-2/k}d\log d}{\zeta^{2-4/k}\kappa_2^2} + \frac{\kappa_2^2 d\log d}{\alpha^{2/k}}\Big)\,, \tag{D.83}$$

such that HPTR *achieves $(1/\sigma)\|\hat{w} - w^*\|_\Sigma = O(k(ka)^a K\kappa_2\alpha^{1-1/k}\zeta^{-1/k})$ with probability $1-\zeta$.*

Note that both of our result in Corollary D.12.4 and Corollary D.12.5 are suboptimal compared to the exponential time algorithm HPTR from Corollary D.12.7. Suppose $K, a, \kappa_2, k, \zeta = \Theta(1)$, HPTR achieves $(1/\sigma)\|w^* - \hat{w}\| = \tilde{O}(\alpha^{1-1/k})$ with sample complexities $n = d/(\alpha^{2(1-1/k)}) + (d + \log(1/\delta))/(\varepsilon n)$. However, in the analysis in Corollary D.12.4, Alg. 25 achieves $(1/\sigma)\|w^* - \hat{w}\| = \tilde{O}(\alpha^{1-2/k})$ with the same sample complexities. In the analysis in Corollary D.12.5, Alg. 25 achieves the same error rate as HPTR but requires extra $\tilde{O}(d/\alpha^2)$ sample complexities. The suboptimality is caused by the gradient truncation step in our algorithm. From Thm. D.12.6, the final error rate of HPTR only depends on the first resilience ρ_1. However in Thm. D.12.3, the final error rate of Alg. 25 depends on $\hat{\rho}(\alpha) = \max\{\rho_1, \rho_2, \rho_4\sqrt{\rho_2/\alpha}\}$. When the noise is heavy-tailed, the bottleneck is the last term $\rho_4\sqrt{\rho_2/\alpha} \approx \alpha^{1-2/k}$, which is due to the truncation threshold from Eq (D.93). This cannot be tightened by using a smaller truncation threshold. Because we can construct y_i, such that there are α-fraction of points that are at the threshold level $\theta_t \approx \alpha^{-1/k}$(line 6 of Alg. 25). If exponential time complexity is allowed, we could robustly and privately estimate the average of the gradients by directly estimating the $x_i y_i$. However, the current best efficient algorithm [160] for estimating the mean of Gaussian with unknown covariance robustly and privately would require $O(d^{1.5})$ samples.

For a fair comparison, we also rewrite the error rates of Corollary D.12.4, Corollary D.12.5, Corollary D.12.7 as the same accuracy level α and different corruption level α_{corrupt} respectively.

Corollary D.12.8. *Under the same hypotheses of Thm. D.12.3 and under α_{corrupt}-corruption model of Asmp. 10, if $1.2\alpha_{\text{corrupt}} \le \alpha^{k/(k-2)}$ and $K, a, \kappa_2, k = O(1)$, then*

$$n = \tilde{O}(d/(\zeta^{2-2/k}\alpha^{2(k-1)/(k-2)}) + \kappa^{1/2}d\log(1/\delta)/(\varepsilon\alpha^{(k-1)/(k-2)}))$$

samples are sufficient for Alg. 25 to achieve an error rate of $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma^2 = \tilde{O}(\zeta^{-2/k}\alpha^2)$ *with probability* $1-\zeta$*, where* $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$*,* $\tilde{O}(\cdot)$ *hides logarithmic terms in* $\sigma, d, n, 1/\varepsilon, \log(1/\delta), \log(1/$ *and* κ*.*

Corollary D.12.9. *Under the same hypotheses of Thm. D.12.3 and under* α_{corrupt}*-corruption model of Asmp. 10, if* $1.2\alpha_{\text{corrupt}} \le \alpha^{k/(k-1)}$ *and* $K, a, \kappa_2, k = O(1)$*, then*

$$n = \tilde{O}(d/(\zeta^{2-2/k}\alpha^2) + \kappa^{1/2} d \log(1/\delta)/(\varepsilon\alpha^{k/(k-1)}) + (d + \log(1/\zeta)/\alpha^{2k/(k-1)}))$$

samples are sufficient for Alg. 25 to achieve an error rate of $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma^2 = \tilde{O}(\zeta^{-2/k}\alpha^2)$ *with probability* $1-\zeta$*, where* $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$*,* $\tilde{O}(\cdot)$ *hides logarithmic terms in* $\sigma, d, n, 1/\varepsilon, \log(1/\delta), \log(1/$ *and* κ*.*

