Towards Robust and Resilient Machine Learning

Adarsh Prasad

April 2021
CMU-ML-21-103

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee

Sivaraman Balakrishnan (Co-Chair)
Pradeep Ravikumar (Co-Chair)
Larry Wasserman
Sujay Sanghavi

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

This research was sponsored by the National Science Foundation under award numbers OAC-1934584, IIS-1909816, IIS-1955532, DMS1713003, CCF1763734 and the Office of Naval Research via N000141812861.

Keywords: machine learning, robust statistics

Acknowledgments

I would like to thank my PhD advisors, Pradeep Ravikumar and Sivaraman Balakrishnan, for providing me the opportunity to pursue this work. Both Pradeep and Siva gave me exceptional freedom and provided valuable guidance throughout the years. Pradeep taught me to zoom out to ask broader questions, connect the dots and to search for beauty and simplicity in ones arguments. On the other hand, Siva kept me grounded and taught me the importance zooming into a problem, to work out the details, paying close attention to detail, and to keep banging my head against the wall. I am forever in their debt for giving me the confidence to tackle new problems, and I hope to make them proud.

I would also like to thank my committee members Larry Wasserman and Sujay Sanghavi for providing excellent feedback and encouraging me to pursue this direction. Larry taught me Intemediate Statistics and Statistical Machine Learning, and has always surprised me with his uncanny ability to explain complex things in a simple manner, which I hope to emulate. Sujay has also always served as a constant source of inspiration.

I would like to thank Diane Stidle for her instantaneous help with administrative work throughout graduate school. Diane makes sure that the department is always warm and welcoming.

Through my PhD, I have had the pleasure of collaborating with multiple people with diverse skills, which has been a transformative learning experience. I want to thank Alexandru Niculescu-Mizil, Arun Sai Suggala, Saurabh Garg, Vishwak Srinivasan, Emilio Parisotto Josh Zhanson and Ainesh Bakshi.

Also, a big shout-out to past and current members of RAIL for constantly providing inspiration and for serving as role-models - Rashish, Harsh, Tianyang, David, Bryon, Ian-En, Justin, Xun, Arun, Chen, Biswa, Kartik, Leqi, Chih-Kuan, Elan, Vishwak and Bingbin. I learned a lot from them both academically and personally. I enjoyed the company of other PhD students in the department - Avi, Mrinmaya, Mariya, Chris, Anthony, Otilia, Ritesh, Chirag, Paul, Ojash. Special thanks to Avi, Ritesh, Xun and Justin for helping me prepare for jobs.

I would also like to thank my officemates Ifi, Greg, Elan and Jeremy for ensuring that there was never a dull day at work.

The squash group at CMU ensured that I stayed sane and maintained some level of fitness throughout this journey. I would like thank the regular gang - Nirmoho, Alex, Heewook, Vyas, Vikram, Siva for all the fun games.

Lastly and most importantly, I would like to thank my parents, brother and girlfriend for their unconditional love and support.

Abstract

Some common *assumptions* when building machine learning pipelines are: (1) the training data is sufficiently "clean" and well-behaved, so that there are few or no outliers, or that the distribution of the data does not have very long tails, (2) the testing data follows the same distribution as the training data, and (3) the data is generated from or is close to a known model class, such as a linear model or neural network.

However, with easier access to computer, internet and various sensor-based technologies, modern data sets that arise in various branches of science and engineering are no longer carefully curated and are often collected in a decentralized, distributed fashion. Consequently, they are plagued with the complexities of heterogeneity, adversarial manipulations, and outliers. As we enter this *age of dirty data*, the aforementioned assumptions of machine learning pipelines are increasingly indefensible.

For the widespread adoption of Machine Learning, we believe that it is imperative that any model should have the following three basic elements:

- **Robustness:** The model can be trained even with noisy and corrupted data.
- **Reliability:** After training and when deployed in the real-world, the model should not break down under benign shifts of the distribution.
- **Resilience:** The modeling procedure should work under *model mis-specification*, *i.e.* even when the modeling assumption breaks down, the model should find the best possible solution.

In this thesis, our goal is modify state of the art ML techniques and design new algorithms so that they work even without the aforementioned assumptions, and are robust, reliable and resilient. Our contributions are as follows:

In chapter 2, we provide a new class of statisically-optimal estimators that are provably robust to a variety of robustness settings, such as arbitrary contamination, and heavy-tailed data, among oth-

ers. In Chapter 3, we complement our statistical optimal estimators with a new class of computationally-efficient estimators for robust risk minimization. These results provide some of the first computationally tractable and provably robust estimators for general statistical models such linear regression, logistic regression, among others. In Chapter 4, we study the problem of learning Ising models in a setting where some of the samples from the underlying distribution can be arbitrarily corrupted. Finally, in Chapter 5, we discuss implications of our results for modern machine learning.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In classical analyses of statistical estimators, statistical guarantees are derived under strong model assumptions, and in most cases these guarantees hold only in the absence of arbitrary outliers, and other deviations from the model assumptions. Strong model assumptions are rarely met in practice, and this has motivated the development of robust inferential procedures, and which has a rich history in statistics with seminal contributions due to Box [2], Tukey [3], Huber [4], Hampel [5] and several others. These have led to rich statistical concepts such as the influence function, the breakdown point, and the Huber ϵ-contamination model, to assess the robustness of estimators. Despite this progress however, the statistical methods with the strongest robustness guarantees are computationally intractable, for instance those based on non-convex M-estimators [4], ℓ_1 tournaments [6, 7, 8] and notions of depth [9, 10, 11].

In this thesis, we aim to design estimators that applicable to a variety of *notions of robustness.* and in particular, we focus on two canonical robustness settings:

(a) **Robustness to arbitrary outliers:** In this setting, we focus on Huber's ϵ-contamination model, where rather than observe samples directly from P in (3.1) we instead observe samples drawn from P_ϵ which for an *arbitrary* distribution Q is defined as:

$$P_\epsilon = (1 - \epsilon)P + \epsilon Q. \tag{1.1}$$

The distribution Q allows for arbitrary outliers, which may correspond to gross corruptions or more subtle deviations from the assumed model. This model can be equivalently viewed as model mis-specfication in the Total Variation (TV) metric.

(b) **Robustness to heavy-tails:** In this setting, we are interested in developing estimators under weak moment assumptions. We assume that the distribution P from which we obtain samples only has finite low-order moments (see Section 3.5.3 for a precise characterization). Such heavy-tailed distributions arise frequently in the analysis of financial data and large-scale biological datasets (see for instance examples in [12, 13]). In contrast to classical analyses of empirical risk minimization [14], in this setting the empirical risk is not uniformly close to the population risk, and methods that directly minimize the empirical risk perform poorly (see Section 5.4).

1.1 Goals of Thesis and Roadmap.

In this thesis, our primary goal is to study statistical estimation under the aforementioned notions of robustness. In particular, for any given parameter estimation problem, we want to give answers to three different questions:

1. **Information Theoretic Limits:** Note that in the ϵ-contamination model, since, we observe samples from a contaminated distribution P_ϵ, we cannot hope to recover the true underlying parameter. Hence, our first goal is to derive information theoretic limits for different parameter estimation tasks in this contaminated setup. In particular, we focus on studying the dependence of bounds on the contamination level ϵ and its interplay with characteristics of different distribution classes.
2. **Statistically Optimal Estimators:** Having studied the limits of estimation, our next goal is to design statistically optimal estimators, which match the information theoretic limits in the asymptotic regime and also come with optimal dependence on the number of samples, dimension and high-probability bounds.
3. **Computationally Efficient Estimators:** Finally, we also want to design estimators which are computationally efficient and come with provable guarantees.

Next, we outline some of the concrete parameter estimation problems studied in this thesis.

1.1.1 Mean and Covariance Estimation under Bounded 2k-moments.

Bounded $2k$-moment Class. Let x be a random vector with mean μ and covariance Σ. We say that x has bounded $2k$-moments if for all $v \in \mathcal{S}^{p-1}$, $\mathbb{E}[(v^T(x-\mu))^{2k}] \leq C_{2k}\left(\mathbb{E}[(v^T(x-\mu))^2]\right)^k$. We let $\mathcal{P}_{2k}^{\sigma^2}$ be the class of distributions with bounded $2k$ moments with covariance matrix $\Sigma \lesssim \sigma^2 \mathcal{I}_p$.

Informal Question. *Suppose P is a multivariate distribution with mean μ, covariance Σ and bounded $2k$-moments. Then given n-samples from the mixture distribution (1.1), can we design statistically optimal estimators for mean and covariance ?*

We answer the above mentioned question affirmatively in Chapter 2. In particular, we design statistically optimal estimators for mean, covariance and other functionals both in the low and high-dimensional regime. Our workhorse is a novel reduction to univariate estimation, which we leverage to provide a *dimension boosting* based meta-estimator that converts any univariate estimator to the multivariate setting while maintaining its optimality.

1.1.2 Risk Minimization

In the setting of risk minimization, we assume that we have access to a differentiable loss function $\bar{\mathcal{L}} : \Theta \times \mathcal{Z} \mapsto \mathbb{R}$, where Θ is a convex subset of $\mathbb{R}^p$. Let $\mathcal{R}(\theta) = \mathbb{E}_{z\sim P}\left[\bar{\mathcal{L}}(\theta; z)\right]$ be the population loss, also known as the *risk*, and let θ^* be the minimizer of the population risk $\mathcal{R}(\theta)$, over the set Θ:

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{R}(\theta).$$

The goal of risk minimization is to minimize the population risk $\mathcal{R}(\theta)$, given only finite samples in order to estimate the unknown parameter θ^*. In this work we assume that the population risk is convex to ensure tractable minimization. The framework of risk minimization is a central paradigm of statistical estimation and is widely applicable and includes canonical tasks such as linear regression and generalized linear models.

Informal Question. *Given n-samples from a contaminated mixture distribution (1.1), can we design a computationally efficient algorithm which minimizes the population risk robustly and hence recovers the true unknown parameter?*

In Chapter 3, we introduce a new class of robust estimators for risk minimization (3.1), which provide some of the first computationally tractable and provably robust estimators for these canonical statistical models. These estimators are based on robustly estimating gradients of the population risk to then plug in to a projected gradient descent algorithm, and are computationally tractable by design. We provide specific consequences of our theory for linear regression, logistic regression and for canonical parameter estimation in an exponential family.

1.1.3 Ising Models

Consider an undirected graph $G = (V, E)$ defined over a set of vertices $V = \{1, 2, \ldots, p\}$ with edges $E \subset \{(s, t) : s, t \in V, s \neq t\}$. The neighborhood of any node $s \in V$ is the subset $\mathcal{N}(s) \subset V$ given by $\mathcal{N}(s) \overset{\text{def}}{=} \{t | (s, t) \in E\}$, and the degree of any vertex s is given by $d_s = |\mathcal{N}(s)|$. Then, the degree of a graph $d = \max_s d_s$ is the maximum vertex degree, and $k = |E|$ is the total number of edges. We obtain an MRF by associating a random variable X_v at each vertex $v \in V$, and then considering a joint distribution $\mathbb{P}$ over the random vector $(X_1, \ldots, X_p)$. An Ising model is a special instantiation of an MRF where each random variable X_s take values in $\{-1, +1\}$, and the joint probability mass function is given by:

$$\mathbb{P}_\theta(x_1, \ldots, x_p) \propto \exp\left(\sum_{1 \leq s < t \leq p} \theta_{st} x_s x_t\right), \tag{1.2}$$

where we view θ as the parameter vector of the distribution. Note that $\theta \in \mathbb{R}^{p \times p}$ is such that $\theta_{ij} = 0 \Leftrightarrow (i, j) \notin E$ and $\theta = \theta^T$.

Informal Question *Suppose P is an ising model distribution. Then, given n-samples from a contaminated mixture distribution (1.1), can one design an estimator which recovers the true underlying graph?*

In Chapter 4, we give the *first* statistically optimal estimator for learning Ising models under the ϵ-contamination model. Our estimators achieve a dimension-independent asymptotic error as a function of the fraction of outliers ϵ, while simultaneously achieving high probability deviation bounds.

1.1.4 Efficient Heavy-Tailed Estimation.

In the heavy tailed model we observe n samples $x_1, \ldots, x_n$ drawn independently from a distribution P, which is only assumed to have low-order moments be finite (for instance, P might only have finite variance). The goal of past work [15, 16, 17, 18] has been to design an estimator $\widehat{\theta}_n$ of the true mean μ of P which has a small ℓ_2-error with high-probability. Formally, for a given $\delta > 0$, we would like an estimator with minimal r_δ such that,

$$P(\|\widehat{\theta}_n - \mu\|_2 \leq r_\delta) \geq 1 - \delta.$$

As a benchmark for estimators in the heavy-tailed model, we observe that when P is the multivariate normal (or sub-Gaussian) distribution with mean μ and covariance Σ, it can be shown (see Hanson and Wright [19]) that the sample mean $\widehat{\mu}_n = (1/n)\sum_i x_i$ satisfies, with probability at least $1 - \delta$

$$\|\widehat{\mu}_n - \mu\|_2 \lesssim \sqrt{\frac{\operatorname{trace}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}. \tag{1.3}$$

The bound is referred to as a *sub-Gaussian*-style error bound. However, for heavy tailed distributions, it is well-known that the sample mea However, for heavy tailed distributions, as for instance showed in Catoni [15], the sample mean only satisfies the sub-optimal bound $r_\delta = \Omega(\sqrt{\operatorname{trace}(\Sigma)/n\delta})$.

Informal Question. *Suppose P is a distribution with only 4 moments. Then, given n-samples can one design a computationally efficient and practical estimator that achieves a sub-gaussian rate?*

In Chapter 5, we propose and study *practical* estimators that in some cases improve achieve a sub-Gaussian error bound. We use our practical mean estimators to design provably near-optimal algorithms for heavy-tailed linear regression and generalized linear models. We also conduct extensive synthetic experiments which backup our theoretical improvements, and as one consequence of our results, show improvement in training GANs using our algorithms.

Chapter 2

A Robust Univariate Estimator is All You Need

We study the problem of designing estimators when the data has heavy-tails and is corrupted by outliers. In such an adversarial setup, we aim to design statistically optimal estimators for flexible non-parametric distribution classes such as distributions with bounded-2k moments and symmetric distributions. Our primary workhorse is a conceptually simple reduction from multivariate estimation to univariate estimation. Using this reduction, we design estimators which are optimal in both heavy-tailed and contaminated settings. Our estimators achieve an optimal dimension independent bias in the contaminated setting, while also simultaneously achieving high-probability error guarantees with optimal sample complexity. These results provide some of the first such estimators for a broad range of problems including Mean Estimation, Sparse Mean Estimation, Covariance Estimation, Sparse Covariance Estimation and Sparse PCA.

2.1 Introduction

Modern data sets that arise in various branches of science and engineering are characterized by their ever increasing scale and richness. This has been spurred in part by easier access to computer, internet and various sensor-based technologies that enable the automated acquisition of such heterogeneous datasets. On the flip side, these large and rich data-sets are no longer carefully curated, are often collected in a decentralized, distributed fashion, and consequently are plagued with the complexities of heterogeneity, adversarial manipulations, and outliers. The analysis of these huge datasets is thus fraught with methodological challenges.

To understand the fundamental challenges and tradeoffs in handling such "dirty data" is precisely the premise of the field of robust statistics. Here, the aforementioned complexities are largely formalized under two different models of robustness: (1) **The heavy-tailed model:** In this model the sampling distribution can have thick tails, for instance, only low-order moments of the distribution are assumed to be finite; and (2) **The ϵ-contamination model:** Here the sampling distribution is modeled as a well-behaved distribution contaminated by an ϵ fraction of arbitrary outliers. In each case, classical estimators of the distribution (based for instance on the maximum likelihood estimator) can behave considerably worse (potentially arbitrarily worse) than under standard settings where the data is better behaved, satisfying various regularity properties. In particular, these classical estimators can be extremely sensitive to the tails of the distribution or to the outliers and the broad goal in robust statistics is to construct estimators that improve on these classical estimators by reducing their sensitivity to outliers.

Heavy Tailed Model. Concretely, focusing on the fundamental problem of robust mean estimation, in the heavy tailed model we observe n samples $x_1, \ldots, x_n$ drawn independently from a distribution P, which is only assumed to have low-order moments finite (for instance, P only has finite variance). The goal of past work [15, 16, 17, 18] has been to design an estimator $\widehat{\theta}_n$ of the true mean μ of P which has a small ℓ_2-error with high-probability. Formally, for a given $\delta > 0$, we would like an estimator with minimal r_δ such that,

$$P(\|\widehat{\theta}_n - \mu\|_2 \leq r_\delta) \geq 1 - \delta. \tag{2.1}$$

As a benchmark for estimators in the heavy-tailed model, we observe that when P is a multivariate normal distribution (or more generally is a sub-Gaussian distribution) with mean μ and covariance Σ, it can be shown (see [19]) that the sample mean $\widehat{\mu}_n = (1/n)\sum_i x_i$ satisfies, with probability at least $1 - \delta$[1],

$$\|\widehat{\mu}_n - \mu\|_2 \lesssim \sqrt{\frac{\text{trace}\,(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}. \tag{2.2}$$

[1] Here and throughout our paper we use the notation $\lesssim$ to denote an inequality with universal constants dropped for conciseness.

where $\|\Sigma\|_2$ denotes the operator norm of the covariance matrix Σ.

The bound is referred to as a *sub-Gaussian*-style error bound. However, for heavy tailed distributions, as for instance showed in [15], the sample mean only satisfies the sub-optimal bound $r_\delta = \Omega(\sqrt{d/n\delta})$. Somewhat surprisingly, recent work [17] showed that the sub-Gaussian error bound is achievable while *only assuming that P has finite variance*, but by a carefully designed estimator. In the univariate setting, the classical median-of-means estimator [20, 21, 22] and Catoni's M-estimator [15] achieve this surprising result but designing such estimators in the multivariate setting has proved challenging. Estimators that achieve truly sub-Gaussian bounds, but which are computationally intractable, were proposed recently by Lugosi and Mendelson [17] and subsequently Catoni and Giulini [18]. Hopkins [23] and Cherapanamjeri et al. [24] developed a sum-of-squares based relaxation of Lugosi and Mendelson [17]'s estimator, thereby giving a polynomial time algorithm which achieves optimal rates.

Huber's ϵ-Contamination Model. In this setting, instead of observing samples directly from the true distribution P, we observe samples drawn from P_ϵ, which for an arbitrary distribution Q is defined as a mixture model,

$$P_\epsilon = (1-\epsilon)P + \epsilon Q. \tag{2.3}$$

The distribution Q allows one to model arbitrary outliers, which may correspond to gross corruptions, or subtle deviations from the true model. There has been a lot of classical work studying estimators in the ϵ-contamination model under the umbrella of robust statistics (see for instance [25] and references therein). However, most of the estimators come that come with strong guarantees are computationally intractable [3], while others are statistically sub-optimal heuristics [26]. Recently, there has been substantial progress [27, 28, 29, 30, 31, 32] designing provably robust which are computationally tractable while achieving near-optimal contamination dependence (i.e. dependence on the fraction of outliers ϵ) for computing means and covariances. In the Huber model, using information-theoretic lower bounds [28, 33, 34], it can be shown that any estimator must suffer a *non-zero* bias (the asymptotic error as the number of samples go to infinity). For example, for the class of distributions with bounded variance, $\Sigma \precsim \sigma^2 \mathcal{I}_p$, the bias of any estimator is lower bounded by $\Omega(\sigma\sqrt{\epsilon})$. Surprisingly, the optimal bias that can be achieved is often independent of the data dimension. In other words, in many interesting cases optimally robust estimators in Huber's model can tolerate a constant fraction

ϵ of outliers, *independent of the dimension.*

While the aforementioned recent estimators for mean estimation under Huber contamination have a polynomial computational complexity, their corresponding sample complexities are only known to be *polynomial* in the dimension p. For example, Kothari et al. [29] and Hopkins and Li [34] designed estimators which achieve optimal bias for distributions with *certifiably* bounded $2k$-moments, but their statistical sample complexity scales as $O(p^k)$. Steinhardt et al. [35] studied mean estimation and presented an estimator which has a sample complexity of $\Omega\left(p^{1.5}\right)$.

Despite their apparent similarity, developments of estimators that are robust in each of these models, have remained relatively independent. Focusing on mean estimation we notice subtle differences, in the heavy-tailed model our target is the mean of the sampling distribution whereas in the Huber model our target is the mean of the *decontaminated* sampling distribution P. Beyond this distinction, it is also important to note that as highlighted above the natural focus in heavy-tailed mean estimation is on achieving strong, high-probability error guarantees, while in Huber's model the focus has been on achieving dimension independent bias.

Contributions. In this work, we aim to design estimators which are statistically optimally robust in both models simultaneously, *i.e.* they achieve a dimension-independent asymptotic bias in the ϵ-contamination model and achieve high probability deviation bounds similar to (5.3). Our main workhorse is a conceptually simple way of reducing multivariate estimation to the univariate setting. Then, by carefully solving mean estimation in the univariate setting, we are able to design optimal estimators for multivariate mean and covariance estimation for non-parametric distribution classes both in the low-dimensional ($n \geq p$) and high-dimensional ($n < p$) setting. We achieve these rates for non-parametric distribution classes such as distributions with bounded $2k$-moments and the class of symmetric distributions.

2.2 Background and Setup

In this section, we formally define two classes of distributions which we work with in this paper, (1) Bounded-$2k$-Moment distributions and (2) Symmetric Distributions.

Bounded $2k$-moment Class. Let x be a random vector with mean μ and covariance Σ. We say that x has bounded $2k$-moments if for all $v \in \mathcal{S}^{p-1}$, $\mathbb{E}[(v^T(x-\mu))^{2k}] \leq C_{2k} \left(\mathbb{E}[(v^T(x-\mu))^2]\right)^k$. We let $\mathcal{P}_{2k}^{\sigma^2}$ be the class of distributions with bounded $2k$ moments with covariance matrix $\Sigma \lesssim \sigma^2 \mathcal{I}_p$.

Symmetric Distributions. There exist several notions of symmetry for multivariate distributions. We discuss these notions briefly, but refer the reader to [36] for a detailed discussion. A random vector in $\mathbb{R}^p$ is centrally symmetric about $\theta \in \mathbb{R}^p$, if, $x - \theta \overset{\mathrm{d}}{=} \theta - x$, where $\overset{\mathrm{d}}{=}$ denotes *equal in distribution.* Equivalently, this corresponds to $u^T(x-\theta) \overset{\mathrm{d}}{=} u^T(\theta - x)$ for all unit vectors $u \in \mathcal{S}^{p-1}$. Liu [37] introduced the broader notion of angular symmetry, where a random vector $x \in \mathbb{R}^p$ is angularly symmetric about θ, if $\frac{x-\theta}{\|x-\theta\|_2} \overset{\mathrm{d}}{=} \frac{\theta - x}{\|x-\theta\|_2}$, or equivalently, $\frac{x-\theta}{\|x-\theta\|_2}$ is centrally-symmetric. Central symmetry about θ implies angular symmetry about θ (see Lemma 2.2 in [37]).

Halfspace(H)-Symmetry. For any unit vector $u \in \mathcal{S}^{p-1}$, let $H_{u,t} = \{x : u^T x \leq t\}$ be a closed halfspace in $\mathbb{R}^p$. Its interior is an open subspace and the boundary $\{x : u^T x = t\}$ is a hyperplane. Recall that for any random variable $y \in \mathbb{R}$, the *median* of the distribution of y (med(y)) is defined to be any number c such that $\Pr(y \leq c) \geq 0.5, \quad \Pr(y \geq c) \geq 0.5$. Then, a random vector in $\mathbb{R}^p$ is H-symmetric about $\theta \in \mathbb{R}^p$ if, $\Pr(x \in H) \geq \frac{1}{2}$ for all closed halfspaces H with θ on boundary. Note that angular symmetry about a point θ implies halfspace-symmetry about it as well (see Lemma 2.4 [36]). Moreover, if we have that x is H-symmetric about θ, then (1) $\mathrm{med}(u^T x) = u^T\theta$, and (2) $\Pr(u^T(x-\theta) \geq 0) \geq \frac{1}{2}$ for all $u \in \mathcal{S}^{p-1}$ (see Theorem 2.1 [36]). Note that till now, our discussion hasn't required the distribution to have bounded moments, in particular, it need not even have bounded first moments (mean). However, if the distribution has a finite mean, then, it is easy to see that $\mathrm{med}(u^T x) = \mathbb{E}[u^T x] = u^T\theta$. Our last assumption ensures that the median is unique and hence identifiable. To this end, let $\mathcal{P}_{\text{sym}}$ be the class of *H-symmetric* distributions with unique center of H-symmetry. Moreover suppose $\mathcal{P}_{\text{sym}}^{t_0,\kappa} \subset \mathcal{P}_{\text{sym}}$ is the class of distributions such that for any $P \in \mathcal{P}_{\text{sym}}^{t_0,\kappa}$ the CDF of the univariate projection$(u^T P)$ given by $F_{u^T P}$ increases at least linearly around

$u^T\theta$. Formally, for all $x_1 \in [\text{med}(u^TP), F^{-1}_{u^TP}(\frac{1}{2}+t_0)]$ we have that

$$F_{u^TP}(x_1) - \frac{1}{2} \geq \frac{1}{\kappa_{u,P}}(x_1 - \text{med}(u^TP)) \tag{2.4}$$

and for all $x_2 \in [F^{-1}_{u^TP}(\frac{1}{2}-t_0), \text{med}(u^TP)]$, we have that $\frac{1}{2} - F_{u^TP}(x_2) \geq \frac{1}{\kappa_{u,P}}(\text{med}(u^TP) - x_2)$ for $\kappa_{u,P} \leq \kappa$. A higher κ corresponds to slower rate of increase in CDF around the median. Note that κ can be thought of as a measure of variance or dispersion. In particular, for example, in the case of univariate Gaussian distribution, *i.e.* $P = \mathcal{N}(\mu, \sigma^2)$, $\kappa(P) = C\sigma$. Similarly for univariate Cauchy distribution with scale γ, $\kappa(P) \approx C\gamma$. Note that any univariate distribution $P \in \mathcal{P}_{\text{sym}}$ with density function $p(x)$ such that $\min\limits_{|t|<t_0} p(P_F^{-1}(\frac{1}{2}+t)) > \frac{1}{\kappa}$ also belongs to $\mathcal{P}^{t_0,\kappa}_{\text{sym}}$.

2.3 Some Candidate Multivariate Estimators

In this section, we study some natural candidate estimators, to see if they achieve an optimal asymptotic bias in the ϵ-contamination model. We assume that the true distribution is a multivariate isotropic gaussian, $P = \mathcal{N}(0, \mathcal{I}_p)$. Observe that it lies in both $\mathcal{P}^{\sigma^2}_{2k}$ and $\mathcal{P}^{t_0,\kappa}_{\text{sym}}$ for $\sigma^2 = 1$, and $\kappa = O(1)$, hence our results for both distribution classes.

Convex M-estimation. M-estimators were originally proposed by Huber [38], and were shown to be robust in one dimension. Subsequent research in 1970s showed that M-estimators perform poorly for multivariate data [39]. In particular, Donoho and Gasko [40] showed that when the data is p-dimensional, the breakdown point of M-estimators scales inversely with the dimension. Lai et al. [28] and Diakonikolas et al. [27] derived similar negative results for the specific case of geometric median. We further extend this observation, and show that even at a very small contamination level,*i.e.* $\epsilon \mapsto 0$, the bias of certain convex M-estimators which are Fisher-consistent for $\mathcal{N}(0, \mathcal{I}_p)$ will necessarily scale polynomially in the dimension.

Lemma 1. *Let $P = \mathcal{N}(0, \mathcal{I}_p)$ and consider the convex risk $R_P(\theta) = \mathbb{E}_{z\sim P}[\ell(\|z - \theta\|_2)]$ where $\ell : \mathbb{R} \mapsto \mathbb{R}$ be any twice differentiable Fisher-consistent convex loss,* i.e. $\theta(P) = \text{argmin}_\theta R_P(\theta) = 0$. *Then, there exists a corruption Q such that $\lim\limits_{\epsilon\mapsto 0} \|\theta(P_\epsilon)\|_2 \geq \epsilon\sqrt{p}$.*

Recall that when the true distribution $P = \mathcal{N}(0, \mathcal{I}_p)$, then, the lower bound on estimation in the Huber Model is $\Theta(\epsilon)$ [11]. Our explicit dimension-dependent lower bound on the bias of M-estimators shows their sub-optimality.

Subset Search. Having ruled out convex estimation to a certain extent, we next turn our attention to non-convex methods. Perhaps the most simple non-convex method is simple search. Intuitively, the squared loss measures the *fit* between a parameter θ and samples $\mathcal{Z}$, and if all samples don't come from the same distribution(*i.e.* have outliers), then the corresponding *fit* should be bad. To capture this intuition algorithmically, one can (1) consider all subsets of size $\lfloor (1-\epsilon)n \rfloor$, (2) minimize the squared loss over these subsets, and then (3) return the estimator corresponding to the subset with least squared loss or best fit. To be precise, given n samples from P_ϵ

$$\begin{aligned} S^* &\overset{\text{def}}{=} \underset{S \text{ s.t. } |S|=(1-\epsilon)n}{\operatorname{argmin}} \min_\theta \frac{1}{(1-\epsilon)n} \sum_{x_i \in S} \|x_i - \theta\|_2^2 \\ \widehat{\theta}_{\text{SRM}} &\overset{\text{def}}{=} \min_\theta \frac{1}{(1-\epsilon)n} \sum_{x_i \in S^*} \|x_i - \theta\|_2^2 \end{aligned} \tag{2.5}$$

Our next result studies the asymptotic performance of this estimator.

Lemma 2. *Let $P = \mathcal{N}(0, \mathcal{I}_p)$, then as $n \mapsto \infty$, we have that*

$$\sup_Q \|\widehat{\theta}_{SRM} - \mathbb{E}_{x \sim P}[x]\|_2 = \frac{\epsilon}{\sqrt{(1-\epsilon)(1-2\epsilon)}} \sqrt{p}. \tag{2.6}$$

The above result shows that, while subset-search has a finite dimension-independent breakdown point(0.5), the bias of this estimator necessarily scales with the dimension $\sqrt{p}$.

2.4 Optimal Univariate Estimation

In the previous section, we studied some natural candidate estimators and showed that they don't achieve the optimal asymptotic bias in ϵ-contamination model for multivariate mean estimation. In this section, we take a step back, and study univariate estimation.

Algorithm 1 Robust Univariate Mean Estimation

```
function INTERVAL1D($\{z_i\}_{i=1}^{2n}$,CORRUPTION LEVEL $\epsilon$, CONFIDENCE LEVEL $\delta$)
    Split the data into two subsets: $\mathcal{Z}_1 = \{z_i\}_{i=1}^{n}$ and $\mathcal{Z}_2 = \{z_i\}_{i=n+1}^{2n}$.
    Let $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$.
    Using $\mathcal{Z}_1$, let $\hat{I} = [a, b]$ be the shortest interval containing $n(1 - 2\alpha - \sqrt{2\alpha\frac{\log(4/\delta)}{n}} - \frac{\log(4/\delta)}{n})$ points.
    Use $\mathcal{Z}_2$ to identify points lying in $[a, b]$.
    return $\frac{1}{\sum_{i=n}^{2n} \mathbb{I}\left\{z_i \in \hat{I}\right\}} \sum_{i=n}^{2n} z_i \mathbb{I}\left\{z_i \in \hat{I}\right\}$
end function
```

2.4.1 Bounded 2k-moments

We study the interval estimator which was initially proposed by [28]. The estimator, presented in Algorithm 1, proceeds by using half of the samples to identify the shortest interval containing at least $(1-\epsilon)n$ fraction of the points, and then the remaining half of the points is used to return an estimate of the mean.

We assume that the contamination level ϵ and confidence level δ are such that,

$$2\epsilon + \sqrt{\epsilon\frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n} < \frac{1}{2}.$$

Then, we have the following Lemma.

Lemma 3. *Suppose P be any $2k$-moment bounded distribution over $\mathbb{R}$ with mean μ with variance bounded by σ^2. Given, n samples $\{x_i\}_{i=1}^{n}$ from the mixture distribution (2.3), Algorithm 1 returns an estimate $\widehat{\theta}_\delta$ such that with probability at least $1-\delta$,*

$$\begin{aligned} |\widehat{\theta}_\delta - \mu| \lesssim & \sigma \max(2\epsilon, \frac{\log(1/\delta)}{n})^{1-\frac{1}{2k}} \\ & + \sigma(\frac{\log n}{n})^{1-\frac{1}{2k}} + \sigma\sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

Observe that Algorithm 1 has an asymptotic bias of $O(\sigma\epsilon^{1-1/2k})$ in the ϵ-contamination setting, which is known to be information theoretically optimal [28, 34].

Moreover, observe that for $\epsilon = 0$, P has atleast bounded 4th moment, *i.e.* $k \geq 2$, $\frac{\log(n)}{n}^{1-1/2k}$ term can be ignored for large enough n. Hence, for $k \geq 2$

Algorithm 2 Sample Median

```
function SAMPLE MEDIAN - 1D ($\{z_i\}_{i=1}^{2n+1}$)
    Let $z_{[k]}$ be $k^{th}$ order-statistic
    return $z_{[n+1]}$
end function
```

and large enough n, Algorithm 1 achieves the deviation rate of $\sigma\sqrt{\frac{\log(1/\delta)}{n}}$.

2.4.2 Symmetric Distributions

In the univariate setting, our estimator presented in Algorithm 2 simply returns the sample median of the observed samples. While this idea is simple and crucially exploits that the mean and median overlap for a symmetric distribution, this leads to profound implications on the effect of contamination in the Huber contamination model. Next, we present the theoretical bound achieved by this estimator, which was also shown in [41].

We further assume that ϵ and δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}} \leq t_0.$$

Then we have that,

Lemma 4. *Let P be any univariate distribution in $\mathcal{P}_{sym}^{t_0,\kappa}$. Given n-samples from the mixture distribution (2.3), we get that with probability at least $1-\delta$, Algorithm 2 returns an estimate $\widehat{\theta}$ such that,*

$$|\widehat{\theta} - \mathbb{E}_{x\sim P}[x]| \leq C_1\kappa\epsilon + C_2\kappa\sqrt{\frac{\log(1/\delta)}{n}}$$

Observe that Algorithm 2 has an asymptotic bias of $O(\kappa\epsilon)$, which is also information theoretically optimal. To see this, observe that $\mathcal{N}(\cdot,\kappa^2)$ lies in $\mathcal{P}_{\text{sym}}^{t_0,\kappa}$ for some constant t_0 and the fact that $TV(\mathcal{N}(\mu_1,\kappa^2),\mathcal{N}(\mu_2,\kappa^2)) = O(|\mu_1-\mu_2|/\kappa)$(Theorem 1.3 [42]).

2.5 From 1D to p-D: A meta-estimator

In this section, we propose a general meta-estimator to extend any univariate estimator to the multivariate setting. For any univariate estimator $f(\cdot)$, suppose that when given n-samples from the mixture model, it returns an estimate $f(\mathcal{X}_n)$ such that with probability $1-\delta$,

$$|f(\mathcal{X}_n, \epsilon, \delta) - \mu(P)| \leq \omega_f(\epsilon, P, \delta).$$

Note that $\omega_f(\epsilon, P, \delta)$ is the error suffered by the univariate estimator at a contamination level ϵ, and confidence level δ, when the true univariate distribution is P.

2.5.1 Mean Estimation

The proposed meta-estimator proceeds by robustly estimating the mean along almost every direction u, and returns an estimate $\widehat{\theta}$, whose projection along $u(u^T\widehat{\theta})$ is close to these univariate robust mean along that direction. In particular, let $\mathcal{N}^{1/2}(\mathcal{S}^{p-1})$ is the half-cover of the unit sphere $\mathcal{S}^{p-1}$, *i.e.* $\forall u \in \mathcal{S}^{p-1}$, there exists a $y \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})$ such that $u = y + z$ for some $\|z\|_2 \leq \frac{1}{2}$. Then, for any point $\theta \in \mathbb{R}^p$ and *any* univariate estimator $f(\cdot)$, consider the following loss,

$$D_{\mathrm{f}}(\theta, \{x_i\}_{i=1}^n) = \sup_{u \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \mathrm{f}(\{u^Tx_i\}_{i=1}^n, \epsilon, \frac{\delta}{5^p})|,$$

Then, we use it to construct the following multivariate meta-estimator, $\widehat{\theta}_f$ which takes in n-samples $\{x_i\}_{i=1}^n$ and a univariate estimator $f(\cdot)$,

$$\widehat{\theta}_{\mathrm{f}}(\{x_i\}_{i=1}^n) = \operatorname*{argmin}_{\theta} D_f(\theta, \{x_i\}_{i=1}^n),$$

Such directional-control based estimators have been previously studied in the context of heavy-tailed mean estimation by [43] and [18]. Joly et al. [43] used the median-of-means estimator, while Catoni and Giulini [18] used Catoni's M-estimator [15] as their univariate estimator. Then, we have the following result.

Lemma 5. *Suppose P is a multivariate distribution with mean μ. Given n-samples from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_f(\mathcal{X}_n) - \mu\|_2 \lesssim \sup_{u \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})} \omega_f(\epsilon, u^TP, \frac{\delta}{5^p}),$$

where u^TP is the univariate distribution of P along u.

Sparse Mean Estimation. In this setting, we further assume that the true mean vector of the distribution P has only a few non-zero co-ordinates, *i.e.* it is sparse. Such sparsity patterns are known to be present in high-dimensional data(see [44] and references therein). Then, the goal is to design estimators which can exploit this sparsity structure, while remaining robust under the ϵ-contamination model. Formally, for a vector $x \in \mathbb{R}^p$, let $\mathrm{supp}(x) = \{i \in [p] \;\text{ s.t. }\; x(i) \neq 0\}$. Then, x is s-sparse if $|\mathrm{supp}(x)| \leq s$. We further assume that $s \leq p/2$. Let Θ_s be the set of s-sparse vectors in $\mathbb{R}^p$, and let $\mathcal{N}_{2s}^{\frac{1}{2}}(\mathcal{S}^{p-1})$ is the half-cover of the set of unit vectors which are 2s-sparse. Then, for any univariate estimator $f(\cdot)$, let

$$D_{f,s}(\theta, \{x_i\}_{i=1}^n) = \sup_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \mathrm{f}(u^T\mathcal{X}_n, \epsilon, \frac{\delta}{(\frac{6ep}{s})^s})|.$$

Then, we can define the following meta-estimator,

$$\widehat{\theta}_{\mathrm{f,s}}(\mathcal{X}_n) = \operatorname*{argmin}_{\theta \in \Theta_s} D_{f,s}(\theta, \mathcal{X}_n),$$

which has the following error guarantee.

Lemma 6. *Suppose P is a multivariate distribution with mean μ such that μ is s-sparse. Given n-samples from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{f,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \sup_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} \omega_f(\epsilon, u^T P, \frac{\delta}{(\frac{6ep}{s})^s}),$$

where $u^T P$ is the univariate distribution of P along u.

2.5.2 Covariance-Estimation

In this section, we study recovering the true covariance matrix, when given samples from a mixture distribution. We first center our observations by defining pseudo-samples $z_i = \frac{x_i - x_{i+n/2}}{\sqrt{2}}$. We can think of z_i as being sampled from the Huber Contamination $\widetilde{P}_{2\epsilon}$, where $\widetilde{P} = \frac{1}{\sqrt{2}}(P - P)$. Let $\mathcal{Z}_n = \{z_i\}_{i=1}^n$ be the set of these pseudo-samples, and let $u^T\mathcal{Z}_n^{\otimes 2} = \{(u^T z_i)^2\}_{i=1}^n$. Let $\mathcal{F} = \{\Sigma = \Sigma^T \in \mathbb{R}^{p \times p} : \Sigma \succeq 0\}$ be the class of positive semi-definite symmetric matrices.

Then, for any matrix $\Theta \in \mathcal{F}$, let,

$$\mathcal{D}_{\mathrm{f}}^{\otimes 2}(\Theta, \mathcal{Z}_n) = \sup_{u \in \mathcal{N}^{1/4}} |u^T \Theta u - \mathrm{f}(u^T \mathcal{Z}_n^{\otimes 2}, \epsilon, \frac{\delta}{9^p})|$$

Then, consider the meta-estimator $\widehat{\Theta}_{\mathrm{f}}(\mathcal{Z}_n)$ given as,

$$\widehat{\Theta}_{\mathrm{f}}(\mathcal{Z}_n) = \operatorname*{argmin}_{\Theta \in \mathcal{F}} \mathcal{D}_{\mathrm{f}}^{\otimes 2}(\Theta, \mathcal{Z}_n)$$

Lemma 7. *Suppose P is a multivariate distribution with covariance Σ. Given n-samples from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_f(\mathcal{Z}_n) - \Sigma\|_2 \lesssim \sup_{u \in \mathcal{N}^{1/4}(\mathcal{S}^{p-1})} \omega_f(2\epsilon, u^T \tilde{P}^{\otimes 2}, \frac{\delta}{9^p}),$$

where $u^T \tilde{P}^{\otimes 2}$ is the univariate distribution of $(u^T z_i)^2$ for $z_i \sim \widetilde{P}$.

Sparse Covariance Estimation. Next, we consider sparse covariance matrices. In particular, we assume that there exists a subset S of $|S| = s$ covariates that are correlated with each other, and the remaining covariates $[p]\backslash S$ are independent from this subset and from each other. Such sparsity patterns arise naturally in various real-world data [45]. More concretely, for a subset of co-ordinates S, define $\mathcal{G}(S) \overset{\text{def}}{=} \{G = (g)_{ij} \in \mathbb{R}^{p \times p}, g_{ij} = 0 \text{ if } i \notin S \text{ or } j \notin S\}$, and let $\mathcal{G}(s) = \bigcup_{S \subset [p]: |S| \leq s} \mathcal{G}(S)$. Consider the class of matrices $\mathcal{F}_s$ such that,

$$\mathcal{F}_s = \{\Sigma = \Sigma^T, \Sigma \succeq 0, \Sigma - \mathrm{diag}\Sigma \in \mathcal{G}(s)\}$$

Then for any matrix Θ and univariate estimator f, let

$$D_{\mathrm{f},s}(\Theta, \mathcal{Z}_n) = \sup_{u \in \mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} |u^T \Theta u - \mathrm{f}(u^T \mathcal{Z}_n^{\otimes 2}, \epsilon, \frac{\delta}{(\frac{9ep}{s})^s})|.$$

Then, we can define the following estimator,

$$\widehat{\Theta}_{\mathrm{f,s}}(\mathcal{X}_n) = \operatorname*{arginf}_{\theta \in \mathcal{F}_s} D_{\mathrm{f},s}(\Theta, \mathcal{Z}_n),$$

Lemma 8. *Suppose P is a multivariate distribution with covariance Σ such that $\Sigma \in \mathcal{F}_s$. Given n-samples from the mixture distribution (2.3), we get that with probability at least $1 - \delta$,*

$$\|\widehat{\Theta}_{f,s}(\mathcal{Z}_n) - \Sigma\|_2 \lesssim \sup_{u \in \mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} \omega_f(2\epsilon, u^T \tilde{P}^{\otimes 2}, \frac{\delta}{(\frac{9ep}{s})^s}),$$

where $u^T \tilde{P}^{\otimes 2}$ is the univariate distribution of $(u^T z_i)^2$ for $z_i \sim \tilde{P}$.

2.6 Consequences for $\mathcal{P}_{2k}^{\sigma^2}$

Next, we study the performance of our meta-estimator for multivariate estimation for the class of distributions with bounded $2k$-moments. In particular, we use the interval estimator(IM) presented in Algorithm 1 as our univariate estimator to instantiate our meta-estimator.

Multivariate Mean Estimation. In the multivariate setting, we further assume that the contamination level ϵ, and confidence are such that,

$$2\epsilon + \sqrt{\epsilon(\frac{p}{n} + \frac{\log(1/\delta)}{n})} + \frac{p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 1. *Suppose P has bounded $2k$ moments with mean μ and covariance Σ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (2.3), we get that with probability at least $1 - \delta$,*

$$\|\widehat{\theta}_{IM}(\mathcal{X}_n) - \mu\|_2 \lesssim \|\Sigma\|_2^{1/2}\epsilon^{1-1/2k} + \|\Sigma\|_2^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} + \|\Sigma\|_2^{1/2}(\sqrt{\frac{p}{n}} + (\frac{\log n}{n})^{1-\frac{1}{2k}})$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\sigma\epsilon^{1-1/2k})$ in the ϵ-contamination model for multivariate mean estimation, with a sample complexity of $O(p)$.

Sparse Mean Estimation. In this setting, we assume that the contamination level ϵ, and confidence are such that,

$$2\epsilon + \sqrt{\epsilon(\frac{s\log p}{n} + \frac{\log(1/\delta)}{n})} + \frac{s\log p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 2. *Suppose P has bounded $2k$ moments with mean μ and covariance Σ, where μ is s-sparse. Then, given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{\theta}_{IM,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \|\Sigma\|_{2,2s}^{1/2}\epsilon^{1-1/2k} + \|\Sigma\|_{2,2s}^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} \\ + \|\Sigma\|_{2,2s}^{1/2}(\sqrt{\frac{s\log p}{n}} + (\frac{\log n}{n})^{1-\frac{1}{2k}}),\end{aligned}$$

where $\|\Sigma\|_{2,2s} = \sup_{u\in\mathcal{S}^{p-1},\|u\|_0\leq 2s} u^T\Sigma u$.

The above result shows that the proposed estimator exploits the underlying sparsity structure, and achieves the near-optimal sample complexity of $O(s\log p)$, while simultaneously achieving the optimal asymptotic bias of $O(\|\Sigma\|_{2,2s}^{1/2}\epsilon^{1-1/2k})$.

Covariance Estimation. We begin by first calculating $\omega_{IM}(2\epsilon, u^T\widetilde{P}^{\otimes 2}, \delta)$. To do this, recall that for fixed u, for the clean samples in z_i, $(u^Tz_i)^2$ has mean $u^T\Sigma(P)u$, and variance $C_4(u^T\Sigma(P)u)^2$. Note that the scalar random variables $(u^Tz_i)^2$ have bounded k moments, whenever x_i has bounded $2k$-moments. Hence, from Lemma 3, we have that

$$\begin{aligned}\omega_{IM}(2\epsilon, u^T\widetilde{P}^{\otimes 2}, \delta) \lesssim & (u^T\Sigma(P)u)\epsilon^{1-1/k} \\ & + u^T\Sigma(P)u\sqrt{\frac{\log 1/\delta}{n}}.\end{aligned}$$

We assume that the contamination level ϵ, and confidence are such that,

$$4\epsilon + \sqrt{2\epsilon(\frac{p}{n} + \frac{\log(1/\delta)}{n})} + \frac{p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 3. *Suppose P has bounded $2k$-moments, then, given $\mathcal{X}_n$ drawn from the mixture model, then, we have that with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{\Theta}_{IM} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2\epsilon^{1-1/k} + \|\Sigma(P)\|_2\sqrt{\frac{p}{n}} \\ + \|\Sigma(P)\|_2\sqrt{\frac{\log 1/\delta}{n}}\end{aligned}$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\sigma^2\epsilon^{1-1/k})$ in the ϵ-contamination model for multivariate covariance estimation, with a sample complexity of $O(p)$.

Sparse Covariance Estimation. In this setting, we assume that the contamination level ϵ, and confidence δ are such that,

$$4\epsilon + \sqrt{2\epsilon(\frac{s\log p}{n} + \frac{\log(1/\delta)}{n})} + \frac{s\log p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 4. *Suppose P has bounded $2k$-moments and $\Sigma(P) \in \mathcal{F}_s$, then, given $\mathcal{X}_n$ drawn from the mixture model, we have that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_{IM,s} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2\epsilon^{1-1/k} + \|\Sigma(P)\|_2\sqrt{\frac{s\log p}{n}} + \|\Sigma(P)\|_2\sqrt{\frac{\log 1/\delta}{n}}$$

As before, even in this case, the proposed estimator achieves a dimension independent bias of $O(\sigma^2\epsilon^{1-1/k})$, with a sample complexity of $O(s\log p)$.

Sparse PCA in Spiked Covariance Model As an application of the sparse-covariance estimation, we consider the following spiked covariance model, where the true distribution $P \in \mathcal{P}_{2k}$ is such that

$$\Sigma(P) = V\Lambda V^T + \mathcal{I}_p, \tag{2.7}$$

where $V \in \mathbb{R}^{p\times r}$ is an orthonormal matrix, and $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix with entries $\Lambda_1 \geq \Lambda_2 \geq \ldots \geq \Lambda_r > 0$. In this setting, suppose we observe samples from a mixture distribution P_ϵ, then the goal is to estimate the subspace projection matrix VV^T, *i.e.* construct $\widehat{V}$ such that $\|\widehat{V}\widehat{V} - VV^T\|_F$ is small. Note that when V has only s non-zero rows, then the corresponding covariance matrix Σ is s-sparse($\Sigma \in \mathcal{F}_s$).