Corollary D.12.10 (HPTR)**.** *Under the same hypotheses of Thm. D.12.3 and under* α_{corrupt}*-corruption model of Asmp. 10, if* $\alpha_{\text{corrupt}} \le \alpha^{k/(k-1)}$ *and* $\alpha^{(k-2)/(k-1)} \le c$ *and* $K, a, \kappa_2, k = O(1)$*, then*

$$n = \tilde{O}(\frac{d}{\zeta^{2-2/k}\alpha^2} + \frac{d + \log(1/(\delta\zeta))}{\varepsilon\alpha^{k/k-1}})$$

samples are sufficient for HPTR to achieve an error rate of $(1/\sigma^2)\|\hat{w} - w^*\|_\Sigma^2 = \tilde{O}(\zeta^{-2/k}\alpha^2)$ *with probability* $1-\zeta$*,* $\tilde{O}(\cdot)$ *hides logarithmic terms in* $\sigma, d, n, 1/\varepsilon, \log(1/\delta), \log(1/\zeta)$ *and* κ*.*

D.12.1 Proof of Thm. D.12.3

Proof. The proof follows similarly as the proof of Thm. D.8.2. We only highlight the difference in the proof.

Let S_{good} be the uncorrupted dataset for S_3 and S_{bad} be the corrupted data points in S_3. Let G denote the clean data that satisfies resilience conditions. We know $|G| \ge (1 - 1.2\alpha_{\text{corrupt}})n \ge (1-\alpha)n$.

Let $\lambda_{\max} = \|\Sigma\|_2$. Define $\hat{\Sigma} := (1/n)\sum_{i\in G} x_i x_i^\top$, $\hat{B} := \mathbf{I}_d - \eta\hat{\Sigma}$. Lemma D.10.4 implies that if $n = O(K^2 d \log(d/\zeta)\log^{2a}(n/\zeta))$, then

$$0.9\Sigma \preceq \hat{\Sigma} \preceq 1.1\Sigma\,. \tag{D.84}$$

We pick step size η such that $\eta \leq 1/(1.1\lambda_{\max})$ to ensure that $\eta \leq 1/\|\hat{\Sigma}\|_2$. Since the covariates $\{x_i\}_{i\in S}$ are not corrupted, from Lemma D.10.3, we know with probability $1-\zeta$, for all $i \in S_3$,

$$\|x_i\|^2 \leq K^2 \operatorname{Tr}(\Sigma) \log^{2a}(n/\zeta) \; . \tag{D.85}$$

The rest of the proof is under Eq (D.84), Eq (D.85) and the resilience conditions.

Let $\phi_t = (\sqrt{2\log(1.25/\delta_0)}\Theta\theta_t)/(\varepsilon_0 n)$. Define $g_i^{(t)} := x_i(x_i^\top w_t - y_i)$. For $i \in S_{\text{good}}$, we know $y_i = x_i^\top w^* + z_i$. Let $\tilde{g}_i^{(t)} = \text{clip}_\Theta(x_i)\text{clip}_{\theta_t}(x_i^\top w_t - y_i)$. Note that under Eq (D.85), $\text{clip}_\Theta(x_i) = x_i$ for all $i \in S_3$.

From Alg. 25, we can write one-step update rule as follows:

$$\begin{aligned}
&w_{t+1} - w^* \\
=&w_t - \eta\left(\frac{1}{n}\sum_{i\in S}\tilde{g}_i^{(t)} + \phi_t\nu_t\right) - w^* \\
=&\left(\mathbf{I} - \frac{\eta}{n}\sum_{i\in G} x_i x_i^\top\right)(w_t - w^*) + \frac{\eta}{n}\sum_{i\in G} x_i z_i + \frac{\eta}{n}\sum_{i\in G}(g_i^{(t)} - \tilde{g}_i^{(t)}) - \eta\phi_t\nu_t - \frac{\eta}{n}\sum_{i\in S_3\setminus G\cup E_t}\tilde{g}_i^{(t)}
\end{aligned} \tag{D.86}$$

Let $E_t := \{i \in G : \theta_t \leq |x_i^\top w_t - y_i|\}$ be the set of clipped clean data points such that $\sum_{i\in G}(g_i^{(t)} - \tilde{g}_i^{(t)}) = \sum_{i\in E_t}(g_i^{(t)} - \tilde{g}_i^{(t)})$. We define $\hat{v} := (1/n)\sum_{i\in G} x_i z_i$, $u_t^{(1)} := (1/n)\sum_{i\in E_t} x_i x_i^\top (w_t - w^*)$, $u_t^{(2)} := (1/n)\sum_{i\in E_t} -x_i z_i$, and $u_t^{(3)} := (1/n)\sum_{i\in S_3\setminus G\cup E_t}\tilde{g}_i^{(t)}$.

We can further write the update rule as:

$$w_{t+1} - w^* = \hat{B}(w_t - w^*) + \eta\hat{v} + \eta u_{t-1}^{(1)} + \eta u_{t-1}^{(2)} - \eta\phi_t\nu_t - \eta u_{t-1}^{(3)} \; . \tag{D.87}$$

Since $G \subset S_{\text{good}}$ and $|G| \geq (1-\alpha)n$, using the resilience property in Eq (5.5), we know

$$\begin{aligned}
\|\Sigma^{-1/2}\hat{v}\| &= |G| \max_{\|v\|=1} \Sigma^{-1/2}\left\langle v, \frac{1}{|G|}\sum_{i\in G} x_i z_i\right\rangle \\
&\leq (1-\alpha)\rho_1\sigma &\text{(D.88)} \\
&\leq \rho_1\sigma \; . &\text{(D.89)}
\end{aligned}$$

Let $\alpha_2 = |E_t|/n$. Following the proof of Lemma D.3.3, we can show following lemma.

Lemma D.12.11. *Under Assumptions 9, if* $\theta_t \geq \sqrt{\max\{8\rho_2/\alpha, 8\rho_3/\alpha\}+1}\cdot(\|w^* - w_t\|_\Sigma + \sigma)$, *then*

$$\left|\left\{i \in G : \left|w_t^\top x_i - y_i\right| \geq \theta_t\right\}\right| \leq \alpha n$$

, *for all* $t \in [T]$.

Similar as Thm. D.3.1, we have following theorem.

Theorem D.12.12. *Alg. 23 is* $(\varepsilon_0, \delta_0)$*-DP. For an* $(\alpha_{\text{corrupt}}, \bar{\alpha}, \rho_1, \rho_2, \rho_3, \rho_4)$*-corrupted good dataset* S_2 *and an upper bound* $\bar{\alpha}$ *on* α_{corrupt} *that satisfy Asmp. 9 and* $\rho_1 + \rho_2 + \rho_3 \leq 1/4$, *for any* $\zeta \in (0,1)$, *if*

$$n \;=\; O\left(\frac{\log(1/\zeta)\log(1/(\delta_0\zeta))}{\bar{\alpha}\varepsilon_0}\right)\,, \tag{D.90}$$

with a large enough constant then, with probability $1-\zeta$, *Alg. 23 returns* ℓ *such that* $\frac{1}{4}(\|w_t - w^*\|_\Sigma^2 + \sigma^2) \;\leq\; \ell \;\leq\; 4(\|w_t - w^*\|_\Sigma^2 + \sigma^2)$.