We follow [11] to use our sparse covariance estimator $\widehat{\Theta}_{\text{IM,s}}(\mathcal{X}_n)$ to construct $\widehat{V} \in \mathbb{R}^{p\times r}$ by setting its j^{th} column to be the j^{th} eigenvector of $\widehat{\Theta}_{\text{IM,s}}(\mathcal{X}_n)$. Then, under the assumption that (ϵ, n) are such that

$$(1+\Lambda_1)\epsilon^{1-1/k} + (1+\Lambda_1)\sqrt{\frac{s\log p}{n}} + (1+\Lambda_1)\sqrt{\frac{\log 1/\delta}{n}} \lesssim \frac{\Lambda_r}{2},$$

we have the following result.

Corollary 5. *Suppose P has bounded $2k$-moments, and $\Sigma(P)$ is of the form of (2.7) and we are given n samples from the mixture distribution. Then, we have that with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{V}\widehat{V}^T - VV^T\|_F^2 \lesssim (\frac{1+\Lambda_1}{\Lambda_r})^2(\epsilon^{2-2/k}) \\ + (\frac{1+\Lambda_1}{\Lambda_r})^2(\frac{s\log p}{n} + \frac{\log 1/\delta}{n})\end{aligned}$$

Discussion. Throughout this section, all our estimators achieve a dimension-independent asymptotic bias. Hence, our proposed meta-estimator allows us to escape the dimension dependence in the ϵ-contamination setting.

Next, we expand on a more subtle aspect of our estimators. Observe that when $\epsilon = 0$, *i.e.* there is no contamination, we see that the typical error rate of our estimators for $k \geq 2 (k \geq 4$ for covariance) is $O(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}})$ is the low dimensional setting, and $O(\sqrt{\frac{s\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}})$ in the high-dimensional setting. Typically, such high-probability bounds are achieved only under the restrictive assumption that the true distribution is Gaussian or sub-gaussian. In contrast, all our results are valid for the much broader class of distributions with bounded $2k$-moment. As discussed in the introduction, while such results have been recently obtained for mean estimation, our simple meta-estimator achieves these high-probability error guarantees for a much broader range of problems. To the best of our knowledge, these are some of the first estimators which get such high-probability deviation bounds for sparse-mean, covariance, sparse-covariance and sparse-PCA.

2.7 Consequences for $\mathcal{P}_{\mathbf{sym}}^{t_0,\kappa}$

Next, we study the performance of our meta-estimator for multivariate estimation for the class of symmetric distributions. In particular, we use the sample median presented in Algorithm 2 as our univariate estimator to instantiate our meta-estimator.

In the multivariate setting, we further assume that the contamination level ϵ, and confidence level δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{p}{n} + \frac{\log(2/\delta)}{n}} \leq t_0.$$

Then, we have the following result.

Corollary 6. *Suppose $P \in \mathcal{P}_{sym}^{t_0,\kappa}$ is a multivariate distribution with mean μ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{Med}(\mathcal{X}_n) - \mu\|_2 \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(1/\delta)}{n}} + \kappa\sqrt{\frac{p}{n}}$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\kappa\epsilon)$ in the ϵ-contamination model for multivariate mean estimation, with a sample complexity of $O(p)$.

Sparse Mean Estimation. In this setting, we assume that the contamination level ϵ, and confidence are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{s\log p}{n} + \frac{\log(2/\delta)}{n}} \lesssim t_0.$$

Then, we have the following result.

Corollary 7. *Suppose $P \in \mathcal{P}_{sym}^{t_0,\kappa}$ is a multivariate distribution with mean μ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (2.3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{Med,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(1/\delta)}{n}} + \kappa\sqrt{\frac{s\log p}{n}}$$

The above result shows that the proposed estimator exploits the underlying sparsity structure, and achieves the near-optimal sample complexity of $O(s\log p)$, while simultaneously achieving the optimal asymptotic bias of $O(\kappa\epsilon)$. Moreover for the case of $\epsilon = 0$, the proposed estimator achieves the near-optimal deviation bound for sparse-mean estimation, for symmetric distributions without moments. Note that similar results can be derived for other higher-order moments.

Discussion. Observe the difference in achievable rates for $\mathcal{P}_{\mathrm{sym}}^{t_0,\kappa}$ and $\mathcal{P}_{2k}^{\sigma^2}$. In particular, for symmetric distributions including those which have no finite variance, the maximum bias introduced by Huber Contamination Model is at most $O(\kappa\epsilon)$. In contrast for distributions with bounded $2k$-moments, the lower bound for mean estimation is $\Omega(\sigma\epsilon^{1-1/2k})$. Note that the depth based estimators of [11] also implicitly assume that the underlying distribution is symmetric, and hence obtain similar rates for elliptical distributions.

2.8 Conclusion and Future Directions.

In this work we provided a conceptually simple way of reducing multivariate estimation to univariate estimation. In particular, we showed how to use any robust univariate estimator to design statistically optimal robust estimators for multivariate estimation. Through this reduction, we derived optimal estimators for non-parametric distribution classes such as distributions with bounded $2k$-moments and symmetrical distributions. Our estimators achieved optimal asymptotic bias in the ϵ-contamination model, and also high-probability deviation bounds in the uncontaminated setting. There are several avenues for future work, some of which we discuss below.

Extension to General Risk Minimization. Consider the setting of risk minimization, where we assume that we have access to a differentiable loss function $\bar{\mathcal{L}} : \mathbb{R}^p \times \mathcal{Z} \mapsto \mathbb{R}$. Let $\mathcal{R}(\theta) = \mathbb{E}_{z \sim P}(\bar{\mathcal{L}}(\theta; z))$ be the population loss. Moreover, let $\theta^*(P)$ be the minimizer of the population risk. Then, in this setting, given n-samples from the mixture model, the goal is to return a parameter $\widehat{\theta}$ which minimizes the population risk. Prasad et al. [1] and Diakonikolas et al. [46] observed that at any point θ, the population gradient $\nabla \mathcal{R}(\theta) = \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta, z)]$ is essentially the *mean* of some distribution. Hence, they proposed a robust-gradient based algorithm, in which, they used a robust multivariate mean estimator to estimate the gradient robustly. In particular the authors' suggest the following update rule

$$\theta^{t+1} = \theta^t - \eta \mathrm{MVMean}(\{\nabla \bar{\mathcal{L}}(\theta^t, z_i)\}_{i=1}^n),$$

where $\mathrm{MVMean}(\cdot)$ is a robust multivariate mean estimator. Our particular p-D to 1-D reduction shows that as long as one has a robust univariate estimator, one can do robust risk minimization for a broad range of problems. While this approach gives near-optimal asymptotic bias, however, such an iterative rule requires sample-splitting at each step and hence is not sample-efficient. Getting past this sample splitting, requires developing mean estimators which are *uniformly* robust over a function class and we leave that as an open problem.

Computationally Efficient Estimators. As noted in the introduction, there has been a flurry of work in the theoretical computer science community on designing polynomial time estimators for robust mean estimation. Designing sample-efficient estimators for sparse-mean estimation for the bounded $2k$-moment class is an open problem. Similarly for covariance estimation, most existing work has focused on Frobenius norm, or Mahalanobis distance, and designing estimators for covariance estimation in operator norm for general bounded $2k$-moment is an open problem. Another important challenge is to design computationally and statistically efficient estimators for the mean of a symmetric distributions.

Chapter 3

Robust Estimation via Robust Gradient Estimation

We provide a new computationally-efficient class of estimators for risk minimization. We show that these estimators are robust for general statistical models, under varied robustness settings, including in the classical Huber ϵ-contamination model, and in heavy-tailed settings. Our workhorse is a novel robust variant of gradient descent, and we provide conditions under which our gradient descent variant provides accurate estimators in a general convex risk minimization problem. We provide specific consequences of our theory for linear regression, logistic regression and for canonical parameter estimation in an exponential family. These results provide some of the first computationally tractable and provably robust estimators for these canonical statistical models. Finally, we study the empirical performance of our proposed methods on synthetic and real datasets, and find that our methods convincingly outperform a variety of baselines.

3.1 Introduction

In classical analyses of statistical estimators, statistical guarantees are derived under strong model assumptions, and in most cases these guarantees hold only in the absence of arbitrary outliers, and other deviations from the model assumptions. Strong model assumptions are rarely met in practice, and this has motivated the development of robust inferential procedures, and which has a rich history in statistics with seminal contributions due to Box [2], Tukey [3], Huber [4], Hampel [5] and several others. These have led to rich statistical concepts such as the influence function, the breakdown point, and the Huber ϵ-contamination model, to assess the robustness of estimators. Despite this progress however, the statistical methods with the strongest robustness guarantees are computationally intractable, for instance those based on non-convex M-estimators [4], ℓ_1 tournaments [6, 7, 8] and notions of depth [9, 10, 11].

In this paper, we present a general class of estimators that are computationally tractable, and have strong robustness guarantees. The estimators we propose are obtained by robustifying iterative updates of risk minimization, and are broadly applicable to a wide-range of parametric statistical models. In the risk minimization framework, the target parameter θ^* is defined as the solution to an optimization problem:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}(\theta) \equiv \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{z \sim P} \left[\bar{\mathcal{L}}(\theta; z) \right], \tag{3.1}$$

where $\bar{\mathcal{L}}$ is an appropriate loss-function, R is the population risk and Θ is the set of feasible parameters. The statistical inference problem within the risk minimization framework is then to compute an approximate minimizer to the above program when given access to samples $\mathcal{D}_n = \{z_1, \ldots, z_n\}$. A classical approach to do so is via *empirical risk minimization* (ERM), where we substitute the empirical expectation given the samples for the population expectation in the specification of the risk objective. While most modern statistical estimators use the above empirical risk minimization framework, a standard assumption that is imposed on $\mathcal{D}_n$ is that the data has no outliers, and has no arbitrary deviations from model assumptions; i.e., it is typically assumed that each of the z_i's are independent and identically distributed according to the distribution P. Moreover, many analyses of risk minimization further assume that P follows a sub-Gaussian distribution, or has otherwise well-controlled tails in order to appropriately control the deviation between the population risk and its empirical counterpart. Due in part to these caveats with ERM, the seminal work of M-estimation replaces the risk minimization objective with a robust counterpart, so that the minimizer of the empirical expectation of the robust counterpart is more robust than the ERM minimizer. As noted above, for strong statistical guarantees, these in turn require solving computationally intractable non-convex optimization programs.

In contrast to this classical work, we propose a class of estimators that have a shift in perspective: rather than specify a robust objective, we consider an algorithm, namely projected gradient descent, that directly optimizes the population risk objective in Eq. (3.1), and focus on making this algorithm robust. Thus, in contrast to specifying the robust parameter estimate as the solution to an optimization program as in M-estimation, which in turn could be computationally intractable, we specify the robust parameter estimate as the limit of a sequence of iterative updates that are individually robust as well as computationally tractable. We find that this shift in perspective leads to estimators that are both computationally tractable as well with strong robustness guarantees, that are as broadly applicable as ERM or M-estimators, and moreover with a unified statistical treatment for varied statistical models.

In addition to being applicable to a variety of statistical models, our general results are also applicable to a variety of *notions of robustness*. In this paper, we derive corollaries in particular for two canonical robustness settings:

(a) **Robustness to arbitrary outliers:** In this setting, we focus on Huber's ϵ-contamination model, where rather than observe samples directly from P in (3.1) we instead observe samples drawn from P_ϵ which for an *arbitrary* distribution Q is defined as:

$$P_\epsilon = (1 - \epsilon)P + \epsilon Q.$$

The distribution Q allows for arbitrary outliers, which may correspond to gross corruptions or more subtle deviations from the assumed model. This model can be equivalently viewed as model mis-specfication in the Total Variation (TV) metric.

(b) **Robustness to heavy-tails:** In this setting, we are interested in developing estimators under weak moment assumptions. We assume that the distribution P from which we obtain samples only has finite low-order moments (see Section 3.5.3 for a precise characterization). Such heavy-tailed distributions arise frequently in the analysis of financial data and large-scale biological datasets (see for instance examples in [12, 13]). In contrast to classical analyses of empirical risk minimization [14], in this setting the empirical risk is not uniformly close to the population risk, and methods that directly minimize the empirical risk perform poorly (see Section 5.4).

While we provide corollaries demonstrating robustness with respect to the above deviations, we emphasize that our framework is more general. Below, we provide an outline of our results and contributions.

1. **Estimators.** Our first contribution is to introduce a new class of robust estimators for risk minimization (3.1). These estimators are based on robustly estimating gradients of the population risk to then plug in to a projected gradient descent algorithm, and are computationally tractable by design. A crucial ingredient of our framework is the design of robust gradient estimators for the population risk in (3.1). Our main insight is that in this general risk minimization setting, the gradient of the population risk is simply a multivariate mean vector, and we can leverage prior work on mean estimation to design robust gradient estimators. Thus, for our two canonical robustness cases, we develop such robust gradient estimators building on prior work for robust mean estimation in the Huber model [28], and in the heavy-tailed model [16]. Another perspective of our framework is that it significantly generalizes the applicability of *mean estimation* methods to general parametric models.
2. **Empirical Investigations.** Our estimators are computationally practical, and accordingly, our second contribution is to conduct extensive numerical experiments on real and simulated data with our proposed class of estimators. We provide guidelines for tuning parameter selection, and compare the proposed estimators with several competitive baselines [4, 47, 48]. We find that our estimators consistently perform well across different settings, and across various metrics.
3. **Statistical Guarantees.** Finally, we provide rigorous robustness guarantees for the estimators we propose for a variety of classical statistical models: linear regression, logistic regression, and exponential family models. Our contributions in this direction are two-fold: building on prior work [49] we provide a general result on the stability of gradient descent for risk minimization, and show that under certain conditions, gradient descent can be quite tolerant to inaccurate gradient estimates. Subsequently, in concrete settings, we provide a careful analysis of the quality of gradient estimation afforded by our proposed gradient estimators, and combine these results to obtain guarantees on our final parameter estimates.

Broadly, as we discuss in the sequel, our work suggests that our class of estimators based on robust gradient estimation offer a variety of practical, conceptual, statistical and computational advantages for robust estimation. They provide the general applicability of classical M-estimators, together with computational practicality even for large-scale models, as well as strong robustness guarantees.

3.1.1 Related Work

There has been extensive work in the broad area of robust statistics (see for instance [5] and references therein); we focus this section on some lines of work that are most related to this paper. For the robustness setting of ϵ-contaminated models, several classical estimators have been developed that are optimally robust for a variety of inferential tasks, including hypothesis testing [38], mean estimation [9], general parametric estimation [7, 8, 33], and non-parametric estimation [6]. However, a major drawback with this classical line of work has been that most of the estimators with strong robustness guarantees are computationally intractable [3], while the remaining ones use heuristics and are consequently not optimal [26]. A complementary line of recent research [10, 11] has focused on providing minimax upper and lower bounds on the performance of estimators under ϵ-contamination model, without the constraint of computational tractability. Recently, there has been a flurry of research in theoretical computer science [27, 28, 50, 51] designing provably robust estimators which are computationally tractable while achieving near-optimal contamination dependence, for special classes of problems such as computing means and covariances. Some of the proposed algorithms are however not computationally practical as they rely on the ellipsoid algorithm or require solving semi-definite programs [27, 50, 51] which can be slow for modern problem sizes. Lecué and Lerasle [52] proposed a median-of-means approach to solve ERM under the ϵ-contaminated setting. While their estimator achieves good statistical rates, it is not computationally efficient and in particular, involves solving a saddle point problem. While in the general ϵ-contamination setting, the contamination distribution could be arbitrary, there has been a lot of work in settings where the contamination distribution is restricted in various ways. For example, recent work in high-dimensional statistics (for instance [53, 54, 55, 56, 57]) have studied problems like principal component analysis and linear regression under the assumption that the corruptions are evenly spread throughout the dataset.

For the robustness setting of heavy-tailed distributions, robust estimators aim to relax the sub-Gaussian or sub-exponential distributional assumptions that are typically imposed on the target distribution, and allow it to be a heavy-tailed distribution. Most approaches in this category substitute the *empirical* *mean* of the risk objective in risk minimization with robust mean estimators such as [15, 58] that exhibit sub-Gaussian type concentration around the true mean for distributions satisfying mild moment assumptions. The median-of-means estimator [58] and Catoni's mean estimator [15] are two popular examples of such robust mean estimators. In particular, Hsu and Sabato [59] use the median-of-means estimator to solve the corresponding robust variant of ERM. Although this estimator has strong theoretical guarantees, and is computationally tractable, as noted by the authors in [59] it performs poorly in practice. In recent work Brownlees et al. [60] use the Catoni's mean estimator to solve the corresponding robust variant of ERM. The authors provide risk bounds similar to bounds one can achieve under sub-Gaussian distributional assumptions. However, their estimator is not easily computable and the authors do not provide a practical algorithm to compute the estimator. Other recent work by Lerasle and Oliveira [58], Lugosi and Mendelson [61] use similar ideas to derive estimators that perform well theoretically, in heavy-tailed situations. However, these approaches involve optimization of complex objectives for which no computationally tractable algorithms exist. We emphasize that in contrast to our work, these works focus on robustly estimating the population risk which does not directly lead to a computable estimator. In contrast, we consider robustly estimating the *gradient* of the population risk, and embedding these estimates within the iterative algorithm of projected gradient descent, which leads naturally to a computionally practical estimator.

3.1.2 Outline

We conclude this section with a brief outline of the remainder of the paper. In Section 4.1.2, we provide some background on risk minimization and the running robustness settings of Huber contamination, and heavy-tailed noise models. In Section 3.3, we introduce our class of robust estimators, and provide concrete algorithms for the ϵ-contaminated and heavy-tailed settings. In Section 5.4 we study the empirical performance of our estimator on a variety of tasks and datasets. We complement our empirical results with theoretical guarantees in Sections 3.5, 3.6 and 3.7. We defer technical details to the Appendix. Finally, we conclude in Section 3.8 with a discussion of some open problems.

3.2 Background and Problem Setup

In this section we provide the necessary background on risk minimization, gradient descent, and introduce two notions of robustness that we consider in this work.

3.2.1 Risk Minimization and Parametric Estimation

In the setting of risk minimization, we assume that we have access to a differentiable loss function $\bar{\mathcal{L}} : \Theta \times \mathcal{Z} \mapsto \mathbb{R}$, where Θ is a convex subset of $\mathbb{R}^p$. Let $\mathcal{R}(\theta) = \mathbb{E}_{z \sim P}\left[\bar{\mathcal{L}}(\theta; z)\right]$ be the population loss, also known as the *risk*, and let θ^* be the minimizer of the population risk $\mathcal{R}(\theta)$, over the set Θ:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}(\theta). \tag{3.2}$$

The goal of risk minimization is to minimize the population risk $\mathcal{R}(\theta)$, given only n samples $\mathcal{D}_n = \{z_i\}_{i=1}^n$, in order to estimate the unknown parameter θ^*.

In this work we assume that the population risk is convex to ensure tractable minimization. Moreover, in order to ensure identifiability of the parameter θ^*, we impose two standard regularity conditions [62] on the population risk. These properties are defined in terms of the error of the first-order Taylor approximation of the population risk, i.e. defining, $\tau(\theta_1, \theta_2) := \mathcal{R}(\theta_1) - \mathcal{R}(\theta_2) - \langle \nabla \mathcal{R}(\theta_2), \theta_1 - \theta_2 \rangle$, we assume that

$$\frac{\tau_\ell}{2}\|\theta_1 - \theta_2\|_2^2 \leq \tau(\theta_1, \theta_2) \leq \frac{\tau_u}{2}\|\theta_1 - \theta_2\|_2^2, \tag{3.3}$$

where the parameters $\tau_\ell, \tau_u > 0$ denote the strong-convexity and smoothness parameters respectively.

3.2.2 Illustrative Examples of Risk Minimization

The framework of risk minimization is a central paradigm of statistical estimation and is widely applicable. In this section, we provide illustrative examples that fall under this framework.

Linear Regression

Here we observe paired samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$. We assume that the (x, y) pairs sampled from the true distribution P

are linked via a linear model:

$$y = \langle x, \theta^* \rangle + w, \tag{3.4}$$

where w is drawn from a zero-mean distribution such as normal distribution with variance σ^2 ($\mathcal{N}(0, \sigma^2)$) or a more heavy-tailed distribution such as student-t or Pareto distribution. We suppose that under P the covariates $x \in \mathbb{R}^p$, have mean 0, and covariance Σ.

For this setting we use the squared loss as our loss function, which induces the following population risk:

$$\bar{\mathcal{L}}(\theta; (x, y)) = \frac{1}{2} \left(y - \langle x, \theta \rangle\right)^2, \quad \text{and} \quad \mathcal{R}(\theta) = \frac{1}{2}(\theta - \theta^*)^T \Sigma (\theta - \theta^*).$$

Note that the true parameter θ^* is the minimizer of the population risk $\mathcal{R}(\theta)$. The strong-convexity and smoothness assumptions from (3.3) in this setting require that $\tau_\ell \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \tau_u$.

Generalized Linear Models

Here we observe paired samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$. We suppose that the (x, y) pairs sampled from the true distribution P are linked via a linear model such that when conditioned on the covariates x, the response variable has the distribution:

$$P(y|x) \propto \exp\left(\frac{y \langle x, \theta^* \rangle - \Phi(\langle x, \theta^* \rangle)}{c(\sigma)}\right) \tag{3.5}$$

Here $c(\sigma)$ is a fixed and known scale parameter and $\Phi : \mathbb{R} \mapsto \mathbb{R}$ is the link function. We focus on the random design setting where the covariates $x \in \mathbb{R}^p$, have mean 0, and covariance Σ. We use the negative conditional log-likelihood as our loss function, i.e.

$$\bar{\mathcal{L}}(\theta; (x, y)) = -y \langle x, \theta \rangle + \Phi(\langle x, \theta \rangle). \tag{3.6}$$

Once again, the true parameter θ^* is the minimizer of the resulting population risk $\mathcal{R}(\theta)$. It is easy to see that Linear Regression with Gaussian Noise lies in the family of generalized linear models. A popular instance of such GLMs is a logistic regression model.

Logistic Regression. In this case the (x, y) pairs are linked as:

$$y = \begin{cases} 1 & \text{with probability} \quad \frac{1}{1+\exp(-\langle x, \theta^* \rangle)}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

This corresponds to setting $\Phi(t) = \log(1 + \exp(t))$ and $c(\gamma) = 1$ in (5.10). The hessian of the population risk is given by

$$\nabla^2 \mathcal{R}(\theta) = \mathbb{E}\left[\frac{\exp\langle x, \theta \rangle}{(1 + \exp\langle x, \theta \rangle)^2} x x^T\right].$$

Note that as θ diverges, the minimum eigenvalue of the hessian approaches 0 and the loss is no longer strongly convex. To prevent this, in this case we take the parameter space Θ to be bounded.

Exponential Families and Canonical Parameters.

Finally we consider the case where the true distribution P is in exponential family with canonical parameters $\theta^* \in \mathbb{R}^p$, and a vector of sufficient statistics obtained from the map $\phi : \mathcal{Z} \mapsto \mathbb{R}^p$. Note that while the linear and logistic regression models are indeed in an exponential family, our interest in those cases was not in the canonical parameters.

In more details, we can write the true distribution P in this case as

$$P(z) = h(z) \exp\left(\langle \phi(z), \theta^* \rangle - A(\theta^*)\right),$$

where $h(z)$ is some base measure. The negative log-likelihood gives us the following loss function:

$$\bar{\mathcal{L}}(\theta; z) = -\langle \phi(z), \theta \rangle + A(\theta). \tag{3.8}$$

The strong-convexity and smoothness assumptions require that there are constants τ_ℓ, τ_u such that $\tau_\ell \leq \nabla^2 A(\theta) \leq \tau_u$, for $\theta \in \Theta$.

3.2.3 Empirical Risk Minimization

Given data $\mathcal{D}_n = \{z_i\}_{i=1}^n$, empirical risk minimization (ERM) substitutes the empirical expectation of the risk for the population risk in the risk minimization objective:

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \bar{\mathcal{L}}(\theta; z_i).$$

Most modern statistical estimators follow this ERM recipe above. When the loss is the log-likelihood of the statistical model, this reduces to the classical Maximum Likelihood Estimation (MLE) principle. The empirical risk minimizer is however a poor estimator of θ^* in the presence of outliers in the data: since ERM depends on the sample mean, outliers in the data can effect the sample mean and lead ERM to sub-optimal estimates. This observation has led to a large body of research that focuses on developing robust M-estimators, where we substitute in the empirical expectation of a robust counterpart of the loss function $\bar{\mathcal{L}}$; the resulting estimators have favorable statistical properties, but are often computationally intractable.

3.2.4 Projected Gradient Descent

A popular approach for solving the empirical risk minimization problem is projected gradient descent. Projected gradient descent generates a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$, by refining an initial parameter $\theta_0 \in \Theta$ via the update:

$$\theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t - \eta \nabla \mathcal{R}_n(\theta^t)\right),$$

where $\eta > 0$ is the step size and $\mathcal{P}_\theta$ is the projection operator onto Θ. While this gradient descent method is simple, it is not however robust to various deviations, for general convex losses. Accordingly, we have a small shift in perspective: instead of performing gradient descent on the empirical risk, we perform gradient descent on the population risk. Our work then relies on the important observation that this gradient of the population risk ($\mathbb{E}_{z \sim P}\left[\nabla \bar{\mathcal{L}}(\theta; z)\right]$) is simply a mean vector: one that can be estimated robustly by leveraging recent advances in robust mean estimation [16, 28]. This leads to a general method for risk minimization based on embedding robust gradient estimation within a projected gradient descent algorithm (see Algorithm 3).

3.2.5 Robust Estimation

One of the goals of this work is to develop general statistical estimation methods that are robust under varied robustness settings. We derive corollaries in particular for two robustness models: Huber's ϵ-contamination model, and the heavy-tailed model. We now briefly review these two notions of robustness.

(a) **Huber's ϵ-contamination model:** Huber [38, 63] proposed the ϵ-contamination model where we observe samples that are obtained from a mixture of the form

$$P_\epsilon = (1-\epsilon)P + \epsilon Q, \tag{3.9}$$

where P is the true distribution, ϵ is the expected fraction of outliers and Q is an *arbitrary* outlier distribution. Given i.i.d. observations drawn from P_ϵ, our objective is to estimate θ^*, the minimizer of the population risk $\mathcal{R}(\theta) = \mathbb{E}_{z\sim P}\left[\bar{\mathcal{L}}(\theta; z)\right]$, robust to the contamination from Q.

(b) **Heavy-tailed model:** In the heavy-tailed model it is assumed that the data follows a heavy-tailed distribution (i.e, P is heavy-tailed). While heavy-tailed distributions have various possible characterizations: in this paper we consider a characterization via gradients. For a fixed $\theta \in \Theta$ we let P_g^θ denote the multivariate distribution of the gradient of population loss, i.e. $\nabla\bar{\mathcal{L}}(\theta; z)$. We refer to a potentially heavy-tailed distribution as one for which our only assumption on P_g^θ is that it has finite second moments for any $\theta \in \Theta$. As we illustrate in Section 3.7, in various concrete examples this translates to relatively weak low-order moment assumptions on the data distribution P.

Given n i.i.d observations from P, our objective is to estimate the minimizer of the population risk. From a conceptual standpoint, the classical analysis of risk-minimization which relies on uniform concentration of the empirical risk around the true risk, fails in the heavy-tailed setting necessitating new estimators and analyses [17, 59, 60, 61].

3.3 Robust Gradient Descent via Gradient Estimation

Gradient descent and its variants are at the heart of modern optimization and are well-studied in the literature. Suppose we have access to the true distribution P_{θ^*}. Then to minimize the population risk $\mathcal{R}(\theta)$, we can use projected gradient descent, where starting at some initial θ^0 and for an appropriately chosen step-size η, we update our estimate according to:

$$\theta^{t+1} \leftarrow \mathcal{P}_\Theta(\theta^t - \eta\nabla\mathcal{R}(\theta^t)). \tag{3.10}$$

However, we only have access to n samples $\mathcal{D}_n = \{z_i\}_{i=1}^n$. The key technical challenges are then to estimate the gradient of $\mathcal{R}(\theta)$ from samples $\mathcal{D}_n$, and to ensure that an appropriate modification of gradient descent is stable to the resulting estimation error.

To address the first challenge we observe that the gradient of the population risk at any point θ is the mean of a multivariate distribution, *i.e.* $\nabla \mathcal{R}(\theta) = E_{z \sim P}\left[\nabla \bar{\mathcal{L}}(\theta; z)\right]$. Accordingly, the problem of gradient estimation can be reduced to a multivariate mean estimation problem, where our goal is to *robustly* estimate the true mean $\nabla \mathcal{R}(\theta)$ from n samples $\{\nabla \bar{\mathcal{L}}(\theta; z_i)\}_{i=1}^n$. For a given sample-size n and confidence parameter $\delta \in (0, 1)$ we define a gradient estimator:

Definition 1. *A function $g(\theta; \mathcal{D}_n, \delta)$ is a gradient estimator, if for functions α and β, with probability at least $1 - \delta$, at any fixed $\theta \in \Theta$, the estimator satisfies the following inequality:*

$$\|g(\theta; \mathcal{D}_n, \delta) - \nabla \mathcal{R}(\theta)\|_2 \leq \alpha(n, \delta)\|\theta - \theta^*\|_2 + \beta(n, \delta). \tag{3.11}$$

In subsequent sections, we will develop conditions under which we can obtain gradient estimators with strong control on the functions $\alpha(n, \delta)$ and $\beta(n, \delta)$ in the Huber and heavy-tailed models. Furthermore, by investigating the stability of gradient descent we will develop sufficient conditions on these functions such that gradient descent with an inaccurate gradient estimator still returns an accurate estimate.

To minimize $\mathcal{R}(\theta)$, we replace $\nabla \mathcal{R}(\theta)$ in equation (3.10) with the gradient estimator $g(\theta; \mathcal{D}_n, \delta)$ and perform projected gradient descent. In order to avoid complex statistical dependency issues that can arise in the analysis of gradient descent, for our theoretical results we consider a sample-splitting variant of the algorithm where each iteration is performed on a fresh batch of samples. We summarize the overall robust gradient descent algorithm via gradient estimation in Algorithm 3. In contrast to M-estimation where we use robust estimates of the overall loss function, here we use robust estimates of the gradient, a small shift in perspective, but which has strong statistical and computational consequences: we obtain a computationally practical algorithm, and moreover with strong robustness guarantees via careful statistical analyses of the stability of the resulting biased and inexact gradient descent iterates.

We further assume that the number of gradient iterations T is specified

Algorithm 3 Robust Gradient Descent

```
function RGD (Gradient Estimator $g(\cdot)$, Data $\{z_1, \ldots, z_n\}$, Step Size $\eta$, Number of Iterations $T$, Confidence $\delta$)
    Split samples into $T$ subsets $\{\mathcal{Z}_t\}_{t=1}^{T}$ of size $\widetilde{n}$.
    for $t = 0$ to $T - 1$ do
        $\theta^{t+1} = \mathrm{argmin}_{\theta \in \Theta} \|\theta - \left(\theta^t - \eta\, g(\theta^t; \mathcal{Z}_t, \widetilde{\delta})\right) \|_2^2$.
    end for
end function
```

a-priori, and accordingly we define:

$$\widetilde{n} = \left\lfloor \frac{n}{T} \right\rfloor \quad \text{and} \quad \widetilde{\delta} = \frac{\delta}{T}.$$

We discuss methods for selecting T, and the impact of sample-splitting in later sections. As confirmed in our experiments (see Section 5.4), sample-splitting should be viewed as a device introduced for theoretical convenience which can likely be eliminated via developing uniformly robust gradient estimators. We provide some partial results along these lines in Appendix B.17, noting that these are more complex in our general setting where we do not even assume smoothness of the robust gradient estimators; see also the work [49].

It can be seen that the key ingredient in the robust gradient descent Algorithm 3 is a robust estimator of the gradients. Next, we consider the two notions of robustness described in Section 4.1.2, and derive specific gradient estimators for each of the models using the framework described above. Although we derive corollaries of our general results for these two settings of Huber contamination and heavy-tailed models, we emphasize that our class of estimators are more general and are not restricted to these two notions of robustness.

3.3.1 Gradient Estimation in Huber's ϵ-contamination model

There has been a flurry of recent interest [27, 28, 50, 51] in designing mean estimators which, under the Huber contamination model, can robustly estimate the mean of a random vector. While some of these results are focused on the case where the uncorrupted distribution is Gaussian, or isotropic, we are interested in robust mean oracles for more general distributions. Diakonikolas et al. [31] and Lai et al. [28] proposed robust mean estimators for general distributions, satisfying weak moment assumptions. Although the estimator proposed by [31] works under weak assumptions, it requires side information about the true distribution, which, makes it hard to tune in practice. Hence, in our methodology and experimental sections, we primarily leverage Lai et al. [28]'s estimator to design a *Huber gradient estimator* $g(\theta; \mathcal{D}_n, \delta)$ which works in the Huber contamination model. However, in our theoretical results, we do provide an analysis of the gradient estimator obtained by using the robust mean estimator of [31].

The estimator of Lai et al. [28] builds upon the fact that with a single dimension, it is relatively easy to estimate the gradient robustly. In higher dimensions, the crucial insight of Lai et al. [28] is that the effect of the contamination distribution Q on the mean of uncontaminated distribution P is effectively one-dimensional provided we can accurately estimate the direction along which the mean is shifted. In our context, if we can compute the gradient shift direction, i.e. the direction of the difference between the sample (corrupted) mean gradient and the true (population) gradient, then the true gradient can be estimated by using a robust 1D-mean algorithm along the gradient-shift direction and a non-robust sample-gradient in the orthogonal direction since the contamination has no effect on the gradient in this orthogonal direction. In order to identify this gradient shift direction, we follow Lai et al. [28] and use a recursive Singular Value Decomposition (SVD) based algorithm. In each stage of the recursion, we first remove gross-outliers via a truncation algorithm (described in more detail in the Appendix, and termed HuberOutlierGradientTruncation in Algorithm 4). We subsequently identify two subspaces using an SVD – a clean subspace where the contamination has a small effect on the mean and another subspace where the contamination has a potentially larger effect. We use a simple sample-mean estimator in the clean subspace and recurse our computation on the other subspace. Building on the work of Lai et al. [28], in Lemma 9 and Appendix B.13 we provide a careful non-asymptotic analysis of this gradient estimator.

Algorithm 4 presents the overall *Huber gradient estimator* $g(\theta; \mathcal{D}_n, \delta)$.

Algorithm 4 Huber Gradient Estimator

```
function HuberGradientEstimator(Sample Gradients $S = \{\nabla\bar{\mathcal{L}}(\theta; z_i)\}_{i=1}^{n}$, Corruption Level $\epsilon$, Dimension $p$, $\delta$)
    $\widetilde{S} = $ HuberOutlierGradientTruncation$(S, \epsilon, p, \delta)$.
    if p=1 then
        return mean($\widetilde{S}$)
    else
        Let $\Sigma_{\widetilde{S}}$ be the covariance matrix of $\widetilde{S}$.
        Let $V$ be the span of the top $p/2$ principal components of $\Sigma_{\widetilde{S}}$ and $W$ be its complement.
        Set $S_1 := P_V(\widetilde{S})$ where $P_V$ is the projection operation on to $V$.
        Let $\widehat{\mu}_V :=$ HuberGradientEstimator$(S_1, \epsilon, p/2, \delta)$.
        Let $\widehat{\mu}_W :=$ mean$(P_W\widetilde{S})$.
        Let $\widehat{\mu} \in \mathbb{R}^p$ be such that $P_V(\widehat{\mu}) = \widehat{\mu}_V$, and $P_W(\widehat{\mu}) = \widehat{\mu}_W$.
      return $\widehat{\mu}$.
    end if
end function
```

3.3.2 Gradient Estimation in the Heavy-Tailed model

To design gradient estimators for the heavy-tailed model, we leverage recent work on designing robust *mean* estimators in this setting. These robust mean estimators build on the classical work of Alon et al. [20], Nemirovski and Yudin [21] and Jerrum et al. [22] on the so-called median-of-means estimator. For the problem of one-dimensional mean estimation, Catoni [15], Lerasle and Oliveira [58] propose robust mean estimators that achieve exponential concentration around the true mean for any distribution with bounded second moment. In this work we require mean estimators for multivariate distributions. Several works ([16, 17, 59]) extend the median-of-means estimator of to general metric spaces. Recently, Hopkins [23] developed a Sum-of-Squares based polynomial-time algorithm(Median-SDP) that achieves optimal error for mean estimation. However, even though Median-SDP is polynomial-time, it is not practically implementable. We explore the theoretical properties of using Median-SDP to design the gradient estimator $g(\theta; \mathcal{D}_n, \delta)$ in Appendix B.15, but given the focus on practicality, we use the geometric median-of-means estimator (Gmom), which was originally proposed and analyzed by Minsker [16], to design the gradient estimator $g(\theta; \mathcal{D}_n, \delta)$.

The basic idea behind the Gmom estimator is to first split the samples into non-overlapping subsamples and estimate the sample mean of each of the subsamples. Then the GMOM estimator is given by the median-of-means of the subsamples. Formally, let $\{x_i \dots x_n\} \in \mathbb{R}$ be n i.i.d random variables sampled from a distribution P. Then the GMOM estimator for estimating the mean of P can be described as follows. Partition the n samples into b blocks $B_1, \dots B_b$ each of size $\lfloor n/b \rfloor$. Let $\{\widehat{\mu}_1, \dots, \widehat{\mu}_b\}$ be the sample means in each block, where $\widehat{\mu}_i = \frac{1}{|B_i|}\sum_{x_j \in B_i} x_j$. Then the GMOM estimator is given by median$\{\widehat{\mu}_1, \dots \widehat{\mu}_b\}$. In high dimensions where different notions of the median have been considered Minsker [16] uses geometric median:

$$\widehat{\mu} = \operatorname*{argmin}_{\mu} \sum_{i=1}^{b} \|\mu - \widehat{\mu}_i\|_2.$$

Algorithm 5 presents the gradient estimator $g(\theta; \mathcal{D}_n, \delta)$ obtained using GMOM as the mean estimator.

Algorithm 5 Heavy Tailed Gradient Estimator

```
function HEAVYTAILEDGRADIENTESTIMATOR(SAMPLE GRADIENTS $S = \{\nabla \bar{\mathcal{L}}(\theta; z_i)\}_{i=1}^{n}$, $\delta$)
    Define number of buckets $b = 1 + \lfloor 3.5 \log 1/\delta \rfloor$.
    Partition $S$ into $b$ blocks $B_1, \dots B_b$ each of size $\lfloor n/b \rfloor$.
    for $i = 1 \dots n$ do
        $\widehat{\mu}_i = \frac{1}{|B_i|} \sum_{s \in B_i} s$.
    end for
    Let $\widehat{\mu} = \operatorname*{argmin}_{\mu} \sum_{i=1}^{b} \|\mu - \widehat{\mu}_i\|_2$.
    return $\widehat{\mu}$.
end function
```

To conclude this section, we note that the gradient estimators described in Algorithm 4 depend on corruption level ϵ, which is typically not known in advance. In Appendix B.1, we briefly discuss some heuristic methods for adapting to the unknown ϵ that we use in our experiments.

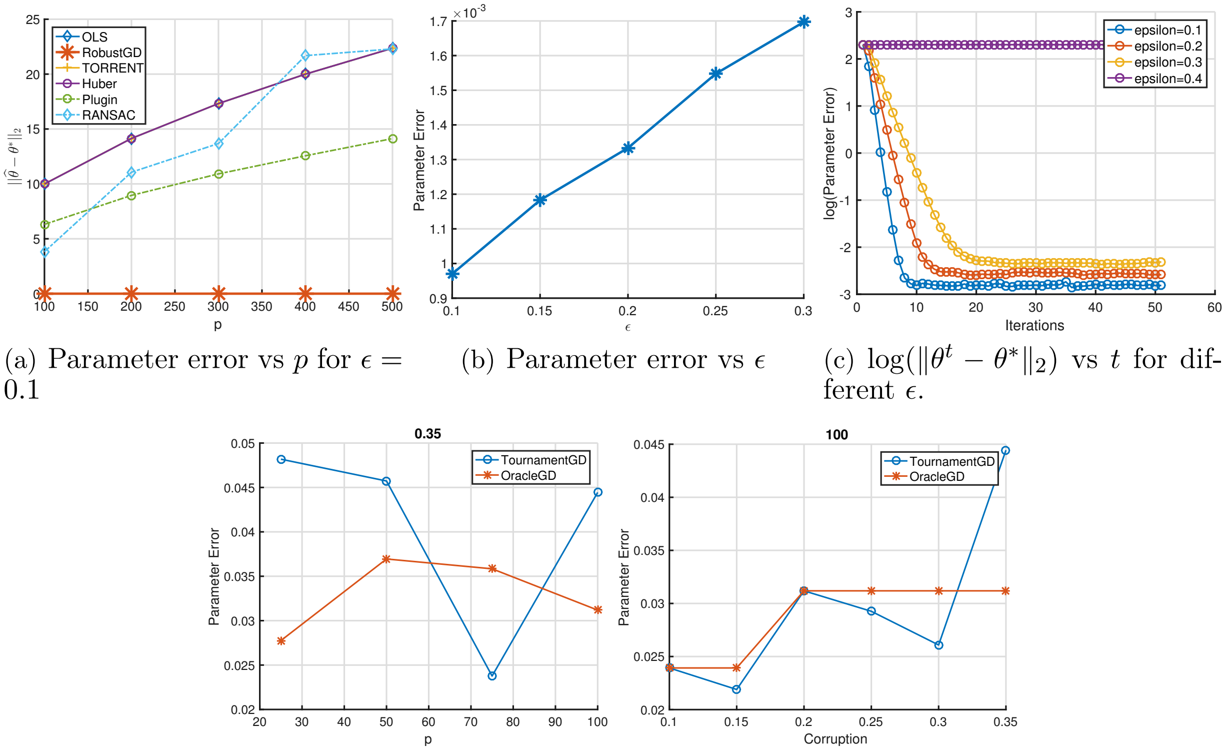
3.4 Experiments

In this section we demonstrate our proposed methods for the Huber contamination and heavy-tailed models, on a variety of simulated and real data examples.

3.4.1 Huber Contamination

We first consider the Huber contamination model and demonstrate the practical utility of gradient-descent based robust estimator described in Algorithms 3 and 4.

Synthetic Experiments: Linear Regression



(a) Parameter error vs p for $\epsilon = 0.1$

(b) Parameter error vs ϵ

(c) $\log(\|\theta^t - \theta^*\|_2)$ vs t for different ϵ.

(d) Hyperparameter Tuning vs p

(e) Hyperparameter Tuning vs ϵ

Figure 3.1: Robust Linear Regression.

Recall the linear regression model described in (B.17) where we observe paired samples $\{(x_i, y_i)\}_{i=1}^n$. We assume that the (x, y) pairs sampled from the true distribution P are linked via a linear model: $y = \langle x, \theta^* \rangle + w$. We now describe the experiment setup, the data model and present the results.

Setup. We fix the contamination level $\epsilon = 0.1$ and $\sigma^2 = 0.1$. Next, we generate $(1 - \epsilon)n$ *clean* covariates from a multivariate Gaussian $x \sim \mathcal{N}(0, \mathcal{I}_p)$, and we generate the corresponding clean responses using $y = \langle x, \theta^* \rangle + w$ where

$\theta^* = [1, \ldots, 1]^T$ and $w \sim \mathcal{N}(0, \sigma^2)$. We simulate an outlier distribution by drawing the covariates from $\mathcal{N}(0, p^2\mathcal{I}_p)$, and setting the responses to 0. The total number of samples is set to be $(10\frac{p}{\epsilon^2})$. We note that the sample size we choose increases with the dimension. This scaling is used to ensure that the statistical (minimax) error, in the absence of any contamination, is roughly 0.001. An optimally robust method should have error close to 0.1 (roughly equal to corruption level), which ours does (see Figure 3.1).

Metric. We measure the parameter error in ℓ_2-norm. We also study the convergence properties of our proposed method, for different contamination levels ϵ. We use code provided by Lai et al. [28] to implement our gradient estimator.

Baselines. As our baselines, we use OLS, TORRENT [47], the Huber-loss M-estimator, RANSAC and a plugin estimator (detailed further in Section 3.6.1, and which in a nutshell robustly estimates the sufficient statistics required for the OLS estimator). TORRENT is an iterative hard-thresholding based alternating minimization algorithm, where in one step, it calculates an active set of examples by keeping only $(1-\epsilon)n$ samples which have the smallest absolute values of residual $r = y - \langle x, \theta^t \rangle$, and in the other step it updates the current estimates by solving OLS on the active set. Bhatia et al. [47] have shown the superiority of TORRENT over a variety of other convex-penalty based outlier techniques, hence, we do not compare against those methods. The plugin estimator is implemented using Algorithm 4 to estimate both the mean vector $\frac{1}{n}\sum_{i=1}^n y_i x_i$ and the covariance matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^T$, which are the required sufficient statistics for the OLS estimator.

Results. We summarize our main findings here.

- **Error vs dimension p:** All estimators except our proposed algorithm perform poorly with increasing dimension, as shown in Figure 3.1(a). Note that the TORRENT algorithm has strong guarantees when only the response y is corrupted but performs poorly in the Huber contamination model where both x and y may be contaminated. We find that the error for the robust plugin estimator increases with dimension. We investigate this theoretically in Section 3.6.1, where we find that the error of the plugin estimator grows with the norm of θ^*. In our experiments, we choose $\|\theta^*\|_2 = \sqrt{p}$, and thus

Figure 3.1(a) corroborates Corollary 10 in Section 3.6.1.

- **Error vs ϵ:** In Figure 3.1(b) we find that the parameter error $\|\widehat{\theta} - \theta^*\|_2$ increases linearly with the contamination rate ϵ. Under a more general setting, we study this in Section 3.6.1, and show that the error scales at most as $\sqrt{\epsilon}$.
- **Error vs iteration t:** Finally, Figure 3.1(c) shows that the convergence rate decreases with increasing contamination ϵ and after ϵ is high enough, the algorithm remains stuck at θ_0, corroborating Lemma 27 (in the Appendix).
- **Hyper-parameter Tuning:** In Figures 3.1(d) and 3.1(e), we find the final solution chosen by our tournament based heuristic for hyper-parameter selection (TournamentGD) has roughly the same performance as the algorithm which knows the true value of ϵ (OracleGD). In particular, our final error does not scale with p.

Next, we study the performance of our proposed method in the context of classification.

Synthetic Experiments: Logistic Regression

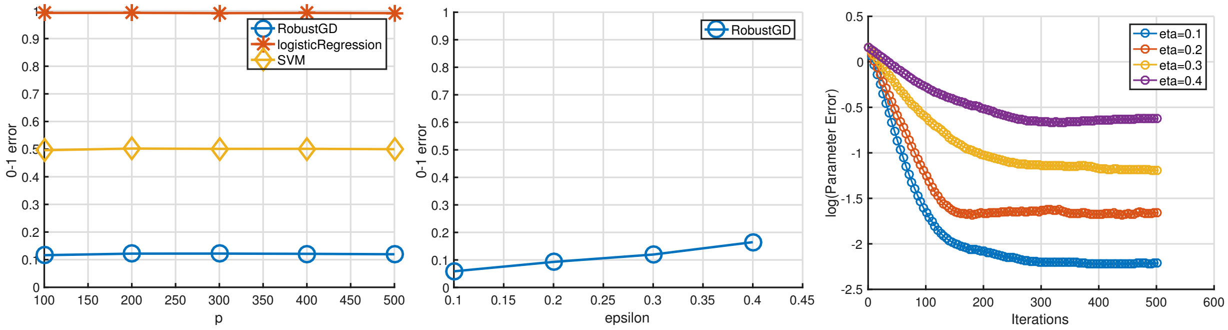
Setup. We simulate a linearly separable classification problem, where the *clean* covariates are sampled from $\mathcal{N}(0, \mathcal{I}_p)$, the corresponding clean responses are computed as $y = \text{sign}(\langle x, \theta^* \rangle)$ where $\theta^* = [1/\sqrt{p}, \ldots, 1/\sqrt{p}]^T$. We set our domain Θ to be the unit ball, *i.e.* $\Theta = \theta$ s.t. $\|\theta\|_2 \leq 1$. Constraining the domain to be the ℓ_2 ball makes the population risk function of logistic loss strongly convex with the optimizer being at θ^*.

We simulate the outlier distribution by adding asymmetric noise, *i.e.* we flip the labels of one class, and increase the variance of the corresponding covariates by multiplying them by p^2. The total number of samples are set to be $(10p/\epsilon^2)$.

Metric. We measure the 0-1 classification error on a held-out (clean) test set. We study how the 0-1 error changes with p and ϵ and the parameter estimation error of our proposed method for different contamination levels ϵ.

Baselines. We use the logistic regression MLE and the linear Support Vector Machine (SVM) as our baselines.

Results. We summarize our main findings below:



(a) 0-1 Error vs p at $\epsilon = 0.1$

(b) 0-1 error vs ϵ

(c) $\log(\|\theta^t - \theta^*\|_2)$ vs t for different ϵ

Figure 3.2: Robust Logistic Regression.