This means $\alpha_2 \leq \alpha$, and we have

$$\|\Sigma^{-1/2} u_t^{(1)}\| = \|\Sigma^{-1/2}\frac{1}{n}\sum_{i\in E_t} x_i x_i^\top (w_t - w^*)\|\,.$$

From Corollary D.10.8, we know

$$
\begin{aligned}
&\left| \|\Sigma^{-1/2}\frac{1}{|E_t|}\sum_{i\in E_t} x_i x_i^\top (w_t - w^*)\| - \|w_t - w^*\|_\Sigma \right| \\
=&\left| \max_{u:\|u\|=1} \frac{1}{|E_t|}\sum_{i\in E_t} u^\top \Sigma^{-1/2} x_i x_i^\top (w_t - w^*)\| - \max_{v:\|v\|=1} v^\top \Sigma^{1/2}(w_t - w^*) \right| \\
\le& \max_{u:\|u\|=1} \left| \frac{1}{|E_t|}\sum_{i\in E_t} u^\top \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2}\Sigma^{1/2}(w_t - w^*)\| - u^\top \Sigma^{1/2}(w_t - w^*) \right| \\
\le& \max_{u:\|u\|=1} \left| \frac{1}{|E_t|}\sum_{i\in E_t} u^\top \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right)\Sigma^{1/2}(w_t - w^*)\| \right| \\
=& \left\| \frac{1}{|E_t|}\sum_{i\in E_t} \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right)\Sigma^{1/2}(w_t - w^*) \right\| \\
\le& \left\| \frac{1}{|E_t|}\sum_{i\in E_t} \left(\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} - \mathbf{I}_d\right) \right\| \cdot \left\|\Sigma^{1/2}(w_t - w^*)\right\| \\
\le& \frac{2-\alpha_2}{\alpha_2}\rho_2 \left\|w_t - w^*\right\|_\Sigma \ .
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\|\Sigma^{-1/2} u_t^{(1)}\| \ &\le \ \|\Sigma^{-1/2}\frac{1}{n}\sum_{i\in E} x_i x_i^\top (w_t - w^*)\| \\
&\le (\alpha_2 + 2\rho_2)\left\|w_t - w^*\right\|_\Sigma \\
&\le 3\rho_2 \left\|w_t - w^*\right\|_\Sigma \ , \qquad \text{(D.91)}
\end{aligned}
$$

where the last inequality follows from the fact that $\alpha_2 \le \alpha$ and our assumption that $\alpha \le \rho_2$ from Asmp. 10. Similarly, we use resilience property in Eq (5.5) instead of Eq (5.6), we can show that

$$
\|\Sigma^{-1/2} u_t^{(2)}\| \le 3\rho_3\sigma \ . \qquad \text{(D.92)}
$$

Next, we consider $u_t^{(3)}$. Since $|S_3 \setminus G| \le 1.2\alpha_{\text{corrupt}} n$ and $|E_t| \le \alpha n$, using Eq (5.8) and

Corollary D.10.8, we have

$$\begin{aligned}\|\Sigma^{-1/2}u_t^{(3)}\| &= \max_{v:\|v\|=1}\frac{1}{n}\sum_{i\in S_{\text{bad}}\cup E_t} v^\top\Sigma^{-1/2}x_i\text{clip}_{\theta_t}(x_i^\top w_t - y_i)\\ &\le 2\rho_4\theta_t\\ &\le 2\rho_4\sqrt{8\max\{\rho_2/\alpha,\rho_3/\alpha\}+1}\cdot(\|w_t-w^*\|_\Sigma+\sigma)\,. \end{aligned} \tag{D.93}$$

The analysis of convergence follows similarly as in Step 3 and Step 4 of the proof of Thm. D.8.2 except we set $\hat{\rho}(\alpha) = \max\{\rho_1, 3\rho_2, 2\rho_4\sqrt{8\max\{\rho_2/\alpha,\rho_3/\alpha\}+1}\}$.