- **0/1 Error vs dimension p:** In Figure 3.2(a) we observe that both the SVM and logistic regression MLE perform poorly with increasing dimension. The logistic regression MLE completely flips the labels and has a 0-1 error close to 1, whereas the linear SVM outputs a random hyperplane classifier that flips the label for roughly half of the dataset.
- **0/1 Error vs ϵ and t:** Figures 3.2(b) and 3.2(c) show qualitatively similar results to the linear regression setting, i.e. that the error of our proposed estimator degrades gracefully (and grows linearly) with the contamination level ϵ and that the gradient descent iterates converge linearly.

Finally, in Appendix B.2, we show the efficacy of our algorithm in a semi-synthetic experiment where we attempt to reconstruct face images (from the Cropped Yale Dataset [64]) that have been corrupted with heavy occlusion. In this experiment, the occluding pixels play the role the outliers, and we show that our proposed algorithm significantly outperforms TORRENT, SCRRR and OLS.

3.4.2 Heavy-tailed Estimation

We now consider the heavy-tailed model and present experimental results on synthetic datasets comparing the gradient descent based robust estimator described in Algorithms 3 and 5 (which we call RobustGD) with ERM and several other recent proposals. In these experiments we focus on the problem of linear regression which is described in Section 3.4.1 and work with heavy-tailed noise distributions.

Synthetic Experiments: Simple Linear Regression

Setup. The covariate $x \in \mathbb{R}^p$ is sampled from a zero-mean isotropic Gaussian distribution. We set each entry of θ^* to $1/\sqrt{p}$. The noise w is sampled from a Pareto distribution, with mean zero, variance σ^2 and tail parameter β. The tail parameter β determines the moments of the Pareto random variable. More specifically, the moment of order k exists only if $k < \beta$, hence, smaller the β the more heavy-tailed the distribution. In this setup, we keep the dimension p fixed to 128 and vary n, σ and β. We always maintain the sample-size n to be at least $4p$.

Methods. We compare RobustGD with several baselines. Since we are always in the low-dimensional ($n \geq p$) setting, the solution to ERM has a closed form expression and is simply the OLS solution. We also study OLS-GD, which performs a gradient descent on ERM and is equivalent to using empirical mean as the gradient oracle in our framework. We also compare against the robust estimation techniques of Hsu and Sabato [59] and Namkoong and Duchi [65], which we refer to as RobustHS, RobustDN and two classical techniques namely the LASSO [66] and ridge regression. In our experiments, all the iterative techniques are run until convergence.

Metrics. We use two metrics to compare the performance of various approaches: a) parameter error which is defined as $\|\theta - \theta^*\|_2$ and b) to compare the performance of two estimators $\widehat{\theta}_1$, $\widehat{\theta}_2$, we define the notion of relative efficiency:

$$\textsc{RelEff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\|\widehat{\theta}_2 - \theta^*\|_2 - \|\widehat{\theta}_1 - \theta^*\|_2}{\|\widehat{\theta}_1 - \theta^*\|_2}.$$

Roughly, this corresponds to the percentage improvement in the parameter error obtained using $\widehat{\theta}_1$ over $\widehat{\theta}_2$. Whenever $\textsc{RelEff}(\widehat{\theta}_1, \widehat{\theta}_2) > 0$, $\widehat{\theta}_1$ has a lower parameter error, and higher the value, the more the fractional improvement.

Results. To reduce the variance in the plots presented here, we averaged results over 20 repetitions. Figure 3.3 shows the benefits of using RobustGD over other baselines.

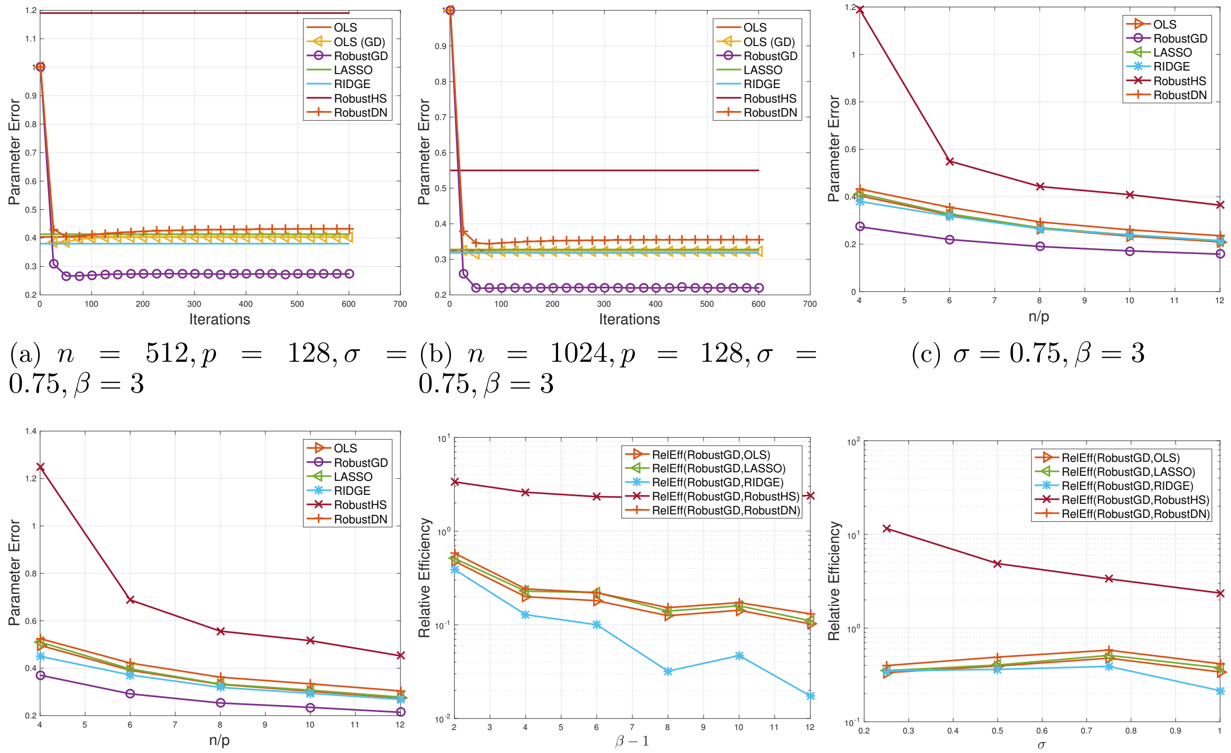
- **Error vs number of iterations:** In Figures 3.3(a), 3.3(b) we plot the excess risk of various approaches against the number of iterations (for OLS, LASSO, ridge regression and the method of Hsu and Sabato [59] we only plot the excess risk of the final iterate). We see that upon convergence RobustGD has a much lower parameter error. As expected, OLS-GD converges to OLS.
- **Error vs number of samples:** Next, in Figures 3.3(c), 3.3(d) we plot the parameter error as n/p increases. We see that RobustGD is always better than other baselines, even when the number of samples is 12 times the dimension p.
- **Relative Efficiency vs β, and σ:** In Figure 3.3(e), we plot the relative efficiency against β, the moment bound of Pareto distribution. This shows that the percentage improvement in the excess risk by RobustGD decreases as the moment bound β increases. This behavior is expected because as we increase the moment bound the tails of the noise distribution become lighter. This shows that there is more benefit in using RobustGD in the heavy-tailed setting. We do a similar study to see the relative efficiency against the variance of the noise distribution. Figure 3.3(f) plots relative efficiency against standard deviation of the noise distribution.

3.5 Theoretical Preliminaries

In this section we develop some theoretical preliminaries. We first develop a general theory on convergence of projected gradient descent in Section 3.5.1. Next we analyze the gradient estimators defined in Algorithms 4 and 5 in Sections 3.5.2 and 3.5.3 respectively. Finally in Sections 3.6 and 3.7 we present consequences of our general theory for the canonical examples of risk minimization described in Section 3.2.2, under Huber contamination and heavy-tailed models.

For some of our examples, we will assume certain mild moment conditions. Concretely, for a random vector $x \in \mathbb{R}^p$, let $\mu = \mathbb{E}[x]$ and Σ be the covariance matrix. Then x has bounded $2k^{\text{th}}$ moments if there exists a constant C_{2k} such that for every unit vector v we have that

$$\mathbb{E}\left[\langle x-\mu, v\rangle^{2k}\right] \leq C_{2k}\left(\mathbb{E}\left[\langle x-\mu, v\rangle^{2}\right]\right)^{k}. \tag{3.12}$$



(a) $n = 512, p = 128, \sigma = 0.75, \beta = 3$

(b) $n = 1024, p = 128, \sigma = 0.75, \beta = 3$

(c) $\sigma = 0.75, \beta = 3$

(d) $\sigma = 1, \beta = 3$

(e) $n = 512, p = 128, \sigma = 0.75$

(f) $n = 512, p = 128, \beta = 3$

Figure 3.3: Linear Regression: Performance comparison of RobustGD against baselines.

3.5.1 Stability of Gradient Descent

In this section we develop a general theory for the convergence of the projected gradient descent described in Algorithm 3. Note that our gradient estimators could be biased and are not guaranteed to be consistent estimators of the true gradient $\nabla\mathcal{R}(\theta)$. This is especially true in the Huber contamination model where it is impossible to obtain consistent estimators of the gradient of the risk because of the non-vanishing bias caused by the contaminated samples. Hence, we turn our attention to understanding the behavior of projected gradient descent with a biased, inexact, gradient estimator of the form in (C.77). Before we present our main result, we define the notion of stability of a gradient estimator, which plays a key role in the convergence of gradient descent.

Definition 2 (Stability)**.** *A gradient estimator is* stable *for a given risk function $\mathcal{R} : \Theta \mapsto \mathbb{R}$ if for some $\phi \in [0, \tau_\ell)$,*

$$\alpha(\widetilde{n}, \widetilde{\delta}) < \tau_\ell - \phi.$$

We denote by κ the following contraction parameter:

$$\kappa := \sqrt{1 - \frac{2\eta\tau_\ell\tau_u}{\tau_\ell + \tau_u}} + \eta\alpha(\widetilde{n}, \widetilde{\delta}), \tag{3.13}$$

and note that $\kappa < 1$. With these definitions in place we state our main result on the stability of gradient descent:

Theorem 8. *Suppose that the gradient estimator satisfies the condition in* (C.77) *and is stable for the risk function $\mathcal{R} : \Theta \mapsto \mathbb{R}$. Then Algorithm 3 initialized at θ^0 with step-size $\eta = 2/(\tau_\ell + \tau_u)$, returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1 - \delta$ for the contraction parameter κ above we have that,*

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\beta(\widetilde{n}, \widetilde{\delta}). \tag{3.14}$$

We defer a proof of this result to the Appendix. For the bound (C.78), the first term is decreasing in T, while the second term is increasing in T. This suggests that for a given n and δ, we need to run just enough iterations for the first term to be bounded by the second. Hence, we can fix the number of iterations T^* as the smallest positive integer such that:

$$T \geq \log_{1/\kappa} \frac{(1-\kappa)\|\theta^0 - \theta^*\|_2}{\beta(\widetilde{n}, \widetilde{\delta})}.$$

Since we obtain linear convergence, i.e. $\kappa < 1$, typically a logarithmic number of iterations suffice to obtain an accurate estimate.

Theorem 56 provides a general result for risk minimization and parameter estimation, and requires bounds on $\alpha(\widetilde{n}, \widetilde{\delta}), \beta(\widetilde{n}, \widetilde{\delta})$ which capture the the error suffered by the gradient estimator for a given risk minimization problem. In any concrete instantiation for a given gradient estimator, risk pair, we first estimate these gradient estimator error bounds by studying the distribution of the gradient of the risk, and then apply Theorem 56. In the next two sections, we provide some general analyses of the gradient estimator in Algorithm 4 for the Huber contamination model, and the gradient estimator in Algorithm 5 for the heavy-tailed model, and which apply to any risk minimization problem. In Sections 3.6,3.7 we then instantiate these gradient estimator error results for various illustrative statistical models such as linear regression, logistic regression, and general exponential families. Plugging these into Theorem 56, we then get consequences of our robustness guarantees for various statistical model, robustness setting pairs.

3.5.2 General Analysis of Huber Contamination Gradient Estimators

We now analyze the gradient estimator described in Algorithm 4 for Huber contamination model and study the error suffered by it. As stated before, Algorithm 4 uses the robust mean estimator of Lai et al. [28]. Hence, while our proof strategy mimics that of Lai et al. [28], we present a different result which is obtained by a more careful non-asymptotic analysis of the algorithm.

We define:

$$\gamma(n,p,\delta,\epsilon) := \left(\frac{p\log p\log\left(n/(p\delta)\right)}{n}\right)^{3/8} + \left(\frac{\epsilon p^2\log p\log\left(\frac{p\log(p)}{\delta}\right)}{n}\right)^{1/4}, \quad (3.15)$$

and with this definition in place we have the following result:

Lemma 9. *Let P be the true probability distribution of z and let P_θ be the true distribution of the gradients $\nabla\bar{\mathcal{L}}(\theta;z)$ on $\mathbb{R}^p$ with mean $\mu_\theta = \nabla\mathcal{R}(\theta)$, covariance Σ_θ, and bounded fourth moments. There exists a positive constant $C_1 > 0$, such that given n samples from the distribution in (3.9), the Huber Gradient Estimator described in Algorithm 4 when instantiated with the contamination level ϵ, with probability at least $1-\delta$, returns an estimate $\widehat{\mu}$ of μ_θ such that,*

$$\|\widehat{\mu} - \mu_\theta\|_2 \le C_1\Big(\sqrt{\epsilon} + \gamma(n,p,\delta,\epsilon)\Big)\|\Sigma_\theta\|_2^{\frac{1}{2}}\sqrt{\log p}.$$

We note in particular, if $n \to \infty$ (with other parameters held fixed) then $\gamma(n,p,\delta,\epsilon) \to 0$ and the error of our gradient estimator satisfies

$$\|\widehat{\mu} - \mu_\theta\|_2 \le C\sqrt{\|\Sigma_\theta\|_2\epsilon\log p},$$

and has only a weak dependence on the dimension p.

We also analyze the gradient estimator obtained by the filtering technique of Diakonikolas et al. [31]. The complete algorithm and is its proof are described in Appendix B.16. Our analysis is obtained by a combination of martingale style arguments along with tight non-asymptotic bounds. As a result, we obtain high-probability bounds which are almost dimension independent (i.e. they depend on the dimension primarily through $\mathrm{tr}(\Sigma_\theta)$).

Lemma 10. *Let P be the true probability distribution of z and let P_θ be the true distribution of the gradients $\nabla\bar{\mathcal{L}}(\theta;z)$ on $\mathbb{R}^p$ with mean $\mu_\theta = \nabla\mathcal{R}(\theta)$, covariance Σ_θ, and bounded second moments. There exists a positive constant*

$C_1 > 0$, *such that given* n *samples from the distribution in* (3.9), *the Huber Gradient Estimator described in Algorithm* 10 *when instantiated with the contamination level* ϵ, *and knowledge of* $\|\Sigma_\theta\|_2$ *and trace* (Σ_θ), *with probability at least* $1 - \delta$, *returns an estimate* $\widehat{\mu}$ *of* μ_θ *such that,*

$$\|\widehat{\mu} - \mu_\theta\|_2 \leq C_1 \|\Sigma_\theta\|_2^{\frac{1}{2}} \max(\epsilon, \frac{\log(1/\delta)}{n})^{\frac{1}{2}} + C_2 \sqrt{\frac{\textit{trace}\,(\Sigma_\theta) \log(p/\delta)}{n}}$$

For Algorithm 10, we see that if $n \to \infty$ (with other parameters held fixed) then the error of our gradient estimator satisfies $\|\widehat{\mu} - \mu_\theta\|_2 \leq C\sqrt{\|\Sigma_\theta\|_2 \epsilon}$. Comparing Algorithms 4 and 10, we see that Algorithm 10 achieves the same asymptotic error rate at weaker assumptions. However, it requires knowledge of an upper bound of the operator norm of the gradients at any point θ. For the general cases of GLMs such as linear regression, $\|\Sigma_\theta\|_2$ depends on $\|\theta^* - \theta\|_2$, *i.e.* it depends on how far the current point is from the true optimal. Since, we don't have this information, we cannot use Algorithm 10 as black-box gradient estimator. We believe that one can typically have an iterative update rule for Σ_{θ^t},(decreasing it after every step), but we don't explore it further. However, in cases such as Exponential Family, one can show that $\|\Sigma_\theta\|_2 < C \ \forall\ \theta$, and we derive bounds achieved by using Algorithm 10 as a gradient estimator.

3.5.3 General Analysis of Heavy-tailed Model Gradient Estimator in Algorithm 5

In this section we analyze the gradient estimator for heavy-tailed setting, described in Algorithm 5. The following result shows that the gradient estimate has exponential concentration around the true gradient, under the mild assumption that the gradient distribution has bounded second moment. Its proof follows directly from the analysis of geometric median-of-means estimator of Minsker [16]. We use trace (Σ_θ) to denote the trace of the matrix Σ_θ.

Lemma 11. *Let* P *be the probability distribution of* z *and* P_θ *be the distribution of the gradients* $\nabla \bar{\mathcal{L}}(\theta; z)$ *on* $\mathbb{R}^p$ *with mean* $\mu_\theta = \nabla \mathcal{R}(\theta)$, *covariance* Σ_θ. *Then the heavy-tailed gradient estimator described in Algorithm* 5 *returns an estimate* $\widehat{\mu}$ *that satisfies the following exponential concentration inequality, with probability at least* $1 - \delta$*:*

$$\|\widehat{\mu} - \mu_\theta\|_2 \leq 11 \sqrt{\frac{\textit{trace}\,(\Sigma_\theta) \log(1.4/\delta)}{n}}.$$

The results of the Lemmas 9 and 11 effectively ensure that under relatively mild moment assumptions we can robustly estimate multivariate mean vectors and in subsequent sections we show how to leverage these strong guarantees for robust parametric estimation.

3.6 Consequences for Estimation under ϵ-Contaminated Model

We now turn our attention to the examples introduced earlier, and present specific applications of Theorem 56, for parametric estimation under Huber contamination model. As shown in Lemma 9, we need the added assumption that the true gradient distribution has bounded fourth moments, which suggests the need for additional assumptions. We make our assumptions explicit and defer the technical details to the Appendix.

3.6.1 Linear Regression

We assume that the covariates $x \in \mathbb{R}^p$ have bounded 8^{th}-moments and the noise w has bounded 4^{th} moments.

Theorem 9 (Robust Linear Regression). *Consider the statistical model in (B.17), and suppose that the number of samples n is large enough such that $\gamma(\widetilde{n}, p, \widetilde{\delta}) < \frac{C_1 \tau_\ell}{\|\Sigma\|_2 \sqrt{\log p}}$ and the contamination level is such that*

$$\epsilon < \left(\frac{C_2 \tau_\ell}{\|\Sigma\|_2 \sqrt{\log p}} - \gamma(\widetilde{n}, p, \widetilde{\delta}) \right)^2,$$

for some constants C_1 and C_2. Then, there are universal constants C_3, C_4, such that if Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 4 as gradient estimator, then it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that for a contraction parameter $\kappa < 1$, with probability at least $1 - \delta$,

$$\|\widehat{\theta}^t - \theta^*\|_2 \le \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{C_3 \sigma \sqrt{\|\Sigma\|_2 \log p}}{1 - \kappa} \Big(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta}) \Big). \tag{3.16}$$

In the asymptotic setting when the number of samples $n \to \infty$ (and other parameters are held fixed), we see that for the Huber Gradient Estimator, the corresponding maximum allowed contamination level is

$$\epsilon < \frac{C_1 \tau_\ell^2}{\tau_u^2 \log p},$$

i.e. the better conditioned the covariance matrix Σ, the higher the contamination level we can tolerate.

Plugin Estimation. For linear regression, the true parameter can be written in closed form as $\theta^* = \mathbb{E}[xx^T]^{-1}\mathbb{E}[xy]$. A non-iterative way to estimate θ^* is to separately estimate $\mathbb{E}[xx^T]$ and $\mathbb{E}[xy]$ using robust covariance and mean oracles respectively. Under the assumption that $x \sim \mathcal{N}(0, \mathcal{I}_p)$, one can reduce the problem to robustly estimating $\mathbb{E}[xy]$. Under this setting, we now present a result using our robust mean estimator (from Lemma 9) to directly estimate $\mathbb{E}[xy]$. Recall, the definition of γ in (3.15). We have the following result:

Corollary 10. *Consider the model in* (B.17) *with the covariates drawn from* $\mathcal{N}(0, \mathcal{I}_p)$ *and* $w \in \mathcal{N}(0,1)$, *then there are universal constants* C_1, C_2 *such that if* $\epsilon < C_1$, *the plugin estimator* $\widehat{\theta}$ *of* $\mathbb{E}[xy]$ *described above with probability at least* $1 - \delta$ *satisfies:*

$$\|\widehat{\theta} - \theta^*\|_2 \leq C_2 \sqrt{(1 + 2\|\theta^*\|_2^2) \log p}\left(\epsilon^{\frac{1}{2}} + \gamma(n, p, \delta, \epsilon)\right). \tag{3.17}$$

Corollary 10 suggests that even when the plugin estimator does not have to estimate the covariance matrix, the error of the plugin estimator depends on $\|\theta^*\|_2$, which would make the estimator vacuous if $\|\theta^*\|_2$ scales with the dimension p. In Appendix B.2, we empirically verify that this upper bound((3.17)) is indeed tight, *i.e.* the asymptotic error of plugin estimator does indeed scale linearly with $\|\theta^*\|_2$. Intuitively, this dependence occurs because the variance of the sufficient statistic xy scales with $\|\theta^*\|_2^2$ and from minimax results for robust mean estimation [11], it is known the dependence on variance is unavoidable in the ϵ-contaminated setting. Next, we apply our estimator to generalized linear models.

3.6.2 Generalized Linear Models

Here we assume that the covariates have bounded 8th moments. Additionally, we assume smoothness of $\Phi'(\cdot)$ around θ^*. In particular, we assume that there exist universal constants $L_{\Phi,2k}$, B_{2k} such that

$$\mathbb{E}_x\left[\left|\Phi'(\langle x,\theta\rangle)-\Phi'(\langle x,\theta^*\rangle)\right|^{2k}\right]\leq L_{\Phi,2k}\|\theta^*-\theta\|_2^{2k}+B_{\Phi,2k},\quad\text{for } k=1,2,4$$

We also assume that $\mathbb{E}_x[\left|\Phi^{(t)}(\langle x,\theta^*\rangle)\right|^k]\leq M_{\Phi,t,k}$ where $\Phi^{(t)}(\cdot)$ is the t^{th}-derivative of $\Phi(\cdot)$.

Theorem 11 (Robust Generalized Linear Models)**.** *Consider the statistical model in (5.10), and suppose that the number of samples n is large enough such that*

$$\gamma(\widetilde{n},p,\widetilde{\delta})<\frac{C_1\tau_\ell}{\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[L_{\Phi,4}^{\frac{1}{4}}+L_{\Phi,2}^{\frac{1}{2}}]},$$

and the contamination level is such that,

$$\epsilon<\left(\frac{C_2\tau_\ell}{\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[L_{\Phi,4}^{\frac{1}{4}}+L_{\Phi,2}^{\frac{1}{2}}]}-\gamma(\widetilde{n},p,\widetilde{\delta})\right)^2,$$

for some constants C_1 and C_2. Then, there are universal constants C_3,C_4, such that if Algorithm 3 is initialized at θ^0 with stepsize $\eta=2/(\tau_u+\tau_\ell)$ and Algorithm 4 as gradient estimator, then it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1-\delta$

$$\begin{aligned}\|\widehat{\theta}^t-\theta^*\|_2\leq&\kappa^t\|\theta^0-\theta^*\|_2\\&+\frac{C_3\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[B_{\Phi,4}^{\frac{1}{4}}+B_{\Phi,2}^{\frac{1}{2}}+c(\sigma)^{\frac{1}{2}}M_{\Phi,2,2}^{\frac{1}{4}}+c(\sigma)^{\frac{3}{4}}M_{\Phi,4,1}^{\frac{1}{4}}]}{1-\kappa}\Big(\epsilon^{\frac{1}{2}}+\gamma(\widetilde{n},p,\widetilde{\delta})\Big),\end{aligned}\tag{3.18}$$

for some contraction parameter $\kappa<1$.

Note that for the case of linear regression with Gaussian noise, it is relatively straightforward to see that $L_{\Phi,2k}=C_{2k}\|\Sigma\|_2^k$, $B_{\Phi,2k}=0$, $M_{\Phi,t,k}=1\ \ \forall(t=2,k\in\mathcal{N})$ and $M_{\Phi,t,k}=0\ \ \forall(t\geq 3,k\in\mathcal{N})$ under the assumption of bounded 8^{th} moments of the covariates; which essentially leads to an equivalence between Theorem 37 and Theorem 11 for this setting. In the following section, we instantiate the above Theorem for logistic regression and compare and contrast our results to other existing methods.

Logistic Regression

By observing that $\Phi^{(t)}(\cdot)$ is bounded for logistic regression for all $t \geq 1$, we can see that $L_{\Phi,2k} = 0$, and that there exists a universal constant $C > 0$ such that $B_{\Phi,2k} < C$ and $M_{\Phi,t,k} < C \ \ \forall (t \geq 1, k \in \mathcal{N})$.

Corollary 12 (Robust Logistic Regression). *Consider the model in* (5.13)*, then there are universal constants C_1, C_2, such that if $\epsilon < C_1$, then Algorithm 3 initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 4 as gradient estimator, returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$, such that with probability at least $1 - \delta$*

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{C_2 \sqrt{\|\Sigma\|_2 \log p}}{1 - \kappa} \left(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta}) \right),$$

for some contraction parameter $\kappa < 1$.

Under the restrictive assumption that $x \sim \mathcal{N}(0, \mathcal{I}_p)$, Balakrishnan et al. [50] exploited Stein's trick to derive a plugin estimator for logistic regression. However, similar to the linear regression, the error of the plugin estimator scales with $\|\theta^*\|_2$, which is avoided in our robust gradient descent algorithm. We also note that our algorithm extends to general covariate distributions.

3.6.3 Exponential Family

Here we assume that the random vector $\phi(z), z \sim P$ has bounded 4th moments.

Theorem 13 (Robust Exponential Family 1). *Consider the model in* (3.8)*, then there are universal constants C_1, C_2, such that if $\epsilon < C_1$, then Algorithm 3 initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 4 as gradient oracle, returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$, such that with probability at least $1 - \delta$*

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{C_2 \sqrt{\tau_u \log p}}{1 - \kappa} \left(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta}) \right),$$

for some contraction parameter $\kappa < 1$.

We also state the results obtained for Exponential Families when using Algorithm 10 as gradient estimator. In this case we only need bounded second moment assumptions on the random vector $\phi(z), z \sim P$.

Theorem 14 (Robust Exponential Family 2). *Consider the model in* (3.8)*, then there are universal constants C_1, C_2, such that if $\epsilon < C_1$, then Algorithm 3* *initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 10 as gradient oracle, returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$, such that with probability at least $1 - \delta$*

$$\|\widehat{\theta}^t - \theta^*\|_2 \le \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{C_2}{1-\kappa}\left(\sqrt{\tau_u}\epsilon^{\frac{1}{2}} + \sqrt{\frac{\mathit{trace}\left(\nabla^2 A(\theta^*)\right)\log 1/\widetilde{\delta}}{\widetilde{n}}}\right),$$

for some contraction parameter $\kappa < 1$.

Plugin Estimation. Since the true parameter θ^* is the minimizer of the negative log-likelihood, we know that $\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta^*)] = 0$, which implies that $\nabla A(\theta^*) = \mathbb{E}_{\theta^*}[\phi(Z)]$. This shows that the true parameter θ^* can be obtained by inverting the ∇A operator, whenever possible. In the robust estimation framework, we can use a robust mean of the sufficient statistics to estimate $\mathbb{E}_{\theta^*}[\phi(Z)]$. We instantiate this estimator using the mean estimator of [28] to estimate $\mathbb{E}_{\theta^*}[\phi(Z)]$:

Corollary 15. *Consider the model in (3.8), then there are universal constants C_1, C_2 such that if $\epsilon < C_1$, then [28] returns an estimate $\widehat{\mu}$ of $\mathbb{E}[\phi(z)]$, such that with probability at least $1 - \delta$*

$$\|\mathcal{P}_\Theta\left[(\nabla A)^{-1}\widehat{\mu}\right] - \theta^*\|_2 \le C_2 \frac{\sqrt{\tau_u \log p}}{\tau_\ell}\left(\epsilon^{\frac{1}{2}} + \gamma(n, p, \delta, \epsilon)\right), \tag{3.19}$$

where $\mathcal{P}_\Theta[\theta] = \operatorname{argmin}_{y\in\Theta}\|y - \theta\|_2^2$ is the projection operator onto the feasible set Θ.

3.6.4 Discussion and Limitations

In the asymptotic setting of $n \to \infty$, Algorithm 3 with Algorithm 4 as gradient estimator converges to a point $\widehat{\theta}$ such that $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon \log p})$. Hence, our error scales only logarithmically with the dimension p. This dependency on the dimension p is a facet of using the estimator from Lai et al. [28] for gradient estimation. Using better oracles will only improve our performance. Next, we would like to point to the difference in the maximum allowed contamination ϵ^* between the three models. For logistic regression and exponential family, $\epsilon^* < C_1$, while for linear regression, $\epsilon^* < \frac{C_1\tau_\ell^2}{\tau_u^2 \log p}$. These differences are in large part due to differing variances of the gradients, which naturally depend on the underlying risk function. This scaling of the variance of gradients for linear regression also provides insights into the limitations of our robust gradient descent approach in Algorithm 3. In the Appendix, we provide an upper bound for the contamination level ϵ based on the initialization point θ^0, above which, Algorithm 3 would not work for any gradient estimator.

3.7 Consequences for Heavy-Tailed Estimation

In this section we present specific applications of Theorem 56 for parametric estimation, under heavy-tailed setting. The proofs of the results can be found in the Appendix.

3.7.1 Linear Regression

We first consider the linear regression model described in (B.17). We assume that the covariates $x \in \mathbb{R}^p$ have bounded 4^{th}-moments and the noise w has bounded 2^{nd} moments. This assumption is needed to bound the error in the gradient estimator (see Lemma 11).

Theorem 16 (Heavy Tailed Linear Regression)**.** *Consider the statistical model in* (B.17)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$\widetilde{n} > \frac{\mathit{trace}\,(\Sigma)\,\tau_u}{\tau_l^2} \log 1/\widetilde{\delta}$$

and if Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 5 as gradient estimator, then it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1-\delta$

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{C_2 \sigma}{1-\kappa} \sqrt{\frac{\mathit{trace}\,(\Sigma) \log 1/\widetilde{\delta}}{\widetilde{n}}}, \tag{3.20}$$

for some contraction parameter $\kappa < 1$.

3.7.2 Generalized Linear Models

In this section we consider generalized linear models described in (5.10), where the covariate x is allowed to have a heavy-tailed distribution. Here we assume that the covariates have bounded 4^{th} moment. Additionally, we assume smoothness of $\Phi'(\cdot)$ around θ^*. Specifically, we assume that there exist universal constants $L_{\Phi,2k}$, B_{2k} such that

$$\mathbb{E}_x\left[\left|\Phi'(\langle x,\theta\rangle)-\Phi'(\langle x,\theta^*\rangle)\right|^{2k}\right]\le L_{\Phi,2k}\|\theta^*-\theta\|_2^{2k}+B_{\Phi,2k},\quad \text{for } k=1,2$$

We also assume that $\mathbb{E}_x\left[\left|\Phi^{(t)}(\langle x,\theta^*\rangle)\right|^k\right]\le M_{\Phi,t,k}$ for $t\in\{1,2,4\}$, where $\Phi^{(t)}(\cdot)$ is the t^{th}-derivative of $\Phi(\cdot)$.

Theorem 17 (Heavy Tailed Generalized Linear Models)**.** *Consider the statistical model in* (5.10)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$\widetilde{n}>\frac{C\,trace\,(\Sigma)\,\sqrt{C_4}\sqrt{L_{\Phi,4}}\log\,(1/\widetilde{\delta})}{\tau_\ell^2},$$

and if Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u+\tau_\ell)$ and Algorithm 5 as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1-\delta$

$$\begin{aligned}\|\widehat{\theta}^t-\theta^*\|_2\le&\kappa^t\|\theta^0-\theta^*\|_2\\&+\frac{C_2\left[B_{\Phi,4}^{\frac{1}{4}}+c(\sigma)^{\frac{1}{2}}M_{\Phi,2,2}^{\frac{1}{4}}+c(\sigma)^{\frac{3}{4}}M_{\Phi,4,1}^{\frac{1}{4}}\right]}{1-\kappa}\left(\sqrt{\frac{trace\,(\Sigma)\log(1/\widetilde{\delta})}{\widetilde{n}}}\right),\end{aligned}\tag{3.21}$$

for some contraction parameter $\kappa<1$.

We now instantiate the above Theorem for the logistic regression model.

Corollary 18 (Heavy Tailed Logistic Regression)**.** *Consider the model in* (5.13)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$\widetilde{n}>\frac{C_1^2\,trace\,(\Sigma)}{\tau_l^2}\log\,(1/\widetilde{\delta}).$$

and if Algorithm 3 initialized at θ^0 with stepsize $\eta = 2/(\tau_u+\tau_\ell)$ and Algorithm 5 as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1-\delta$

$$\|\widehat{\theta}^t-\theta^*\|_2\le\kappa^t\|\theta^0-\theta^*\|_2+\frac{C_2}{1-\kappa}\left(\sqrt{\frac{trace\,(\Sigma)\log(1/\widetilde{\delta})}{\widetilde{n}}}\right),$$

for some contraction parameter $\kappa<1$.

3.7.3 Exponential Family

We now instantiate Theorem 56 for parameter estimation in heavy-tailed exponential family distributions. Here we assume that the random vector $\phi(z), z \sim P$ has bounded 2^{nd} moments, and we obtain the following result:

Theorem 19 (Heavy Tailed Exponential Family)**.** *Consider the model in* (3.8)*. If Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 5 as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$, such that with probability at least $1 - \delta$*

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa} C \sqrt{\frac{trace\left(\nabla^2 A(\theta^*)\right) \log 1/\widetilde{\delta}}{\widetilde{n}}},$$

for some contraction parameter $\kappa < 1$ and universal constant C.

Recently, Hopkins [23] developed a Sum-of-Squares based polynomial-time algorithm that achieves optimal error for mean estimation. Although polynomial time, their algorithm is not practically implementable. However, when we use their algorithm to robustly estimate gradients, the theoretical results obtained in this section can be improved substantially, and are near-optimal. In particular, instead of the $O(\sqrt{\frac{\text{trace}(\Sigma)\log(1/\widetilde{\delta})}{\widetilde{n}}})$ term obtained in this section, we get results of the form $O(\sqrt{\frac{\|\Sigma\|_2 \log(1/\widetilde{\delta})}{\widetilde{n}}} + \sqrt{\frac{\text{trace}(\Sigma)}{\widetilde{n}}})$. Note that our proposed estimators are the first polynomial time estimators, which achieve these exponential concentration in heavy-tailed models. We present these results in Appendix B.15.

3.8 Discussion

In this paper we introduced a broad class of robust estimators, that leverage the inherent robustness of gradient descent, together with the observation that for risk minimization in most statistical models, the gradient of the risk takes the form of a simple multivariate mean, which can be robustly estimated using recent work on robust mean estimation. In contrast to classical M-estimators that use robust estimates of the risk, our class of estimators employ a shift in perspective, and use robust estimates of gradients of the risk instead, which can then be embedded into a simple projected gradient descent iterative algorithm. Our class of robust gradient descent estimators work well in practice and in many cases outperform other robust (and non-robust) estimators. We also show that these estimators have strong robustness guarantees under Huber's ϵ-contamination model and for heavy-tailed distributions.

There are several avenues for future work, including a better understanding of robust mean estimation, any improvement in which would immediately translate to improved guarantees for our robust gradient descent estimators. For example, our current algorithm requires sample-splitting which is wasteful. One way to get around this sample-splitting is to develop uniformly robust gradient estimators. We provide some partial results along these lines in Appendix B.17. Finally, it would also be of interest to understand the extent to which we can relax our assumption of strong convexity of the population risk. In particular, our analysis relies on the linear convergence of the population iterates (a consequence of strong convexity and smoothness but not equivalent to it). Hence, we can, for instance, in a straightforward way analyze robust gradient descent in certain non-convex problems, for instance those arising in the estimation of a mixture of two Gaussians (under suitable initialization) [49] or settings in which the Polyak-Lojasiewicz condition (a weaker condition allowing interesting non-convex functions but still implying the linear convergence of gradient descent iterates) holds [67]. Completely eliminating this assumption, to instead consider cases where the risk is convex but not strongly-convex (for instance) poses identifiability issues, and warrants further investigation and is an interesting direction. In particular, it might necessitate focusing on prediction error (in linear or logistic regression, and analogues in the other models) as opposed to parameter error.

3.9 Acknowledgements

A.P., A.S., P.R. acknowledge the support of PNC, and NSF via IIS-1149803, IIS-1664720, DMS-1264033. S.B. acknowledges the support of NSF via DMS-1713003. We thank Larry Wasserman and Ankit Pensia for helpful comments on the paper.

Chapter 4

On Learning Ising Models under Huber's Contamination Model

We study the problem of learning Ising models in a setting where some of the samples from the underlying distribution can be arbitrarily corrupted. In such a setup, we aim to design statistically optimal estimators in a high-dimensional scaling in which the number of nodes p, the number of edges k and the maximal node degree d are allowed to increase to infinity as a function of the sample size n. Our analysis is based on exploiting moments of the underlying distribution, coupled with novel reductions to univariate estimation. Our proposed estimators achieve an optimal dimension independent dependence on the fraction of corrupted data in the contaminated setting, while also simultaneously achieving high-probability error guarantees with optimal sample-complexity. We corroborate our theoretical results by simulations.

4.1 Introduction

Undirected graphical models (also known as Markov random fields (MRFs)) have gained significant attention as a tool for discovering and visualizing dependencies among covariates in multivariate data. Graphical models provide compact and structured representations of the joint distribution of multiple random variables using graphs that represent conditional independences between the individual random variables. They are used in domains as varied as natural language processing[68], image processing [69, 70, 71], spatial statistics [72] and computational biology [73], among others. Given samples drawn from the distribution, a key problem of interest is to recover the underlying dependencies represented by the graph. A slew of recent results [74, 75, 76] have shown that it is possible to learn such models even in domains and settings where the number of samples is potentially smaller than the number of variables. These results however make the common assumption that the sample data is clean, and have no corruptions. However, modern data sets that arise in various branches of science and engineering are no longer carefully curated. They are often collected in a decentralized and distributed fashion, and consequently are plagued with the complexities of outliers, and even adversarial manipulations.

Huber [63] proposed the ϵ-contamination model as a framework to study such datasets with potentially arbitrary corruptions. In this setting, instead of observing samples directly from the true distribution $\mathbb{P}^\star$, we observe samples drawn from $\mathbb{P}_\epsilon$, which for an arbitrary distribution Q is defined as a mixture model,

$$\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q. \tag{4.1}$$

Then, given n samples from $\mathbb{P}_\epsilon$, the goal is to recover functionals of $\mathbb{P}^\star$. There has been a lot of classical work on estimators for the ϵ-contamination model setting that largely trade off computational versus statistical efficiency (see [77] and references therein). Moreover, there has been substantial progress [27, 28, 29, 30, 31, 32, 46, 78] on designing provably robust estimators which are computationally tractable while achieving near-optimal contamination dependence (*i.e.* dependence on the fraction of outliers ϵ). However, to the best of our knowledge, there are no known results for learning general graphical models robustly.

4.1.1 Related Work

In this work, we focus on the specific undirected graphical model sub-class of Ising models [79]. There has been a lot of work for learning Ising models in the uncontaminated setting dating back to the classical work of Chow and Liu [80]. Csiszár and Talata [81] discuss pseudo likelihood based approaches for estimating the neighborhood at a given node in MRFs. Subsequently, a simple search based method is described in [82] with provable guarantees. Later, Ravikumar et al. [76] showed that under an incoherence assumption, node-wise (regularized) estimators provably recover the correct dependency graph with a small number of samples. Recently, there has been a flurry of work [83, 84, 85, 86, 87] to get computationally efficient estimators which recover the true graph structure without the incoherence assumption, including extensions to identity and independence testing [88]. However, all the aforementioned results are in the uncontaminated setting. Recently, Lindgren et al. [89] derived preliminary results for learning Ising models robustly. However, their upper and lower bounds do not match. Moreover, their analysis primarily focuses on the robustness of the Sparsitron algorithm in [84], and they do not explore the effect of the underlying graph and correlation structures comprehensively.

Contributions. In this work, we give the *first* statistically optimal estimator for learning Ising models under the ϵ-contamination model. Our estimators achieve a dimension-independent asymptotic error as a function of the fraction of outliers ϵ, while simultaneously achieving high probability deviation bounds. As an important special case of our results, we also close known sample complexity gaps in the uncontaminated setting for some classes of Ising models. We finally corroborate our theoretical findings with simulation studies.

4.1.2 Background and Problem Setup

We begin with some background on Ising models and then provide the precise formulation of the problem. We follow the notation of Santhanam and Wainwright [90] very closely.

Consider an undirected graph $G = (V, E)$ defined over a set of vertices $V = \{1, 2, \ldots, p\}$ with edges $E \subset \{(s,t) : s,t \in V, s \neq t\}$. The neighborhood of any node $s \in V$ is the subset $\mathcal{N}(s) \subset V$ given by $\mathcal{N}(s) \overset{\text{def}}{=} \{t|(s,t) \in E\}$, and the degree of any vertex s is given by $d_s = |\mathcal{N}(s)|$. Then, the degree of a graph $d = \max_s d_s$ is the maximum vertex degree, and $k = |E|$ is the total number of edges. We obtain an MRF by associating a random variable X_v at each vertex $v \in V$, and then considering a joint distribution $\mathbb{P}$ over the random vector $(X_1, \ldots, X_p)$. An Ising model is a special instantiation of an MRF where each random variable X_s take values in $\{-1, +1\}$, and the joint probability mass function is given by:

$$\mathbb{P}_\theta(x_1, \ldots, x_p) \propto \exp\left(\sum_{1 \leq s < t \leq p} \theta_{st} x_s x_t \right), \tag{4.2}$$

where we view θ as the parameter vector of the distribution. Note that $\theta \in \mathbb{R}^{p \times p}$ is such that $\theta_{ij} = 0 \Leftrightarrow (i, j) \notin E$ and $\theta = \theta^T$.

Graph Classes. In this work, we consider two classes of Ising models (4.2) based on the conditions imposed on the edge set:

1. $\mathcal{G}_{p,d}$: the collection of graphs G with p vertices such that each vertex has at most d neighbors for some $d \geq 1$, and
2. $\mathcal{G}_{p,k}$: the collection of graphs G with p vertices such that the total number of edges in the graph is at most k for some $k \geq 1$.

In addition to these structural properties, we also consider some subclasses based on the parameters of the Ising model. We define the *model width* as:

$$\omega^*(\theta(G)) \stackrel{\text{def}}{=} \max_{u \in V} \sum_{v \in V} |\theta_{uv}|.$$

It is well-known (see for instance [90]) that estimation in Ising models becomes harder with increasing value of edge parameters, since, large values of edge parameters may hide the contributions of other edges. Similarly, we define the *minimum edge weight* as:

$$\lambda^*(\theta(G)) \stackrel{\text{def}}{=} \min_{(s,t) \in E} |\theta_{st}| \,.$$

With these structural and parameter properties in place, we define the classes of Ising models that we will be studying in the rest of the paper. Given a pair of positive numbers (λ, ω):

1. $\mathcal{G}_{p,d}(\lambda, \omega)$: the set of all Ising models defined over a graphs G with p vertices, with each vertex having degree at most d and parameters satisfying

$$\lambda^*(\theta(G)) \geq \lambda \text{ and } \omega^*(\theta(G)) \leq \omega.$$

2. $\mathcal{G}_{p,k}(\lambda, \omega)$: the set of all Ising models defined over a graphs G with p vertices, with total number of edges at most k and parameters satisfying

$$\lambda^*(\theta(G)) \geq \lambda \text{ and } \omega^*(\theta(G)) \leq \omega.$$

Furthermore, we work in the **high temperature regime** where we assume that the model width bound $\omega^*(\theta(G)) \leq 1 - \alpha$ for some $\alpha > 0$. Note that this assumption implies the Dobrushin condition [91], which in case of Ising models is given by

$$\max_{u \in V} \sum_{v \in V} \tanh(|\theta_{uv}|) \leq 1 - \alpha, \qquad \alpha \in (0, 1). \tag{4.3}$$

While this may seem restrictive, this assumption is widely popular for studying Ising models, for example, see related works in statistical physics [92, 93], mixing times of Glauber dynamics [94, 95], correlation decay [96] and more recently in estimation and testing problems [88, 97].

Notation: Given a matrix M of dimensions $l \times m$, we will denote the i^{th} row of matrix by M_i and the $(i, j)^{th}$ element by M_{ij} or $M(i, j)$. M_{-i} denotes the sub-matrix formed by all rows except i, and analogously $M_{:,-j}$ denotes the sub-matrix formed by all columns except j. $M(i)$ denotes the vector $[M_i]_{-i}$ i.e., the i^{th} row of M excluding element M_{ii}. Given a vector v, $\|v\|_p = \sqrt[p]{\sum_i |v_i|^p}$ denotes its ℓ_p-norm, and its ℓ_∞-norm is given by $\|v\|_{\max} = \max_i |v_i|$. For a matrix M, $\|M\|_{p,q}$ denotes the mixed $\ell_{p,q}$-norm, which is the q-norm of the collection of p-norms of the rows of M. We also use the shorthand $[d] = \{1, 2, \dots, d\}$. We denote the total variation (TV) distance between two discrete distributions p, q with support $\mathcal{X}$ by $d_{\mathrm{TV}}(p, q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|$.

4.2 Information-theoretic bounds for the ϵ-contamination model

Recall that in the ϵ-contamination model (4.1), we observe n samples from $\mathbb{P}_\epsilon = (1 - \epsilon)\mathbb{P}^\star + \epsilon Q$. In this model, even in the asymptotic setting as $n \to \infty$, we cannot expect to recover the true parameters exactly. To see this, suppose that $\mathbb{P}_1^\star, \mathbb{P}_2^\star$ are such that there exist two distributions Q_1 and Q_2 such that

$$\mathbb{P}_\epsilon = (1 - \epsilon)\mathbb{P}_1^\star + \epsilon Q_1 = (1 - \epsilon)\mathbb{P}_2^\star + \epsilon Q_2,$$

then, we cannot hope to distinguish between the two distributions. It is easy to show (see [98]) that the above condition is equivalent to assuming that $d_{\mathrm{TV}}(\mathbb{P}_1^\star, \mathbb{P}_2^\star) = \frac{\epsilon}{1-\epsilon}$. Thus, for any given contaminated distribution $\mathbb{P}_\epsilon$, there is a set of possible uncontaminated distributions (including the ground truth uncontaminated distribution among others) within a ball of some fixed radius with respect to the TV distance, any of which could give rise to the given contaminated distribution $\mathbb{P}_\epsilon$. Thus, when estimating the uncontaminated distribution with respect to some loss function, in the worst case we could incur loss corresponding to the farthest pair of distributions in the ball of some fixed radius with respect to TV distance. This is captured by the geometric notion of modulus of continuity [99], which can then be used to derive sharp bounds on estimation in such a setting:

Definition 3 (TV modulus of continuity)**.** *Given a loss function $L : \Theta \times \Theta \to \mathbb{R}^+$ defined over the parameter space Θ, a class of distributions $\mathcal{D}$, a functional $f : \mathcal{D} \to \Theta$ and a proximity parameter ϵ, the modulus of continuity $\omega(f, \mathcal{D}, L, \epsilon)$ is defined as*

$$\omega(f, \mathcal{D}, L, \epsilon) \overset{\text{def}}{=} \sup_{\substack{\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{D} \\ d_{\text{TV}}(\mathbb{P}_1, \mathbb{P}_2) \leq \epsilon}} L(f(\mathbb{P}_1), f(\mathbb{P}_2)). \tag{4.4}$$

Intuitively, this quantity controls how far the functionals of two distributions can be, subject to the constraint that the TV distance between them is ϵ. Note that for general Ising models, there do not exist *any* results that directly relate the total variation distance to the difference in parameters *i.e.* which study the TV modulus of continuity for the parameters of an Ising model.

A key contribution of our work is to establish sharp upper bounds on the TV modulus of continuity for parameter error in the high temperature regime. The loss function is considered to be the $\ell_{2,\infty}$ norm *i.e.* for matrices $x, y \in \mathbb{R}^{p \times p}$, $L(x, y) = \max_i \|x_i - y_i\|_2$.

Theorem 20. *Consider two Ising models defined over two graphs $G^{(1)}$ and $G^{(2)}$ with p vertices with parameters $\theta^{(1)}$ and $\theta^{(2)}$ respectively, each of which satisfy the high temperature condition (4.3) with constant α. If $d_{TV}\left(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}\right) \leq \epsilon$, then we have that:*

$$\|\theta^{(1)}(i) - \theta^{(2)}(i)\|_2 \lesssim {}^{1}\epsilon\sqrt{C_1(\alpha) \log\left(\frac{2}{\epsilon}\right)} \quad \textit{for all} \quad i \in [p],$$

where $C_1(\alpha)$ is a constant depending on α.

Observe that Theorem 20 shows that the parameter error is *independent* of the dimension p, degree d and the number of edges k. Furthermore, it is also independent of the minimum edge weight λ. As expected, when $\epsilon \to 0$, we see

[1] Here and throughout our paper we use the notation $\lesssim$ to denote an inequality with universal constants dropped for conciseness.

that the parameters are equal providing an alternate route to showing that the parameters of an Ising model are identifiable in the high temperature setting. We also establish that the dependence on ϵ is tight upto logarithmic factors by providing a complementary lower bound – proofs of which are made available in the appendix (Sections D.3.1 and D.3.2).

Lemma 12. *There exists two Ising models satisfying the properties in Theorem 20 whose parameters $\theta^{(1)}$ and $\theta^{(2)}$ satisfy:*

$$\|\theta^{(1)}(i) - \theta^{(2)}(i)\|_2 \gtrsim \epsilon \ \textit{ for all } \ i \in [p].$$

4.3 TV Projection Estimators

Recall the geometric picture of TV contamination discussed in the previous section: given the contaminated distribution, there is a set of possible uncontaminated distributions within a ball of some fixed radius with respect to TV. It is thus natural to consider the TV projection of the contaminated distribution onto the set of all possible uncontaminated distributions. These are also called *minimum distance estimators* and were proposed by Donoho and Liu [8], which we consider for our setting to learn Ising models robustly, leveraging our Theorem 20.

4.3.1 Population Robust Estimators for $\mathcal{G}_p$

Let us first consider the population setting *i.e.*, in which we have distribution access to the contaminated distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^*} + \epsilon Q$, where $\mathbb{P}_{\theta^*} \in \mathcal{G}_p(\lambda, \omega)$ [2]. In this setting, we use the minimum distance estimator [8] to construct robust estimators. In particular, let $\mathbb{P}_{\widehat{\theta}_{\text{MDE}}}$ be the minimum distance estimate defined as

$$\mathbb{P}_{\widehat{\theta}_{\text{MDE}}} = \operatorname*{argmin}_{\mathbb{P}_\theta \in \mathcal{G}_p} d_{\text{TV}}(\mathbb{P}_\theta, \mathbb{P}_\epsilon). \tag{4.5}$$

This estimator is effectively the TV projection of the contaminated distribution onto the set of all Ising model distributions whose underlying graph lies in $\mathcal{G}_p$.

Noting that $d_{\text{TV}}(\mathbb{P}_{\theta^*}, \mathbb{P}_{\widehat{\theta}_{\text{MDE}}}) \le \epsilon$, by an application of the triangle inequality we have that $d_{\text{TV}}(\mathbb{P}_{\theta^*}, \mathbb{P}_{\widehat{\theta}_{\text{MDE}}}) \le 2\epsilon$. Combining this with Theorem 20, we get

[2] We define the class $\mathcal{G}_p(\lambda, \omega)$ as the set of Ising models defined over p vertices with minimum edge weight λ and model width ω

that,

$$\|\widehat{\theta}_{\mathrm{MDE}}(i) - \theta^*(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{2}{\epsilon}\right)} \text{ for all } \ i \in [p].$$

Corollary 21. *Let $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE},\lambda}}$ be the TV projection of the contaminated distribution $\mathbb{P}_\epsilon$ onto the class of Ising models $\mathcal{G}_{p,d}$ with minimum edge weight at least λ. Define the edge set of $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}}$ as $E(\widehat{\theta}_{\mathrm{MDE},\lambda}) = \{(i,j) : |\widehat{\theta}_{\mathrm{MDE},\lambda}(i,j)| > \frac{\lambda}{2}\}$. When $\epsilon\sqrt{C(\alpha)\log\left(\frac{2}{\epsilon}\right)} \leq \frac{\lambda}{2C_1}$, where C_1 is a universal constant, the edge sets of $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE},\lambda}}$ and $\mathbb{P}_{\theta^*}$ coincide i.e.,*

$$E(\widehat{\theta}_{\mathrm{MDE},\lambda}) = E(\theta^*).$$

Observe that this result is interesting and surprising, because one would generally not expect to be able to recover the true edge $E(\theta^*)$ under contamination. Additionally, as mentioned earlier, there is no dependence on p, d or k, which means that the irrespective of the size of graph, if the minimum edge weight is sufficiently large or the level of contamination is sufficiently small, we would be able to recover the true edge set in the infinite sample limit.

4.3.2 Empirical Robust Estimators for $\mathcal{G}_{p,k}$

The minimum distance estimator is not suitable for non-asymptotic settings since we do not have access to the population contaminated distribution, but only to its discrete empirical counterpart, obtained via samples from the contaminated distribution. It would thus be ideal if there were an approximation to the TV distance that is amenable to projections of discrete distributions, and that preserves the optimality properties of the full TV projections.

Remarkably, Yatracos [100] proposed just such an approximation to TV projections. Consider a class of distributions $\mathcal{P}$. It is known that $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$, where the supremum is over all possible measurable sets $A \subseteq \mathrm{supp}(P)$. While uniform convergence fails over all sets, Yatracos [100] showed that we can consider a much smaller collection of clevely chosen sets. In particular, Yatracos [100] suggested approximating the TV distance between distribution $P, Q \in \mathcal{P}$ as

$$d_{\mathrm{TV}}(P,Q) \approx \sup_{A\in\mathcal{A}} |P(A) - Q(A)|,$$

where $\mathcal{A}$ are sets of the form

$$\mathcal{A} = \{A(\mathbb{P}_1, \mathbb{P}_2) : \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}\}, \tag{4.6}$$

and $A(\mathbb{P}_1, \mathbb{P}_2) = \{x : \mathbb{P}_1(x) > \mathbb{P}_2(x)\}$. This approximation allows us to construct statistically optimal estimators for $\mathcal{G}_{p,k}$.

Non-Asymptotic Robust Estimators for $\mathcal{G}_{p,k}$

Given samples $\{x^{(i)}\}_{i=1}^{n}$ from the mixture model $\mathbb{P}_\epsilon$ defined in (4.1), define $\widehat{\mathbb{P}}_{n,\epsilon}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{x^{(i)} \in A\right\}$ for all $A \in \mathcal{A}$, where $\mathcal{A}$ is the same as defined in (4.6) with the class of distributions $\mathcal{G}_{p,k}$. Our estimator is defined as

$$\mathbb{P}_{\widehat{\theta}} = \operatorname*{argmin}_{\mathbb{P}_\theta \in \mathcal{G}_{p,k}} \sup_{A \in \mathcal{A}} \left| \mathbb{P}_\theta(A) - \widehat{\mathbb{P}}_{n,\epsilon}(A) \right|. \tag{4.7}$$

The following lemma characterizes the performance of our estimator.

Lemma 13. *Given n samples from a contaminated distribution P_ϵ, the Yatracos estimate (4.7) satisfies with probability least $1-\delta$:*

$$d_{\mathrm{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^\star}) \leq 2\epsilon + \mathcal{O}\left(\sqrt{\frac{k\log(p^2e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

The lemma above shows that the Yatracos estimate is close to the true Ising model in TV distance with high-probability. Combining Lemma 13 and Theorem 20, we get parameter error guarantees for the Yatracos estimate.

Corollary 22. *Given n samples from $\mathbb{P}_\epsilon$, the Yatracos' estimator returns a $\widehat{\theta}$ such that with probability at least $1-\delta$,*

$$\|\widehat{\theta}(i) - \theta^\star(i)\|_2 \lesssim 2\epsilon\sqrt{\log(1/\epsilon)} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{k\log(p^2e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \quad \text{for all } i \in [p], \tag{4.8}$$

where $\tilde{\mathcal{O}}(.)$ hides logarithmic factors involving its argument.

Remarks. Note that the proposed estimator achieves the same (asymptotic) dimension-independent error as the Minimum Distance Estimate discussed in

Section 4.3.1, while simultaneously achieving an $\widetilde{\mathcal{O}}\left(\sqrt{\frac{k\log p}{n}}\right)$ error rate. Moreover, observe that in the uncontaminated setting, i.e., when $\epsilon = 0$, this is the *first* estimator to get an $\widetilde{O}\left(\sqrt{\frac{k\log p}{n}}\right)$ error rate. As a consequence, Yatracos' estimator followed by an additional thresholding step gives the first estimator to recover the true edge set $E(\theta^*)$ with only $\widetilde{\mathcal{O}}\left(\frac{k\log(p)}{\lambda^2}\right)$ samples. In contrast, the estimator proposed by [90] posit that the sample size should satisfy $\mathcal{O}(1/\lambda^4)$ when the parameters are unknown. In the contaminated case, note that we show a better dependence on ϵ – $\mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$ vs. $\sqrt{\epsilon}$ in [89]. The proof for Lemma 13 is presented in Section D.4.2 of the appendix. A similar analysis was conducted in [101], however [101] study density estimation, and not parameter estimation. The bound on the modulus of continuity obtained in Theorem 20 allows us to relate the TV distance between the estimated distribution and the true distribution to the parameter error, thus giving us bounds for parameter estimation.

Non-Asymptotic Robust Estimators for $\mathcal{G}_{p,d}$

Under the same setting as considered for $\mathcal{G}_{p,k}$, we see that directly employing the estimator (4.7) would lead to a sub-optimal rate. Our guarantee for (4.7) for $\mathcal{G}_{p,k}$ relies on the fact that parameters for Ising models in $\mathcal{G}_{p,k}$ contain at most k non-zero elements, hence the subset $A(\theta^{(1)}, \theta^{(2)}) = \{x : \mathbb{P}_{\theta^{(1)}}(x) > \mathbb{P}_{\theta^{(2)}}(x)\}$ is a half-space defined by a vector with at most $2k+1$ non-zero elements. However, these subsets defined with parameters $\theta^{(1)}, \theta^{(2)}$ of two Ising models in $\mathcal{G}_{p,d}$ is a half-space defined by a vector that have at most $pd+1$ non-zero elements. This leads to a rate term that is proportional to $\sqrt{pd\log(p)/n}$, which does not scale well in high-dimensional settings.

4.4 Robust Conditional Likelihood Estimators

In the previous section, we have seen that the estimator based on Yatracos classes [100] provides an approximate TV projection for $\mathcal{G}_{p,k}$ but not for $\mathcal{G}_{p,d}$. The main caveat with this estimator is that it is not tractable and takes infinite time. To circumvent this issue, we consider a more direct approach to robust estimation: we "robustify" the gradient samples obtained from samples $\{x^{(i)}\}_{i=1}^{n}$ of the contaminated distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^*} + \epsilon Q$.

Neighborhood-based logistic regression. In a classical paper, Besag [102] made the key structural observation that under model (4.2), the conditional distribution of node X_i given the other variables $X_{-i} = x_{-i}$ is given by

$$\mathbb{P}_{\theta^*}(X_i = x_i | X_{-i} = x_{-i}) = \frac{\exp(2x_i \sum_{t\in\mathcal{N}(i)} \theta^*_{it} x_t)}{\exp(2x_i \sum_{t\in\mathcal{N}(i)} \theta^*_{it} x_t) + 1} = \sigma(x_i \langle 2\theta^*(i), x_{-i}\rangle).$$

Thus the variable X_i can be viewed as the response variable in a logistic regression model with X_{-i} as the covariates and $2\theta^*(i)$ as the regression vector. In particular, this implies that $\mathbb{E}_{x\sim\mathbb{P}_{\theta^*}}[\nabla l_i(2\theta^*(i); x)] = \mathbf{0}$ where $\ell_i(\theta(i); x) = \log\sigma(x_i \langle \theta(i), x_{-i}\rangle)$ is the conditional log-likelihood of x under $\mathbb{P}_\theta$. Note that for graphs with maximum degree at most d, the parameter vector $\theta^*(i)$ has at most d non-zero entries, and for graphs with at most k edges, the parameter vector $\theta^*(i)$ has at most k non-zero entries. Ravikumar et al. [76] solved an ℓ_1-regularized logistic regression to recover the node parameters for graphs with bounded maximum degree. However, in our setting, the data is contaminated with outliers, and hence the minimizer of the likelihood can be arbitrarily bad. While there has been recent work giving provably optimal algorithms for robust logistic regression [78], all of these results are in the low-dimensional setting. We propose the *first* statistically optimal estimator for sparse logistic regression, and use that to provide estimators for learning Ising models.

Robust Sparse Logistic Regression. Our approach is based on a reduction to robust univariate estimation initially proposed by [103]. In particular, note that when we have clean data, then, in the population setting, $\theta^*(i)$ is the unique solution to the equation $\|\mathbb{E}_{x\sim\mathbb{P}_{\theta^*}}[\nabla\ell_i(\theta(i); x)]\|_2 = \mathbf{0}$ or equivalently, it is the unique minimizer for the following optimization problem:

$$\theta^*(i) = \underset{w:\|w\|_0\le s}{\operatorname{argmin}} \sup_{u\in\mathcal{S}^{p-2}} \left|\mathbb{E}_{x\sim\mathbb{P}_{\theta^*}}[u^T \nabla\ell_i(w; x)]\right|,$$

where we have simply used the variational form of the norm of a vector. Observe that $\mathbb{E}_{x\sim\mathbb{P}_{\theta^*}}[u^T\nabla\ell_i(w; x)]$ is simply the population (uncontaminated) mean of the gradients, when projected along the direction u. Unfortunately, we only

Algorithm 6 Robust1DMean - Robust univariate mean estimator

```
function Interval1D($\{z_i\}_{i=1}^{2n}$, Corruption Level $\epsilon$, Confidence Level $\delta$)
    Split the data into two subsets: $\mathcal{Z}_1 = \{z_i\}_{i=1}^{n}$ and $\mathcal{Z}_2 = \{z_i\}_{i=n+1}^{2n}$.
    Let $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$.
    Using $\mathcal{Z}_1$, let $\hat{I} = [a, b]$ be the shortest interval containing $n(1 - 2\alpha - \sqrt{2\alpha\frac{\log(4/\delta)}{n}} - \frac{\log(4/\delta)}{n})$ points.
    Use $\mathcal{Z}_2$ to identify points lying in $[a, b]$.
    return $\frac{1}{\sum_{i=n}^{2n} \mathbb{I}\{z_i \in \hat{I}\}} \sum_{i=n}^{2n} z_i \mathbb{I}\left\{z_i \in \hat{I}\right\}$
end function
```

have finite samples which are moreover contaminated. We can pass these univariate projections of the gradient through a *robust* univariate mean estimator, and return a point which has the *smallest* (robust) mean along any direction. This leads to the following program,

$$\widehat{\theta}(i) = \operatorname*{argmin}_{w \in \mathcal{N}_s^{\gamma}(\mathcal{S}^{p-2})} \sup_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{Robust1DMean}(\{u^T \nabla \ell_i(w; x^{(j)})\}_{j=1}^{n}) \right|, \qquad (4.9)$$

where $\mathcal{N}_s^{\gamma}(\mathcal{S}^{p-2})$ is a γ-cover of the unit sphere over $p-1$ dimensions with s non-zero entries i.e., for every $x \in \mathcal{S}^{p-2}$ that has s non-zero entries, there exists $y \in \mathcal{N}_s^{\gamma}(\mathcal{S}^{p-2})$ such that $\|x - y\|_2 \leq \gamma$. Our robust univariate mean estimator is based on the shortest interval estimator (Shorth) studied in [28, 103, 104]. The estimator, presented in Algorithm 6, proceeds by using half of the samples to identify the shortest interval containing roughly $(1-\epsilon)n$ fraction of the points, and then the remaining half of the points is used to return an estimate of the mean. Intuitively, this estimator effectively trims distant outliers, thereby limiting their influence on the estimate.

We assume that the contamination level ϵ, confidence parameter δ, and sparsity s are such that,

$$2\epsilon + \sqrt{\epsilon\left(\frac{s\log(p)}{n} + \frac{\log({}^{p}\!/\!{}_{\delta})}{n}\right)} + \frac{s\log(p)}{n} + \frac{\log({}^{4p}\!/\!{}_{\delta})}{n} < c, \qquad (4.10)$$

for some small constant $c > 0$. As noted earlier, the sparsity parameter s is the maximum degree d for $\mathcal{G}_{p,d}$ and the maximum number of edges k for $\mathcal{G}_{p,k}$.

Theorem 23 (Guarantees for $\mathcal{G}_{p,d}$)**.** *Under the setting considered in 4.4 along with Assumption (4.3), the estimator in (4.9) returns estimates* $\{\widehat{\theta}(i)\}_{i=1}^{p}$ *with*

$\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{np}\right\}$ *returns with probability at least* $1 - \delta$

$$\|\widehat{\theta}(i)-\theta^\star(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)}+\sqrt{C(\alpha)\frac{d}{n}\log\left(\frac{3ep^2}{d\gamma}\right)}+\max\left(\epsilon, \frac{\log(1/\delta)}{n}\right) \quad \textit{for all } i \in [$$

Corollary 24 (Guarantees for $\mathcal{G}_{p,k}$)**.** *Under the setup considered in Theorem 23, the estimator in (4.9) returns estimates* $\{\widehat{\theta}(i)\}_{i=1}^p$ *with* $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{np}\right\}$ *returns with probability at least* $1 - \delta$

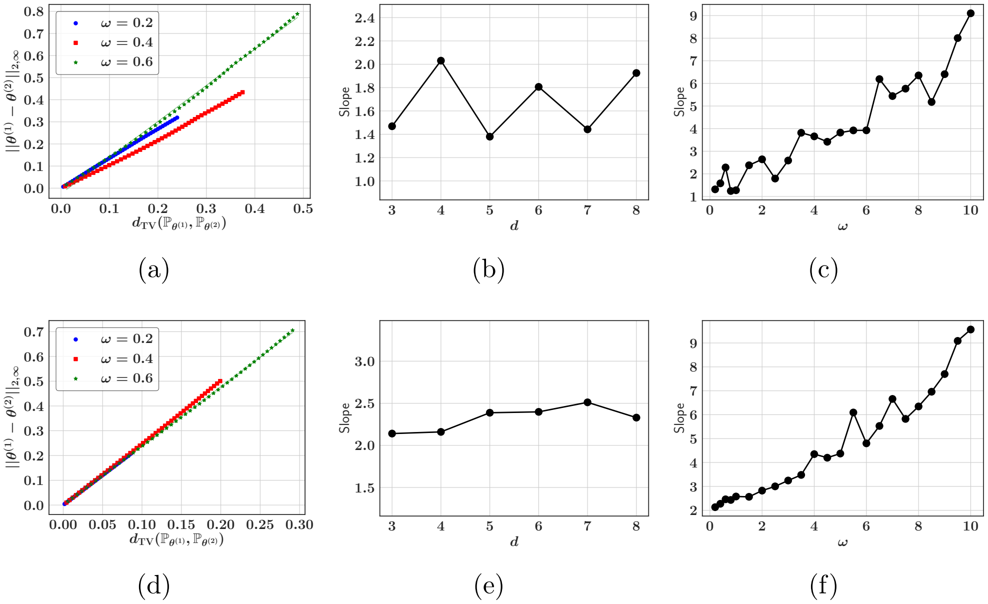
$$\|\widehat{\theta}(i)-\theta^\star(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)}+\sqrt{C(\alpha)\frac{k}{n}\log\left(\frac{3ep^2}{k\gamma}\right)}+\max\left(\epsilon, \frac{\log(1/\delta)}{n}\right) \quad \textit{for all } i \in [$$

Remarks. Observe that our estimator achieves the same (asymptotic) bias as the Minimum Distance Estimator, previously discussed in Section 4.3.1. Define the recovered edge set as those edges (i, j) satisfying $|\widehat{\theta}_{ij}| \geq \lambda/2$. When $\epsilon = 0$, i.e., no contamination, for $\mathcal{G}_{p,d}$, we require the number of samples $n \geq \mathcal{O}\left(\frac{d\log(p)}{\lambda^2}\right)$ to recover the true edge set $E(\theta^*)$. Even in the uncontaminated setting, there is *no* known estimator which achieves the same optimal sample complexity as ours. In particular, Santhanam and Wainwright [90] achieve similar rates when they assume that the structure is known, while other approaches of [76, 86] have worse dependence on the degree d. Hence, our proposed estimator has an optimal (asymptotic) bias and optimal high probability bounds. For $\mathcal{G}_{p,k}$, we obtain the same rate and sample complexity as Yatracos' estimator (4.7), which we remarked is optimal. The proof of Theorem 23 is presented in Section D.5.1 of the appendix.

4.5 Synthetic Experiments

Our theoretical results crucially hinge on bounds on the TV modulus of continuity derived in Theorem 20, and we devote this section to corroborating these bounds.

Setup. We consider two different ensembles. A graph $G \in \mathcal{G}_{p,d}^{\text{star}}$ when one of the p nodes is connected d other vertices, and no other edges are present in the



(a) (b) (c)

(d) (e) (f)

Figure 4.1: Left: Variation of $\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}$ with $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})$ for $G^{(1)}, G^{(2)} \in \mathcal{G}^{\text{clique}}_{15,4}$ (top) and $G^{(1)}, G^{(2)} \in \mathcal{G}^{\text{star}}_{15,4}$ (bottom) graphs with varying ω. Middle: Variation of slope with d for cliques (top) and star (bottom) with $p = 12$ and $\omega = 0.4$. Right: Variation of slope with ω for cliques (top) and star (bottom) with $p = 15$ and $d = 5$. The slope is defined as $\frac{\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}}{d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})}$.

graph, resembling a star. A graph $G \in \mathcal{G}_{p,d}^{\text{clique}}$ contains $\lfloor \frac{p}{d+1} \rfloor$ cliques of size $d+1$, and the remainder of the nodes $p \mod (d+1)$ fully connected amongst themselves. We generate our plots in the following manner: first we construct two graphs with the same structure - either from $\mathcal{G}_{p,d}^{\text{clique}}$ of $\mathcal{G}_{p,d}^{\text{star}}$. We instantiate parameters for the first graph with $\theta^{(1)}$ with model width ω and then vary the parameters for the second graph as $\theta^{(2)} = \theta^{(1)} \cdot \frac{i}{25}$ for i ranging from 1 to 50. We vary $p \in \{12, 15\}$, $d \in \{3 : 8 : 1\}$ and $\omega \in \{0.2 : 1.0 : 0.2\} \cup \{1.5 : 10 : 0.5\}$ where $\{a : b : c\}$ denotes values between a and b (both inclusive) with consecutive values differing by c.

Results. Figures 4.1(a) and 4.1(d) exhibits a linear relationship between $d_{\text{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})$ and $\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}$, as suggested by our theoretical results from previous sections. Furthermore, we notice that the slope is not drastically affected by ω, which also suggests that the constant $C(\alpha)$ appearing in our results is $O(1)$. We also note from Figures 4.1(b) and 4.1(e), that the slope is unaffected by a change in degree. Finally, in Figures 4.1(c) and 4.1(f), we notice the variation in the slope with increasing model width ω. While our current result study the case when $\omega < 1$, it is also interesting to note an increasing trend when $\omega \geq 1$ suggesting an explicit dependence on ω in the low-temperature regime.

4.6 Discussion and Future Work

In this work we provided the first statistically optimal robust estimators for learning Ising models in the high temperature regime. Our estimators achieved optimal asymptotic error in the ϵ-contamination model, and also high-probability deviation bounds in the uncontaminated setting. There are several avenues for future work, some of which we discuss below.

Beyond Dobrushin's conditions. In the low-temperature setting, Lindgren et al. [89] showed the existence of an estimator which gets an $O(\sqrt{\epsilon})$ error. In Appendix D.1, we tighten this for edge-bounded graphs by providing estimators which achieve $O(\min(\sqrt{\epsilon}, \epsilon\sqrt{k}))$ error, where k is the maximum number of edges in the graph. However, giving matching lower bounds in this setting is an open problem. Our synthetic experiments surprisingly show that one may expect similar rates in the two temperature regimes.

Computationally Efficient Estimators. While in this work, we designed statistically optimal estimators that achieve an $O(\epsilon\sqrt{\log(1/\epsilon)})$ parameter error, whereas, existing computationally efficient approaches [84, 89] achieve a suboptimal error of $O(\sqrt{\epsilon})$. Developing computationally efficient algorithms which close this gap is an interesting open problem.

Other Contamination Models. In this work, our focus was on designing estimators for the ϵ-contaminated model, i.e., where a fraction of the data is arbitrarily corrupted. Another model of corruption - motivated by sensor networks and distributed computation where node failures are common - is when only a few features(nodes) get corrupted, and we still want to learn the appropriate graph structure for the uncontaminated nodes. Recent work by Goel et al. [105] discusses results for this model of contamination.

However, if used without prior analysis of the data presented, this could potentially reduce the effect of outlier samples, which in the case of voting patterns, are representative of a minority groups.

Acknowledgements

AP, VS and PR acknowledge the support of NSF via IIS-1955532, OAC-1934584, DARPA via HR00112020006, and ONR via N000141812861. SB and AP acknowledge the support of NSF via DMS-17130003 and CCF-1763734. We would also like to thank an anonymous reviewer for pointing out related work by Devroye et al. [101].

Chapter 5

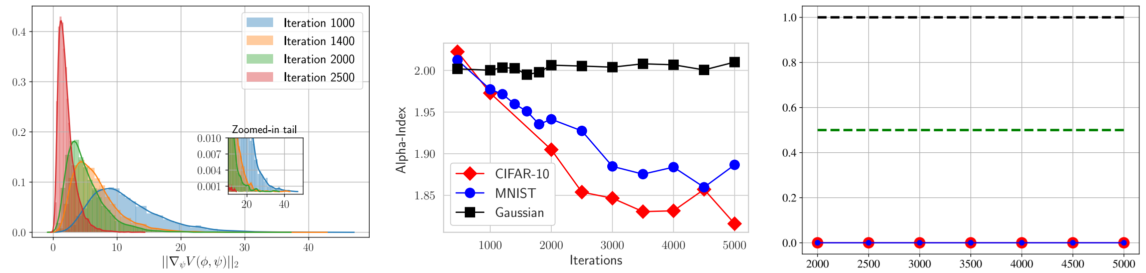
Efficient Estimators for Heavy-Tailed Machine Learning

With a dramatic improvement in data collection technologies, our era has witnessed a massive explosion in unstructured and heterogeneous data sets. This has led to a prevalence of heavy tailed distributions across a broad range of tasks in machine learning. In this work, we aim to develop estimators which can handle such ill-behaved distributions. Our workhorse is a novel and computationally-efficient estimator for mean estimation, which is both practical and provably near-statistically-optimal. We provide specific consequences of our theory for both supervised learning tasks such as linear regression, generalized linear models and generative modeling tasks such as Generative Adversarial Networks. We study the empirical performance of our proposed estimators on synthetic and real-world data sets, and find that our methods convincingly outperform a variety of practical baselines.

5.1 Introduction

Existing estimators in machine learning are largely designed for thin-tailed data, such as those coming from a Gaussian distribution. In particular, it is well known that in the absence of light tails, classical estimators based on minimizing the empirical error perform poorly [15, 106].

Modern datasets however are frequently heavy-tailed, see for instance [12, 13, 107] and references therein for examples from domains ranging from large scale biological datasets, and financial datasets, among others.



(a) Distribution of norms of generator gradients trained on CIFAR10

(b) Variation of α-Index for generator gradients trained on MNIST and CIFAR10

(c) Results of Gaussianity Tests of generator gradients trained on MNIST

Figure 5.1: Non-Gaussianity of Generator Gradients at Different Iterations.

There has also been a line of recent work showing that heavy-tailed distributions occur even in intermediate outputs of machine learning algorithms. In particular, Simsekli et al. [108] and Zhang et al. [109] recently showed that the distribution of noise in stochastic gradients is heavy-tailed for popular deep learning architectures such as attention models. We further validate this in our experiments via an empirical investigation into the distribution of gradients generative adversarial networks.

5.1.1 Empirical Study of Gradients in Generative Adversarial Networks.

Generative Adversarial Networks (GANs) [110] have risen to prominence in machine learning as an unsupervised method for learning and efficient sampling from complex distributions. At the population level, the GAN objective is based on the following minimax problem:

$$\min_{G\sim\mathcal{F}_G}\max_{D\sim\mathcal{F}_D}\mathbb{E}_{Z\sim p_Z}[f(1-D(G(Z)))]+\mathbb{E}_{X\sim\nu}[f(D(X))]$$

Given a target distribution ν, the goal is to learn a map G from the *generator class* $\mathcal{F}_G$ that transforms samples from P_Z (known as a prior distribution) and minimizes the loss of the best test function D inside the *discriminator class* $\mathcal{F}_D$. f is any monotone function. The loss can be minimized by obtaining samples that are "similar" to those sampled from ν. In practice, deep neural networks are used to represent both discriminator and generator classes. In particular, suppose (ϕ, ψ) represent the parameters of the discriminator and generator respectively, then the classical instantiation of the GAN framework is given by,

$$V(\phi, \psi) \quad = \quad \mathbb{E}_{x \sim p_{\text{data}}}[\log(D_\phi(x))] \; + \; \mathbb{E}_{Z \sim p_Z}[\log(1 \; - \; D_\phi(G_\psi(Z))] \tag{5.1}$$

where p_{data} represents the given samples. Then, we finally estimate our parameters ψ_t and ϕ_t via an alternating procedure as,

$$\phi_t = \underset{\phi}{\operatorname{argmax}} V(\phi_{t-1}, \psi_{t-1}) \quad \psi_t = \underset{\psi}{\operatorname{argmin}} V(\phi_t, \psi_{t-1})$$

The above optimization problem is typically solved using first-order methods such as stochastic gradient descent or adaptive methods such as ADAM [111].

In this section, our goal is to study the distribution of the gradients of the generator $\nabla_\psi V(\phi, \psi)$ at different iterations. For these experiments, we train a deep convolutional GAN (DCGAN) [112] on the MNIST and CIFAR10 datasets using SGD with a fixed learning rate of $3 \cdot 10^{-4}$. We update our sequence of estimates as described above and elaborate on the setup in Section 5.4.3.

Our investigation begins by plotting the distribution of ℓ_2 norms of the generator gradients over several iterations in Figure 5.1(a) for CIFAR10. Visually, the distribution of the norms of the gradients clearly exhibit heavy-tailedness with the degree of heavy-tailedness increasing as the iterations increase.

α-Index of Gradient Norms. To further quantify this effect, we use the α-index estimator for α-stable distributions (which are a broad category of heavy-tailed distributions) proposed by Mohammadi et al. [113] and used by Simsekli et al. [108] as a heuristic measure of heavy-tailedness of the generated gradients. Concretely, the α-index estimator for *i.i.d.* α-stable random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N$ where $N = mn$ is given as follows. Construct $\mathbf{Y}_i = \sum_{j=1}^m \mathbf{X}_{j+(i-1)m}$ for $i = 1, \ldots, n$ from $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Then, the α-index estimator is given by

$$\frac{1}{\widehat{\alpha}(m,n)} = \frac{1}{n \log m} \sum_{i=1}^{n} \log |\mathbf{Y}_i| - \frac{1}{N \log m} \sum_{i=1}^{N} \log |\mathbf{X}_i|$$

Intuitively, this α-estimator is given by splitting the samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ into m blocks of size n each, summing each block, and computing the discrepancy between the average log norms of the blocks and the average log norms of the samples. In order to satisfy the preconditions of this alpha-estimator, we center the data by subtracting the mean. A value of $\alpha = 2.0$ corresponds to a Gaussian distribution, while $\alpha = 1.0$ corresponds to a Cauchy distribution. As a soft signal, the lower the α-index is, the more heavy-tailed the distribution is. In other words, as the α-index decreases, the central peak of the distribution gets higher, the valley before the central peak gets deeper, and the tails get heavier.

For calibration, we measure the α-index of the norms of random vectors drawn from a multivariate normal distribution with zero mean and identity covariance. In Figure 5.1(b), we can see that the our estimator returns an estimate of 2.0 for random gaussian vectors. Moreover, Figure 5.1(b) also plots the α-index of the gradient norms as iterations proceed. The plots clearly show a decreasing trend of the α-index with increasing iterations, thereby confirming our hypothesis regarding the heavy-tailedness of gradients in GANs.

Non-Gaussianity via Random Projections. Another step in our investigation is to confirm our hypothesis of heavy-tailed gradients (and non-Gaussianity) via hypothesis tests. We borrow the empirical framework of Panigrahi et al. [114] and (1) draw 10000 stochastic gradients every 500 iterations, (2) project the data onto 1000 random directions, and (3) conduct Anderson-Darling [1954] and Shapiro-Wilk [1965] normality tests for these univariate projections of the gradients to find the proportion of accepted projection directions and average confidence that the gradients are Gaussian respectively. Figure 5.1(c) plots the results of the Gaussianity tests run on random projections of the stochastic gradients every 1000 iterations. The red circles depict the proportion of accepted random directions out of 1000 as given by the Anderson-Darling test, and the blue dots indicate the average confidence of the Shapiro-Wilk test. The black (top) and green (middle) lines at indicate the proportion of accepted directions and average confidence in the ideal scenario of Gaussianity. The tests for the gradients fail extremely, thereby validating with very high probability that the gradients are indeed non-Gaussian.

5.1.2 Theoretical Background and Setup

In the previous section, we showed that the stochastic gradients in certain practical scenarios are not necessarily Gaussian and could be also heavy-tailed. Informally, for such gradient distributions which do not enjoy Gaussian-like concentration, the empirical expectation based estimates of the gradient do not necessarily point in the right descent direction leading to bad solutions, prolonged training time, or a mixture of both.

More formally, this question can be studied by looking at concentration of empirical averages to their population quantities, when the data has only finite-order moments *i.e.* it is heavy-tailed. There has been a long line of work in theoretical statistics focusing on designing efficient estimators which work which for distributions with finite-order moments, which we describe next. Concretely, focusing on the fundamental problem of robust mean estimation, in the heavy tailed model we observe n samples $x_1, \dots, x_n$ drawn independently from a distribution P, which is only assumed to have low-order moments be finite (for instance, P might only have finite variance). The goal of past work [15, 16, 17, 18] has been to design an estimator $\widehat{\theta}_n$ of the true mean μ of P which has a small ℓ_2-error with high-probability. Formally, for a given $\delta > 0$, we would like an estimator with minimal r_δ such that,

$$P(\|\widehat{\theta}_n - \mu\|_2 \leq r_\delta) \geq 1 - \delta. \tag{5.2}$$

As a benchmark for estimators in the heavy-tailed model, we observe that when P is the multivariate normal (or sub-Gaussian) distribution with mean μ and covariance Σ, it can be shown (see Hanson and Wright [19]) that the sample mean $\widehat{\mu}_n = (1/n) \sum_i x_i$ satisfies, with probability at least $1 - \delta$[1],

$$\|\widehat{\mu}_n - \mu\|_2 \lesssim \sqrt{\frac{\text{trace}\,(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}. \tag{5.3}$$

The bound is referred to as a *sub-Gaussian*-style error bound. However, for heavy tailed distributions, as for instance showed in Catoni [15], the sample mean only satisfies the sub-optimal bound $r_\delta = \Omega(\sqrt{\text{trace}\,(\Sigma)\,/n\delta})$. Only recently did work by Lugosi and Mendelson [17] show that the sub-Gaussian error bound is achievable while *only assuming that P has finite variance*, but by a carefully designed impractical estimator. In the univariate setting, the classical median-of-means estimator [20, 21, 22] and Catoni's M-estimator [15] achieve this surprising result but designing such estimators in the multivariate setting

[1] Here and throughout our paper we use the notation $\lesssim$ to denote an inequality with universal constants dropped for conciseness.

has proved challenging. Minsker [16] proved results for the geometric median-of-means (GMOM), which, (1) partitions the data into $k = \lceil 3.5 \log(1/\delta) \rceil$ blocks, (2) computes sample mean within each block $\{\widehat{\mu}_i\}_{i=1}^k$ and (3) and returns the geometric median $\widehat{\theta}_{\mathrm{MOM},\delta} = \mathrm{argmin}_\theta \sum_i \|\theta - \widehat{\mu}_i\|_2$. In particular, the paper [16] showed that $\widehat{\theta}_{\mathrm{MOM},\delta}$ is such that, with probability at least $1 - \delta$,

$$\|\widehat{\theta}_{\mathrm{MOM},\delta} - \mu\|_2 \lesssim \sqrt{\frac{\mathrm{trace}\,(\Sigma) \log(1/\delta)}{n}}. \tag{5.4}$$

Note that the GMOM estimator does not match the true sub-Gaussian bound (5.3). Estimators that achieve truly sub-Gaussian bound, but which are computationally intractable, were proposed recently by Lugosi and Mendelson [17] and subsequently Catoni and Giulini [18]. Hopkins [23] and later Cherapanamjeri et al. [24] developed a sum-of-squares based relaxation of Lugosi and Mendelson's estimator, thereby giving a polynomial time algorithm which achieves optimal rates. However, while polynomial-time, these estimators are still far from being implementable and/or practical.

Contributions. In this paper, we propose and study *practical* estimators that in some cases improve on GMOM and in some cases achieve a sub-Gaussian error bound. We use our practical mean estimators to design provably near-optimal algorithms for heavy-tailed linear regression and generalized linear models. We also conduct extensive synthetic experiments which backup our theoretical improvements, and as one consequence of our results, show improvement in training GANs using our algorithms.

Notation and some definitions. Let x be a random vector with mean μ and covariance Σ. We say that the x has bounded $2k$-moments if for all $v \in \mathcal{S}^{p-1}$, $\mathbb{E}[(v^T(x-\mu))^{2k}] \le C_{2k} \left(\mathbb{E}[(v^T(x-\mu))^2]\right)^k$. We let,

$$\mathrm{OPT}_{n,\Sigma,\delta} \overset{\mathrm{def}}{=} \sqrt{\frac{\mathrm{trace}\,(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}} \tag{5.5}$$

denote the sub-Gaussian deviation bound in (5.3), satisfied by the sample mean of a sub-Gaussian distribution, at a confidence level δ. Let $r(\Sigma) \overset{\mathrm{def}}{=} \frac{\mathrm{trace}(\Sigma)}{\|\Sigma\|_2}$ be

Algorithm 7 Heavy Tailed Mean Estimator

```
function FilterPD(S = $\{z_i\}_{i=1}^n$, Num Steps: $T^*$)
    $t = 1$
    while $t \leq T^*$ do
        Let $\widehat{\theta}_S = \frac{1}{|S|}\sum_{i=1}^{|S|} z_i$ be the sample mean.
        Let $\Sigma_S = \frac{1}{|S|}\sum_{i=1}^{|S|}(z_i - \widehat{\theta}_S)(z_i - \widehat{\theta}_S)^T$ be the sample covariance matrix.
        Let $(\lambda, v)$ be the largest eigenvalue,eigenvector of $\Sigma_S$.
        For each $z_i$, let $\tau_i \overset{\text{def}}{=} \left(v^T(z_i - \widehat{\theta}_S)\right)^2$ to be its score
        Randomly sample a point $z$ from $S$ according to
            $$\Pr(z_i \text{ chosen}) = \frac{\tau_i}{\sum_j \tau_j}$$
        Remove sample and update $S = S \backslash \{z\}$.
        $t = t + 1$
    end while
end function
```

the *effective rank* of Σ. Note that $1 \leq r(\Sigma) \leq r$, where r is the rank of Σ. Throughout the paper, we use $c, c_1, c_2, \ldots, C, C_1, C_2, \ldots$ to denote positive universal constants.

5.2 Efficient and Near-Optimal Mean Estimation.

In this section, we present our algorithm for near-optimal heavy-tailed mean estimation. Our algorithm formally presented in Algorithm 7 is primarily based on the SVD-based filtering algorithm, which has appeared in different forms [117, 118] and was recently reused by Diakonikolas et al. [27, 31] for adversarial mean estimation. However, the previous versions and their analysis, while suited to bounds on the expected deviation, do not give tight high-probability non-asymptotic rates. It proceeds in an iterative fashion, by (1) computing the principal eigenvector of the empirical covariance matrix, (2) projecting points along the the principal eigenvector, and (3) randomly sampling points based on their projection scores. This procedure is repeated for a fixed number of steps.

Diakonikolas et al. [31] follow a similar procedure, but remove a subset of points at a step, depending on if their projection score is above or below a randomly chosen threshold. While only a modest difference from ours, deriving high-probability results for their algorithm is not clear, and in particular, the bounds provided by Diakonikolas et al. [31] are in expectation. In contrast our variant of this iterative sample-and-remove procedure allows us to borrow tools from martingale analysis [119, 120], and we are able to get tight non-asymptotic high-probability bounds for mean estimation.

We present our first result for heavy-tailed mean estimation for the distributions with bounded 4-moments. Given $\delta \in (0, 0.5)$, suppose the number of samples n satisfies:

$$n \geq Cr^2(\Sigma)\frac{\log^2(p/\delta)}{\log(1/\delta)}, \tag{5.6}$$

for some small constant $C > 0$. Then, we have the following result.

Theorem 25. *Suppose P has bounded 4^{th} moment. Then, Algorithm 7 when instantiated for $T^* = \lceil C\log(1/\delta)\rceil$ steps returns an estimate $\widehat{\theta}_\delta$ such that, with probability at least $1-4\delta$,*

$$\|\widehat{\theta}_\delta - \mu\|_2 \lesssim OPT_{n,\Sigma,\delta}$$

Remark: If the number of samples n satisfies the condition in Eqn. (5.6), then Algorithm 7 achieves the *the optimal sub-Gaussian deviation bound.* The above presented result shows that it is possible to *prune* samples to get high-probability bounds for the heavy-tailed problem. In comparison to the SDP based algorithms of Hopkins [23], Cherapanamjeri et al. [24], our algorithm is easy to implement and practical. In particular, our estimator can also be computed in linear-time, requiring an overall runtime of $O(np\log(1/\delta))$ compared to $O(n^4 + np)$ runtime of Cherapanamjeri et al. [24].

Next, we present our result for heavy-tailed mean estimation for distributions with bounded 2nd moment.

Corollary 26. *Suppose P has bounded 2nd moment. Then, Algorithm 7 when instantiated for $T^* = C(\log(1/\delta))$ steps returns an estimate $\widehat{\theta}_\delta$ such that, with probability at least $1-4\delta$,*

$$\|\widehat{\theta}_\delta - \mu\|_2 \lesssim \sqrt{\frac{trace\,(\Sigma)\log(p/\delta)}{n}}$$

Remark: In the univariate setting, Corollary 26 shows that Algorithm 7 achieves the optimal sub-Gaussian deviation bound. As discussed in the introduction, even for the univariate setting, Catoni's M-estimation [15] and Median-of-Means [20, 21, 22] are the only known estimators to achieve these rates for any 2^{nd} moment bounded distribution. Algorithm 7 is the *first sample-pruning* based estimator, which achieves these optimal bounds, without any further assumptions.

In the multivariate setting, while our theoretical upper bounds are weaker than the guarantees of GMOM, we conduct extensive simulations in Section 5.4 which suggest otherwise.

5.2.1 Proof Sketch.

In this section, we present a brief sketch of the proof of Theorem 25 and highlight the key technical contributions. The detailed proofs for each of the presented Lemmas can be found in Appendix C.1.

- Our first contribution is showing that given an arbitrary collection of points S, and information about the size of an unknown subset $G \subset S$, then Algorithm 7 approximates the mean of the points in G efficiently with *high probability.*

Lemma 14. *Let S be any arbitrary collection of points, and let $G^0 \subset S$ be an unknown subset of size n_G^0 such that $8\frac{n-n_{G^0}}{n} + 36\frac{\log(1/\delta)}{n} < \frac{1}{4}$. Then, when Algorithm 7 is run for $T^* = \lceil 3(n - n_{G^0}) + 18\log(1/\delta) \rceil$ steps on S, it returns an estimate $\widehat{\theta}_\delta$ such that with probability at least $1 - \delta$,*

$$\left\| \widehat{\theta}_\delta - \frac{1}{n_{G^0}} \sum_{x_i \in G^0} x_i \right\|_2 \lesssim \|\Sigma_{G^0}\|_2^{\frac{1}{2}} \left(\frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n} \right)^{\frac{1}{2}},$$

where Σ_{G^0} is the covariance of the unknown subset of points.

- Our second contribution is showing that when given n samples from a distribution with bounded moments, there *exists* a good subset of points. This subset satisfies: such that (1) The size of the subset is big, (2) the mean of the points within the subset concentrates strongly around the true mean of P, and (3) the covariance of the points is well-behaved. In particular, given n-samples from a distribution P, we define a *good point selector* $\mathcal{O} : \mathbb{R}^p \mapsto \{0, 1\}$ by

$$\mathcal{O}(x) = \mathbb{I}\left\{ \|x - \mu(P)\|_2 \leq R \right\}, \tag{5.7}$$

and let $G = \{x_i | \mathcal{O}(x_i) = 1\}$ to be the set of points chosen by $\mathcal{O}$. Note that this (unknown) subset of points chosen by the ℓ_2-radius based point selector, is precisely our unknown subset from the previous subsection. Let

$$\widehat{\mu}_n = \Big(\sum_{i=1}^{n} \mathcal{O}(x_i)\Big)^{-1} \sum_{i=1}^{n} x_i \mathcal{O}(x_i),$$

be the sample mean of the points within the subsets, and let

$$\widehat{\Sigma}_n^{\mathcal{O}} = (\sum_{i=1}^{n} \mathcal{O}(x_i))^{-1} \sum_{i=1}^{n} (x_i - \widehat{\mu}_n)(x_i - \widehat{\mu}_n)^T \mathcal{O}(x_i)$$

Then, we have that

Lemma 15. *Let P be any distribution with mean μ and covariance Σ and bounded $2k$-moments for $k \in \{1, 2\}$. For any $\delta \in (0, 0.5)$ such that $\left(\frac{\sqrt{trace(\Sigma)}}{R}\right)^{2k} + \frac{\log(1/\delta)}{n} < c$ with probability at least $1 - 3\delta$,*

$$\frac{n - |G|}{n} \leq C_1 \frac{\log(1/\delta)}{n} + \frac{(\sqrt{trace\,(\Sigma)})^{2k}}{R^{2k}}$$

$$\begin{aligned} \|\widehat{\mu}_n - \mu\|_2 \lesssim & OPT_{n,\Sigma,\delta} + \frac{R\log(1/\delta)}{n} \\ & + \|\Sigma\|_2^{\frac{1}{2}} \left(\frac{\sqrt{trace\,(\Sigma)}}{R}\right)^{2k-1}. \end{aligned}$$

$$\|\widehat{\Sigma}_n^{\mathcal{O}}\|_2 \lesssim \|\Sigma\|_2 + R\|\Sigma\|_2^{\frac{1}{2}} \sqrt{\frac{\log(p/\delta)}{n}} + \frac{R^2 \log(p/\delta)}{n}.$$

- For distributions with bounded 4^{th}-moment, we choose $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\log(1/\delta)/n)^{1/4}}$ and recover Theorem 25. Similarly, for distributions with bounded 2^{nd} moment, we choose $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\log(1/\delta)/n)^{1/2}}$ and recover Corollary 26.

Algorithm 8 Robust Gradient Descent [1]

```
function RGD (Data $\{z_1, \dots, z_n\}$, Loss Function $\bar{\mathcal{L}}$, Step Size $\eta$, Number of Iterations $T$, Confidence $\delta$,)
    Split samples into $T$ subsets $\{\mathcal{Z}_t\}_{t=1}^{T}$ of size $\widetilde{n}$.
    for $t = 0$ to $T - 1$ do
        Let $S_t = \{\nabla \bar{\mathcal{L}}(\theta; z_i) | z_i \in \mathcal{Z}_t\}$.
        Set $T^* = C \log(\delta/T)$.
        Let $g^t = \text{FilterpD}(S_t, T*)$.
        Update $\theta^{t+1} = \text{argmin}_{\theta \in \Theta} \|\theta - (\theta^t - \eta g^t)\|_2^2$.
    end for
end function
```

5.3 Consequences for Generalized Linear Models

At this stage, with an optimal mean estimator in hand, we explore its consequences for general supervised learning tasks. As before, our goal is to design efficient estimators which work well in the presence of heavy-tailed data. To this end, we borrow the *robust gradient* framework of Prasad et al. [1] and present it in Algorithm 8. In particular, the algorithm presented in Algorithm 8 proceeds by passing the gradients at the current iterate θ^t through our heavy-tailed mean estimator. Note that Prasad et al. [1] used a similar algorithm in their work, but used GMOM [16] as their mean estimator, which lead to weaker results in the heavy-tailed setting. As our results show next, using Algorithm 7 as the mean estimator automatically leads to better bounds and are also near-optimal. The proofs for the technical results appearing in this section can be found in Appendix C.2.