The second term in Eq (D.80) ensures the added Gaussian noise is small enough such that $\phi_t^2\|v_t\|^2 \le \hat{\rho}(\alpha)^2(\mathbb{E}[\|w_t-w^*\|_\Sigma^2]+\sigma^2)$, which is similar as in Eq (D.51)

□

D.12.2 Proof of Lemma D.12.2

Proof. For any x that is (K,a)-sub-Weibull from Definition D.8.1, Eq (D.68) implies that for any $k \ge 1$,

$$\mathbb{E}[|\langle v,x\rangle|^k] = \int_0^\infty \mathbb{P}(|\langle v,x\rangle| \ge t^{1/k})dt \tag{D.94}$$

$$\le \int_0^\infty 2\exp\left(-\frac{t^{\frac{1}{ka}}}{(K^2\mathbb{E}[\langle v,x\rangle^2])^{\frac{1}{2a}}}\right)dt \tag{D.95}$$

$$= 2K^k(\mathbb{E}[\langle v,x\rangle^2])^{k/2}ka\int_0^\infty e^{-u}u^{ka-1}du \tag{D.96}$$

$$= 2K^k(\mathbb{E}[\langle v,x\rangle^2])^{k/2}\Gamma(ka+1) \tag{D.97}$$

$$\le 2K^k(\mathbb{E}[\langle v,x\rangle^2])^{k/2}(ka)^{ka} \tag{D.98}$$

This implies that x_i is also $((ka)^aK,k)$-hypercontractive. Since x_i and z_i are independent, we have

$$\mathbb{E}\left[\left|\left\langle v,\sigma^{-1}\Sigma^{-1/2}x_iz_i\right\rangle\right|^k\right] = \mathbb{E}\left[\left|\left\langle v,\Sigma^{-1/2}x_i\right\rangle\right|^k\right]\mathbb{E}\left[\left|\sigma^{-1}z_i\right|^k\right] \le 2(ka)^{ka}K^k\kappa_2^k\,. \tag{D.99}$$

This means x_iz_i is also $((ka)^aK\kappa_2,k)$-hypercontractive. From [217, Lemma G.10], we know with probability $1-\zeta$, there exists $S_1 \subset S_{\text{good}}$ with $|S_1| \ge (1-0.1\alpha)|S_{\text{good}}|$, such that for

any $T \subset S_1$ with $|T| \geq (1-\alpha)|S_1|$, we have

$$\Big|\frac{1}{|T|}\sum_{(x_i,y_i)\in S}\big\langle v, \sigma^{-1}\Sigma^{-1/2}x_i(y_i - x_i^\top w^*)\big\rangle\Big| \leq C_2 k(ka)^a K\kappa_2\alpha^{1-1/k}\zeta^{-1/k}\,. \tag{D.100}$$

Similarly, there exists $S_2 \subset S_{\text{good}}$ with $|S_2| \geq (1-0.1\alpha)|S_{\text{good}}|$, such that for any $T \subset S_2$ with $|T| \geq (1-\alpha)|S_2|$, we have

$$\Big|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}(\sigma^{-1}(y_i - x_i^\top w^*))^2 - 1\Big| \leq C_2 k^2\kappa_2^2\alpha^{1-2/k}\zeta^{-2/k}\,. \tag{D.101}$$

From Lemma D.10.7, for any $T \subset S_{\text{good}}$ with $|T| \geq (1-\tilde{\alpha})|S_{\text{good}}|$, we have

$$\Big|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\big\langle v, \Sigma^{-1/2}x_i\big\rangle^2 - 1\Big| \leq C_2 K\tilde{\alpha}\log^{2a}(1/\tilde{\alpha})\,. \tag{D.102}$$

and

$$\Big|\frac{1}{|T|}\sum_{(x_i,y_i)\in T}\big\langle v, \Sigma^{-1/2}x_i\big\rangle\Big| \leq C_2 K\tilde{\alpha}\log^{a}(1/\tilde{\alpha})\,. \tag{D.103}$$

Set $S = S_1 \cap S_2$, we know $|S| \geq (1-0.2\alpha)|S_{\text{good}}|$ and S is $(0.2\alpha, \alpha, C_2k(ka)^aK\kappa_2\alpha^{1-1/k}\zeta^{-1/k}, C_2K^2\tilde{\alpha}\log^{2a}(1/\tilde{\alpha}), C_2k^2\kappa_2^2\alpha^{1-2/k}\zeta^{-2/k}, C_2K\tilde{\alpha}\log^a(1/\tilde{\alpha}))$-corrupt good with respect to (w^*, Σ, σ). This completes the proof. □