Linear Regression. In this setting, we observe paired samples $\{(x_1, y_1), \dots (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$. We assume that the (x, y) pairs sampled from the true distribution P are linked via a linear model:

$$y = x^T \theta^* + w, \tag{5.8}$$

where w is drawn from a zero-mean distribution with bounded 4th moment with variance σ^2. We suppose that under P the covariates $x \in \mathbb{R}^p$, have mean 0, covariance $\tau_\ell \mathcal{I}_p \preceq \Sigma_x \preceq \tau_u \mathcal{I}_p$ and bounded 8th moments.

Corollary 27 (Heavy Tailed Linear Regression)**.** *Consider the statistical model in* (B.17)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$n \geq d^2 \frac{T \log^2(pT/\delta)}{C_1 \log(T/\delta)}$$

and if Algorithm 8 is initialized at 0 *with stepsize* $\eta = 2/(\tau_u + \tau_\ell)$ *and confidence* δ *then, it returns iterates* $\{\widehat{\theta}^t\}_{t=1}^T$ *such that with probability at least* $1 - \delta$

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^*\|_2 + \frac{C_2\sigma}{1-\kappa}\left(\sqrt{\frac{trace\left(\Sigma_x\right)}{(n/T)}}\right) + \frac{C_2\sigma}{1-\kappa}\left(\sqrt{\frac{\|\Sigma_x\|_2 \log T/\delta}{(n/T)}}\right) \tag{5.9}$$

for some contraction parameter $\kappa < 1$.

Remark: Note that setting $T \approx \log_{1/\kappa}\left(\sigma\sqrt{\frac{\mathrm{trace}(\Sigma_x)}{n}} + \sqrt{\frac{\|\Sigma_x\|_2 \log 1/\delta}{n}}\right)$ suggests that upto logarithmic factors, at a large enough sample size, we get an error rate of

$$\widetilde{\mathcal{O}}\left(\sigma\left(\sqrt{\frac{\mathrm{trace}(\Sigma_x)}{n}} + \sqrt{\frac{\|\Sigma_x\|_2 \log(1/\delta)}{n}}\right)\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. Note that even allowing for logarithmic factors, our estimator is the first efficient and practical approach which gets these rates for our assumptions. In particular, Prasad et al. [1], Hsu and Sabato [59] get an error of $\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathrm{trace}(\Sigma_x)\log(1/\delta)}{n}}\right)$. Other previous works in statistics such as Fan et al. [107], Sun et al. [121] achieve similar rates, but under the additional assumption that the covariates are sub-Gaussian. Recently, Cherapanamjeri et al. [122] also studied the problem of heavy-tailed linear regression, when the covariates are isotropic and have *certifiably* bounded 8th moments. In this setting, barring logarithmic factors, they achieve the same rate as us, but at a better sample complexity of $d^{3/2}$. However, the proposed estimator is based on a degree 8 sum-of-squares program and is not yet practical. We next present results for the case of generalized linear models.

Generalized Linear Models. In this setting, we observe data $\{(x_1, y_1), \dots (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$. We suppose that the (x, y) pairs sampled from the true distribution P_{θ^*} are linked via a linear model such that when conditioned on the covariates x, the response variable has the distribution:

$$P(y|x) \propto \exp\left(\frac{y\langle x, \theta^*\rangle - \Phi(\langle x, \theta^*\rangle)}{c(\sigma)}\right) \tag{5.10}$$

Here $c(\sigma)$ is a fixed and known scale parameter and $\Phi : \mathbb{R} \mapsto \mathbb{R}$ is the link function. We focus on the random design setting where the covariates $x \in \mathbb{R}^p$, have mean 0, and covariance Σ. We use the negative conditional log-likelihood as our loss function, i.e.

$$\bar{\mathcal{L}}(\theta; (x, y)) = -y\langle x, \theta\rangle + \Phi(\langle x, \theta\rangle) \tag{5.11}$$

Here we assume that the covariates have bounded 8th moment and that $\Phi'(\cdot)$ is smooth around θ^*. Specifically, we assume that there exist universal constants $L_{\Phi,2k}$, B_{2k} such that

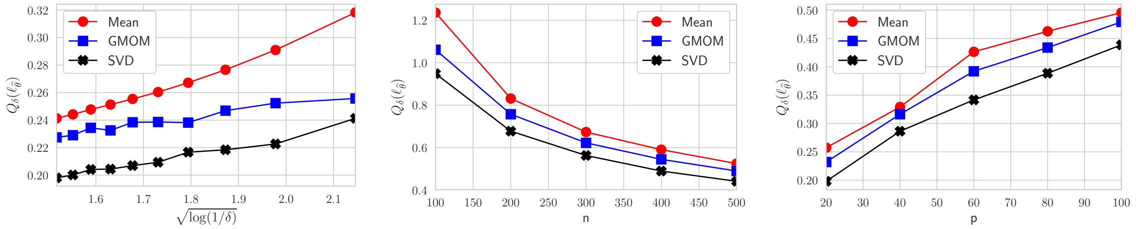
$$\begin{aligned}\mathbb{E}_x\left[\left|\Phi'(\langle x, \theta\rangle) - \Phi'(\langle x, \theta^*\rangle)\right|^{2k}\right] \le{} & L_{\Phi,2k}\|\theta^* - \theta\|_2^{2k} \\ & + B_{\Phi,2k}, \quad \text{for } k = 1, 2\end{aligned}$$

We also assume that $\mathbb{E}_x[\left|\Phi^{(t)}(\langle x, \theta^*\rangle)\right|^k] \le M_{\Phi,t,k}$ for $t \in \{1, 2, 4\}$, where $\Phi^{(t)}(\cdot)$ is the t^{th}-derivative of $\Phi(\cdot)$.

Corollary 28 (Heavy Tailed Generalized Linear Models)**.** *Consider the statistical model in* (5.10)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$n \ge d^2 \frac{T\log^2(pT/\delta)}{C_1\log(T/\delta)}$$

and if Algorithm 8 is initialized at 0 *with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and confidence δ then, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least* $1 - \delta$



(a) $Q_\delta(\ell_{\widehat{\theta}})$ vs $\sqrt{\log(1/\delta)}$ (b) $Q_\delta(\ell_{\widehat{\theta}})$ vs n (c) $Q_\delta(\ell_{\widehat{\theta}})$ vs p

Figure 5.2: Mean Estimation for Multivariate Pareto Distribution

$$
\begin{aligned}
\|\widehat{\theta}^t - \theta^*\|_2 \leq & \kappa^t \|\theta^*\|_2 \\
& + C_* \left(\sqrt{\frac{\mathrm{trace}(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\widetilde{\delta})}{\widetilde{n}}} \right),
\end{aligned} \tag{5.12}
$$

where $C_* = \frac{C_2\left[B_{\Phi,4}^{\frac{1}{4}} + c(\sigma)^{\frac{1}{2}} M_{\Phi,2,2}^{\frac{1}{4}} + c(\sigma)^{\frac{3}{4}} M_{\Phi,4,1}^{\frac{1}{4}}\right]}{1-\kappa}$ *for some contraction parameter* $\kappa < 1$.

Logistic Regression. In this case the (x, y) pairs are linked as:

$$
P(y = 1 | X = x) = \frac{1}{1 + \exp(-\langle x, \theta^* \rangle)} \tag{5.13}
$$

This corresponds to setting $\Phi(t) = \log(1 + \exp(t))$ and $c(\gamma) = 1$ in (5.10). The hessian of the population risk is given by

$$
\nabla^2 \mathcal{R}(\theta) = \mathbb{E}\left[\frac{\exp(\langle x, \theta \rangle)}{(1 + \exp(\langle x, \theta \rangle))^2} x x^T \right].
$$

Note that as θ diverges, the minimum eigenvalue of the hessian approaches 0 and the loss is no longer strongly convex. To prevent this, in this case we take the parameter space Θ to be bounded in an ℓ_2 sense

Corollary 29. *[Heavy-Tailed Logistic Regression] Consider the statistical model in* (5.13)*. There are universal constants* $C_1, C_2 > 0$ *such that if*

$$
n \geq d^2 \frac{T \log^2(pT/\delta)}{C_1 \log(T/\delta)}
$$

and if Algorithm 8 is initialized at 0 *with stepsize* $\eta = 2/(\tau_u+\tau_\ell)$ *and confidence* δ *then, it returns iterates* $\{\widehat{\theta}^t\}_{t=1}^T$ *such that with probability at least* $1-\delta$

$$\begin{aligned} \|\widehat{\theta}^t - \theta^*\|_2 \leq & \kappa^t \|\theta^*\|_2 \\ & + \left(\sqrt{\frac{\mathit{trace}\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\widetilde{\delta})}{\widetilde{n}}} \right), \end{aligned} \tag{5.14}$$

for some contraction parameter $\kappa < 1$*.*

5.4 Experiments

5.4.1 Mean Estimation

In this section, we conduct synthetic experiments to study the performance of our proposed estimators for heavy-tailed mean estimation.

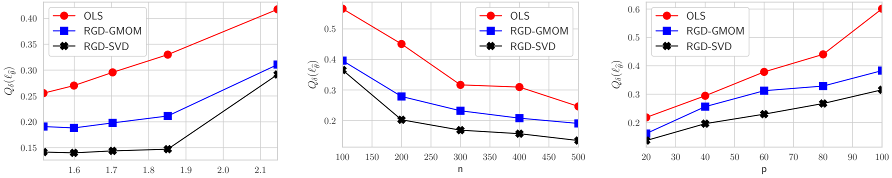
Setup. We generate $x \in \mathbb{R}^p$ from an isotropic zero-mean heavy-tailed distribution. We experiment with multivariate Pareto Distribution. For Pareto-distribution with tail-parameter β, the k^{th} order moments exists only if $k < \beta$, hence, smaller the β, the more heavy-tailed the distribution. We fix $k = 3$. In this setup, we experiment with different n, p and δ. For each setting of (n, p, δ), cumulative metrics are reported over 2000 trials. We vary n from 100 to 500, and p from 20 to 100.

Methods. We compare the filtering estimator with two baselines: (1) Sample mean, (2) Geometric Median of Means [16] which we refer to as GMOM.

Metric. For any estimator$(\widehat{\theta}_{n,\delta})$, we use $\ell(\widehat{\theta}_{n,\delta}) = \|\widehat{\theta} - \mu(P)\|_2$ as our primary metric. For each setting of (n, p, δ), we run the experiment for 2000 trials, which gives us access to the distribution of $\ell_{\widehat{\theta}_{n,\delta}}$. Since, we care about the deviation performance, we also measure the quantile error of the estimator, *i.e.* $Q_\delta(\widehat{\theta}) = \inf\{\alpha : \Pr(\ell(\widehat{\theta}) > \alpha) \leq \delta\}$. This can also be thought of as the length of confidence interval for a confidence level of $1-\delta$.

Hyperparameter Tuning. Apart from sample mean, all other estimators take into knowledge of δ, which is the desired confidence level. For GMOM, we follow the recommendation of Minsker [16] and set the number of blocks k is set to $\lceil 3.5 \log(1/\delta) \rceil$. We also set the number of iterations for the filtering estimator to $\lceil 3.5 \log(1/\delta) \rceil$.

Results. Figure 5.2 shows that our filtering estimator clearly outperforms both baselines across several metrics. Figure5.2(a) show that for any confidence level $1-\delta$, the length of the oracle confidence interval $(Q_\delta(\widehat{\theta}))$ for our estimator is better than all baselines. We also see better sample dependence in Figure 5.2(b), and better dimension dependence in Figure 5.2(c).



(a) $Q_\delta(\ell_{\widehat{\theta}})$ vs $\sqrt{\log(1/\delta)}$ (b) $(Q_\delta(\ell_{\widehat{\theta}}))$ vs n (c) $(Q_\delta(\ell_{\widehat{\theta}}))$ vs p

Figure 5.3: Linear Regression with Pareto Noise

5.4.2 Linear Regression

Setup. We generate $x \in \mathbb{R}^p$ from an isotropic zero-mean heavy-tailed distribution and set the true regression parameter $\theta^* = [1, 1, \ldots, 1] \in \mathbb{R}^p$. The response y is generated by $y = x^T\theta^* + w$ where w is drawn from a Pareto-distribution with tail-parameter β, $\beta = 3$. In this setup, we experiment with different n, p and δ. For each setting of (n, p, δ), cumulative metrics are reported over 2000 trials. We vary n from 100 to 500, and p from 20 to 100.

Methods. We compare the filtering based gradient descent estimator with two baselines: (1) Ordinary Least Squares (OLS), (2) gradient descent estimator which uses Algorithm 8 with GMOM as used in Prasad et al. [1]. Note that Prasad et al. [1] had previously shown that RGD-GMOM outperformed several other estimators like Hsu and Sabato [59] and ridge regression, hence, we skip them in our comparison.

Metric and Hyperparameter Tuning For any estimator$(\widehat{\theta}_{n,\delta})$, we use $\ell(\widehat{\theta}_{n,\delta}) = \|\widehat{\theta} - \theta^*\|_2$ as our primary metric. As before in the mean setting, for each setting of (n, p, δ), we run the experiment for 2000 trials, which gives us access to the distribution of $\ell_{\widehat{\theta}_{n,\delta}}$ and use the same quantile error in Section 5.4.1 as our metric. We use the same setting as our experiments for mean estimation and set the number of blocks k is set to $\lceil 3.5 \log(1/\delta) \rceil$ to estimate gradient robustly. Similarly, for RGD-SVD, we set the number of iterations for the filtering estimator to $\lceil 3.5 \log(1/\delta) \rceil$.

Results. Figure 5.3 shows that our filtering estimator clearly outperforms both baselines across several metrics. Figure 5.3(a) show that for any confidence level $1-\delta$, the length of the oracle confidence interval $(Q_\delta(\widehat{\theta}))$ for our estimator is better than all baselines. We also see better sample dependence in Figure 5.3(b), and better dimension dependence in Figure 5.3(c).

5.4.3 Generative Adversarial Networks

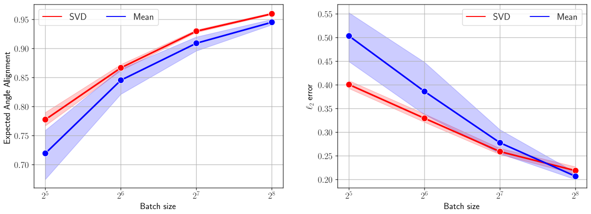
In Section 5.1.1, we found conclusive evidence that the gradients of the generator are certainly non-Gaussian and in particular are heavy-tailed. In this section, we study the effect of using our heavy-tailed mean estimator in training GANs, which is motivated by the prior analyses in the paper.

Setup. The setup is the same as the one considered for our initial investigation. As a baseline, we train a DCGAN on the MNIST and CIFAR10 datasets using mini-batch stochastic gradient descent. We use a batch size of 32 and a learning rate of $3 \cdot 10^{-4}$. As shown in Figure 5.1, we begin noticing heavy-tailed characteristics of the gradients at iteration 1000 and while training on CIFAR10, we noticed heavy-tailed behavior at around iteration 500. We refer to refer iteration numbers as points of heavy-tailedness.

Methods. We run our filtering estimator on a copy of the GANs trained on MNIST and CIFAR10 from the points of heavy-tailedness. We set the number of steps in the heavy tailed mean estimator to 5, and retain the same batch size and learning rate for a fair comparison. Both threads (using mini-batch SGD and heavy-tailed mean estimator) are trained for 5000 iterations in total.

Metric. We keep note of the variation in objectives. Additionally, we also compute the expected angle alignment with the true gradient *i.e.* $\cos \angle \hat{\theta}, \theta^*$. We use a large sample gradient computed using 10^5 samples of stochastic gradients as a proxy for the true gradient, which is intractable to compute in this case. As done in the previous experiments, we provide the ℓ_2 error as well. Finally, we also compute the log-likelihood on the test dataset as given by the discriminator for both models trained using mini-batch SGD (referred to as *Mean* and the heavy tailed mean estimator (referred to as *SVD*).

Results. First, we observe that that the gradients returned by the heavy-tailed mean estimator are better aligned with the large sample gradient, and achieves considerably lower ℓ_2-error for smaller batch sizes in Figure 5.4.3. Consistent with this observation, in Figure 5.4.3, we see that our algorithm allows achieving a higher objective function, both for the MNIST and CIFAR10 datasets. Table 5.1 shows that SVD achieves a considerably large log-likelihood at the end of 5000 iterations for both the MNIST and CIFAR10 datasets.



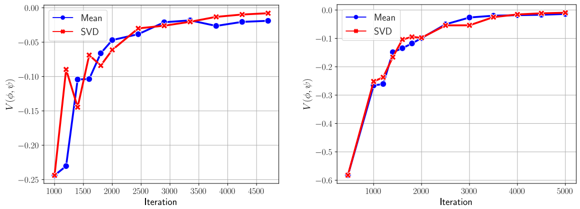
(a) Expected Align Alignment

(b) ℓ_2 Distance

Figure 5.4: Metrics computed on the sample mean and heavy-tailed mean estimator w.r.t. Expected Align Alignment and ℓ_2 distance metrics at a specific iteration while training MNIST.

Dataset	Model	Train	Test
2*MNIST	Mean	-438.69	-60.21
	SVD	**-221.25**	**-32.75**
2*CIFAR10	Mean	-341.65	-69.97
	SVD	**-237.19**	**-48.55**

Table 5.1: Log-likelihood on the train and test datasets for MNIST and CIFAR10 after 5000 iterations



(a) MNIST

(b) CIFAR10

Figure 5.5: Variation of GAN objective (5.1) with iterations

Chapter 6

Conclusion

In this thesis, we studied some fundamental problems in robust estimation. In particular, we gave (1) first statistically optimal estimators for mean and covariance estimation in the contaminated setting, (2) first computationally efficient estimator for robust linear regression, (3) first statistically optimal estimator for robustly learning Ising models, and (4) gave efficient and practical estimators which give sub-gaussian rates under weak moment assumptions. We conclude this thesis by highlighting some interesting open problems and directions for future work.

Learning Symmetric Distributions Efficiently. In Chapter 2, we showed that for symmetric distributions, we can recover the true mean upto an ℓ_2 error of $\Theta(\epsilon)$. This is a much faster rate than as compared to the worse rate of $\Omega(\epsilon^{1-1/2k})$ for 2k-moment bounded distributions. However, our proposed estimator is computationally inefficient. It is an interesting open problem to give a *computationally efficient* estimator for mean estimation for symmetric distributions which gives an error of $\Theta(\epsilon)$.

Robust Stochastic Optimization under Memory Constraints. In chapter 3, we studied robust estimators for risk minimization via robust gradient descent. However, our estimator relies on doing a full-batch gradient descent, which is computationally prohibitive. Moreover, our estimator requires constructing a covariance matrix of gradients, which requires storing $\Omega(p^2)$ entries where p is the dimensionality of the target parameter. It is an interesting problem to design robust stochastic algorithms that have low memory-footprints. In particular, even the seemingly benign task of giving a robust mean estimator that works with $O(p)$ memory is an open problem.

Robust Modern Machine Learning. Training of Modern Machine Learning algorithms such as Deep Reinforcement Learning based systems depend crucially on a cleverly selection of hyper parameters. Recently, Garg et al. [123] attributed the prevalence of clipping-style heuristics to the presence of inherently heavy-tailed gradients in such reinforcement learning tasks. It is an interesting line of future work to transfer some of the algorithms proposed in this thesis to the training of modern machine learning systems.

Bibliography

[1] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[2] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40(3-4):318–335, 1953.

[3] 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.

[4] P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[5] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[6] L. Devroye and L. Györfi. *Nonparametric density estimation: the L1 view*. Wiley series in probability and mathematical statistics. Wiley, 1985.

[7] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *Ann. Statist.*, 13(2):768–774, 06 1985.

[8] David L Donoho and Richard C Liu. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, pages 552–586, 1988.

[9] Ivan Mizera. On depth and deep points: a calculus. *Ann. Statist.*, 30(6):1681–1736, 12 2002. doi: 10.1214/aos/1043351254. URL https://doi.org/10.1214/aos/1043351254.

[10] Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 05 2020. doi: 10.3150/19-BEJ1144. URL https://doi.org/10.3150/19-BEJ1144.

[11] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *ArXiv e-prints, to appear in the Annals of Statistics*, 2015.

[12] Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery, 2016.

[13] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *arXiv preprint arXiv:1711.05381*, 2017.

[14] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[15] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

[16] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21 (4):2308–2335, 2015.

[17] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

[18] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[19] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[20] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.

[21] A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[22] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169 – 188, 1986.

[23] Samuel B Hopkins. Sub-Gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

[24] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-Gaussian rates. *arXiv preprint arXiv:1902.01998*, 2019.

[25] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics*. Wiley Online Library, 1986.

[26] Cecil Hastings, Frederick Mosteller, John W Tukey, and Charles P Winsor. Low moments for small samples: a comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426, 1947.

[27] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[28] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

[29] 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. ACM, 2018.

[30] 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. ACM, 2017.

[31] 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, 2017.

[32] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient

robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

[33] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's epsilon-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

[34] 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. ACM, 2018.

[35] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

[36] Yijun Zuo and Robert Serfling. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference*, 84(1-2):55–79, 2000.

[37] Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

[38] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

[39] Ricardo Antonio Maronna. Robust M-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67, 1976.

[40] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

[41] Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *arXiv preprint arXiv:1802.09514*, 2018.

[42] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[43] Emilien Joly, Gábor Lugosi, and Roberto Imbuzeiro Oliveira. On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11(1):440–451, 2017.

[44] Irina Rish, Guillermo A Cecchi, Aurelie Lozano, and Alexandru Niculescu-Mizil. *Practical applications of sparse modeling*. MIT Press, 2014.

[45] Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[46] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[47] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[48] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6): 381–395, 1981. doi: 10.1145/358669.358692. URL http://doi.acm.org/10.1145/358669.358692.

[49] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[50] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 169–212. PMLR, 2017. URL http://proceedings.mlr.press/v65/balakrishnan17a.html.

[51] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60. ACM, 2017. doi: 10.1145/3055399.3055491. URL https://doi.org/10.1145/3055399.3055491.

[52] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means : theory and practice. 2017.

[53] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[54] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[55] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4152–4160, 2016. URL http://papers.nips.cc/paper/6445-fast-algorithms-for-robust-pca-via-gradient-descent.

[56] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *Ann. Statist.*, 45(2):866–896, 04 2017.

[57] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 774–782. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/chen13h.html.

[58] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators, 2011.

[59] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

[60] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[61] Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, December 2019. doi: 10.4171/jems/937. URL https://doi.org/10.4171/jems/937.

[62] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. doi: 10.1561/2200000050. URL https://doi.org/10.1561/2200000050.

[63] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[64] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

[65] Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2971–2980, 2017. URL http://papers.nips.cc/paper/6890-variance-based-regularization-with-convex-objectives.

[66] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[67] Hamed Karimi, Julie Nutini, and Mark W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50. URL https://doi.org/10.1007/978-3-319-46128-1_50.

[68] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. 1999.

[69] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

[70] George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–39, 1983.

[71] Martin Hassner and Jack Sklansky. The use of markov random fields as models of texture. In *Image Modeling*, pages 185–198. Elsevier, 1981.

[72] Brian D Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2005.

[73] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, 2004.

[74] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[75] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[76] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[77] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[78] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, Pradeep Ravikumar, et al. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, 82(3):601–627, 2020.

[79] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258,

1925.

[80] C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[81] Imre Csiszár and Zsolt Talata. Consistent estimation of the basic neighborhood of markov random fields. *The Annals of Statistics*, pages 123–145, 2006.

[82] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 343–356. Springer, 2008.

[83] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782, 2015.

[84] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

[85] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[86] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

[87] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8069–8079, 2019.

[88] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Transactions on Information Theory*, 65(11):6829–6852, 2019.

[89] Erik M Lindgren, Vatsal Shah, Yanyao Shen, Alexandros G Dimakis, and Adam Klivans. On robust learning of ising models. *NeurIPS Workshop on Relational Representation Learning*, 2019.

[90] Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58 (7):4117–4134, 2012.

[91] PL Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

[92] Roland L Dobrushin and Senya B Shlosman. Completely analytical interactions: constructive description. *Journal of Statistical Physics*, 46(5-6):983–1014, 1987.

[93] Daniel W Stroock and Boguslaw Zegarlinski. The logarithmic sobolev inequality for discrete spin systems on a lattice. *Communications in Mathematical Physics*, 149(1):175–193, 1992.

[94] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Ensuring rapid mixing and low bias for asynchronous gibbs sampling. In *JMLR workshop and conference proceedings*, volume 48, page 1567. NIH Public Access, 2016.

[95] Christof Külske. Concentration inequalities for functions of gibbs fields with application to diffraction and random gibbs measures. *Communications in mathematical physics*, 239(1-2): 29–51, 2003.

[96] H Künsch. Decay of correlations under dobrushin's uniqueness condition and its applications.

Communications in Mathematical Physics, 84(2):207–222, 1982.

[97] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Estimating ising models from one sample. *arXiv preprint arXiv:2004.09370*, 2020.

[98] 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.

[99] David L Donoho and Richard C Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, pages 668–701, 1991.

[100] Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

[101] Luc Devroye, Abbas Mehrabian, Tommy Reddad, et al. The minimax learning rates of normal and ising undirected graphical models. *Electronic Journal of Statistics*, 14(1):2338–2361, 2020.

[102] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[103] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

[104] DF Andrews, PJ Bickel, FR Hampel, PJ Huber, WH Rogers, and JW Tukey. Robust estimates of location: Survey and advances, 1972.

[105] Surbhi Goel, Daniel M. Kane, and Adam R. Klivans. Learning ising models with independent failures. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1449–1469, Phoenix, USA, 25–28 Jun 2019. PMLR.

[106] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019.

[107] Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

[108] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/simsekli19a.html.

[109] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why adam beats sgd for attention models. *arXiv preprint arXiv:1912.03194*, 2019.

[110] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[111] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[112] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[113] Mohammad Mohammadi, Adel Mohammadpour, and Hiroaki Ogata. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika: International Journal for Theoretical and Applied Statistics*, 78(5):549–561, July 2015. doi: 10.1007/s00184-014-0515-7. URL https://ideas.repec.org/a/spr/metrik/v78y2015i5p549-561.html.

[114] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

[115] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. doi: 10.1080/01621459.1954.10501232.

[116] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

[117] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

[118] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458. ACM, 2014.

[119] Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: the high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2013.

[120] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

[121] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, pages 1–24, 2019.

[122] Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. *arXiv preprint arXiv:1912.11071*, 2019.

[123] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J Zico Kolter, Sivaraman Balakrishnan, Zachary C Lipton, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. *arXiv preprint arXiv:2102.10264*, 2021.

[124] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[125] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[126] Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

[127] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[128] Roman Vershynin. On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE, 2009.

[129] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[130] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer Science & Business Media, 2012.

[131] Yin Wang, Caglayan Dicle, Mario Sznaier, and Octavia Camps. Self scaled regularized robust regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3261–3269, 2015.

[132] S Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *CoRR, abs/0910.0610*, 2009.

[133] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[134] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[135] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[136] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Higher order concentration for functions of weakly dependent random variables. *Electron. J. Probab.*, 24:19 pp., 2019.

[137] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[138] Mehmet Eren Ahsen and Mathukumalli Vidyasagar. An approach to one-bit compressed sensing based on probably approximately correct learning theory. *The Journal of Machine Learning Research*, 20(1):408–430, 2019.

Appendix A

Supplementary Material for Chapter 2

A.1 Proofs

A.1.1 Proof of Lemma 1

Proof. Let $P = \mathcal{N}(0, \mathcal{I}_p)$ be the isotropic normal distribution. Let $R_P(\theta) = \mathbb{E}_{z \sim P}[\ell(\|z - \theta\|_2)]$, where $\ell : \mathbb{R} \mapsto \mathbb{R}$ is a convex loss, and let $\theta(P) = \operatorname{argmin}_\theta R_P(\theta)$ be the minimizer of the population risk. We assume that $\psi(\cdot) = \ell'(\cdot) < C$ is bounded. Note that when the derivative is unbounded, it is easy to argue that the corresponding risk will be non-robust. We also assumed that this risk is fisher-consistent for the Gaussian-distribution, *i.e.* $\theta(P) = 0$. For notational convenience, let $u(t) = \frac{\psi(t)}{t}$. Then,

$$\nabla R_P(\theta) = -\mathbb{E}_{z \sim P}\left[\underbrace{\frac{\psi(\|z-\theta\|_2)}{\|z-\theta\|_2}}_{u(\|z-\theta\|_2)}(z-\theta)\right].$$

As before, let $P_\epsilon = (1-\epsilon)P + \epsilon Q$. Then, we are interested in studying $\widehat{\theta}(P_\epsilon)$. To do this, by first order optimality, we know that $\theta(P_\epsilon)$ is a solution to the following equation:

$$(1-\epsilon)\nabla R_P(\theta(P_\epsilon)) + \epsilon \nabla R_Q(\theta(P_\epsilon)) = 0$$

First we calculate the derivative of $\theta(P_\epsilon)$ w.r.t. ϵ using the fixed point above. Taking derivative of the above equation w.r.t. ϵ

$$(1-\epsilon)\nabla^2 R_P(\theta(P_\epsilon))\dot{\theta}(P_\epsilon) - \nabla R_P(\theta(P_\epsilon)) + \epsilon\nabla^2 R_Q(\theta(P_\epsilon))\dot{\theta}(P_\epsilon) + \nabla R_Q(\theta(P_\epsilon)) = 0 \tag{A.1}$$

Under our assumption that ψ is continuous, we get that at $\epsilon = 0$,

$$\dot{\theta}(P_\epsilon)_{|\epsilon=0} = (-\nabla^2 R_P(\theta(P)))^{-1}\nabla R_Q(\theta(P)) \tag{A.2}$$

By fisher consistency of ℓ for $\mathcal{N}(0, \mathcal{I}_p)$, we have that $\theta(P) = 0$. Suppose that Q is a point mass distribution with all mass on θ_Q. Then, we have that,

$$\nabla R_Q(0) = -u(\|\theta_Q\|_2)\theta_Q$$

Our next step is to lower bound the operator norm of $-\nabla^2 R_P(\theta(P))$. To do this we show that for any unit vector $v \in \mathcal{S}^{p-1}$, $v^T(-\nabla^2 R_P(\theta(P)))v \leq \frac{C_2}{\sqrt{p}}$.

$$\nabla^2 R_P(\theta) = -\mathbb{E}_{z\sim P}\left[u(\|z-\theta\|_2)\mathcal{I}_p + \frac{u'(\|z-\theta\|_2)}{\|z-\theta\|_2}((z-\theta)(z-\theta)^T)\right]$$

Now, by definition $u(t) = \psi(t)/t$, so $u'(s) = (\psi'(s) - u(s))/s$. Plugging this above,

$$\nabla^2 R_P(\theta) = -E_{z\sim P}\left[u(\|z-\theta\|_2)(\mathcal{I}_p - \frac{(z-\theta)(z-\theta)^T)}{\|z-\theta\|_2^2}) + \frac{\psi'(\|z-\theta\|_2)}{\|z-\theta\|_2^2}(z-\theta)(z-\theta)^T))\right.$$

Hence, we get that

$$v^T\nabla^2 R_P(0)v = -\mathbb{E}_{z\sim N(0,I_p)}\left[u(\|z\|_2)(\|v\|_2^2 - (v^T(z/\|z\|_2))^2) + \psi'(\|z\|_2)(v^T(z/\|z\|_2))^2\right]$$

Further for Isotropic Gaussian, $\|z\|_2$ and $z/\|z\|_2$ are independent random variables. Also, since, $z/\|z\|_2$ is uniformly distributed on unit sphere, we get that $\mathbb{E}_{z\sim N(0,I)}[(v^T z/\|z\|_2)^2)] = \|v\|_2^2/p$.

$$(v^T(-\nabla^2 R_P(0))v) = \underbrace{\mathbb{E}_{z\sim N(0,I_p)}\left[u(\|z\|_2)\right](1-1/p)}_{\mathbf{T1}} + \underbrace{\mathbb{E}_{z\sim N(0,I_p)}\left[\psi'(\|z\|_2)\right]/p}_{\mathbf{T2}}$$

- **Controlling T1**

$$\begin{aligned}\mathbb{E}_{z\sim N(0,I_p)}[u(\|z\|_2)] &= \mathbb{E}_{z\sim\mathcal{N}(0,I_p)}\left[\frac{\psi(\|z\|_2)}{\|z\|_2}\right]\\ &\leq \sqrt{C\mathbb{E}\frac{1}{\|z\|_2^2}}\\ &\leq \frac{\sqrt{C_1}}{\sqrt{p-2}}, \qquad \text{(A.3)}\end{aligned}$$

  where we use that ψ is bounded by constant C. The last inequality is combination of Jensen's Inequality and plugging the mean of reciprocal of inverse chi-squared random variable [124].

- **Controlling T2.** Under our assumption that $\psi'(\cdot)$ exists and is bounded, we get that $T2 \leq \frac{C_1}{p}$ and can be ignored.

Hence, for large p, we get that $(v^T(-\nabla^2 R_P(0))v) \leq \sqrt{C_1/p}$. Now, if we put θ_Q at ∞, and use that $\psi(\infty) = C_1$, we get that,

$$\|\dot{\theta}(P_\epsilon)\|_2 = \psi(\|\theta_Q\|_2)\|\nabla^2 R_P(0)\frac{\theta_Q}{\|\theta_Q\|_2}\|_2 \geq C_2\sqrt{p}$$

□

A.1.2 Proof of Lemma 2

Proof. Let $P = N(0, \mathcal{I}_p)$. Every subset of size $(1-\epsilon)n$ can be thought of as samples from a mixture distribution defined in (2.3), where the mixture proportion η, ranges from $[0, \epsilon/(1-\epsilon)]$. In the asymptotic setting of $n \mapsto \infty$, the empirical squared loss over each subset corresponds to the population risk with the sampling distribution as P_η. For a given contamination distribution Q, let $R_{P_\eta}(\theta) = \mathbb{E}_{x \sim P_\eta}\left[\|x-\theta\|_2^2\right]$ and let $\theta(P_\eta) \overset{\text{def}}{=} \operatorname{argmin}_\theta R_{P_\eta}(\theta)$, then subset risk minimization returns,

$$\begin{aligned} &\widehat{\theta}_{\text{SRM}} = \theta(P_{\eta^*}) \\ &\text{where } \eta^* = \underset{\eta \in [0, \frac{\epsilon}{1-\epsilon}]}{\operatorname{argmin}} R_{P_\eta}(\theta(P_\eta)) \end{aligned} \tag{A.4}$$

We are interested in bounding the bias of SRM *i.e.*

$$\sup_Q \|\widehat{\theta}_{\text{SRM}} - \theta^*\|_2$$

To do this, we know that for any contamination distribution Q, the solution of SRM necessarily satisfies the following conditions.
Condition 1: Local Stationarity. $\theta(P_\eta) = \operatorname{argmin}_\theta R_{P_\eta}(\theta)$ is the minimizer of the risk with respect to a mixture distribution iff

$$\begin{aligned} \nabla R_{P_\eta}(\theta(P_\eta)) = (1-\eta)\nabla R_{P_\theta^*}(\theta(P_\eta)) \\ + \eta \nabla R_Q(\theta(P_\eta)) = 0. \end{aligned} \tag{A.5}$$

Condition 2: Global Fit Optimality. $\widehat{\theta}_{\text{SRM}} = \theta(P_{\eta^*})$ is the global minimizer of the population risk over all mixture distributions iff

$$\begin{aligned} R_{P_{\eta^*}}(\theta(P_{\eta^*})) &= (1-\eta^*) R_{P_0}(\theta(P_{\eta^*})) + \eta^* R_Q(\theta(P_{\eta^*})) \\ &\le R_{P_\eta}(\theta(P_\eta)) \quad \forall \eta \in \left[0, \frac{\epsilon}{1-\epsilon}\right] \end{aligned} \tag{A.6}$$

Using Conditions 1 and 2, we next derive the bias of SRM for mean estimation. We make a few simple observations.

- **Observation 1.** For any distribution P, we have,

$$R_P(\theta) = \operatorname{trace}(\Sigma(P)) + \|\theta - \mu(P)\|_2^2$$

- **Observation 2.** Condition 1 reduces to,
$$\mu(P_\eta) = \theta_\eta = (1-\eta)\mu(P) + \eta\mu(Q),$$
where $\mu(\cdot)$ is the Expectation functional.

Lemma 16. *Under the mixture model in Equation (2.3), for the squared error, we have that,*

$$R_{P_\eta}(\theta_\eta) = trace\,(\Sigma(P_\eta)) = (1-\eta)trace\,(\Sigma(P^*)) + \eta trace\,(\Sigma(Q)) + \eta(1-\eta)\|\mu(P^*) - \mu(Q)\|_2^2.$$

Now, from Lemma 16, we know that

$$R_{P_\eta}(\theta_\eta) = (1-\eta)\text{trace}\,(\Sigma(P)) + \eta\text{trace}\,(\Sigma(Q)) + \eta(1-\eta)\|\mu(P) - \mu(Q)\|_2^2$$

As a function of η, $R_{P_\eta}(\theta_\eta)$ is a concave quadratic function. Hence, it is always minimized at the end points of the interval $[0, \epsilon/(1-\epsilon)]$, which implies that $\eta^* \in \{0, \frac{\epsilon}{1-\epsilon}\}$.

Hence, we have that,

$$\widehat{\theta}_{\text{SRM}} = \begin{cases} \theta_{\frac{\epsilon}{1-\epsilon}}, & \text{if } R_{P_{\frac{\epsilon}{1-\epsilon}}}(\theta_{\frac{\epsilon}{1-\epsilon}}) \leq R_{P_0}(\theta_0). \\ \theta^*, & \text{otherwise.} \end{cases}$$

From Lemma 16, $R_{P_{\frac{\epsilon}{1-\epsilon}}}(\theta_{\frac{\epsilon}{1-\epsilon}}) \leq R_{P_0}(\theta_0)$ iff

$$\left(1 - \frac{\epsilon}{1-\epsilon}\right)\|\mu(P) - \mu(Q)\|_2^2 \leq \text{trace}\,(\Sigma(P)) - \text{trace}\,(\Sigma(Q))$$

Moreover, from Observation 2, we have that,

$$\|\theta_{\frac{\epsilon}{1-\epsilon}} - \mu(P)\|_2 = \frac{\epsilon}{1-\epsilon}\|\mu(P) - \mu(Q)\|_2$$

Combining the above two, we get that,

$$\|\widehat{\theta}_{\text{SRM}} - \mu(P)\|_2 = \left[\frac{\epsilon}{1-\epsilon}\|\mu(P) - \mu(Q)\|_2\right].\mathbf{1}\left\{\|\mu(P) - \mu(Q)\|_2^2 \leq \left(\frac{1-\epsilon}{1-2\epsilon}\right)(\text{trace}\,(\Sigma(P)) - \text{trace}\,(\Sigma(Q)))\right\}. \quad \text{(A.7)}$$

Equation 2.6 follows from it.

□

Proof of Lemma 16

Proof. We give two alternate proofs of the Lemma.

- Proof 1: This proceeds by expanding on the definition of risk.

  $$\begin{aligned} R_{P_\eta}(\theta_\eta) &= E_{z\sim P_\eta}[\|z-\theta_\eta\|_2^2] \\ &= (1-\eta)E_{z\sim P_0}[\|z-\theta_\eta\|_2^2] + \eta E_{z\sim Q}[\|z-\theta_\eta\|_2^2] \quad \text{Expectation by conditioning.} \\ &= (1-\eta)\left[\operatorname{trace}\left(\Sigma(P^*)\right) + \|\theta_\eta - \mu(P^*)\|_2^2\right] \\ &+ \eta\left[\operatorname{trace}\left(\Sigma(Q)\right) + \|\theta_\eta - \mu(Q)\|_2^2\right] \quad \text{From Observation 1.} \end{aligned}$$

  Now, using Observation 2 we get that,

  $$\|\theta_\eta - \mu(Q)\|_2 = (1-\eta)\|\mu(P^*) - \mu(Q)\|_2$$

  $$\|\theta_\eta - \mu(P^*)\|_2 = \eta\|\mu(P^*) - \mu(Q)\|_2$$

  Plugging this into above, we get,

  $$R_{P_\eta}(\theta_\eta) = (1-\eta)\operatorname{trace}\left(\Sigma(P^*)\right) + \eta\operatorname{trace}\left(\Sigma(Q)\right) + \|\mu(P^*) - \mu(Q)\|_2^2 \left(\eta^2(1-\eta) + (1 -\right.$$

  which recovers the statement of the Lemma.

- Proof 2: This proceeds by Law of Total Variance, or the Law of Total Cummulants. We know that $R_{P_\eta} = \operatorname{trace}\left(\Sigma(P_\eta)\right)$. Let $Z \sim P_\eta$, and let $Y \sim \text{Bernoulli}(1-\eta)$ be the indicator if the sample is from the true distribution. Then $Z|Y = 1 \sim P^*$, while $Z|Y = 0 \sim Q$.

  $$\operatorname{trace}\left(\Sigma(P_\eta)\right) = \underbrace{(1-\eta)\operatorname{trace}\left(\Sigma(P^*)\right) + \eta\operatorname{trace}\left(\Sigma(Q)\right)}_{\text{Var}(E[Z|Y])} + \underbrace{\eta(1-\eta)\|\mu(P^*) - \mu(Q)\|_2^2}_{E[\text{Var}(Z|Y)]}.$$

□

A.1.3 Proof of Lemma 3

Proof. Let $P_\epsilon = (1-\epsilon)P^* + \epsilon Q$. Let I^* be the interval $\mu \pm \frac{\sigma}{\delta_1^{\frac{1}{2k}}}$, where $\mu = \mathbb{E}_{x\sim P^*}[x]$. Moreover for notational convenience, let $f_n(u,v) = \sqrt{u(1-u)}\sqrt{\frac{\log(2/v)}{n}} + \frac{2}{3}\frac{\log(2/v)}{n}$. Let $\hat{I} = [a,b]$ be the interval obtained using $\mathcal{Z}_1$, *i.e.* the shortest interval containing $n(1-(\delta_1+\epsilon+f_n(\epsilon+\delta_1,\delta_3)))$ points of $\mathcal{Z}_1$. Note that in the algorithm, we have $\delta_1 = \epsilon$, and $\delta_3 = \delta/4$. As a first step, we bound the length of $\hat{I}$ and show that $\hat{I}$ and I^* must necessarily intersect.

Claim 1. *Let $\hat{I}$ be the shortest interval containing $1-\delta_4$ fraction of points, where $\delta_4 = (\delta_1+\epsilon) + f_n(\epsilon+\delta_1,\delta_3)$. Further assume that $\delta_4 < \frac{1}{2}$. Then with probability at least $1-\delta_3$,*

$$\mathit{length}(\hat{I}) \le \mathit{length}(I^*) \le \frac{2\sigma}{\delta_1^{\frac{1}{2k}}},$$

Moreover, if $\delta_4 < \frac{1}{2}$, then $\hat{I} \cap I^* \neq \phi$, which implies

$$|z-\mu| \le \frac{4\sigma}{\delta_1^{\frac{1}{2k}}} \forall z \in \hat{I}$$

Proof. We first show that with probability at least $1-\delta_3$, I^* contains at least $n(1-\delta_4)$ points(Claim 5). Hence, since our algorithm chooses the shortest interval($\hat{I}$) containing $1-\delta_4$ fraction of points, length of $\hat{I}$ is less than length of I^*. Next, if δ_4 is less than $\frac{1}{2}$, then there are two intervals $\hat{I}$ and I^* respectively, which contain at least $n/2$ points. Hence, they must necessarily intersect. □

Next, we control the final error of our estimator. Let $|\hat{I}| = \sum_{z\in\mathcal{Z}_2} \mathbb{I}\left\{z_i \in \hat{I}\right\}$ be the number of points which lie in $\hat{I}$. Similarly, let $|\hat{I}_Q|$ and $|\hat{I}_{P^*}|$ number of points which lie in $\hat{I}$, which are distributed according to Q and P^* respectively.

$$\left|\frac{1}{|\hat{I}|}\sum_{x_i\in\hat{I}} x_i - \mu\right| \le \underbrace{\left|\frac{1}{|\hat{I}|}\sum_{\substack{x_i\in\hat{I}\\x_i\sim Q}}(x_i-\mu)\right|}_{T1} + \underbrace{\left|\frac{1}{|\hat{I}|}\sum_{\substack{x_i\in\hat{I}\\x_i\sim P^*}}(x_i-\mu)\right|}_{T2} \tag{A.8}$$

Control of T1. To control T1, we can write it as:

$$\begin{aligned} T1 &= \left| \frac{1}{|\hat{I}|} \sum_{\substack{x_i \in \hat{I} \\ x_i \sim Q}} (x_i - \mu) \right| \\ &\leq \underbrace{\frac{|\hat{I}_Q|}{|\hat{I}|}}_{T1a} \underbrace{\max_{\substack{x_i \in \hat{I} \\ x_i \sim Q}} |x_i - \mu|}_{T1b} \end{aligned} \tag{A.9}$$

where $\hat{I}_Q$ is the number of points in $\hat{I}$ distributed according to Q. To control T1a, we use Bernsteins inequality. To control T1b, we use Claim 1. The claim below formally controls T1.

Claim 2. *Let $\hat{I}$ be the shortest interval containing $n(1 - \delta_4)$ of the points, where $\delta_4 = (\delta_1 + \epsilon) + f_n(\epsilon + \delta_1, \delta_3)$. Further assume that $\delta_4 < \frac{1}{2}$. Then, with probability at least $1 - \delta_3 - \delta_5$, we have that,*

$$T1 \leq \frac{|\hat{I}_Q|}{|\hat{I}|} \max_{\substack{x_i \in \hat{I} \\ x_i \sim Q}} |x_i - \mu| \leq \frac{\epsilon + f_n(\epsilon, \delta_5)}{1 - \delta_4} \frac{4\sigma}{\delta_1^{1/2k}} \tag{A.10}$$

Proof. Using Bernstein's bound, we know that wp at least $1 - \delta_5$,

$$|\hat{I}_Q| \leq n(\epsilon + \sqrt{\epsilon(1 - \epsilon)} \sqrt{\frac{\log(1/\delta_5)}{n}} + \frac{2}{3} \frac{\log(1/\delta_5)}{n}),$$

This follows from the fact that number of points drawn from Q which lie in $\hat{I}$ is less than the total number of points drawn according to Q. In Claim 1, we showed that when $\delta_4 < \frac{1}{2}$, then, with probability at least $1 - \delta_3$, we get that $\hat{I} \cap I^* \neq \phi$, *i.e.* the intervals intersect, and that $length(\hat{I}) < length(I^*)$. Hence, we get,

$$\max_{\substack{x_i \in \hat{I} \\ x_i \sim Q}} |x_i - \mu| \leq \frac{4\sigma}{\delta_1^{1/2k}}$$

□

Control of T2. To control T2, we write it as

$$T2 = \left| \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \left[\frac{1}{|\hat{I}_{P^*}|} \sum_{\substack{x_i \in \hat{I} \\ x_i \sim P^*}} (x_i - \mu) \right] \right| \tag{A.11}$$

$$\leq \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \underbrace{\left| \left(\frac{1}{|\hat{I}_{P^*}|} \sum_{\substack{x_i \in \hat{I} \\ x_i \sim P^*}} x_i\right) - E[x|x \in \hat{I}, x \sim P^*] \right|}_{T2a} + \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \underbrace{\left| E[x|x \in \hat{I}, x \sim P^*] - \mu \right|}_{T2b} \tag{A.12}$$

- **Control of T2a:** To bound the distance between the mean of the points from P^* within $\widehat{I}$ and $E[x|x \sim P^*, x \in \hat{I}]$, we will use Bernsteins bound(Lemma 17) for bounded random variables. We know that the random variables are in a bounded interval $b = length(\widehat{I}) \leq \frac{\sigma}{\delta^{\frac{1}{2k}}}$, and that conditional variance of the random variables, when conditioned on them lying in $\hat{I}$ is controlled using Lemma 20. In particular, Lemma 20 shows that for any event E, which occurs with probability $P(E) \geq \frac{1}{2}$,

  $$E_{x \sim P^*}[(x - E[x|x \in E])^2 | x \in E] \leq \sigma^2 / P(E).$$

  Using these arguments, we get that with probability at least $1 - \delta_7$,

  $$T2a \leq \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} + \frac{2\sigma}{\delta_1^{1/2k}} \frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|}, \tag{A.13}$$

  where $P^*(\hat{I})$ is the probability that a random variable drawn according to P^* lies in $\hat{I}$.
- **Control of T2b:** To control $T2b$, we use the general mean shift lemma (Lemma 6), which controls how far the mean can move when conditioned on an event. We get that,

  $$T2b \leq 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} \tag{A.14}$$

Combining the bounds in (A.13) and (A.14), we get

$$T2 \leq 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} + \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} + \frac{2\sigma}{\delta_1^{1/2k}} \frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|} \tag{A.15}$$

Combining the upper bound on T1 in (A.10) with (A.15), we get that with probability at least $1 - \delta_3 - \delta_5 - \delta_6 - \delta_7$

$$T1{+}T2 \leq \frac{\epsilon + f_n(\epsilon, \delta_5)}{1 - \delta_4}\frac{4\sigma}{\delta_1^{1/2k}} {+} 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} {+} \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} {+} \frac{2\sigma}{\delta_1^{1/2k}}\frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|}$$

We rearrange terms and use our assumption that ϵ is small enough that $\hat{I}_{P^*} \geq n/2$. We also plugin the upper bound on $(P^*(\hat{I})^c)^{1-1/(2k)}$ from Claim 3 and set $\delta_1 = \epsilon$, and $\delta_5 = \delta_6 = \delta_3 = \delta_7 = \delta/4$. Hence, we get that with probability at least $1 - \delta$

$$T1 + T2 \leq C_1\sigma\epsilon^{1-1/2k} + C_2\sigma(\frac{\log n}{n})^{1-\frac{1}{2k}} + C_3\sigma\sqrt{\frac{\log(1/\delta)}{n}} + C_4\sigma\frac{\log(1/\delta)}{n\epsilon^{\frac{1}{2k}}} \tag{A.16}$$

Since, we ensure that $\epsilon = \max(\epsilon, \frac{\log(1/\delta}{n})$ hence, $\frac{\log(1/\delta)}{n\epsilon^{\frac{1}{2k}}} \leq \epsilon^{1-\frac{1}{2k}}$. Note that our assumption of $\delta_4 < \frac{1}{2}$ boils down to ϵ being small enough such that $2\epsilon + \sqrt{\epsilon\frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n} < \frac{1}{2}$. Hence, we recover the final statement of the theorem.

□

Auxillary Proofs

Claim 3. *Let $\hat{I}$ be the shorted interval containing $n(1 - \delta_4)$ points from $\mathcal{Z}_1$. Let $P^*(\hat{I})$ is the probability that a random variable drawn according to P^* lies in $\hat{I}$. Then, there exists universal constants $C_1, C_2 > 0$ such that wp at least $1 - \delta_6$, we have that*

$$(P^*(\hat{I})^c)^{1-\frac{1}{2k}} \leq C_1\epsilon^{1-\frac{1}{2k}} + C_2\delta_1^{1-\frac{1}{2k}} + C_3(\frac{\log n}{n})^{1-\frac{1}{2k}} + C_4(\frac{\log(1/\delta_6)}{n})^{1-\frac{1}{2k}} + C_5(\frac{\log(1/\delta_3)}{n})^{1-} \tag{A.17}$$

Proof. Note that $\hat{I}$ is obtained by choosing the shortest interval containing $n(1 - \delta_4)$ points from $\mathcal{Z}_1$. We first bound $P_n^*(\hat{I})$, *i.e.* the empirical probability of samples distributed according to P^* which lie in $\hat{I}$. To do this, note that in $\mathcal{Z}_1$, number of points drawn from Q which lie in $\hat{I}$, say $\hat{n}_Q$ is less than the total number of points drawn according to Q. Using Bernstein's bound, we know that wp at least $1-\delta_6$,

$$|\hat{n}_Q| \leq n(\epsilon + \sqrt{\epsilon(1-\epsilon)}\sqrt{\frac{\log(1/\delta_6)}{n}} + \frac{2}{3}\frac{\log(1/\delta_6)}{n})$$

Let $\hat{n}_{P^*}$ be the number of points in $\mathcal{Z}_1$, which are drawn from P^* and which lie in $\hat{I}$. Since $|\hat{n}_Q| + |\hat{n}_{P^*}| = |\hat{I}| = n(1-\delta_4)$, hence the above implies that with probability at least $1-\delta_6$,

$$|\hat{n}_{P^*}| \geq n(1-\delta_4) - n(\epsilon + \sqrt{\epsilon(1-\epsilon)}\sqrt{\frac{\log(1/\delta_6)}{n}} + \frac{2}{3}\frac{\log(1/\delta_6)}{n}),$$

Note that $P^*_n(\hat{I}) = \frac{|\hat{n}_{P^*}|}{\sum_i \mathbb{I}\{x_i \sim P^*\}}$. Hence, we get that,

$$\begin{aligned} P^*_n(\hat{I}) &\geq \frac{|\hat{n}_{P^*}|}{n} \\ &\geq 1 - (\epsilon + \delta_4) - f_n(\epsilon, \delta_6) \end{aligned} \tag{A.18}$$

This implies that,

$$\begin{aligned} P^*_n(\hat{I})^c &\leq (\epsilon + \delta_4) + f_n(\epsilon, \delta_6) \\ &\leq 2\epsilon + \delta_1 + f_n(\epsilon, \delta_6) + f_n(\epsilon + \delta_1, \delta_3) \\ &\leq 4\epsilon + 2\delta_1 + C_1\frac{\log(1/\delta_6)}{n} + C_2\frac{\log(1/\delta_3)}{n} \end{aligned} \tag{A.19}$$

To finally bound the probability of a sample drawn from P^* to lie in $\hat{I}$, we use the relative deviations VC bound(Lemma 18), which gives us,

$$P^*(\hat{I})^c \leq \underbrace{P^*_n(\hat{I})^c}_{A_1} + 4\sqrt{(\frac{P^*_n(\hat{I})^c \log \mathcal{S}[2n]}{n}) + (\frac{P^*_n(\hat{I})^c \log(4/\delta_6)}{n}) + \frac{\log \mathcal{S}[2n]}{n} + \frac{\log(4/\delta_6)}{n}} \tag{A.20}$$

where $\mathcal{S}[2n] = O(n^2)$. Using that $\sqrt{ab} \leq a + b, \forall a, b \geq 0$, we get that,

$$P^*(\hat{I})^c \leq C_1 P^*_n(\hat{I})^c + C_2(\frac{\log \mathcal{S}[2n]}{n} + \frac{\log(4/\delta_6)}{n}) \tag{A.21}$$

Hence, we get that,

$$(P^*(\hat{I})^c)^{1-\frac{1}{2k}} \leq C_1\epsilon^{1-\frac{1}{2k}} + C_2\delta_1^{1-\frac{1}{2k}} + C_3(\frac{\log n}{n})^{1-\frac{1}{2k}} + C_4(\frac{\log(1/\delta_6)}{n})^{1-\frac{1}{2k}} + C_5(\frac{\log(1/\delta_3)}{n})^{1-} \tag{A.22}$$

□

Claim 4. *Let $P^*(I^*)$ be the probability that a sample drawn according from P_ϵ is distributed according to P^* and lies in I^*.*

$$P^*(I^*) \geq (1-\epsilon)(1-\delta_1) = 1-(\epsilon+\delta_1-\epsilon\delta_1) \geq 1-\underbrace{(\epsilon+\delta_1)}_{\delta_2} = 1-\delta_2$$

Proof. For any $x \sim P_\epsilon$, define, $z_i = 1$ if $x \sim P^*$. Now, for any $x \sim P^*$, we know that, by chebyshevs we know that,

$$P(|x-\mu| \geq t) = P((x-\mu)^{2k} \geq t^{2k}) \leq E[(x-\mu)^{2k}]/t^{2k} \leq C_{2k}\sigma^{2k}/t^{2k}$$

Hence, we get that wp at least $1-\delta_1$, $x \in \mu \pm \sigma/(\delta_1)^{1/2k}$ □

The following claim lower bounds the empirical fraction of samples which are distributed according to P^* and lie in I^*, when n samples are drawn from P_ϵ.

Claim 5. *Let $P_n^*(I^*)$ be the empirical fraction of points which are distributed according to P^* and lie in I^*, when n samples are drawn from P_ϵ. Then, with probability at least $1-\delta_3$,*

$$P_n^*(I^*) \geq 1 - \underbrace{(\delta_2 + \sqrt{(\delta_2(1-\delta_2))}\sqrt{\frac{\log(1/\delta_3)}{n}} + \frac{2}{3}\frac{\log(1/\delta_3)}{n})}_{\delta_4=(\delta_1+\epsilon)+f_n(\epsilon+\delta_1,\delta_3)},$$

Proof. This follows from Bernstein's inequality(Lemma 17). □

Lemma 17. *[Bernsteins bound,] Let $X \sim P^*$ be a scalar random variable such that $|X-E[x]| \leq b$ with variance σ^2. Then, given n samples $\{x_1, x_2, \ldots, x_n\} \sim P^*$, the empirical mean, $\bar{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is such that,*

$$P(|\bar{x_n} - E[x]| > t) \leq 2\exp(\frac{-nt^2}{2\sigma^2 + 2bt/3})$$

which can be equivalently re-written as. With probability at least $1 - \delta$,

$$|\bar{x_n} - E[x]| \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2b\log(1/\delta)}{3n}$$

Lemma 18. *[Relative deviations, [125]] Let $\mathcal{F}$ be a function class consisting of binary functions f. Then, with probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} |P(f) - P_n(f)| \leq 4\sqrt{P_n(f)\frac{\log(S_{\mathcal{F}}(2n)) + \log(4/\delta)}{n}} + C_1\frac{\log(S_{\mathcal{F}}(2n)) + \log(4/\delta)}{n},$$

where $S_{\mathcal{F}}(n) = \sup\limits_{z_1,z_2,\ldots,z_n} |\{(f(z_1), f(z_2), \ldots, f(z_n)) : f \in \mathcal{F}\}|$ is the growth function, i.e. *the maximum number of ways into which n-points can be classified the function class.*

Lemma 19. *[General Mean shift, [126]] Suppose that a distribution P^* has mean μ and variance σ^2 with bounded $2k^{th}$-moments. Then, for any event A which occurs with probability at least $1 - \epsilon \geq \frac{1}{2}$,*

$$|\mu - E[x|A]| \leq 2\sigma\epsilon^{1-\frac{1}{2k}}$$

In particular, for just bounded second moments, we get that $|\mu - E[x|A]| \leq 2\sigma\sqrt{\epsilon}$.

Proof. For any event E, Let $\mathbb{I}\{E\}$ denote the indicator variable for E.

$$|E_{x\sim P^*}[x|E] - \mu]| = \frac{|E_{x\sim P^*}((x-\mu)\mathbb{I}\{E\})|}{P(E)} \leq \frac{E[|x-\mu|^p]^{\frac{1}{p}}(E[\mathbb{I}\{E\}^q]^{1/q})}{P(E)}, \tag{A.23}$$

where $p, q > 1$ are such that $1/p + 1/q = 1$. Put $p = 2k$, we get,

$$|E_{x\sim P^*}[x|E] - \mu]| \leq \frac{\sigma}{(P(E))^{1/2k}}$$

Now, we know that, $|E[X|A] - \mu| = \frac{1-P(A)}{P(A)}|E[X|A^c] - \mu|$. Putting $E = A^c$, we get,

$$|E[X|A] - \mu| \leq \frac{1-P(A)}{P(A)}\frac{\sigma}{(1-P(A))^{1/2k}} \leq 2\sigma\epsilon^{(1-\frac{1}{2k})}.$$

□

Lemma 20. *[Conditional Variance Bound] Suppose that a distribution P^* has mean μ and variance σ^2. Then, for any event A which occurs with probability at least $1 - \epsilon$, the variance of the conditional distribution is bounded as:*

$$(E[(x - E[x|A])^2|A]) \leq \frac{\sigma^2}{(1 - \epsilon)}$$

Proof. Let $\mu_A = E[y|A]$, $d = \mu_A - \mu$. From Lemma 6, we know, $d \leq \sigma 2\sqrt{\epsilon}$. Observe the following,

$$E[(y - \mu_A)^2|A] = E[(y - \mu - d)^2|A] = E[((y - \mu)^2 - 2d(y - \mu) + d^2)|A] \tag{A.24}$$

$$= E[(y - \mu)^2|A] - d^2 \tag{A.25}$$

$$\leq E[(y - \mu)^2|A] \tag{A.26}$$

$$\leq \frac{\sigma^2}{1 - \epsilon}, \tag{A.27}$$

□

A.1.4 Proof of Lemma 5

Proof. For brevity, let $\widehat{\theta}_\delta = \operatorname*{argmin}_\theta \sup_{u \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \mathrm{f}(\{u^Tx_i\}_{i=1}^n, \epsilon, \frac{\delta}{5^p})|$, where f is our univariate estimator. Let $\theta^* = \mathbb{E}[x]$ be the true mean. Then, we can write the ℓ_2 error in its variational form.

$$\|\widehat{\theta}_\delta - \theta^*\|_2 = \sup_{u \in \mathcal{S}^{p-1}} |u^T(\widehat{\theta}_\delta - \theta^*)| \tag{A.28}$$

Suppose $\{y_i\}$ is a $\frac{1}{2}$-cover of the net, so there exist a y_j such that $u = y_j + v$, where $\|v\|_2 \leq \epsilon$.

$$\begin{aligned}
\|\widehat{\theta}_\delta - \theta^*\|_2 &\leq \sup_{u \in \mathcal{S}^{p-1}} |y_j^T(\widehat{\theta}_\delta - \theta^*)| + |v^T(\widehat{\theta}_\delta - \theta^*)| \\
&\leq \sup_{y_j \in \mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})} |y_j^T(\widehat{\theta}_\delta - \theta^*)| + \|v\|_2\|\widehat{\theta}_\delta - \theta^*\|_2 \\
&\leq 2 \sup_{y_j \in \mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})} |y_j^T(\widehat{\theta}_\delta - \theta^*)|
\end{aligned}$$

$$\begin{aligned}
\|\widehat{\theta}_\delta - \theta^*\|_2 &\leq 2 \sup_{u \in \mathcal{N}^{1/2}} |u^T(\widehat{\theta} - \theta^*)| && \text{(A.29)} \\
&\leq 2\left[\sup_{u \in \mathcal{N}^{1/2}} |u^T\widehat{\theta} - f(u^TP_n, \epsilon; \tilde{\delta})| + \sup_{u \in \mathcal{N}^{1/2}} |u^T\theta^* - f(u^TP_n, \epsilon; \tilde{\delta})|\right] && \text{(A.30)} \\
&\leq 4 \sup_{u \in \mathcal{N}^{1/2}} |u^T\theta^* - f(u^TP_n, \epsilon; \tilde{\delta})| && \text{(A.31)}
\end{aligned}$$

For a fixed u, the distribution u^TP has mean $u^T\theta^*$, where θ^* is the mean of the multivariate distribution P. Hence, we get that, for a confidence level $\tilde{\delta}$, when the univariate estimator is applied to the projection of the data long u, it returns a real number such that, with probability at least $1 - \tilde{\delta}$

$$|f(u^TP_n; \epsilon; \tilde{\delta}) - u^T\theta^*| \leq C_1\omega_f(\epsilon, u^TP, \tilde{\delta})$$

Taking a union bound over the elements of the cover, and using the fact that $|\mathcal{N}^{1/2}(\mathcal{S}^{p-1})| \leq 5^p$ [127], we substitute $\tilde{\delta} = \delta/(5^p)$ and recover the statement of the Lemma.

□

A.1.5 Proof of Lemma 6

Proof. Let $\widehat{\theta}_\delta = \operatorname*{argmin}_{\theta\in\Theta_s} \sup_{u\in\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \mathrm{f}(\{u^Tx_i\}_{i=1}^n, \epsilon, \frac{\delta}{(6ep/s)^s})|$, where $f(\cdot)$ is our univariate estimator. Observe that since $\widehat{\theta}_\delta$ and the true mean θ^* are both s-sparse. Hence, the error vector $\widehat{\theta} - \theta^*$ is atmost $2s$-sparse. Then, we can write the ℓ_2 error in its variational form,

$$\|\widehat{\theta}_\delta - \theta^*\|_2 = \sup_{u\in\mathcal{S}^{p-1}\cap\mathcal{B}_{2s}} |u^T(\widehat{\theta}_\delta - \theta^*)|, \tag{A.32}$$

where $\mathcal{S}^{p-1}\cap\mathcal{B}_{2s}$ is the set of unit vectors which are $2s$-sparse. The remaining of the proof follows along the lines of proof of Lemma 5, coupled with the fact that the cardinality of the half-cover of an $2s$-sparse ball, *i.e.* $\left|\mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})\right| \leq (\frac{6ep}{s})^s$ [128].

□

A.1.6 Proof of Lemma 7

Let $\widehat{\Theta}_{\mathrm{f}} = \operatorname{argmin}_{\Theta\in\mathcal{F}} \sup_{u\in\mathcal{N}^{1/4}(\mathcal{S}^{p-1})} |u^T\Theta u - f(\{(u^Tz_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p})|$, where f is a univariate estimator, and z_i are the pseudo-samples obtained by $z_i = (x_{i+n/2} - x_i)/\sqrt{2}$. We begin by first using one-step discretization,

$$\begin{aligned}\|\widehat{\Theta}_{\mathrm{f}} - \Sigma(P)\|_2 &= \sup_{u\in\mathcal{S}^{p-1}} |u^T(\widehat{\Theta}_{\mathrm{f}} - \Sigma(P))u| \\ &\leq \frac{1}{1-2\gamma} \sup_{y\in\mathcal{N}^\gamma} |y^T(\widehat{\Theta}_{\mathrm{IM}} - \Sigma(P))y|,\end{aligned}$$

where $\mathcal{N}^\gamma$ is the γ-cover of the unit sphere. We set $\gamma = 1/4$.

$$\begin{aligned}\|\widehat{\Theta}_{\mathrm{f}} - \Sigma(P)\|_2 &\leq 2 \sup_{u\in\mathcal{N}^{1/4}} |u^T(\widehat{\Theta}_{\mathrm{f}} - \Sigma(P))u| \\ &\leq 2 \sup_{u\in\mathcal{N}^{1/4}} |u^T\widehat{\Theta}_{\mathrm{IM}}u - f(\{(u^Tz_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p}))| \\ &+ 2 \sup_{u\in\mathcal{N}^{1/4}} |u^T\Sigma(P)u - f(\{(u^Tz_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p})| \\ &\leq 4 \sup_{u\in\mathcal{N}^{1/2}} |u^T\Sigma(P)u - f(u^T\mathcal{X}_n, \epsilon; \tilde{\delta})|\end{aligned}$$

For a fixed u, for the clean samples in z_i, $(u^T z_i)^2$ has mean $u^T \Sigma(P) u$, and variance $C_4 (u^T \Sigma(P) u)^2$. Note that the scalar random variables $(u^T z_i)^2$ have bounded k moments, whenever x_i has bounded $2k$-moments. Hence, for a fixed u, we get that with probability at least $1 - \delta$,

$$|f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p}) - u^T \Sigma(P) u| \lesssim \omega_f(2\epsilon, u^T P^{\otimes 2}, \tilde{\delta})$$

Taking a union bound over the elements of the cover, and using the fact that $|\mathcal{N}^{1/4}(\mathcal{S}^{p-1})| \leq 9^p$ [127], we substitute $\tilde{\delta} = \delta/(9^p)$ and recover the statement of the Lemma.

A.1.7 Proof of Lemma 8

Let $\widehat{\Theta}_{\mathrm{f,s}} = \mathrm{argmin}_{\Theta \in \mathcal{F}_s} \sup_{u \in \mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} |u^T \Theta u - f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{(9ep/s)^s})|$, where f is a univariate estimator, and z_i are the pseudo-samples obtained by $z_i = (x_{i+n/2} - x_i)/\sqrt{2}$.

Observe that since $\widehat{\Theta}_{\mathrm{f,s}}$ and the true covariance $\Sigma(P)$ are both in $\mathcal{F}_s$. Hence, the difference matrix $\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)$ has atmost $2s$ non-zero off diagonal elements. Hence, we can write that $\|\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)\|_2 = \sup_{u \in \mathcal{B}_{2s} \cap \mathcal{S}^{p-1}} |u^T(\widehat{\Theta}_{\mathrm{IM}}^{(s)} - \Sigma(P))u|$, where $\mathcal{B}_{2s} \cap \mathcal{S}^{p-1}$ is the set of unit vectors which are atmost $2s$-sparse. Using the one-step discretization, we get that,

$$\|\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)\|_2 \leq 2 \sup_{u \in \mathcal{N}^{1/4}(\mathcal{B}_{2s} \cap \mathcal{S}^{p-1})} |u^T(\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P))u|$$

The remainder of the proof follows from the proof of Lemma 7 coupled with the fact that the cardinality of the 1/4-cover of an $2s$-sparse ball $\left|\mathcal{N}^{1/4}(\mathcal{S}^{p-1})\right| \leq (\frac{9ep}{s})^s$ [128].

A.1.8 Proof of Corollary 5

Proof. From Corollary 4, we know that the with probability at least $1 - \delta$ sparse covariance estimator satisfies,

$$\underbrace{\|\widehat{\Theta}_{\mathrm{IM,s}} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2 \epsilon^{1-1/k} + \|\Sigma(P)\|_2 \sqrt{\frac{s \log p}{n}} + \|\Sigma(P)\|_2 \sqrt{\frac{\log 1/\delta}{n}}}_{T1}$$

Let $\widehat{\Theta}_{\text{IM,s}} - \Sigma(P) = \Delta$, then, we have that $\|\Delta\|_2 \leq T1$. Using Weyl's Inequality, we know that,

$$|\lambda_{r+1}(\widehat{\Theta}_{\text{IM,s}}) - \lambda_{r+1}(\Sigma(P))| \leq \|\Delta\|_2$$

We know that $\lambda_{r+1}(\Sigma(P)) = 1$. Hence, we have that $\lambda_{r+1}(\widehat{\Theta}_{\text{IM,s}}) \in 1 \pm T1$. We also know that $\lambda_r(\Sigma(P)) = 1 + \Lambda_r$. Hence, we can now lower bound the eigengap, *i.e.*

$$|\lambda_r(\Sigma) - \lambda_{r+1}(\widehat{\Theta}_{\text{IM,s}})| \geq \Lambda_r - T1$$

Under the assumption that $T1 < \frac{1}{2}\Lambda_r$, and using Davis-Kahan Theorem [129], we get that,

$$\|VV^T - \hat{V}\hat{V}^T\|_F \leq \frac{\|\Theta_\delta - \Sigma\|_2}{\Lambda_r - T1} \leq C\frac{T1}{\Lambda_r}$$

□

A.1.9 Proof of Lemma 4

Note that the proof of this follows from Lemma 6 [41], but we provide it for completeness. Let F be a CDF and let $Q_{L,F}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ and $Q_{R,F}(p) = \inf\{x \in \mathbb{R} : F(x) > p\}$ be the left and right quantile functions. Let

$$R_F(t) \geq \max\{Q_{R,F}(\frac{1}{2} + t) - m, m - Q_{L,F}(\frac{1}{2} - t)\},$$

where m is the median. Then, given n-samples from the mixture model, let $\hat{m}(\{x_i\}_{i=1}^n)$ be the empirical median. Then, we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \leq R(\frac{\epsilon}{2(1-\epsilon)} + \sqrt{\frac{2\log(2/\delta)}{n}}).$$

To see this, for each sample x_i define an indicator variable $L_i \in \{0, 1\}$.

$$L_i = \mathbb{I}\left\{x_i \sim Q, or(x_i \sim P \text{ and } x_i \geq Q_{R,F}(\frac{1}{2(1-\epsilon)} + a))\right\},$$

for $a = \frac{\sqrt{\log(2/\delta)}}{(1-\epsilon)\sqrt{n}}$. Note that

$$\begin{aligned}\Pr(L_i = 1) &\leq \epsilon + (1-\epsilon)(1 - (a + \frac{1}{2(1-\epsilon)})) \\ &\equiv \frac{1}{2} - (1-\epsilon)a\end{aligned}$$

$$\hat{m} \geq Q_{R,F}(\frac{1}{2(1-\epsilon)} + a) \implies \sum_i L_i \geq n/2$$

Hence, we have that,

$$\Pr(\hat{m} > Q_{R,F}(\frac{1}{2(1-\epsilon)} + a)) \leq \Pr(\sum_i L_i \geq n/2) \leq \exp(-2n(1-\epsilon)^2 a^2) = \frac{\delta}{2}$$

The other side is also symmetric. Hence, we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \leq R(\frac{\epsilon}{2(1-\epsilon)} + a),$$

where $a = \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}}$. Note that under our assumption that $P \in \mathcal{P}_{\text{sym}}^{t_0,\kappa}$, we have that $R(t) \leq \kappa t$ for all $t \leq t_0$. Hence, as long as the contamination level ϵ, and confidence level δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}} \leq t_0,$$

we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(2/\delta)}{n}}$$

Appendix **B**

Supplementary Material for Chapter 3

B.1 Choice of Hyper-Parameters

In this section, we discuss how to tune the hyperparameters for our algorithms. In particular, note that the gradient estimators described in Algorithms 4, 5 depend on corruption level ϵ, and on confidence δ, which are not known in advance.

Since the standard hyper-parameter selection techniques such as cross validation, hold-out validation, pick hyper-parameters that minimize the empirical mean of the loss on validation data, they cannot be used in the presence of outliers in the data. One criteria we could use in such cases is to choose hyper-parameters that minimize a robust estimate of the population risk on validation data. However, we cannot use any of the existing robust mean estimators to estimate the population risk because they themselves depend on hyper-parameters such as corruption level ϵ.

Huber Contamination. We now consider the Huber contamination model and propose a heuristic based on Scheffe's tournament estimator [7, 130] for hyper-parameter selection. In particular, we consider the gradient descent procedure described in Algorithm 4 and explain our technique for choosing ϵ, δ using hold out cross validation. Note that our goal is to pick hyper-parameters that minimize the population risk $\mathcal{R}(\theta)$. Under the assumption of strong convexity of $\mathcal{R}(\theta)$, this is equivalent to picking hyper-parameters that minimize the parameter error $\|\theta - \theta^*\|_2$.

We begin with the problem of density estimation, where we are given n

i.i.d samples $\{z_i\}_{i=1}^n$ from $(1-\epsilon)P_{\theta^*} + \epsilon Q$, where P_{θ^*} belongs to the class of distributions $\{P_\theta : \theta \in \Theta\}$, and Q is an arbitrary distribution. Our goal is to estimate $\theta^* \in \Theta$ from the samples. Suppose $\left\{P_{\widehat{\theta}_1}, P_{\widehat{\theta}_2}, \dots, P_{\widehat{\theta}_m}\right\}$ are the candidate solutions returned by Algorithm 4 for different settings of ϵ, δ. Consider the following pairwise test function:

$$\phi_{jk} = \mathbb{I}\left\{\left|\frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} \mathbb{I}\left\{p_{\widehat{\theta}_j}(z_i') > p_{\widehat{\theta}_k}(z_i')\right\} - P_{\widehat{\theta}_j}(p_{\widehat{\theta}_j}(z) > p_{\widehat{\theta}_k}(z))\right| > \right.$$
$$\left.\left|\frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} \mathbb{I}\left\{p_{\widehat{\theta}_j}(z_i') > p_{\widehat{\theta}_k}(z_i')\right\} - P_{\widehat{\theta}_k}(p_{\widehat{\theta}_j}(z) > p_{\widehat{\theta}_k}(z))\right|\right\},$$

where $p_{\widehat{\theta}_j}$ is the probability density of $P_{\widehat{\theta}_j}$, $\{z_i'\}_{i=1}^{n_{\text{val}}}$ is the validation data and $\mathbb{I}\{.\}$ is the indicator function. When $\phi_{jk} = 1$, then $\widehat{\theta}_k$ is favored over $\widehat{\theta}_j$ and when $\phi_{jk} = 0$, then $\widehat{\theta}_j$ is favored over $\widehat{\theta}_k$. Then, the final estimate $P_{\widehat{\theta}_{j^*}}$ is given by

$$j^* = \operatorname*{argmin}_{j \in [m]} \sum_{\substack{k=1\\k\neq j}}^{m} \phi_{jk}$$

It can be shown, using standard techniques [130], that the above procedure picks a j^* such that $P_{\widehat{\theta}_{j^*}}$ is close to P_{θ^*} in TV metric. For distributions $\{P_\theta : \theta \in \Theta\}$ whose TV metric is roughly equivalent to the parameter error, the above procedure results in hyper-parameters which minimize the parameter error. This procedure can be extended to supervised learning problems such as regression and classification.

Heavy-Tailed Distribution. For the heavy-tailed setting we experimented with (a) empirical mean of the risk on validation data: $\frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}} \bar{\mathcal{L}}(\theta; z_i')$ where $\{z_i'\}_{i=1}^{n_{\text{val}}}$ is the validation data, which does not require any tuning parameters, as well as (b) median of means based mean of the risk on validation data, for various confidence levels δ. However, both the techniques in the context of hold-out validation resulted in models with similar performance. So, in our experiments with heavy-tailed distributions, we present results obtained using the empirical risk as in (a) on hold-out validation data.

Table B.1: Fitting to original image error.

	Best Possible	Proposed	TORRENT	OLS	SCRRR
Mean RMSE	0.05	0.09	0.175	0.21	0.13

B.2 Experiments with Semi-Synthetic Data: Robust Face Reconstruction

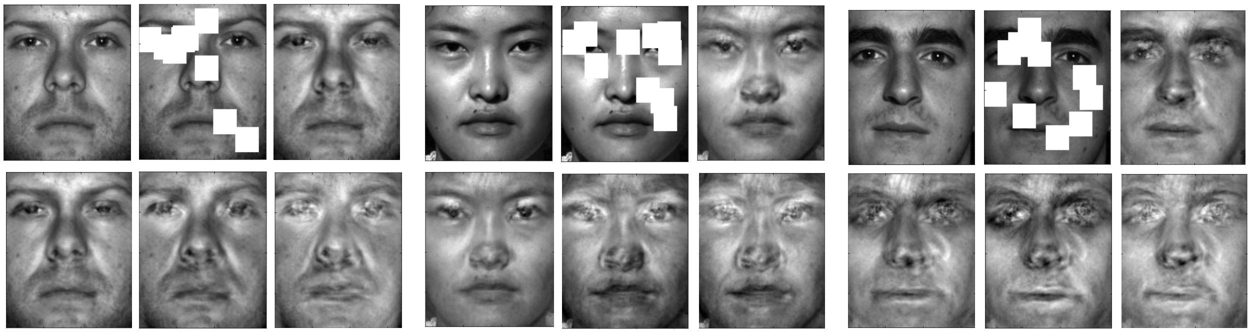
Setup. In this experiment, we show the efficacy of our algorithm by attempting to reconstruct face images that have been corrupted with heavy occlusion, where the occluding pixels play the role the outliers. We use the data from the Cropped Yale Dataset [64] . The dataset contains 38 subjects, and each image has 192×168 pixels. Following the methodology of Wang et al. [131], we choose 8 face images per subject, taken under mild illumination conditions and computed an eigenface set with 20 eigenfaces. Then given a new corrupted face image of a subject, the goal is to get the best reconstruction/approximation of the true face. To remove scaling effects, we normalized all images to $[0, 1]$ range. One image per person was used to test reconstruction. Occlusions were simulated by randomly placing 10 blocks of size 30×30. We repeated this 10 times for each test image. In this setting, each image of a subject corresponds to a different task; *i.e.* X is a common fixed eigenface basis, y is an observed(occluded) image, and the goal is to reconstruct(de-noise) the given image using the given basis. Note that in this example, we use a linear regression model as the uncontaminated statistical model, which is almost certainly not an exact match for the unknown ground truth distribution. Despite this model misspecification, as our results show, that robust mean based gradient algorithms do well.

Metric. We use Root Mean Square Error (RMSE) between the original and reconstructed image to evaluate the performance of the algorithms. We also compute the best possible reconstruction of the original face image by using the 20 eigenfaces.

Methods. Wang et al. [131] implemented popular robust estimators such as RANSAC, Huber Loss *etc.* and showed their poor performance. Wang et al. [131] then proposed an alternate robust regression algorithm called Self Scaled

Regularized Robust Regression (SCRRR). Hence, use TORRENT, SCRRR and OLS as baselines. We also compare against the best possible RMSE obtained by reconstructing the un-occluded image using the eigenfaces.

Results. Table B.1 shows that the mean RMSE is best for our proposed gradient descent based method and that the recovered images are in most cases closer to the un-occluded original image. (Figure B.2). Figure B.1(c) shows a case when none of the methods succeed in reconstruction.



(a) Successful Reconstruction (b) Successful Reconstruction (c) Failed Reconstruction

Figure B.1: Robust Face recovery results: Top; in order from L to R: original image, occluded image, best possible recovery with given basis. Bottom; in order from L to R: Reconstructions using our proposed algorithm, TORRENT and ordinary least squares (OLS).

B.3 Proof of Theorem 56

In this section, we present the proof of our main result on projected gradient descent with an inexact gradient estimator. To ease the notation we will often omit $\{D_n, \delta\}$ from $g(\theta; D_n, \delta)$. At any iteration step $t \in \{1, 2, \dots, T\}$, by assumption we have that with probability at least $1 - \frac{\delta}{T}$,

$$\|g(\theta^t; D_n, \delta/T) - \nabla\mathcal{R}(\theta^t)\|_2 \leq \alpha(n/T, \delta/T)\|\theta - \theta^*\|_2 + \beta(n/T, \delta/T). \tag{B.1}$$

Taking union bound, (B.1) holds over all iteration steps $t \in \{1 \dots T\}$, with probability at least $1 - \delta$. For the remainder of the analysis, we assume this event to be true.

Notation. Let $g(\theta^k) = \nabla\mathcal{R}(\theta^k) + e_k$ be the noisy gradient. Let $\alpha = \alpha(n/T, \delta/T)$ and $\beta = \beta(n/T, \delta/T)$ for brevity.

We have the following Lemma from Bubeck [62].

Lemma 21. *[Lemma 3.11 [62]] Let f be M-smooth and m-strongly convex, then for all $x, y \in \mathbb{R}^p$, we have:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{mM}{m+M}\|x-y\|_2^2 + \frac{1}{m+M}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

By assumptions we have that: $\|\nabla \mathcal{R}(\theta^k) - g(\theta^k)\|_2 = \|e_k\|_2 \leq \alpha\|\theta^k - \theta^*\|_2 + \beta$. Our update rule is $\theta^{k+1} = \mathbb{P}_\Theta\left[\theta^k - \eta g(\theta^k)\right]$. Then we have that:

$$\begin{aligned}
\|\theta^{k+1} - \theta^*\|_2^2 = \|\mathbb{P}_\Theta[\theta^k - \eta g(\theta^k)] - \theta^*\|_2^2 &= \|\mathbb{P}_\Theta[\theta^k - \eta g(\theta^k)] - \mathbb{P}_\Theta[\theta^* - \eta \nabla R(\theta^*)]\|_2^2 \\
&\leq \|\theta^k - \eta g(\theta^k) - (\theta^* - \eta \nabla R(\theta^*))\|_2^2 \qquad \text{(B.2)} \\
&= \|\theta^k - \theta^* - \eta(\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*)) - \eta e_k\|_2^2 \\
&\leq \|\theta^k - \theta^* - \eta(\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*))\|_2^2 + \eta^2\|e_k\|_2^2 \\
&\qquad + 2\|e_k\|_2\|\theta^k - \theta^* - \eta(\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*))\|_2, \qquad \text{(B.3)}
\end{aligned}$$

where (B.2) follows from contraction property of projections. Now, we can write $\|\theta^k - \theta^* - \eta(\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*))\|_2$ as

$$\begin{aligned}
&\|\theta^k - \theta^* - \eta(\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*))\|_2^2 = \|\theta^k - \theta^*\|_2^2 + \eta^2\|\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*)\|_2^2 - 2\eta\left\langle \nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*), \theta^k - \theta^* \right\rangle \\
&\leq \|\theta^k - \theta^*\|_2^2 + \eta^2\|\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*)\|_2^2 - 2\eta\left(\frac{\tau_\ell \tau_u}{\tau_\ell + \tau_u}\|\theta^k - \theta^*\|_2^2 + \frac{1}{\tau_\ell + \tau_u}\|\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*)\|_2^2\right) \\
&= \|\theta^k - \theta^*\|_2^2(1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)) + \eta\|\nabla \mathcal{R}(\theta^k) - \nabla R(\theta^*)\|_2^2(\eta - 2/(\tau_u + \tau_\ell)) \\
&= \|\theta^k - \theta^*\|_2^2(1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)), \qquad \text{(B.4)}
\end{aligned}$$

where the second step follows from Lemma 21 and the last step follows from the step size $\eta = 2/(\tau_\ell + \tau_u)$.

Now, combining Equations (B.3) and (B.4), and using our assumption that $\|e_k\|_2 \leq \alpha\|\theta^k - \theta^*\|_2 + \beta$, we get:

$$\|\theta^{k+1} - \theta^*\|_2^2 \leq \left(\|\theta^k - \theta^*\|_2\sqrt{(1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u))} + \eta\|e_k\|_2\right)^2$$

$$\|\theta^{k+1} - \theta^*\|_2 \leq \left[\sqrt{1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)} + \eta\alpha\right]\|\theta^k - \theta^*\|_2 + \eta\beta.$$

Let $\kappa = \sqrt{1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)} + \eta\alpha$. By the assumption on stability we have $\alpha < \tau_\ell$.

$$\begin{aligned}
\kappa &= \sqrt{1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)} + \eta\alpha \\
&< \sqrt{1 - 2\eta\tau_\ell\tau_u/(\tau_\ell + \tau_u)} + \eta\tau_\ell.
\end{aligned}$$

Since $\eta = 2/(\tau_\ell + \tau_u)$, we get that

$$\begin{aligned}
\kappa &< \sqrt{1 - 4\tau_u^2\tau_\ell^2/(\tau_\ell + \tau_u)^2} + 2\tau_\ell/(\tau_u + \tau_\ell) \\
\kappa &< \frac{\tau_u - \tau_\ell}{\tau_u + \tau_\ell} + 2\tau_\ell/(\tau_u + \tau_\ell) \\
\kappa &< 1
\end{aligned}$$

Therefore, we have that,

$$\|\theta^{k+1} - \theta^*\|_2 \leq \kappa\|\theta^k - \theta^*\|_2 + \eta\beta.$$

for some $\kappa < 1$. Solving the induction,we get:

$$\|\theta^k - \theta^*\|_2 \leq \kappa^k\|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\eta\beta.$$

B.4 Proof of Theorem 37

The proof of Theorem 37 follows from Theorem 11, where we study Generalized Linear Models, which includes linear regression as a special case. For the case of linear regression with Gaussian noise, it is relatively straightforward to see that the smoothness parameters satisfy $L_{\Phi,2k} = C_{2k}\|\Sigma\|_2^k$, $B_{\Phi,2k} = 0$, $M_{\Phi,t,k} = 1 \;\; \forall (t \geq 2, k \in \mathcal{N})$ and $M_{\Phi,t,k} = 0 \;\; \forall (t \geq 3, k \in \mathcal{N})$ under the assumption of bounded 8^{th} moments of the covariates. Substituting these values in Theorem 11 gives us the required result.

B.5 Proof of Theorem 11

To prove our result on Robust Generalized Linear Models, we first study the distribution of gradients of the corresponding risk function.

Lemma 22. *Consider the model in (5.10), then there exist universal constants $C_1, C_2 > 0$ such that*

$$\begin{aligned}
\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 \leq & C_1\|\Delta\|_2^2\|\Sigma\|_2\left(\sqrt{C_4}\sqrt{L_{\Phi,4}} + L_{\Phi,2}\right) \\
& + C_2\|\Sigma\|_2\left(B_{\Phi,2} + \sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3M_{\Phi,4,1}}\right),
\end{aligned}$$

and

$$\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^Tv\right]^4\right] \leq C_2(\mathrm{Var}[\nabla\bar{\mathcal{L}}(\theta)^Tv])^2.$$

Proof. The gradient $\nabla\bar{\mathcal{L}}(\theta)$ and it's expectation can be written as:

$$\begin{aligned}
\nabla\bar{\mathcal{L}}(\theta) &= -y.x + u(\langle x, \theta\rangle).x \\
\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] &= \mathbb{E}[x\left(u(x^T\theta) - u(x^T\theta^*)\right)],
\end{aligned}$$

where $u(t) = \Phi'(t)$.

$$\begin{aligned}
\|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2 &= \sup_{y \in \mathbb{S}^{p-1}} y^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \mathbb{E}[(y^T x) \left(u(x^T\theta) - u(x^T\theta^*) \right)] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}[(y^T x)^2]} \sqrt{\mathbb{E}[(u(x^T\theta) - u(x^T\theta^*))^2]} \\
&\leq C_1 \|\Sigma\|_2^{\frac{1}{2}} \sqrt{L_{\Phi,2} \|\Delta\|_2^2 + B_{\Phi,2}}
\end{aligned}$$

where the last line follows from our assumption of smoothness.

Now, to bound the maximum eigenvalue of the $\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))$,

$$\begin{aligned}
\|\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 &= \sup_{z \in \mathbb{S}^{p-1}} z^T \left(\mathbb{E}\left[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T \right] - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T \right) z \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} z^T \left(\mathbb{E}\left[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T \right] \right) z + \sup_{z \in \mathbb{S}^{p-1}} z^T \left(\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T \right) z \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} z^T \left(\mathbb{E}\left[x x^T \left(u(x^T\theta) - y \right)^2 \right] \right) z + \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2^2 \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \mathbb{E}\left[z^T \left(x x^T \left(u(x^T\theta) - y \right)^2 \right) z \right] + \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2^2 \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}\left[(z^T x)^4 \right]} \sqrt{\mathbb{E}\left[(u(x^T\theta) - y)^4 \right]} + \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2^2
\end{aligned}$$

To bound $\mathbb{E}\left[\left(u(x^T\theta) - y \right)^4 \right]$, we make use of the C_r inequality.

C_r **inequality**. If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^r < \infty$ where $r \geq 1$ then:

$$\mathbb{E}|X + Y|^r \leq 2^{r-1} \left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r \right)$$

Using the C_r inequality, we have that

$$\begin{aligned}
\mathbb{E}\left[\left(u(x^T\theta) - y \right)^4 \right] &\leq 8 \left(\mathbb{E}\left[\left(u(x^T\theta) - u(x^T\theta^*) \right)^4 \right] + \mathbb{E}\left[\left(u(x^T\theta^*) - y \right)^4 \right] \right) \\
&\leq C \left(L_{\Phi,4} \|\Delta\|_2^4 + B_{\Phi,4} + c(\sigma)^3 M_{\Phi,4,1} + 3c(\sigma)^2 M_{\Phi,2,2} \right)
\end{aligned}$$

where the last line follows from our assumption that $P_{\theta^*}(y|x)$ is in the exponential family, hence, the cumulants are higher order derivatives of the log-normalization function.

$$\begin{aligned}
&\|\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 \leq \sqrt{C} \sqrt{C_4} \|\Sigma\|_2 \left(\sqrt{L_{\Phi,4}} \|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3 M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right) + \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2^2 \\
&\leq \sqrt{C} \sqrt{C_4} \|\Sigma\|_2 \left(\sqrt{L_{\Phi,4}} \|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3 M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right) + C_1^2 \|\Sigma\|_2 \left(L_{\Phi,2} \|\Delta\|_2^2 + B_{\Phi,2} \right) \\
&\leq C \|\Delta\|_2^2 \|\Sigma\|_2 \left(\sqrt{C_4} \sqrt{L_{\Phi,4}} + L_{\Phi,2} \right) + C_6 \|\Sigma\|_2 \left(B_{\Phi,2} + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3 M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right)
\end{aligned}$$

Bounded Fourth Moment. To show that the fourth moment of the gradient distribution is bounded, we have

$$\begin{aligned}\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] &\leq \mathbb{E}\left[\left[|(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v|\right]^4\right] \\ &\leq 8\left[\underbrace{\mathbb{E}[|\nabla\bar{\mathcal{L}}(\theta))^T v|^4]}_{A} + \underbrace{\mathbb{E}[|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T v|^4]}_{B}\right].\end{aligned}$$

Control of A.

$$\begin{aligned}\mathbb{E}[|\nabla\bar{\mathcal{L}}(\theta))^T v|^4] &= \mathbb{E}[(x^T v)^4 (u(x^T\theta) - y)^4] \\ &\leq \sqrt{\mathbb{E}[(x^T v)^8]}\sqrt{\mathbb{E}[(u(x^T\theta) - y)^8]} \\ &\leq \sqrt{C_8}\|\Sigma\|_2^2 \sqrt{\mathbb{E}[(u(x^T\theta) - u(x^T\theta^*))^8] + \mathbb{E}[(u(x^T\theta^*) - y)^8]} \\ &\leq \sqrt{C_8}\|\Sigma\|_2^2 \sqrt{L_{\Phi,8}\|\Delta\|_2^8 + B_{\Phi,8} + \sum_{t,k=2}^{8} g_{t,k} M_{\Phi,t,k}} \\ &\leq \sqrt{C}\|\Sigma\|_2^2 \sqrt{L_{\Phi,8}}\|\Delta\|_2^4 + \sqrt{B_{\Phi,8}} + \sqrt{\sum_{t,k=2}^{8} g_{t,k} M_{\Phi,t,k}}\end{aligned}$$

where the last step follows from the fact that the 8th central moment can be written as a polynomial involving the lower cumulants, which in turn are the derivatives of the log-normalization function.

Control of B.

$$\mathbb{E}[|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T v|^4] \leq \|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\|_2^4 \leq C_1\|\Sigma\|_2^2 \left(L_{\Phi,2}^2\|\Delta\|_2^2 + B_{\Phi,2}^2\right)$$

By assumption $L_{\Phi,k}, B_{\Phi,k}, M_{\Phi,t,k}$ are all bounded for $k, t \leq 8$, which implies that there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] \leq c_1\|\Sigma\|_2^2\|\Delta\|_2^4 + c_2$$

Previously, we say that $\|\mathrm{Cov}\nabla\bar{\mathcal{L}}(\theta)\|_2 \leq c_3\|\Sigma\|_2\|\Delta\|_2^2 + c_4$, for some universal constants $c_3, c_4 > 0$, hence the gradient $\nabla\bar{\mathcal{L}}(\theta)$ has bounded fourth moments. □

Having studied the distribution of the gradients, we use Lemma 9 to characterize the stability of Huber Gradient estimator. Using Lemma 9, we know that at any point θ, the Huber Gradient Estimator $g(\theta, \delta/T)$ satisfies that with probability $1 - \delta/T$,

$$\|g(\theta, \delta/T) - \nabla\mathcal{R}(\theta)\|_2 \leq C_2\Big(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta})\Big)\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2^{\frac{1}{2}}\sqrt{\log p}.$$

Substituting the upper bound on $\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2$ from Lemma 58, we get that there are universal constants C_1, C_2 such that with probability at least 1 - δ/T

$$\|g(\theta) - \nabla\mathcal{R}(\theta)\|_2 \leq \underbrace{C_1\Big(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta})\Big)\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[L_{\Phi,4}^{\frac{1}{4}} + L_{\Phi,2}^{\frac{1}{2}}]}_{\alpha(\widetilde{n},\widetilde{\delta})}\|\Delta\|_2$$

$$+ \underbrace{C_2\Big(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta})\Big)\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[B_{\Phi,4}^{\frac{1}{4}} + B_{\Phi,2}^{\frac{1}{2}} + c(\sigma)^{\frac{1}{2}}M_{\Phi,2,2}^{\frac{1}{4}} + c(\sigma)^{\frac{3}{4}}M_{\Phi,4,1}^{\frac{1}{4}}]}_{\beta(\widetilde{n},\widetilde{\delta})} \tag{B.5}$$

To ensure stability of gradient descent, we need that $\alpha(\widetilde{n}, \widetilde{\delta}) < \tau_\ell$. Using (B.5), we obtain that gradient descent is stable as long as the number of samples n is large enough such that $\gamma(\widetilde{n}, p, \widetilde{\delta}) < \frac{C_1\tau_\ell}{\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[L_{\Phi,4}^{\frac{1}{4}} + L_{\Phi,2}^{\frac{1}{2}}]}$, and the contamination level is such that $\epsilon < \left(\frac{C_2\tau_\ell}{\sqrt{\log p}\|\Sigma\|_2^{\frac{1}{2}}[L_{\Phi,4}^{\frac{1}{4}} + L_{\Phi,2}^{\frac{1}{2}}]} - \gamma(\widetilde{n}, p, \widetilde{\delta})\right)^2$ for some constants C_1 and C_2. Plugging the corresponding ϵ and $\beta(\widetilde{n}, \widetilde{\delta})$ into Theorem 56, we get back the result of Theorem 11.

B.6 Proof of Corollary 10

We begin by studying the distribution of the random variable $xy = xx^T\theta^* + x.w$.

Lemma 23. *Consider the model in* (B.17)*, with $x \sim \mathcal{N}(0, \mathcal{I}_p)$ and $w \sim \mathcal{N}(0, 1)$ then there exist universal constants C_1, C_2 such that*

$$\begin{aligned} \mathbb{E}[xy] &= \theta^* \\ \|\mathrm{Cov}(xy)\|_2 &= 1 + 2\|\theta^*\|_2^2 \\ \textit{Bounded fourth moments} \quad \mathbb{E}\left[\left[(xy - \mathbb{E}[xy])^T v\right]^4\right] &\leq C_2(\mathrm{Var}[(xy)^T v])^2. \end{aligned}$$

Proof. **Mean.**

$$\begin{aligned} xy &= xx^T\theta^* + x.w \\ \mathbb{E}[xy] &= \mathbb{E}[xx^T\theta^* + x.w] \\ \mathbb{E}[xy] &= \theta^*. \end{aligned}$$

Covariance.

$$\begin{aligned} \mathrm{Cov}(xy) &= \mathbb{E}[(xx^T - I)\theta^* + x.w)((xx^T - I)\theta^* + x.w)^T)] \\ \mathrm{Cov}(xy) &= \mathbb{E}[(xx^T - I)\theta^*\theta^{*T}(xx^T - I)] + I_p. \end{aligned}$$

Now, $Z = (xx^T - I)\theta^*$ can be written as:

$$(xx^T - I)\theta^* = \begin{bmatrix} (x_1^2-1) & x_1x_2 & \dots & x_1x_p \\ x_1x_2 & (x_2^2-1) & \dots & x_2x_p \\ \vdots & \vdots & \vdots & \vdots \\ x_1x_p & x_2x_p & \dots & (x_p^2-1) \end{bmatrix} \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_p^* \end{bmatrix} = \begin{bmatrix} \theta_1^*(x_1^2-1) + x_1x_2\theta_2^* + \dots + x_1x_p\theta_p^* \\ x_1x_2\theta_1^* + (x_2^2-1)\theta_2^* + \dots + x_2x_p\theta_p^* \\ \vdots \\ x_1x_p\theta_1^* + x_2x_p\theta_2^* + \dots + (x_p^2-1)\theta_p^* \end{bmatrix}.$$

Then,

$$\mathbb{E}\left[ZZ^T\right] = \begin{bmatrix} 2\theta_1^{*2} + \theta_2^{*2} + \dots + \theta_p^{*2} & \theta_1^*\theta_2^* & \dots & \theta_1^*\theta_p^* \\ \theta_1^*\theta_2^* & \theta_1^{*2} + 2\theta_2^{*2} + \dots + \theta_p^{*2} & \dots & \theta_2^*\theta_p^* \\ \vdots & \vdots & \ddots & \vdots \\ \theta_p^*\theta_1^* & \theta_2^*\theta_p^* & \dots & \theta_1^{*2} + \theta_2^{*2} + \dots + 2\theta_p^{*2} \end{bmatrix}.$$

Hence the covariance matrix can be written as:

$$\mathrm{Cov}(xy) = I_p(1 + \|\theta^*\|_2^2) + \theta^*\theta^{*T}.$$

Therefore $\|\mathrm{Cov}(xy)\|_2 = 1 + 2\|\theta^*\|_2^2$.

Bounded Fourth Moment. We start from the LHS

$$\begin{aligned} \mathbb{E}\left[\left[(xy - \mathbb{E}[xy])^T v\right]^4\right] &\le \mathbb{E}\left[\left[\left|(xy - \mathbb{E}[xy])^T v\right|\right]^4\right] \\ &= \mathbb{E}\left[\left|((xx^T - I)\theta^* + wx)^T v\right|\right]^4 \\ &= \mathbb{E}\left[\left|(\theta^{*T}x)(x^Tv) - \theta^{*T}v + wv^Tx\right|\right]^4 \\ &\le 8\left[8\left[\underbrace{\mathbb{E}\left|(\theta^{*T}x)(x^Tv)\right|^4}_{A} + \underbrace{\mathbb{E}\left|\theta^{*T}v\right|^4}_{B}\right] + \mathbb{E}\underbrace{\left|w(x^Tv)\right|^4}_{C}.\right] \end{aligned}$$

The last line follows from two applications of the following inequality:

C_r **inequality**. If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^4 < \infty$ where $r \ge 1$ then:

$$\mathbb{E}|X + Y|^r \le 2^{r-1}\left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r\right).$$

Now to control each term on:

- **Control of** A. Using Cauchy Schwartz, and normality of 1D projections of normal distribution

$$\begin{aligned} A &\le \sqrt{\mathbb{E}[|\theta^{*T}x|^8]}\sqrt{\mathbb{E}[|x^Tv|^8]} \\ &\lesssim \|\theta^*\|_2^4. \end{aligned}$$

- **Control of** B, $B \leq \|\theta^*\|_2^4$.
- **Control of** C, $C = O(1)$, using independence of w and normality of 1D projections of normal distribution.

Therefore the $\mathbb{E}\left[\left[(xy - \mathbb{E}[xy])^T v\right]^4\right] \lesssim c + \|\theta^*\|_2^4$.

For the RHS:

$$\mathrm{Var}((xy)^T v)^2 = (v^T \mathrm{Cov}(xy) v)^2 \leq \|\mathrm{Cov}(xy)\|_2^2.$$

We saw that the $\|\mathrm{Cov}(xy)\|_2 \lesssim c + \|\theta^*\|_2^2$, so both the LHS and RHS scale with $\|\theta^*\|_2^4$. Hence, xy has bounded fourth moments. □

Now that we've established that xy has bounded fourth moments implies that we can use [28] as a mean estimation oracle. Using Theorem 1.3 [28], we know that the oracle of [28] outputs an estimate $\widehat{\theta}$ of $\mathbb{E}[xy]$ such that with probability at least $1 - 1/p^{C_1}$, we have:

$$\|\widehat{\theta} - \theta^*\|_2 \leq C_2 \sqrt{\|\mathrm{Cov}(xy)\|_2 \log p}\left(\epsilon^{\frac{1}{2}} + \gamma(n, p, \delta, \epsilon)\right)$$

Using Lemma 23 to subsitute $\|\mathrm{Cov}(xy)\|_2 \leq 1 + 2\|\theta^*\|_2^2$), we recover the statement of Corollary 10.

B.7 Proof of Theorem 14

To prove our result on Robust Exponential Family, we first study the distribution of gradients of the corresponding risk function.

Lemma 24. *Consider the model in (3.8), then there exists a universal constant C_1 such that*

$$\begin{aligned} \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] &= \nabla A(\theta) - \nabla A(\theta^*) \\ \|\mathrm{Cov}[\nabla \bar{\mathcal{L}}(\theta)]\|_2 &= \|\nabla^2 A(\theta^*)\|_2 \\ \textit{Bounded fourth moments} \quad \mathbb{E}\left[\left[(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v\right]^4\right] &\leq C_1 (\mathrm{Var}[\nabla \bar{\mathcal{L}}(\theta)^T v])^2. \end{aligned}$$

Proof. By Fisher Consistency of the negative log-likelihood, we know that

$$\begin{aligned} \mathbb{E}_{\theta^*}[\nabla \bar{\mathcal{L}}(\theta^*)] &= 0 \\ \implies \nabla A(\theta^*) - \mathbb{E}_{\theta^*}[\phi(z)] &= 0 \\ \implies \nabla A(\theta^*) &= \mathbb{E}_{\theta^*}[\phi(z)]. \end{aligned}$$

For the mean,

$$\begin{aligned} \nabla \bar{\mathcal{L}}(\theta) &= \nabla A(\theta) - \phi(z) \\ \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] &= \nabla A(\theta) - \mathbb{E}_{\theta^*}[\phi(z)] \\ \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] &= \nabla A(\theta) - \nabla A(\theta^*). \end{aligned}$$

Now, for the covariance:

$$\begin{aligned}\mathrm{Cov}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)] &= \mathbb{E}_{\theta^*}\left[\left(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)]\right)\left(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}_{\theta^*}[\nabla(\bar{\mathcal{L}}(\theta)]\right)^T\right] \\ &= \mathbb{E}_{\theta^*}\left[\left(\nabla A(\theta^*) - \phi(z)\right)\left(\nabla A(\theta^*) - \phi(z)\right)^T\right] \\ &= \mathrm{Cov}_{\theta^*}\left[\nabla\bar{\mathcal{L}}(\theta^*)\right] = \nabla^2 A(\theta^*).\end{aligned}$$

Bounded moments follows from our assumption that the sufficient statistics have bounded 4th moments. □

Having studied the distribution of the gradients, we use Lemma 9 to characterize the stability of Huber Gradient estimator. Using Lemma 9, we know that at any point θ, the Huber Gradient Estimator $g(\theta, \delta/T)$ satisfies that with probability $1 - \delta/T$,

$$\|g(\theta, \delta/T) - \nabla\mathcal{R}(\theta)\|_2 \le C_2\left(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta})\right)\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2^{\frac{1}{2}}\sqrt{\log p}.$$

Substituting the upper bound on $\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2$ from Lemma 24, we get that there are universal constants C_1, C_2 such that

$$\|g(\theta) - \nabla\mathcal{R}(\theta)\|_2 \le \underbrace{C_1\left(\epsilon^{\frac{1}{2}} + \gamma(\widetilde{n}, p, \widetilde{\delta})\right)\sqrt{\log p}\sqrt{\tau_u}}_{\beta(\widetilde{n}, \widetilde{\delta})}.$$

In this case we have that $\alpha(\widetilde{n}, \widetilde{\delta}) = 0 < \tau_\ell$ by assumption. Therefore we just have that $\epsilon < C_1$ for some universal constant C_1. Plugging the corresponding ϵ and $\beta(\widetilde{n}, \widetilde{\delta})$ into Theorem 56, we get back the result of Corollary 14.

B.8 Proof of Corollary 15

Using the contraction property of projections, we know that

$$\|\mathcal{P}_\Theta\left[(\nabla A)^{-1}\widehat{\mu}\right] - \theta^*\|_2 = \|\mathcal{P}_\Theta\left[(\nabla A)^{-1}\widehat{\mu}\right] - \mathcal{P}_\Theta\left[\theta^*\right]\|_2 \le \|(\nabla A)^{-1}\widehat{\mu} - \theta^*\|_2.$$

By Fisher Consistency of the negative log-likelihood, we know that

$$\nabla A(\theta^*) = \mathbb{E}_{\theta^*}[\phi(z)].$$

The true parameter θ^* can be obtained by inverting the ∇A operator whenever possible.

$$\begin{aligned}\|(\nabla A)^{-1}\widehat{\mu} - \theta^*\|_2 &= \|(\nabla A)^{-1}\widehat{\mu} - (\nabla A)^{-1}\mathbb{E}_{\theta^*}[\phi(z)]\|_2 \\ &= \|\nabla A^*\widehat{\mu} - \nabla A^*\mathbb{E}_{\theta^*}[\phi(z)]\|_2.\end{aligned}$$

where A^* is the convex conjugate of A. We can use the following result to control the Lipschitz smoothness A^*.

Theorem 30. *(Strong/Smooth Duality) Assume $f(\cdot)$ is closed and convex. Then $f(\cdot)$ is smooth with parameter M if and only if its convex conjugate $f(\cdot)$ is strongly convex with parameter $m = \frac{1}{M}$.*

A proof of the above theorem can be found in [132]. Hence, we have that:

$$\|\mathcal{P}_\Theta\left[(\nabla A)^{-1}\widehat{\mu}\right] - \theta^*\|_2 \leq \frac{1}{\tau_\ell}\|\widehat{\mu} - \mathbb{E}_{\theta^*}[\phi(z)]\|_2 \tag{B.6}$$

By assumption, we have that the fourth moments of the sufficient statistics are bounded. We also know that $\mathrm{Cov}(\phi(z) = \nabla^2 A(\theta^*)$ which implies that we can use [28] as our oracle. Using Lemma 9, we get that, there exists universal constants C_1, C_2 such that with probability at least $1 - 1/p^{C_1}$,

$$\|\widehat{\mu} - \mathbb{E}_{\theta^*}[\phi(z)]\|_2 \leq C_2\sqrt{\tau_u \log p}\Big(\epsilon^{\frac{1}{2}} + \gamma(n, p, \delta, \epsilon)\Big).$$

Combining the above with (B.6) recovers the result of Corollary 15.

B.9 Proof of Theorem 16

Before we present the proof of Theorem 16, we first study the distribution of gradients of the loss function. This will help us bound the error in the gradient estimator.

Lemma 25. *Consider the model in (B.17). Suppose the covariates $x \in \mathbb{R}^p$ have bounded 4^{th}-moments and the noise w has bounded 2^{th} moments. Then there exist universal constants C_1, C_2 such that*

$$\begin{aligned}\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] &= \Sigma\Delta \\ \|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 &\leq \sigma^2\|\Sigma\|_2 + \|\Delta\|_2^2 C_4\|\Sigma\|_2^2 \\ \mathit{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &\leq \sigma^2 \mathit{trace}\,(\Sigma) + C_4 \mathit{trace}\,(\Sigma)\,\|\Sigma\|_2\|\Delta\|_2^2,\end{aligned}$$

where $\Delta = \theta - \theta^*$ and $E[xx^T] = \Sigma$.

Proof. We start by deriving the results for $\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]$.

$$\begin{aligned}\bar{\mathcal{L}}(\theta) = \frac{1}{2}(y - x^T\theta)^2 &= \frac{1}{2}(x^T(\Delta) - w)^2 \\ \nabla\bar{\mathcal{L}}(\theta) &= xx^T\Delta - x.w \\ \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] &= \Sigma\Delta.\end{aligned}$$

Next, we bound the operator norm of the covariance of the gradients $\nabla\bar{\mathcal{L}}(\theta)$ at any point θ.

Covariance.

$$\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) = \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T$$

For any unit vector $z \in \mathcal{S}^{p-1}$, we have that,

$$\begin{aligned} z^T \mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta)) z &= z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z - (\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T z)^2 \\ &\leq z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \\ \implies \sup_{z \in \mathcal{S}^{p-1}} z^T \mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta)) z &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \end{aligned}$$

Hence, we have that

$$\begin{aligned} \lambda_{\max}(\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \\ &= \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[(xx^T \Delta - x.w)(xx^T \Delta - x.w)^T] z \\ &= \sup_{z \in \mathcal{S}^{p-1}} z^T (\mathbb{E}[xx^T \Delta \Delta^T xx^T] + \sigma^2 \mathbb{E}[xx^T]) z \\ &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T (\mathbb{E}[xx^T \Delta \Delta^T xx^T]) z + \sigma^2 \|\Sigma\|_2 \\ &\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y, z \in \mathcal{S}^{p-1}} \mathbb{E}[(z^T x)^2 (y^T z)^2] \\ &\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y, z \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}\left[(y^T x)^4\right]} \sqrt{\mathbb{E}\left[(z^T x)^4\right]} \\ &\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 C_4 \|\Sigma\|_2^2 \end{aligned}$$

where the second last step follows from Cauchy-Schwartz and the last step follows from our assumption of bounded 4^{th} moments (see (3.12)). Now to bound the trace of the covariance matrix,

$$
\begin{aligned}
\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) &= \mathbb{E}[(xx^T-\Sigma)\Delta - xw)(xx^T-\Sigma)\Delta - xw)^T] \\
\text{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &= \mathbb{E}[\|(xx^T-\Sigma)\Delta - xw)\|_2^2] \\
&= \underbrace{\mathbb{E}[\|(xx^T-\Sigma)\Delta\|_2^2]}_{T1} + \underbrace{\mathbb{E}[\|x\|_2^2 w^2]}_{\sigma^2\text{trace}(\Sigma)} \\
T1 &= \mathbb{E}[\|(xx^T-\Sigma)\Delta\|_2^2] = \Delta^T\mathbb{E}[(xx^T-\Sigma)^2]\Delta \\
&= \Delta^T\mathbb{E}[(x^Tx)xx^T + \Sigma^2 - \Sigma xx^T - xx^T\Sigma]\Delta \\
&= \Delta^T\mathbb{E}[(x^Tx)xx^T]\Delta - \Delta^T\Sigma^2\Delta \\
&\leq \Delta^T\mathbb{E}[(x^Tx)xx^T]\Delta \\
&\leq \|\Delta\|_2^2\mathbb{E}[(x^Tx)(x^Tu)^2], \quad \text{where } u = \frac{\Delta}{\|\Delta\|_2} \in \mathcal{S}^{p-1} \\
&\leq \|\Delta\|_2^2\mathbb{E}[(x^Tx)^2]^{\frac{1}{2}}\underbrace{\mathbb{E}[(x^Tu)^4]^{\frac{1}{2}}}_{\leq\sqrt{C_4}\|\Sigma\|_2} \\
x &\overset{\text{def}}{=} \sum_{i=1}^{p}\underbrace{(x^Tq_i)}_{\nu_i}q_i, \quad \text{where } \{q_i\}_{i=1}^p \text{ are eigenvectors of } \Sigma \\
\mathbb{E}[(x^Tx)(x^Tx)] &= \mathbb{E}[(\sum_i \nu_i^2)(\sum_i \nu_i^2)] \\
&= \mathbb{E}[\sum_i \nu_i^4 + 2\sum_{i<j}\nu_i^2\nu_j^2] \\
\mathbb{E}[\nu_i^4] &= \mathbb{E}[(x^Tq_i)^4] \leq C_4\mathbb{E}[(x^Tq_i)^2]^2 = C_4\lambda_i^2 \\
\mathbb{E}[\nu_i^2\nu_j^2] &\leq \sqrt{\mathbb{E}[\nu_i^4]}\sqrt{\mathbb{E}[\nu_j^4]} = C_4\lambda_i\lambda_j \\
\mathbb{E}[(x^Tx)(x^Tx)] &\leq C_4(\sum_i \lambda_i^2 + 2\sum_{i<j}\lambda_i\lambda_j) = C_4\text{trace}\left(\Sigma\right)^2 \\
\text{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &\leq \sigma^2\text{trace}\left(\Sigma\right) + C_4\text{trace}\left(\Sigma\right)\|\Sigma\|_2\|\Delta\|_2^2
\end{aligned}
$$

□

We now proceed to the proof of Theorem 16. From Lemma 11, we know that at any point θ, the gradient estimator described in Algorithm 5, $g(\theta; D_{\widetilde{n}}, \widetilde{\delta})$, satisfies the following with probability at least $1-\delta$,

$$\|g(\theta; D_{\widetilde{n}}, \widetilde{\delta}) - \nabla\mathcal{R}(\theta)\|_2 \;\;\leq\;\; C\sqrt{\tfrac{\mathrm{tr}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)))\log 1/\widetilde{\delta}}{\widetilde{n}}}.$$

We substitute the upper bound for $\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2$ from Lemma 57 in the above equation

$$\begin{aligned}
\|g(\theta; D_{\widetilde{n}}, \widetilde{\delta}) - \nabla\mathcal{R}(\theta)\|_2 &\leq C\sqrt{\tfrac{\mathrm{tr}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)))\log 1/\widetilde{\delta}}{\widetilde{n}}} \\
&\leq C\sqrt{\tfrac{(\sigma^2\mathrm{trace}(\Sigma)+C_4\mathrm{trace}(\Sigma)\|\Sigma\|_2\|\Delta\|_2^2)\log 1/\widetilde{\delta}}{\widetilde{n}}} \\
&\leq \underbrace{C_1\sqrt{\frac{\mathrm{trace}\,(\Sigma)\,\|\Sigma\|_2 \log 1/\widetilde{\delta}}{\widetilde{n}}}}_{\alpha(\widetilde{n},\widetilde{\delta})}\|\theta - \theta^*\|_2 \\
&\quad \underbrace{+C_2\sigma\sqrt{\frac{\mathrm{trace}\,(\Sigma)\log 1/\widetilde{\delta}}{\widetilde{n}}}}_{\beta(\widetilde{n},\widetilde{\delta})}.
\end{aligned}$$

To complete the proof of this theorem, we use the results from Theorem 56. Note that the gradient estimator satisfies the stability condition if $\alpha(\widetilde{n}, \widetilde{\delta}) < \tau_l$. This holds when

$$\widetilde{n} > \frac{\mathrm{trace}\,(\Sigma)\,\tau_u}{\tau_l^2}\log 1/\widetilde{\delta}.$$

Now suppose $\widetilde{n}$ satisfies the above condition, then plugging $\beta(\widetilde{n}, \widetilde{\delta})$ into Theorem 56 gives us the required result.

B.10 Proof of Theorem 17

To prove the Theorem we first derive a useful Lemma 26.

Lemma 26. *Consider the model in (5.10), then there exist universal constants $C_1, C_2 > 0$ such that*

$$\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 \leq \sqrt{C}\sqrt{C_4}\|\Sigma\|_2(\sqrt{L_{\Phi,4}}\|\Delta\|_2^2) + \sqrt{C}\sqrt{C_4}\|\Sigma\|_2\left(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}\right.$$

$$\mathit{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\right) \leq \sqrt{C_4}\,\mathit{trace}\,(\Sigma)\,\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{C_4}\,\mathit{trace}\,(\Sigma)\,(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}})$$

Proof. The gradient $\nabla\bar{\mathcal{L}}(\theta)$ and it's expectation can be written as:

$$\begin{gathered}
\nabla\bar{\mathcal{L}}(\theta) = -y.x + u(\langle x, \theta\rangle).x \\
\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] = \mathbb{E}[x\left(u(x^T\theta) - u(x^T\theta^*)\right)],
\end{gathered}$$

where $u(t) = \Phi'(t)$.

$$\begin{aligned}
\|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2 &= \sup_{y \in \mathbb{S}^{p-1}} y^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \mathbb{E}[(y^T x) \left(u(x^T \theta) - u(x^T \theta^*) \right)] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}[(y^T x)^2]} \sqrt{\mathbb{E}[(u(x^T \theta) - u(x^T \theta^*))^2]} \\
&\leq C_1 \|\Sigma\|_2^{\frac{1}{2}} \sqrt{L_{\Phi,2} \|\Delta\|_2^2 + B_{\Phi,2}}
\end{aligned}$$

where the last line follows from our assumption of smoothness.

Now, to bound the maximum eigenvalue of the $\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))$,

$$\begin{aligned}
\lambda_{\max}(\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \\
&= \sup_{z \in \mathbb{S}^{p-1}} z^T \left(\mathbb{E}\left[x x^T \left(u(x^T \theta) - y) \right)^2 \right] \right) z \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \mathbb{E}\left[z^T \left(x x^T \left(u(x^T \theta) - y \right)^2 \right) z \right] \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}\left[(z^T x)^4 \right]} \sqrt{\mathbb{E}\left[(u(x^T \theta) - y)^4 \right]}
\end{aligned}$$

To bound $\mathbb{E}\left[\left(u(x^T \theta) - y \right)^4 \right]$, we make use of the C_r inequality.

C_r **inequality**. If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^4 < \infty$ where $r \geq 1$ then:

$$\mathbb{E}|X + Y|^r \leq 2^{r-1} \left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r \right)$$

Using the C_r inequality, we have that

$$\begin{aligned}
\mathbb{E}\left[\left(u(x^T \theta) - y \right)^4 \right] &\leq 8 \left(\mathbb{E}\left[\left(u(x^T \theta) - u(x^T \theta^*) \right)^4 \right] + \mathbb{E}\left[\left(u(x^T \theta^*) - y \right)^4 \right] \right) \\
&\leq C \left(L_{\Phi,4} \|\Delta\|_2^4 + B_{\Phi,4} + c(\sigma)^3 M_{\Phi,4,1} + 3c(\sigma)^2 M_{\Phi,2,2} \right)
\end{aligned}$$

where the last line follows from our assumption that $P_{\theta^*}(y|x)$ is in the exponential family, hence, the cumulants are higher order derivatives of the log-normalization function.

$$\|\mathrm{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 \leq \sqrt{C} \sqrt{C_4} \|\Sigma\|_2 \left(\sqrt{L_{\Phi,4}} \|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3 M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right)$$

Now, to control the trace. We have that,

$$\begin{aligned}
\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) &= \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T \\
\text{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &= \text{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]\right) - \text{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T\right) \\
&\leq \text{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]\right) \\
&\leq \text{trace}\left(\mathbb{E}\left[xx^T\left(u(x^T\theta) - y)\right)^2\right]\right) \\
&= \mathbb{E}\left[\text{trace}\left(xx^T\left(u(x^T\theta) - y)\right)^2\right)\right] \\
&= \mathbb{E}[\text{trace}\left((xx^T)\right) u(x^T\theta) - y)^2] \quad \text{Because } (u(x^T\theta) - y)^2 \in \mathbb{R} \\
&\leq \sqrt{\mathbb{E}[\text{trace}\left((xx^T)\right)^2]}\sqrt{\mathbb{E}[(u(x^T\theta) - y)^4]} \\
&\leq \sqrt{C_4}\text{trace}\left(\Sigma\right)\left(\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}\right) \\
&= \sqrt{C_4}\text{trace}\left(\Sigma\right)\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{C_4}\text{trace}\left(\Sigma\right)(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}
\end{aligned}$$

□

From Lemma 11, we know that at any point θ, the gradient estimator described in Algorithm 5, $g(\theta; D_{\widetilde{n}}, \widetilde{\delta})$, satisfies the following with probability at least $1-\delta$,

$$\begin{aligned}
\|g(\theta; D_{\widetilde{n}}, \widetilde{\delta}) - \nabla\mathcal{R}(\theta)\|_2 &\leq C\sqrt{\tfrac{\mathrm{tr}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)))\log 1/\widetilde{\delta}}{\widetilde{n}}} \\
&\leq C\sqrt{\tfrac{(A\|\Delta\|_2^2 + B)\log 1/\widetilde{\delta}}{\widetilde{n}}} \\
&\leq \underbrace{C_1\sqrt{\frac{A\log 1/\widetilde{\delta}}{\widetilde{n}}}}_{\alpha(\widetilde{n},\widetilde{\delta})}\|\theta - \theta^*\|_2 \\
&\quad + \underbrace{C_2\sqrt{\frac{B\log 1/\widetilde{\delta}}{\widetilde{n}}}}_{\beta(\widetilde{n},\widetilde{\delta})}.
\end{aligned}$$

Substituting $A = \sqrt{C_4}\text{trace}\left(\Sigma\right)\sqrt{L_{\Phi,4}}$, $B = \sqrt{C_4}\text{trace}\left(\Sigma\right)(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}})$.

Note that the gradient estimator satisfies the stability condition if $\alpha(\widetilde{n}, \widetilde{\delta}) < \tau_l$. This holds when

$$\widetilde{n} > \frac{C_1^2 A \log 1/\widetilde{\delta}}{\tau_\ell^2} = \frac{C\text{trace}\left(\Sigma\right)\sqrt{C_4}\sqrt{L_{\Phi,4}}\log 1/\widetilde{\delta}}{\tau_\ell^2}.$$

Now suppose $\widetilde{n}$ satisfies the above condition, then plugging $\beta(\widetilde{n}, \widetilde{\delta})$ into Theorem 56 gives us the required result.

B.11 Proof of Theorem 19

The proof proceeds along similar lines as the proof of Theorem 17. To prove the Theorem we utilize the result of Lemma 24, where we showed that $\mathrm{Cov}[\nabla\bar{\mathcal{L}}(\theta)] = \nabla^2 A(\theta^*)$. Combining this result with Lemma 11 we get that with probability at least $1-\delta$

$$\begin{aligned}\|g(\theta; D_{\widetilde{n}}, \widetilde{\delta}) - \nabla\mathcal{R}(\theta)\|_2 &\leq C\sqrt{\tfrac{\mathrm{tr}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)))\log 1/\widetilde{\delta}}{\widetilde{n}}} \\ &\leq \underbrace{C\sqrt{\frac{\mathrm{trace}\left(\nabla^2 A(\theta^*)\right)\log 1/\widetilde{\delta}}{\widetilde{n}}}}_{\beta(\widetilde{n},\widetilde{\delta})}.\end{aligned}$$

Since $\alpha(\widetilde{n}, \widetilde{\delta}) = 0$, the stability condition is always satisfied, as long as $\tau_l > 0$. Substituting $\beta(\widetilde{n}, \widetilde{\delta})$ into Theorem 56 gives us the required result.

B.12 Upper bound on Contamination Level

We provide a complementary result, which gives an upper bound for the contamination level ϵ based on the initialization point θ^0, above which, Algorithm 1 would not work. The key idea is that the error incurred by any mean estimation oracle is lower bounded by the variance of the distribution, and that if the zero vector lies within that error ball, then any mean oracle can be forced to output $\mathbf{0}$ as the mean. For Algorithm 1, this implies that, in estimating the mean of the gradient, if the error is high, then one can force the mean to be $\mathbf{0}$ which forces the algorithm to converge. For the remainder of the section we consider the case of linear regression with $x \sim \mathcal{N}(0, \mathcal{I}_p)$ in the asymptotic regime of $n \to \infty$.

Lemma 27. *Consider the model in (B.17) with $x \sim \mathcal{N}(0, \mathcal{I}_p)$ and $w \sim \mathcal{N}(0, 1)$, then there exists a universal constant C_1 such that if $\epsilon > C_1 \frac{\|\theta^0 - \theta^*\|_2}{\sqrt{1 + 2\|\theta^0 - \theta^*\|_2^2}}$, then for every gradient oracle, there exists a contamination distribution Q such that, Algorithm 1 will converge to θ^0 even when the number of samples $n \to \infty$.*

Proof. Using Lemma 23, we know that for any point θ,

$$\begin{aligned}\nabla\bar{\mathcal{L}}(\theta) &= xx^T\Delta - x.w \\ \mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)] &= (\theta - \theta^*) = \Delta \\ \|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 &= 1 + 2\|\Delta\|_2^2,\end{aligned}$$

where $\Delta = \theta - \theta^*$.
Let $P_{\nabla\bar{\mathcal{L}}(\theta)}$ represent the distribution $\nabla\bar{\mathcal{L}}\theta$. Similarly, let $P_{\epsilon,\nabla\bar{\mathcal{L}}(\theta),Q}$ represent the corresponding ϵ-contaminated distribution. Then, using Theorem 2.1 [11], we know that the minimax rate for estimating the mean of the distribution of gradients is given by:

$$\inf_{\widehat{\mu}} \sup_{\theta\in\mathbb{R}^p, Q} P_{\epsilon,\nabla\bar{\mathcal{L}}(\theta),Q}\left\{\|\widehat{\mu} - \mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)]\|_2^2 \geq C\epsilon^2(1 + 2\|\Delta\|_2^2)\right\} \geq c.$$

The above statement says that at any point θ, any mean oracle Ψ will always incur an error of $\Omega(\sqrt{C\epsilon^2(1+2\|\Delta\|_2^2)})$ in estimating the gradient $\mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)]$.

$$\|\Psi(\theta) - \mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)]\|_2 \geq C\epsilon\sqrt{(1+2\|\Delta\|_2^2)} \quad \forall \quad \Psi$$

For any oracle Ψ, there exists some adversarial contamination Q, such that whenever $\|\mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta)]\|_2 < C\epsilon\sqrt{(1+2\|\Delta\|_2^2)}$, then $\|\Psi(\theta)\|_2 = 0$.

Suppose that the contamination level ϵ is such that,

$$\epsilon > \frac{1}{C}\frac{\|\mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta^0)]\|_2}{\sqrt{(1+2\|\theta^0-\theta^*\|_2^2)}},$$

then for every oracle there exists a corresponding Q such that Algorithm 1 will remain stuck at θ^0.

Plugging $\mathbb{E}_{\theta^*}[\nabla\bar{\mathcal{L}}(\theta^0)] = \theta^0 - \theta^*$, we recover the statement of the lemma. □

Chen et al. [11] provide a general minimax lower bound of $\Omega(\epsilon)$ for ϵ-contamination models in this setting. In contrast, using Algorithm 1 with [28] as oracle, we can only $O(\sqrt{\epsilon\log p})$ close to the true parameter even when the contamination is small, which implies that our procedure is not minimax optimal. Our approach is nonetheless the only practical algorithm for robust estimation of general statistical models.

B.13 Proof of Lemma 9

In this section we present a refined, non-asymptotic analysis of the robust mean estimator of Lai et al. [28], described in Algorithm 4. We begin by introducing some preliminaries. We subsequently analyze the algorithm in 1-dimension and finally turn our attention to the general algorithm.

B.13.1 Preliminaries

Unless otherwise stated, we assume throughout that the random variable X has bounded fourth moments, i.e. for every unit vector v,

$$\mathbb{E}\left[\langle X-\mu, v\rangle^4\right] \leq C_4\left[\mathbb{E}\left[\langle X-\mu, v\rangle^2\right]\right]^2.$$

We summarize some useful results from [28], which bound the deviation of the conditional mean/covariance from the true mean/covariance.

Lemma 28. *[Lemma 3.11 [28]] Let X be a univariate random variable with bounded fourth moments, and let A be any with event with probability $\mathbb{P}(A) = 1-\gamma \geq \frac{1}{2}$. Then,*

$$|\mathbb{E}(X|A) - \mathbb{E}(X)| \leq \sigma\sqrt[4]{8C_4\gamma^3}.$$

Lemma 29. *[Lemma 3.12 [28]] Let X be a univariate random variable with $\mathbb{E}[X] = \mu$, $\mathbb{E}\left((X-\mu)^2\right) = \sigma^2$ and let $\mathbb{E}((X-\mu)^4) \leq C_4\sigma^4$. Let A be any with event with probability $\mathbb{P}(A) = 1-\gamma \geq \frac{1}{2}$. Then,*

$$(1-\sqrt{C_4\gamma})\sigma^2 \leq \mathbb{E}((X-\mu)^2|A) \leq (1+2\gamma)\sigma^2.$$

Corollary 31. *[Corollary 3.13 [28]] Let A be any event with probability $\mathbb{P}(A) = 1-\gamma \geq \frac{1}{2}$, and let X be a random variable with bounded fourth moments. We denote $\Sigma|_A = \mathbb{E}(XX^T|A) - (\mathbb{E}(X|A))(\mathbb{E}(X|A))^T$ to be the conditional covariance matrix. We have that,*

$$(1-\sqrt{C_4\gamma}-\sqrt{8C_4\gamma^3})\Sigma \preceq \Sigma|_A \preceq (1+2\gamma)\Sigma.$$

For random variables with bounded fourth moments we can use Chebyshev's inequality to obtain tail bounds.

Lemma 30. *[Lemma 3.14 [28]] Let X have bounded fourth moments, then for every unit vector v we have that,*

$$\mathbb{P}(|\langle X, v\rangle - \mathbb{E}[\langle X, v\rangle]| \geq t\sqrt{\left[\mathbb{E}\left[\langle X-\mu, v\rangle^2\right]\right]}) \leq \frac{C_4}{t^4}.$$

Our proofs also use the matrix Bernstein inequality for rectangular matrices. As a preliminary, we consider a finite sequence $\{Z_k\}$ of independent, random matrices of size $d_1 \times d_2$. We assume that each random matrix satisfies $\mathbb{E}(Z_k) = 0$, and $\|Z_k\|_{\text{op}} \leq R$ almost surely. We define:

$$\sigma^2 := \max\left\{\|\sum_k \mathbb{E}(Z_k Z_k^T)\|_{\text{op}}, \|\sum_k \mathbb{E}(Z_k Z_k^T)\|_{\text{op}}\right\}.$$

With these preliminaries in place we use the following result from [133].

Lemma 31. *For all $t \geq 0$,*

$$\mathbb{P}\Big(\Big\|\sum_k Z_k\Big\|_{op} \geq t\Big) \leq (d_1+d_2)\exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Equivalently, with probability at least $1-\delta$,

$$\Big\|\sum_k Z_k\Big\|_{op} \leq \sqrt{2\sigma^2 \log\left(\frac{d_1+d_2}{\delta}\right)} + \frac{2R}{3}\log\left(\frac{d_1+d_2}{\delta}\right).$$

We let $\mathcal{I}$ denote the set of all intervals in $\mathbb{R}$. The following is a standard uniform convergence result.

Lemma 32. *Suppose $X_1, \ldots, X_n \sim \mathbb{P}$, then with probability at least $1-\delta$,*

$$\sup_{I\in\mathcal{I}} \left|\mathbb{P}(I) - \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \in I)\right| \leq 2\sqrt{\frac{4\log(en) + 2\log(2/\delta)}{n}}.$$

Algorithm 9 Huber Outlier Gradients Truncation

```
function HuberOutlierGradientTruncation(Sample Gradients $S$, Corruption Level $\epsilon$, Dimension $p$,$\delta$)
    if p=1 then
        Let $[a,b]$ be smallest interval containing $\left(1-\epsilon-C_5\left(\sqrt{\frac{1}{|S|}\log\left(\frac{|S|}{\delta}\right)}\right)\right)(1-\epsilon)$ fraction of points.
        $\widetilde{S} \leftarrow S \cap [a,b]$.
        return $\widetilde{S}$
    else
        Let $[S]_i$ be the samples with the $i^{th}$ co-ordinates only, $[S]_i = \{\langle x, e_i\rangle \,|x \in S\}$
        for $i = 1$ to $p$ do
            $a[i] = $ HuberGradientEstimator($[S]_i, \epsilon, 1, \delta/p$).
        end for
        Let $B(r,a)$ be the ball of smallest radius centered at $a$ containing $\left(1-\epsilon-C_p\left(\sqrt{\frac{p}{|S|}\log\left(\frac{|S|}{p\delta}\right)}\right)(1-\epsilon)\right.$ fraction of points in $S$.
        $\widetilde{S} \leftarrow S \cap B(r,a)$.
        return $\widetilde{S}$
    end if
end function
```

We now turn our attention to an analysis of Algorithm 4 for the 1-dimensional case.

B.13.2 The case when $p = 1$

Firstly, we analyze Algorithm 4 when $p = 1$.

Lemma 33. *Suppose that, P_θ^* is a distribution on $\mathbb{R}^1$ with mean μ, variance σ^2, and bounded fourth moments. There exist positive universal constants $C_1, C_2, C_8 > 0$, such that given n samples from the distribution in (3.9), the algorithm with probability at least $1-\delta$, returns an estimate $\widehat{\mu}$ such that,*

$$\|\widehat{\mu}-\mu\|_2 \leq C_1 C_4^{\frac{1}{4}} \sigma \left(\epsilon + \sqrt{\frac{\log 3/\delta}{2n}} + t\right)^{\frac{3}{4}} + C_2\sigma\left(\epsilon + \sqrt{\frac{\log 3/\delta}{2n}} + t\right)^{\frac{1}{2}} \sqrt{\frac{\log(3/\delta)}{n}}$$

where $t = C_8\sqrt{\frac{1}{n}\log\left(\frac{n}{\delta}\right)}$. which can be further simplified to,

$$\|\widehat{\mu}-\mu\|_2 \leq C_1 C_4^{\frac{1}{4}} \sigma \left(\epsilon + C_8\sqrt{\frac{1}{n}\log\left(\frac{n}{\delta}\right)}\right)^{\frac{3}{4}} + C_2\sigma\left(\epsilon + C_8\sqrt{\frac{1}{n}\log\left(\frac{n}{\delta}\right)}\right)^{\frac{1}{2}} \sqrt{\frac{\log(1/\delta)}{n}}$$

Proof. By an application of Hoeffding's inequality we obtain that with probability at least $1-\delta/3$, the fraction of corrupted samples (i.e. samples from the distribution Q) is less than

$\epsilon + \sqrt{\frac{\log(3/\delta)}{2n}}$. We condition on this event through the remainder of this proof. We let η denote the fraction of corrupted samples. Further, we let S_P be the samples from the true distribution. Let n_P be the cardinality of this set, i.e. $n_P := |S_P|$.

Let $I_{1-\eta}$ be the interval around μ containing $1-\eta$ mass of P_θ^*. Then, using Lemma 30, we have that:

$$\text{length} I_{1-\eta} \leq \frac{C_4^{\frac{1}{4}} \sigma}{\eta^{\frac{1}{4}}}.$$

Using Lemma 32 we obtain that with probability at least $1 - \delta/3$ the number of samples from the distribution P that fall in the interval $I_{1-\eta}$ is at least $1 - \eta - t$ where t is upper bounded as:

$$t \leq 2\sqrt{\frac{4\log(en) + 2\log(6/\delta)}{n}}.$$

Now we let $\widetilde{S}$ be the set of points in the smallest interval containing $(1-\eta-t)(1-\eta)$ fraction of all the points.

- Using VC theory, we know that for every interval $I \subset \mathbb{R}$, there exists some universal constant C_3 such that

$$\mathbb{P}\left(|(P(x \in I | x \sim D) - P(x \in I | x \in_u S_D))| > t/2\right) \leq n_D^2 \exp(-n_D t^2/8) \tag{B.7}$$

  This can be re-written as, that with probability at least $(1 - \delta/3)$, there exists a universal constant C_0 such that,

$$\sup_I |(P(x \in I | x \sim D) - P(x \in I | x \in_u S_D))| \leq C_0 \sqrt{\frac{1}{n_D} \log\left(\frac{n_D}{\delta}\right)} \leq \underbrace{C_5 \sqrt{\frac{1}{n} \log\left(\frac{n}{\delta}\right)}}_{t}$$

- Using (B.7), we know that $(1 - \eta - t)$ fraction of S_D lie in $\mathcal{I}_{1-\eta}$.
  Let $\widetilde{S}$ be the set of points in the smallest interval containing $(1-\eta-t)(1-\eta)$ fraction of the points.
- We know that the length of minimum interval containing $(1-\eta-t)(1-\eta)$ fraction of the points of S is less than length of smallest interval containing $(1-\eta-t)$ fraction of points of S_D, which in turn is less than length of $I_{1-\eta}$.
- Now, $I_{1-\eta}$ and minimum interval containing $(1-\eta-t)$ fraction of points of S_D need to overlap. This is because, n is large enough such that $t < \frac{1}{2} - \eta$ hence, the extreme points for such an interval can be atmost $2\text{length} I_{1-\eta}$ away.
- Hence, the distance of all chosen noise-points from μ will be within the $\text{length} I_{1-\eta}$.
- Moreover, the interval of minimum length with $(1-\eta-t)(1-\eta)$ fraction of S will contain at least $1 - 3\eta - t$ fraction of S_D.
- Hence, we can bound the error of $mean(\widetilde{S})$ by controlling the sources of error.

  - All chosen noise points are within length$I_{1-\eta}$, and there are atmost η of them, hence the maximum error can be ηlength$I_{1-\eta}$.
  - Next, the mean of chosen good points will converge to the mean of the conditional distribution. *i.e.* points sampled from D but conditioned to lie in the minimum length interval. The variance of these random variables is upper bounded using Lemma 29.
  - To control the distance between the mean($E(X)$ and the conditional mean($E(X|A)$), where A is the event that a sample x is in the chosen interval. We know that $P(A) \geq 1 - 3\eta - t$, hence, using Lemma 3.11[28], we get that there exists a constant C_{13} such that,

$$|E[X] - E[X|A]| \leq C_{13} C_4^{\frac{1}{4}} \sigma (\eta + t)^{\frac{3}{4}}$$

- Hence, with probability at least $1 - \delta/3$, the mean of $\widetilde{S}$ will be within

$$\eta \times \text{length} I_{1-\eta} + C_{13} C_4^{\frac{1}{4}} \sigma (\eta + t)^{\frac{3}{4}} + C_6 \sigma (1 + 2\eta)^{\frac{1}{2}} \sqrt{\frac{\log(3/\delta)}{n}}$$

- Taking union-bound over all conditioning statements, and upper bounding, η with $\epsilon + \sqrt{\frac{\log(3/\delta)}{2n}}$, we recover the statement of the lemma.

□

B.13.3 The case when $p > 1$

To prove the case for $p > 1$, we use a series of lemmas. Lemma 34 proves that the outlier filtering constrains the points in a ball around the true mean. Lemma 36 controls the error in the mean and covariance the true distribution after outlier filtering ($\widetilde{D}$). Lemma 37 controls the error for the mean of $\widetilde{S}$ when projected onto the bottom span of the covariance matrix $\Sigma_{\widetilde{S}}$.

Lemma 34. *Suppose that, P_θ^* is a distribution on $\mathbb{R}^p$ with mean μ, covariance Σ, and bounded fourth moments. There exist positive universal constants $C_1, C_2, C_8 > 0$, such that given n samples from the distribution in (3.9), we can find a vector $a \in \mathbb{R}^p$ such that with probability at least $1 - \delta$,*

$$\begin{aligned}\|a - \mu\|_2 \leq & C_1 C_4^{\frac{1}{4}} \sqrt{\mathit{trace}\,(\Sigma)} \left(\epsilon + C_8 \sqrt{\frac{1}{n} \log\left(\frac{np}{\delta}\right)} \right)^{\frac{3}{4}} \\ & + C_2 \left(\epsilon + C_8 \sqrt{\frac{1}{n} \log\left(\frac{np}{\delta}\right)} \right)^{\frac{1}{2}} \sqrt{\mathit{trace}\,(\Sigma)} \sqrt{\frac{\log(p/\delta)}{n}}\end{aligned}$$

Proof. Pick n orthogonal directions $v_1, v_2, \ldots, v_n$, and use method for one-dimensions, and using union bound, we can recover the result. □

Next, we prove the case when $p > 1$. Firstly, we prove that after the outlier step,

Lemma 35. *After the outlier removal step, there exists universal constants $C_{11} > 0$ such that with probability at least $1 - \delta$, every remaining point x satisfies,*

$$\|x - \mu\|_2 \leq r_1^* + 2r_2^*$$

where $r_1^* = C_{10} \frac{C_4^{\frac{1}{4}} \sqrt{p\|\Sigma\|_2}}{\eta^{\frac{1}{4}}}$ and $r_2^* = C_1 C_4^{\frac{1}{4}} \sqrt{p\|\Sigma\|_2}(\eta + t)^{\frac{3}{4}} + C_2 \sqrt{p\|\Sigma\|_2}(\eta + t)^{\frac{1}{2}} \sqrt{\frac{\log(1/\delta)}{n}}$ and $t = C_8 \sqrt{\frac{1}{n} \log\left(\frac{np}{\delta}\right)}$. Here $\eta \leq \epsilon + \sqrt{\frac{\log(1/\delta)}{2n}}$ is the fraction of samples corrupted.

Proof.

- Let $\widetilde{S}$ be the set of points chosen after the outlier filtering. Let $\widetilde{S}_D$ be set of good points chosen after the outlier filtering. Let $\widetilde{S}_N$ be the set of bad points chosen after the outlier filtering.
- Using VC theory we know that for every closed ball $\mathcal{B}(\mu, r) = \{x | \|x - \mu\|_2 \leq r\}$, there exists a constant C_9 such that with probability at least $1 - \delta$

$$\sup_B |P(x \in B | x \sim D) - P(x \in B | x \in_u S_D)| \leq \underbrace{C_9 \sqrt{\frac{p}{n} \log\left(\frac{n}{p\delta}\right)}}_{t_2}$$

- Let $B^* = B(\mu, r^*)$ for $r_1^* = C_{10} \frac{C_4^{\frac{1}{4}}}{(\eta)^{\frac{1}{4}}} \sqrt{p\|\Sigma\|_2}$. Then, we claim that

$$P(x \in B^* | x \sim D) \geq 1 - \eta$$

  - To see this, suppose we have some $x \in D$. Let $z = x - \mu$. Let $z_i = z^T v_i$ for some orthogonal directions $v_1, v_2, \ldots, v_p$. Let $Z^2 = \sum z_i^2 = \|z\|_2^2$.
  -

$$P\left(Z^2 \geq \frac{C_4^{\frac{1}{2}} p \|\Sigma\|_2}{(\eta)^{\frac{1}{2}}}\right) = P\left(Z^4 \geq \frac{C_4 p^2 \|\Sigma\|_2^2}{(\eta)}\right) \leq \frac{(\eta) E(Z^4)}{C_4 p^2 \|\Sigma\|_2^2}$$

  - Now, $E(Z^4) \leq p^2 \max_i E(z_i^4) \leq C_4 p^2 \|\Sigma\|_2^2$. Plugging this in the above, we have that $P(x \in B^* | x \sim D) \geq 1 - \eta$.
- Hence, we have that $P(x \in B^* | x \in_u S_D) \geq 1 - \eta - t_2$.
- Using Lemma 34, we have that at least $(1 - \eta - t_2)$ fraction of good points are $r_1^* + r_2^*$ away from a. Hence, we have that the minimum radius of the ball containing all the $(1 - \eta - t_2)(1 - \eta)$ has a radius of atmost $r_1^* + r_2^*$, which when combined with the triangle inequality recovers the statement of lemma.

□

As before, let $\widetilde{S}$ be the set of points after outlier filtering. Let $\mu_{\widetilde{S}} = mean(\widetilde{S})$, $\mu_{\widetilde{S}_D} = mean(\widetilde{S}_D)$, $\mu_{\widetilde{S}_N} = mean(\widetilde{S}_N)$.

Lemma 36. *Let $\widetilde{S}_D$ be the set of clean points remaining after the outlier filtering. Then, with probability at least $1-\delta$, we have that*

$$\|\mu_{\widetilde{S}_D}-\mu\|_2 \le C_1 C_4^{\frac{1}{4}}(\eta+t_2)^{\frac{3}{4}}\sqrt{\|\Sigma\|_2}\left(1+\frac{\log(p/\delta)}{n}\right)+\sqrt{\|\Sigma\|_2}\sqrt{(1+2(\eta+t_2)}\sqrt{\frac{1}{n}\log(p/\delta)} \\ + C_{15}\frac{(r_1^*+2r_2^*)}{n}\log(p/\delta)$$

and

$$\|\Sigma_{\widetilde{S}_D}\|_2 \le \beta(n,\delta)\|\Sigma\|_2,$$

where

$$\beta(n,\delta)=\left(1+2C(\eta+t_2)+\left(1+\frac{p\sqrt{C_4}}{\sqrt{\eta}}+\sqrt{C_4}p(\eta+t)^{\frac{3}{2}}+(\eta+t_2)^{\frac{3}{2}}\right)\left(\sqrt{\frac{\log(p/\delta}{n}}+\frac{\log(p/\delta)}{n}\right)\right)$$

Proof. We first prove the bounds on the mean shift.

$$\|\mu_{\widetilde{S}_D}-\mu\|_2 \le \underbrace{\|\mu_{\widetilde{S}_D}-\mu_{\widetilde{D}}\|_2}_{A}+\underbrace{\|\mu_{\widetilde{D}}-\mu\|_2}_{B}$$

- **Control of B.** We use Lemma 28 on $X=x^T\frac{\mu_{\widetilde{D}}-\mu}{\|\mu_{\widetilde{D}}-\mu\|_2}$ for $x\sim D$, and A be the event that x is not removed by the outlier filtering.

$$\|\mu_{\widetilde{D}}-\mu\|_2 \le C_1 C_4^{\frac{1}{4}}(\eta+t_2)^{\frac{3}{4}}\sqrt{\|\Sigma\|_2}$$

- **Control of A**. Using Lemma 29,we have that $\|\Sigma_{\widetilde{D}}\|_2 \le (1+2(\eta+t_2))\|\Sigma\|_2$. Now, we use Bernstein's inequality . Lemma 31 with $R=C(r_1^*+2r_2^*+B)$, we get that, with probability at least $1-\delta$,

$$\|\mu_{\widetilde{S}_D}-\mu_{\widetilde{D}}\|_2 \le C_{14}\sqrt{\|\Sigma\|_2}\sqrt{(1+2(\eta+t_2)}\sqrt{\frac{1}{n}\log(p/\delta)}+C_{15}\frac{(r_1^*+2r_2^*+B)}{n}\log(p/\delta)$$

Next, we prove the bound for covariance matrix.

$$\|\Sigma_{\widetilde{S}_D}\|_2 \le \|\Sigma_{\widetilde{S}_D}-\Sigma_{\widetilde{D}}\|_2+\underbrace{\|\Sigma_{\widetilde{D}}-\Sigma\|_2}_{\le 2C(\eta+t_2)\|\Sigma\|_2(\text{By Corollary 31}))}+\|\Sigma\|_2$$

To control $\|\Sigma_{\widetilde{S}_D}-\Sigma_{\widetilde{D}}\|_2$, we use Bernstein's inequality, with $Z_k=\frac{(x_k-\mu_{\widetilde{D}})(x_k-\mu_{\widetilde{D}})^T-\Sigma_{\widetilde{D}}}{n}$. From, Lemma 35, we know that the points are constrained in a ball. Plugging this into Lemma 31,

$$\|\Sigma_{\widetilde{S}_D}-\Sigma_{\widetilde{D}}\|_2 \le C(\|\Sigma\|_2+R^2)\left(\sqrt{\frac{\log(p/\delta}{n}}+\frac{\log(p/\delta)}{n}\right)$$

where $R^2=C\left({r_1^*}^2+{r_2^*}^2+B^2\right)$. □

Plugging in the values, we get that,

$$\|\Sigma_{\widetilde{S}_D} - \Sigma_{\widetilde{D}}\|_2 \leq C\|\Sigma\|_2 \left(1 + \frac{p\sqrt{C_4}}{\sqrt{\eta}} + \sqrt{C_4}p(\eta+t)^{\frac{3}{2}} + (\eta+t_2)^{\frac{3}{2}}\right)\left(\sqrt{\frac{\log(p/\delta}{n}} + \frac{\log(p/\delta)}{n}\right)$$

Finally, we have that,

$$\|\Sigma_{\widetilde{S}_D}\|_2 \leq \|\Sigma\|_2 \underbrace{\left(1 + 2C(\eta+t_2) + \left(1 + \frac{p\sqrt{C_4}}{\sqrt{\eta}} + \sqrt{C_4}p(\eta+t)^{\frac{3}{2}} + (\eta+t_2)^{\frac{3}{2}}\right)\left(\sqrt{\frac{\log(p/\delta}{n}} + \frac{\log(p/\delta)}{n}\right)\right)}_{\beta(n,\delta)}$$

Lemma 37. *Let W be the bottom $p/2$ principal components of the covariance matrix after filtering $\Sigma_{\widetilde{S}}$. Then there exists a universal constant $C > 0$ such that with probability at least $1-\delta$, we have that*

$$\|\eta P_W \delta_\mu\|_2^2 \leq C\eta\left(\beta(n,\delta) + \gamma(n,\delta)C_4^{\frac{1}{2}})\|\Sigma\|_2\right),$$

where $\delta_\mu = \mu_{\widetilde{S}_N} - \mu_{\widetilde{S}_D}$, P_W is the projection matrix on the bottom $p/2$-span of $\Sigma_{\widetilde{S}}$, $\beta(n,\delta)$ is as defined in Lemma 36 and $\gamma(n,\delta) = \left(\eta^{\frac{1}{2}} + (\eta+t)^{5/2} + \eta(\eta+t)\frac{\log(1/\delta)}{n}\right)$

Proof. We have

$$\Sigma_{\widetilde{S}} = \underbrace{(1-\eta)\Sigma_{\widetilde{S}_D}}_{E} + \underbrace{\eta\Sigma_{\widetilde{S}_N} + (\eta-\eta^2)\delta_\mu\delta_{\mu^T}}_{F}$$

By Weyl's inequality we have that,

$$\lambda_{p/2}(\Sigma_{\widetilde{S}}) \leq \lambda_1(E) + \lambda_{p/2}(F)$$

- **Control of $\lambda_{p/2}(F)$.**

$$\begin{aligned}\lambda_{p/2}(F) &\leq \frac{\text{trace}\,(F)}{p/2} \\ &\leq C_{15}\eta\frac{((r_1^*)^2 + (r_2^*)^2) + B^2}{p/2} \\ &\leq C_{16}C_4^{\frac{1}{2}}\|\Sigma\|_2 \underbrace{\left(\eta^{\frac{1}{2}} + (\eta+t)^{5/2} + \eta(\eta+t)\frac{\log(1/\delta)}{n}\right)}_{\gamma(n,\delta)}\end{aligned}$$

  where $t = C_8\sqrt{\frac{1}{n}\log\left(\frac{np}{\delta}\right)}$.

- **Control of $\lambda_1(E)$.**

$$\lambda_1(E) \leq (1-\eta)\beta\|\Sigma\|_2$$

Hence, we have that:

$$\lambda_{p/2}(\Sigma_{\widetilde{S}}) \leq (1-\eta)\beta\|\Sigma\|_2 + C_{16}\gamma\sqrt{C_4}\|\Sigma\|_2$$

Using that W is the space spanned by the bottom $p/2$ eigenvectors of $\Sigma_{\widetilde{S}}$ and P_W is corresponding projection operator, we have that:

$$P_W^T \Sigma_{\widetilde{S}} P_W \preceq \left[(1-\eta)\beta + C_{16}\gamma\sqrt{C_4}\right] \|\Sigma\|_2 I_p$$

Following some algebraic manipulation in [28], we get that,

$$\|\eta P_W \delta_\mu\|_2^2 \leq \eta\left((\beta(n,\delta) + \gamma C_4^{\frac{1}{2}})\|\Sigma\|_2\right)$$

□

Having established all required results, we are now ready to prove Lemma 9. We first present a result for general mean estimation. The proof of Lemma 9 then follows directly from this result.

Lemma 38. *Suppose that, P_θ^* is a distribution on $\mathbb{R}^p$ with mean μ, covariance Σ and bounded fourth moments. There exist positive universal constant $C > 0$, such that given n samples from the distribution in (3.9), the algorithm with probability at least $1-\delta$, returns an estimate $\widehat{\mu}$ such that,*

$$\|\widehat{\mu} - \mu\|_2 \leq C\|\Sigma\|_2^{\frac{1}{2}}(1+\sqrt{\log p})\left(\sqrt{\eta} + C_4^{\frac{1}{4}}(\eta + t_2)^{\frac{3}{4}} + \left(\sqrt{\eta}pC_4^{\frac{1}{2}}\sqrt{\frac{\log p \log(p\log(p/\delta))}{n}}\right)^{\frac{1}{2}}\right)$$

where $\eta = \epsilon + \sqrt{\frac{\log(p)\log(\log p/\delta)}{2n}}$ and $t_2 = \sqrt{\frac{p\log(p)\log(n/(p\delta))}{n}}$.

Proof. We divide n samples into $\lfloor \log(p) \rfloor$ different sets. We choose the first set and keep that as our active set of samples. We run our outlier filtering on this set, and let the remaining samples after the outlier filtering be $\widetilde{S}_D$. By orthogonality of subspaces spanned by eigenvectors, coupled with triangle inequality and contraction of projection operators, we have that

$$\|\widehat{\mu} - \mu\|_2^2 \leq 2\|P_W(\widehat{\mu} - \mu_{\widetilde{S}_D})\|_2^2 + 2\|P_W(\mu_{\widetilde{S}_D} - \mu)\|_2^2 + \|\widehat{\mu}_V - P_V\mu\|_2^2$$

$$\|\widehat{\mu} - \mu\|_2^2 \leq 2\|P_W(\widehat{\mu} - \mu_{\widetilde{S}_D})\|_2^2 + 2\|(\mu_{\widetilde{S}_D} - \mu)\|_2^2 + \|\widehat{\mu}_V - P_V\mu\|_2^2$$

where V is the span of the top $p/2$ principal components of $\Sigma_{\widetilde{S}}$ and where $\widehat{\mu}_V$ is the mean vector of returned by the running the algorithm on the reduced dimensions $dim(V) = p/2$. From Lemma 37, both $\beta(n,\delta)$ and $\gamma(n,\delta)$ are monotonically increasing in the dimension; moreover the upper bound in Lemma 36 is also monotonically increasing in the dimension p, hence, the error at each step of the algorithm can be upper bounded by error incurred when running on dimension p, with $n/\log(p)$ samples, and probability of $\delta/\log p$. Hence, the overall error for the recursive algorithm can be upper bounded as,

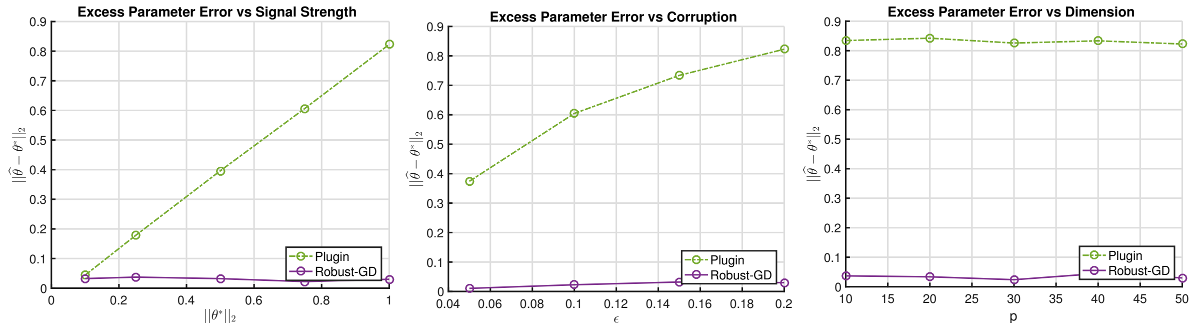
$$\|\widehat{\mu} - \mu\|_2^2 \leq \left(2\|P_W(\widehat{\mu} - \mu_{\widetilde{S}_D})\|_2^2 + 2\|\mu_{\widetilde{S}_D} - \mu\|_2^2\right)(1 + \log p)$$

Combining Lemma 36 and Lemma 37 which are instantiated for $n/\log p$ samples and probability $\delta/\log p$, we get,

$$\|\widehat{\mu} - \mu\|_2 \leq C\|\Sigma\|_2^{\frac{1}{2}} \sqrt{\log p} \left(\sqrt{\eta} + C_4^{\frac{1}{4}} (\eta + t_2)^{\frac{3}{4}} + \left(\sqrt{\eta} p C_4^{\frac{1}{2}} \sqrt{\frac{\log p \log(p \log(p/\delta))}{n}} \right)^{\frac{1}{2}} \right)$$

□

B.14 Empirical Comparison to Robust Plugin Estimation



(a) $\|\widehat{\theta}-\theta^*\|_2$ vs $\|\theta^*\|_2$ (b) $\|\widehat{\theta}-\theta^*\|_2$ vs ϵ (c) $\|\widehat{\theta}-\theta^*\|_2$ vs p

Figure B.2: Empirical Evaluation comparing Robust-GD and Robust Plugin Estimation

In this section, we conduct experiments to compare the performance of robust plugin-estimator. In particular, we empirically show that the upper bound of Corollary 10 is tight, *i.e.* the error of the estimator indeed scales with the signal strength $\|\theta^*\|_2$.

Setup. We follow the setup of Section 3.4.1. At a given contamination level ϵ, we generate $(1-\epsilon)n$ clean covariates from $\mathcal{N}(0,\mathcal{I}_p)$, and we generate the corresponding clean responses using $y = \langle x, \theta^* \rangle + w$ where $\theta^* = \kappa[1/\sqrt{p}, \ldots, 1/\sqrt{p}]^T$ and $w \sim \mathcal{N}(0, \sigma^2)$. Note that $kappa = \|\theta^*\|_2$ measures the signal strength. As before, we simulate an outlier distribution by drawing the covariates from $\mathcal{N}(0, p^2\mathcal{I}_p)$, and setting the responses to 0. The total number of samples is set to be $(10\frac{p}{\epsilon^2})$ to that statistical error, in the absence of contamination, is constant.

Plugin Estimator. The plugin estimator is given by $\widehat{\theta}_{PG} = \widetilde{\Sigma}^{-1}\widetilde{\nu}$, where we use Algorithm 4 to estimate the mean vector $\frac{1}{n}\sum_{i=1}^n x_i y_i$ robustly($\widetilde{\nu}$) and the covariance matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^T$ robustly($\widetilde{\Sigma}$).

Results. We summarize our main findings here.

- **Error vs $\|\theta^*\|_2$.** Figure B.2(a) shows that the error of the robust plugin algorithm scales linearly with the signal strength. This figure confirms that the upper bound derived in Corollary 10 is tight. On the other hand, the error of the robust robust-gd estimator doesn't depend on the signal strength.
- **Error vs ϵ.** Figure B.2(b) shows that at any given contamination level ϵ, the error of the plugin estimator is strictly larger than the error of the Robust-GD algorithm.
- **Error vs p.** Figure B.2(c) shows that the error of both estimators doesn't seem to depend on the dimension p.

B.15 Improved Rates for Heavy-Tailed Estimation

In this section we analyze the Median-SDP gradient estimator for heavy-tailed settings. The following result shows that the gradient estimate has exponential concentration around the true gradient, under the mild assumption that the gradient distribution has bounded second moment. Its proof follows directly from the analysis of Median-SDP of Hopkins [23].

Lemma 39. *Let P be the probability distribution of z and P_θ be the distribution of the gradients $\nabla\bar{\mathcal{L}}(\theta; z)$ on $\mathbb{R}^p$ with mean $\mu_\theta = \nabla\mathcal{R}(\theta)$, covariance Σ_θ. Then the heavy-tailed Median-SDP gradient estimator returns an estimate $\widehat{\mu}$ that satisfies the following exponential concentration inequality, with probability at least $1 - \delta$:*

$$\|\widehat{\mu} - \mu_\theta\|_2 \leq C_1\sqrt{\frac{trace\,(\Sigma_\theta)}{n}} + C_2\sqrt{\frac{\|\Sigma_\theta\|_2 \log{(1/\delta)}}{n}}$$

Next, we present results for parametric estimation under heavy-tailed distributions when Median-SDP is used to estimate gradients. Our assumptions are similar to that of Section 3.7 but we state them for completeness.

B.15.1 Linear Regression

We consider the linear regression model described in (B.17). We assume that the covariates $x \in \mathbb{R}^p$ have bounded 4^{th}-moments and the noise w has bounded 2^{nd} moments. This assumption is needed to bound the error in the gradient estimator (see Lemma 39).

Corollary 32 (Heavy Tailed Linear Regression)**.** *Consider the statistical model in* (B.17)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$C_1\frac{\sqrt{\|\Sigma\|_2}}{\tau_\ell}(\sqrt{\frac{trace\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2 \log{1/\widetilde{\delta}}}{\widetilde{n}}}) < 1$$

and if Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Median-SDP as gradient estimator, then it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1 - \delta$

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{C_2\sigma}{1-\kappa}(\sqrt{\frac{trace\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2 \log{1/\widetilde{\delta}}}{\widetilde{n}}}), \tag{B.8}$$

for some contraction parameter $\kappa < 1$.

B.15.2 Generalized Linear Models

In this section we consider generalized linear models described in (5.10), where the covariate x is allowed to have a heavy-tailed distribution. Here we assume that the covariates have bounded 4th moment. Additionally, we assume smoothness of $\Phi'(\cdot)$ around θ^*. Specifically, we assume that there exist universal constants $L_{\Phi,2k}$, B_{2k} such that

$$\mathbb{E}_x\left[|\Phi'(\langle x, \theta\rangle) - \Phi'(\langle x, \theta^*\rangle)|^{2k}\right] \leq L_{\Phi,2k}\|\theta^* - \theta\|_2^{2k} + B_{\Phi,2k}, \quad \text{for } k = 1, 2$$

We also assume that $\mathbb{E}_x[\left|\Phi^{(t)}(\langle x, \theta^* \rangle)\right|^k] \leq M_{\Phi,t,k}$ for $t \in \{1, 2, 4\}$, where $\Phi^{(t)}(\cdot)$ is the t^{th}-derivative of $\Phi(\cdot)$.

Corollary 33 (Heavy Tailed Generalized Linear Models)**.** *Consider the statistical model in* (5.10)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$C_1(\sqrt{\frac{trace\,(\Sigma)\,\sqrt{L_{\Phi,4}}}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2\sqrt{L_{\Phi,4}}\log 1/\widetilde{\delta}}{\widetilde{n}}}) < \tau_\ell,$$

and if Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Median-SDP as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1 - \delta$

$$\begin{aligned}\|\widehat{\theta}^t - \theta^*\|_2 \leq& \kappa^t\|\theta^0 - \theta^*\|_2 \\ &+ \frac{C_2\left[B_{\Phi,4}^{\frac{1}{4}} + c(\sigma)^{\frac{1}{2}}M_{\Phi,2,2}^{\frac{1}{4}} + c(\sigma)^{\frac{3}{4}}M_{\Phi,4,1}^{\frac{1}{4}}\right]}{1-\kappa}\left(\sqrt{\frac{trace\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2\log(1/\widetilde{\delta})}{\widetilde{n}}}\right),\end{aligned} \tag{B.9}$$

for some contraction parameter $\kappa < 1$.

We now instantiate the above Theorem for logistic regression model.

Corollary 34 (Heavy Tailed Logistic Regression)**.** *Consider the model in* (5.13)*. There are universal constants $C_1, C_2 > 0$ such that if*

$$C_1(\sqrt{\frac{trace\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2\log 1/\widetilde{\delta}}{\widetilde{n}}}) < \tau_\ell$$

and if Algorithm 3 initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Median-SDP as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1 - \delta$

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{C_2}{1-\kappa}\left(\sqrt{\frac{trace\,(\Sigma)}{\widetilde{n}}} + \sqrt{\frac{\|\Sigma\|_2\log 1/\widetilde{\delta}}{\widetilde{n}}}\right),$$

for some contraction parameter $\kappa < 1$.

B.15.3 Exponential Family

We now instantiate Theorem 56 for parameter estimation in heavy-tailed exponential family distributions. Here we assume that the random vector $\phi(z), z \sim P$ has bounded 2nd moments, and we obtain the following result:

Corollary 35 (Heavy Tailed Exponential Family)**.** *Consider the model in* (3.8)*. If Algorithm 3 is initialized at θ^0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$ and Algorithm 5 as gradient estimator, it returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$, such that with probability at least $1 - \delta$*

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}C_1(\sqrt{\frac{tr(\nabla^2 A(\theta^*)}{\widetilde{n}}} + \sqrt{\frac{\|\nabla^2 A(\theta^*)\|_2\log 1/\widetilde{\delta}}{\widetilde{n}}}),$$

for some contraction parameter $\kappa < 1$ and universal constant C.

B.16 Proof of Lemma 10

In this section, we present a Filtering based gradient estimator. The estimator (and its analysis) are primarily based on the Filtering Algorithm of Diakonikolas et al. [31] but we do a careful non-asymptotic analysis. In particular, we obtain high-probability bounds using a martingale style analysis and our results are (almost) dimension independent (i.e. they depend on the dimension primarily through $\mathrm{tr}(\Sigma_\theta)$).

Algorithm 10 Huber Gradient Filtering Estimator

```
function FilterGradientEst(Sample Gradients $S = \{\nabla\bar{\mathcal{L}}(\theta; x_i)\}_{i=1}^n$, Corruption Level $\epsilon$,$\|\Sigma\|_2$,trace$(\Sigma)$,Confidence Level $\delta$)
    Let $\widehat{\theta}_S = \frac{1}{|S|}\sum_{i=1}^{|S|} z_i$ be the sample mean.
    Let $\Sigma_S = \frac{1}{|S|}\sum_{i=1}^{|S|}(z_i - \widehat{\theta}_S)(z_i - \widehat{\theta}_S)^T$ be the sample covariance matrix.
    Let $(\lambda, v)$ be the largest eigenvalue,eigenvector of $\Sigma_S$.
    Let $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$
    if $\lambda < C(\|\Sigma\|_2 + \frac{\mathrm{trace}(\Sigma)\log(p/\delta)}{n\alpha})$ then
        return $\widehat{\theta}_S$
    else
        For each $z_i$, let $\tau_i \overset{\text{def}}{=} (v^T(z_i - \widehat{\theta}_S))^2$ to be its score
        Randomly sample a point $z$ from $S$ according to
            $$\Pr(z_i \text{ chosen}) = \frac{\tau_i}{\sum_j \tau_j}$$
        return FilterGradientEst($S\backslash\{z\}$ $\epsilon$,$\|\Sigma\|_2$,trace$(\Sigma)$, $\delta$)
    end if
end function
```

We first restate the result to ease readability.

Lemma 40. *Let P be the true probability distribution of z and let P_θ be the true distribution of the gradients $\nabla\bar{\mathcal{L}}(\theta; z)$ on $\mathbb{R}^p$ with mean $\mu_\theta = \nabla\mathcal{R}(\theta)$, covariance Σ_θ, and bounded second moments. There exists a positive constant $C_1 > 0$, such that given n samples from the distribution in (3.9), the Huber Gradient Estimator described in Algorithm 10 when instantiated with the contamination level ϵ, and knowledge of $\|\Sigma_\theta\|_2$ and trace(Σ_θ), with probability at least $1-\delta$, returns an estimate $\widehat{\mu}$ of μ_θ such that,*

$$\|\widehat{\mu} - \mu_\theta\|_2 \leq C_1\|\Sigma_\theta\|_2^{\frac{1}{2}}\max(\epsilon, \frac{\log(1/\delta)}{n})^{\frac{1}{2}} + \sqrt{\frac{\mathit{trace}\,(\Sigma_\theta)\log(p/\delta)}{n}}$$

Proof. We go over the main steps of the proof first and defer the intermediate lemmas.

1. We first show that at the start of the algorithm, there is a good set G^0 which satisfies three properties,
    (a) A lot of the empirical mass lies in it.(n_{G^0} is big.)

(b) The operator norm of the covariance of G^0 given by $\|\Sigma_{G^0}\|_2$ is controlled.

(c) The empirical mean of G^0 is close to the true mean. $\|\widehat{\theta}_{G^0} - \theta^*\|_2$ is small.

To this end, let $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$. $G^0 \overset{\text{def}}{=} \left\{ x_i \in S^0, \text{ s.t. } x_i \sim P^*, |x_i - \theta^*| \leq \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}} \right\}$. In Lemma 41, we prove the following results hold with probability at least $1 - 3\delta$.

$$n_{G^0} \geq n(1 - (\epsilon + \alpha) - C_1\sqrt{(\epsilon + \alpha)\frac{\log(1/\delta)}{n}} - C_2\frac{\log(1/\delta)}{n})) \tag{B.10}$$

$$\|\widehat{\theta}_{G^0} - \theta^*\|_2 \leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + C_2\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_3\sqrt{\frac{\text{trace}\,(\Sigma)\log(1/\delta)}{n_{G^0}}} \tag{B.11}$$

$$\|\Sigma_{G^0}\|_2 \leq C_1\|\Sigma\|_2 + \|\Sigma\|_2^{\frac{1}{2}}\text{trace}\,(\Sigma)^{\frac{1}{2}}\log(p/\delta)^{\frac{1}{2}}\frac{1}{\sqrt{n\alpha}} + \text{trace}\,(\Sigma)\log(p/\delta)\frac{1}{n\alpha} \tag{B.12}$$

2. In Lemma 49, we show that the with probability at least $1-\delta$, the algorithm terminates in at most $T^*_\delta = \lceil 18\log(1/\delta) + 3(n - n_{G^0})\rceil$.
3. In Lemma 50 we prove a result which shows that when the algorithm stops in $m = T^*_\delta$ steps, the sample mean of points, $\widehat{\theta}_{S^m}$ is close to the mean of G^0. In particular,

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq C_1(8\frac{n - n_{G^0}}{n} + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}} \tag{B.13}$$

Equipped with the above results, the final theorem statement follows from some algebra which we show below.

$$\begin{aligned}\|\theta^* - \widehat{\theta}_{S^m}\|_2 &\leq \|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 + \|\widehat{\theta}_{G^0} - \theta^*\|_2 \\ &\leq C_1(8\frac{n - n_{G^0}}{n} + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}} + \|\widehat{\theta}_{G^0} - \theta^*\|_2\end{aligned} \tag{B.14}$$

From Equation B.10 we know that

$$(n - n_{G^0})/n \leq (\epsilon + \alpha) + C_1\sqrt{(\epsilon + \alpha)\frac{\log(1/\delta)}{n}} \leq C\alpha$$

where we used that $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$. Using this, we get that,

$$C_1(8\frac{n - n_{G^0}}{n} + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}} \leq (C\alpha)^{\frac{1}{2}}$$

From Equation B.12 we know that,

$$
\begin{aligned}
\|\Sigma_{G^0}\|_2 \leq C_1\|\Sigma\|_2 + \|\Sigma\|_2^{\frac{1}{2}}\text{trace}\,(\Sigma)^{\frac{1}{2}}\log(p/\delta)^{\frac{1}{2}}\frac{1}{\sqrt{n\alpha}} + \text{trace}\,(\Sigma)\log(p/\delta)\frac{1}{n\alpha} \\
\implies \|\Sigma_{G^0}\|_2^{\frac{1}{2}}\sqrt{\alpha} \leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + \underbrace{(\|\Sigma\|_2^{\frac{1}{2}}\alpha^{\frac{1}{2}})^{\frac{1}{2}}(\frac{\text{trace}\,(\Sigma)^{\frac{1}{2}}\log(p/\delta)^{\frac{1}{2}}}{n^{\frac{1}{2}}})^{\frac{1}{2}}}_{\|\Sigma\|_2^{\frac{1}{2}}\alpha^{\frac{1}{2}}+\sqrt{\frac{\text{trace}(\Sigma)\log(p/\delta)}{n}}} + \frac{\text{trace}\,(\Sigma)^{\frac{1}{2}}\log(p/\delta)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \\
\leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + C_2\sqrt{\frac{\text{trace}\,(\Sigma)\log(p/\delta)}{n}}
\end{aligned}
$$

The final statement follows by plugging the above and (B.11) into (B.14). □

Lemma 41. *Let* $G^0 \overset{\text{def}}{=} \left\{x_i \in S^0,\ \text{s.t. } x_i \sim P^*, \|x_i - \theta^*\|_2 \leq \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}}\right\}$ *be some good set, where* $\alpha \overset{\text{def}}{=} \max(\epsilon, \frac{\log(1/\delta)}{n})$. *Then, with probability at least* $1-3\delta$.

$$
n_{G^0} \geq n(1-(\epsilon+\alpha) - C_1\sqrt{(\epsilon+\alpha)\frac{\log(1/\delta)}{n}} - C_2\frac{\log(1/\delta)}{n}))
$$

$$
\|\widehat{\theta}_{G^0} - \theta^*\|_2 \leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + C_2\sqrt{\frac{\textit{trace}\,(\Sigma)}{n_{G^0}}} + C_3\sqrt{\frac{\textit{trace}\,(\Sigma)\log(1/\delta)}{n_{G^0}}} \tag{B.15}
$$

$$
\|\Sigma_{G^0}\|_2 \leq C_1\|\Sigma\|_2 + \|\Sigma\|_2^{\frac{1}{2}}\textit{trace}\,(\Sigma)^{\frac{1}{2}}\log(p/\delta)^{\frac{1}{2}}\frac{1}{\sqrt{n\alpha}} + \textit{trace}\,(\Sigma)\log(p/\delta)\frac{1}{n\alpha} \tag{B.16}
$$

Proof. Consider the event $E_1 = x \sim P$, then $P_\epsilon(E_1) = 1-\epsilon$. Consider the event $E_2 = \|x_i - \theta^*\|_2 \leq \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}}$, We have that $P_\epsilon(E_2|E_1) = 1 - \Pr(\|x-\theta^*\|_2 > \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}}|x \sim P)$. Using Chebyshev's inequality, we have that,

$$
P^*(\|x-\mu\|_2 > \sqrt{\frac{\text{trace}\,(\Sigma)}{\alpha}}) \leq \frac{\mathbb{E}[\|x-\mu\|_2^2]}{(\sqrt{\frac{\text{trace}(\Sigma)}{\alpha}})^2} = \alpha
$$

.

$$
\begin{aligned}
P_\epsilon(E_1 \cap E_2) &\geq (1-\epsilon)(1-\alpha) \\
&= 1 + \epsilon\alpha - (\epsilon+\alpha) \\
&\geq 1 - (\epsilon+\alpha)
\end{aligned}
$$

Now, given n-samples from P_ϵ, we use Bernsteins bound to get the empirical probability, *i.e.* we get that with probability at least $1-\delta$

$$
P_\epsilon(E_1 \cap E_2) - P_{n,\epsilon}(E_1 \cap E_2) \leq C_1\sqrt{(\epsilon+\alpha)}\sqrt{\frac{\log(1/\delta)}{n}} + C_2\frac{\log(1/\delta)}{n}
$$

$$\implies n_{G^0} \geq n(1 - (\epsilon + \alpha) - C_1\sqrt{(\epsilon+\alpha)}\sqrt{\frac{\log(1/\delta)}{n}} - C_2\frac{\log(1/\delta)}{n})$$

This proves the first claim in the Lemma.

Controlling $\|\widehat{\theta}_{G^0} - \theta^*\|_2$.

1. Controlling $\|\theta^* - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. This is a deterministic statement and essentially quantifies the amount the mean can shift, when the random variable is conditioned on an event. This Lemma has appeared in [28, 51].

   Claim 6. *[Mean shift] Suppose that a distribution P has mean μ and covariance Σ. Then, for any event $\mathcal{A}$ which occurs with probability at least $1-\epsilon \geq \frac{1}{2}$,*

   $$\|\mu - E[x|\mathcal{A}]\|_2 \leq 2\|\Sigma\|_2^{\frac{1}{2}}\epsilon^{\frac{1}{2}}$$

   Proof. For any event $\mathcal{A}$, Let $\mathbb{I}\{\mathcal{A}\}$ denote the corresponding indicator variable.

   $$\|E_{x\sim P}[x|\mathcal{A}] - \mu]\|_2 = \frac{1}{P(\mathcal{A})}\|E_{x\sim P^*}((x-\mu)\mathbb{I}\{\mathcal{A}\})\|_2 \leq 2\|E_{x\sim P^*}((x-\mu)\mathbb{I}\{\mathcal{A}\})\|_2,$$

   $$\begin{aligned}\mathbb{E}_{x\sim P}[(x-\mu)\mathbb{I}\{x\in\mathcal{A}^c\} + (x-\mu)\mathbb{I}\{x\in\mathcal{A}\}] &= \mathbb{E}_{x\sim P}[(x-\mu)] = 0\\ \implies \|\mathbb{E}_{x\sim P}[(x-\mu)\mathbb{I}\{x\in\mathcal{A}^c\}]\|_2 &= \|\mathbb{E}_{x\sim P}[(x-\mu)\mathbb{I}\{x\in\mathcal{A}\}]\|_2\end{aligned}$$

   $$\begin{aligned}\|\mathbb{E}_{x\sim P}[(x-\mu)\mathbb{I}\{x\in\mathcal{A}^c\}]\|_2 &= \sup_{u\in\mathcal{S}^{p-1}} |\mathbb{E}_{(x\sim P}[u^T(x-\mu)\mathbb{I}\{x\in\mathcal{A}^c\}]|\\ &\overset{(i)}{\leq} \sup_{u\in\mathcal{S}^{p-1}} \sqrt{E_{x\sim P}[u^T(x-\mu)(x-\mu)^Tu]}\sqrt{E_{x\sim P}[\mathbb{I}\{x\in\mathcal{A}^c\}^2]}\\ &= \|\Sigma\|_2^{\frac{1}{2}}P(A^c)^{\frac{1}{2}} \leq \|\Sigma\|_2^{\frac{1}{2}}\sqrt{\epsilon}\end{aligned}$$

   □

   Now using this Claim, with $\mathcal{A}$ being the event that when $x \sim P^*$, $\|x - \theta^*\|_2 \leq \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}}$.

   $$\|\theta^* - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq \|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha}$$

2. **Controlling** $\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. This term measures how quickly the samples within G^0 converge to their true mean. To show this we use vector version of Bernstein's

inequality. Let $z_i \overset{\text{def}}{=} x_i - \mathbb{E}[\widehat{\theta}_{G^0}]$ be the centered random variables. Then, we have that

$$\begin{aligned}
\|z_i\|_2 &\leq \|\theta^* - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 + \|x_i - \theta^*\|_2 \\
&\leq 2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + \sqrt{\frac{\text{trace}\,(\Sigma)}{\alpha}} \\
&\leq \sqrt{\text{trace}\,(\Sigma)}(2\sqrt{\alpha} + \frac{1}{\sqrt{\alpha}}) \\
&\leq \underbrace{2\sqrt{\frac{\text{trace}\,(\Sigma)}{\alpha}}}_{B}, \quad \forall \alpha < \frac{1}{2}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[\|z_i\|_2^2] &= \mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 | x \in \mathcal{A}] \\
&= \frac{\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 \mathbb{I}\left\{x \in \mathcal{A}\right\}]}{P(\mathcal{A})} \\
&\leq 2\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2] \\
&\leq 2\mathbb{E}[\|x - E[x]\|_2^2] + 2\|\theta^* - E[x|\mathcal{A}]\|_2^2 \\
&\leq 2\text{trace}\,(\Sigma) + 4\|\Sigma\|_2\epsilon \\
&\leq 4\text{trace}\,(\Sigma)
\end{aligned}$$

Now, we first state the vector version of Bernstein's inequality.

Lemma 42. *(Vector Bernstein [134]) Let $(y_k)_{k=1}^n$ be a finite sequence of i.i.drandom vectors. Suppose $\mathbb{E}y_k = 0$, and $\|y_k\|_2 \leq K$ a.s. for dome constant K>0. Let $Z = \|\sum_{k=1}^n y_k\|_2$. Then for any $t > 0$, we have that,*

$$P(Z - \sqrt{E[Z^2]} > t) \leq \exp(-\frac{t^2/2}{EZ^2 + 2K\sqrt{EZ^2} + Kt/3})$$

where $EZ^2 = \sum_{k=1}^n E[\|y_k\|_2^2]$.

We instantiate the above Lemma with $y_k = \frac{z_k}{n_{G^0}}$. Hence, we have that $Z = \|\sum_{k=1}^{n_{G^0}} y_k\|_2 = \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. We also have that, $EZ^2 \leq n_{G^0}\frac{1}{n_{G^0}^2}4\text{trace}\,(\Sigma) \leq C\text{trace}\,(\Sigma)/n_{G^0}$ and $K = \frac{B}{n_{G^0}}$. Using vector Bernstein's we get that with probability at least $1 - \delta$,

$$\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq C_1\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_2\sqrt{\frac{\text{trace}\,(\Sigma)\log(1/\delta)}{n_{G^0}}} + C_3\frac{B\log(1/\delta)}{n_{G^0}} + C_4\sqrt{K\sqrt{EZ^2}}\sqrt{\log(1/\delta}$$

Now, using that $\sqrt{ab} < a + b, \forall a, b \geq 0$. We get that,

$$\sqrt{K\sqrt{EZ^2}}\sqrt{\log(1/\delta)} \leq K\sqrt{\log(1/\delta)} + \sqrt{EZ^2}\sqrt{\log(1/\delta)}$$

Substituting, $K = \frac{B}{n_{G^0}} = C\sqrt{\text{trace}\,(\Sigma)}\frac{1}{\sqrt{\alpha}}\frac{1}{n_{G^0}} = \sqrt{\frac{\text{trace}(\Sigma)}{n_G^0}}\frac{1}{\sqrt{\alpha n_{G^0}}}$. Since, $\alpha \geq \log(1/\delta)/n$, hence, we have that, $\frac{1}{\sqrt{\alpha n_{G^0}}} \leq \frac{1}{\sqrt{\log(1/\delta)}}\sqrt{\frac{n}{n_{G^0}}} \leq \frac{C}{\sqrt{\log(1/\delta)}}$. Hence, we get that, $K = \frac{B}{n_{G^0}} = C\sqrt{\frac{\text{trace}(\Sigma)}{n_{G^0}}}\frac{1}{\sqrt{\log(1/\delta)}}$. Plugging it above, we get that,

$$\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq C_1\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_2\sqrt{\frac{\text{trace}\,(\Sigma)\log(1/\delta)}{n_{G^0}}}$$

The final claim of (B.15) follows from triangle inequality.

$$\begin{aligned}\|\widehat{\theta}_{G^0} - \theta^*\|_2 &\leq \|\mathbb{E}[\widehat{\theta}_{G^0}]\|_2 + \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \\ &\leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\alpha} + C_2\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_3\sqrt{\frac{\text{trace}\,(\Sigma)\log(1/\delta)}{n_{G^0}}}\end{aligned}$$

Controlling $\|\Sigma_{G^0}\|_2$. Using Triangle inequality, we have that

$$\|\Sigma_{G^0}\|_2 \leq \underbrace{\|\Sigma_{G^0} - \mathbb{E}[\Sigma_{G^0}]\|_2}_{T1} + \underbrace{\|\mathbb{E}[\Sigma_{G^0}]\|_2}_{T2}$$

1. **Controlling T2.** Let μ_G be $\mathbb{E}[x|G^0]$ be the mean of the points which lie in G^0.

$$\begin{aligned}\mathbb{E}[\Sigma_{G^0}] = \mathbb{E}[(x-\mu_G)(x-\mu_G)^T|x \in G^0] &= \mathbb{E}[(x-\theta^*)(x-\theta^*)^T|x \in G^0] + (\theta^* - \mu_G)(\theta^* - \mu_G)^T \\ &= \frac{\mathbb{E}[(x-\theta^*)(x-\theta^*)^T\mathbb{I}\{x \in G^0\}]}{P(x \in G^0)} + (\theta^* - \mu_G)(\theta^* - \mu_G)^T\end{aligned}$$

$$\|\mathbb{E}[\Sigma_{G^0}]\|_2 \leq \frac{\|\Sigma\|_2}{1-\alpha} + \|\theta^* - \mu_G\|_2^2 \leq \frac{\|\Sigma\|_2}{1-\alpha} + C_1\|\Sigma\|_2\alpha \leq C_2\|\Sigma\|_2$$

$$T2 \leq C_2\|\Sigma\|_2$$

2. **Controlling T1.** Note that T1 essentially measures how far the sample covariance of the points in G^0 is from there true covariance. This is again a concentration of measure argument, and in particular exploits concentration of covariance for bounded random vectors.

Lemma 43. *[Theorem 5.44 [135]] Let $\{y_i\}_{i=1}^n$ samples such that $y_i \in \mathbb{R}^p$ and $\|y_i\|_2 \leq \sqrt{m}$ and $\mathbb{E}[yy^T] = \Sigma$. Then, with probability at least $1-\delta$,*

$$\|\frac{1}{n}\sum_{i=1}^n y_i y_i^T - \Sigma\|_2 \leq \max(\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\log(p/\delta)}\sqrt{\frac{m}{n}}, \log(p/\delta)\frac{m}{n})$$

$$T1 = \|\frac{1}{n_{G^0}}\sum_{i=1}^{n_{G^0}}(x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T - \mathbb{E}[\Sigma_{G^0}]\|_2$$
$$\leq \underbrace{\|\frac{1}{n_{G^0}}\sum_{i=1}^{n_{G^0}}(x_i - \mu_G)(x_i - \mu_G)^T - \mathbb{E}[\Sigma_{G^0}]\|_2}_{T1a} + \underbrace{\|\widehat{\theta}_{G^0} - \mu_G\|_2^2}_{T1b}$$

(a) **Controlling T1a.** We use Lemma 55 with $y_i = x_i - \mu_G$. Observe that

$$\|x_i - \mu_G\|_2 \leq \|x_i - \theta^*\|_2 + \|\mu_G - \theta^*\|_2 \leq \sqrt{\frac{\text{trace}\,(\Sigma)}{\alpha}} + \sqrt{\|\Sigma\|_2}\sqrt{\alpha} \leq 2\sqrt{\frac{\text{trace}\,(\Sigma)}{\alpha}} = \sqrt{m}$$

$$\sqrt{\frac{m}{n_{G^0}}} = C\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}}\frac{1}{\sqrt{\alpha}}$$

Also note that in controlling T2, we showed that $\|\mathbb{E}[\Sigma_{G^0}]\|_2 \leq C\|\Sigma\|_2$ and when proving the bound on mean, we showed that $\text{trace}\,(\mathbb{E}[\Sigma_{G^0}]) \leq C\text{trace}\,(\Sigma)$. This means that with probability $1 - \delta$,

$$T1a \leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\text{trace}\,(\Sigma)\log(p)}{n\alpha}} + C_2\text{trace}\,(\Sigma)\log(p/\delta)\frac{1}{n\alpha}$$

Note, we can ignore $T1b$ as they are $O(1/n)$ terms.

The final claim of (B.15) follows from triangle inequality.

□

Lemma 44. *Given n samples from P_ϵ. Then, with probability at least $1 - 4\delta$, Algorithm 10 stops in at most $T_\delta^* = \left\lceil 8\log(1/\delta)\frac{\gamma^2}{(\gamma-1)^2} + 2Y^0\frac{\gamma}{\gamma-1}\right\rceil$ steps.*

Proof. At each step of Algorithm 10, we remove one sample based on the probability distribution of the scores. Let $l = 1, 2, \ldots, n$ be the steps of the algorithm. Note that the steps of the Algorithm are dependent, hence to obtain a high probability statement, we will have to use martingale style analysis. The martingale analysis in the proof mostly follows from [119, 120].

Let $\mathcal{F}^l$ be the filtration generated by the sets of events until step l. At step l, let S^l be the set of samples, G^l be the set of good samples, *i.e.* $\{x_i \in S^l | x_i \sim P^* \& \|x_i - \theta^*\|_2 \leq \sqrt{\frac{\text{trace}(\Sigma)}{\alpha}}\}$ Let $B^l = S^l \backslash G^l$ be the remaining(bad) samples. Note that $|S^l| = n_l = n - l$, and $S^l, G^l, B^l \in \mathcal{F}^l$.

Let τ_i be some score for each point. Define $\mathcal{E}^l$ be an event variable at step l which is True if

$$\sum_{i \in G^l}\tau_i \geq \frac{1}{(\gamma - 1)}\sum_{j \in B^l}\tau_j, \equiv \sum_{i \in G^l}\tau_i \geq \frac{1}{\gamma}\sum_{j \in S^l}\tau_j$$

for say $\gamma = 3$. Intuitively, this means the event is true when the sum of the scores of the good points is larger compared to the bad points.

Now, when $\mathcal{E}^l$ is false, we sample a point j according τ_j and remove it. Some algebra shows, that when $\mathcal{E}^l$ is false, then with constant probability of 2/3, we throw a point from B^l.

$$\Pr(\text{sample removed at Step } l \in B^l | \mathcal{F}^l) = \frac{\sum\limits_{i \in B^l} \tau_i}{\sum_{j \in S^l} \tau_j} \geq \frac{\gamma - 1}{\gamma} = 2/3$$

Essentially, our argument shows that whenever $\mathcal{E}^l$ is false, then we are more likely to throw a point from the bad set. This means, that in the next iteration the fraction of bad points will reduce. To argue more formally, let $T \overset{\text{def}}{=} \min\{l : \mathcal{E}^l \text{ is true}\}$ be the first time that $\mathcal{E}^l$ is True. Then, our goal is to show that T is small.

To show this, based on T, define Y^l, as

$$Y^l = \begin{cases} |B^{T-1}| + \frac{\gamma-1}{\gamma}(T-1), \text{ if } l \geq T \\ |B^l| + \frac{\gamma-1}{\gamma} l, \text{ if } l < T \end{cases}$$

Now, we show that $\{Y^l, \mathcal{F}^l\}$ is a supermartingale, *i.e.* $\mathbb{E}[Y^l | \mathcal{F}^{l-1}] \leq Y^{l-1}$. To see this, we split it into three cases:

- **Case 1.** $l < T$. This means that $\mathcal{E}^l$ is false.

  $$Y^l - Y^{l-1} = |B^l| - |B^{l-1}| + \frac{\gamma - 1}{\gamma},$$

  Now, $|B^l| = |B^{l-1}|$ if no bad point is thrown, and $|B^l| = |B^{l-1}| - 1$ if the point thrown is bad. Since, $\mathcal{E}^{l-1}$ is false, hence, we have that,

  $$\mathbb{E}[Y^l - Y^{l-1} | \mathcal{F}^{l-1}] = -1(\Pr(\text{sample removed at Step } l-1 \in B^{l-1})) + \frac{\gamma-1}{\gamma} \overset{(i)}{\leq} 0$$

  where (i) is true because $\mathcal{E}^{l-1}$ is false.
- **Case 2.** $l = T$, This follows by construction, because at $l = T$, $Y^l = Y^{l-1}$.
- **Case 3.** $l > T$, This also follows by construction.

So, we have that $Y^l, \mathcal{F}^l$ is a supermartingale. Now, we need to bound the steps T_δ such that the probability that the algorithm doesn't stop in T_δ steps is less than δ, *i.e.*

$$\Pr(\bigcap_{l=1}^{T_\delta} (\mathcal{E}^l)^c) \leq \delta$$

Note, that,

$$\Pr(\bigcap_{l=1}^{T_\delta} (\mathcal{E}^l)^c) = \Pr(T \geq T_\delta) \overset{(ii)}{\leq} \Pr(Y^{T_\delta} \geq \frac{\gamma-1}{\gamma} T_\delta)$$

where (ii) follows because, if $T > T_\delta \implies Y^{T_\delta} = |B^{T_\delta}| + \frac{\gamma-1}{\gamma}T_\delta \geq \frac{\gamma-1}{\gamma}T_\delta$. Now,

$$\Pr(Y^{T_\delta} \geq \frac{\gamma-1}{\gamma}T_\delta) = \Pr(Y^{T_\delta} - Y^0 \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0)$$

Now, defining $D^l = Y^l - Y^{l-1}$, and let $Z^l = D^l - \mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}]$. Then,

$$Y^{T_\delta} - Y^0 = \sum_{l=1}^{T_\delta} D^l = \sum_{l=1}^{T_\delta} Z^l + \sum_{l=1}^{T_\delta} \mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}]$$

Since, we know that $\{Y^l, \mathcal{F}^l\}$ is a supermartingale, hence the difference process is such that

$$\mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}] \leq 0$$

This implies that

$$Y^{T_\delta} - Y^0 \leq \sum_{l=1}^{T_\delta} Z^l \implies \Pr(Y^{T_\delta} - Y^0 \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0) \leq \Pr(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0)$$

Since, $|D^l| \leq 1$, and $Z^l \leq 2$ are bounded, hence we can use Azuma-Hoeffding to bound the above probability. In particular,

$$\Pr(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0) \leq \exp(-\frac{(\frac{\gamma-1}{\gamma}T_\delta - Y_0)^2}{8T_\delta})$$

Now, we want a T_δ such that, $\exp(-\frac{(\frac{\gamma-1}{\gamma}T_\delta - Y_0)^2}{8T_\delta}) \leq \delta$. Solving the quadratic, we need a T_δ such that,

$$(\frac{\gamma-1}{\gamma})^2 T_\delta^2 - (8\log(1/\delta) + 2Y^0\frac{\gamma-1}{\gamma})T_\delta + Y_0^2 \geq 0$$

Some algebra shows that $T_\delta^* = \left\lceil 8\log(1/\delta)\frac{\gamma^2}{(\gamma-1)^2} + 2Y^0\frac{\gamma}{\gamma-1} \right\rceil$ satisifies the above equation. Hence, we know that with probability at least $1-\delta$, there exists at least one good event in 1 to T_δ^* iterations. Note than $Y^0 = n_{B^0} = n - n_{G^0}$.

While we have established that there is at least one good event in 1 to T_δ^* iterations, we need to show that whenever $\mathcal{E}^l$ is True then Algorithm 10 stops, *i.e.* our checking condition is violated. To show this, we first prove that for $m \leq T_{\delta^*}$, when $\mathcal{E}^m$ is true then $\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$(See Claim 12). Coupling this with Claim 11, which shows that $\|\Sigma_{G^m}\|_2 \leq 2\|\Sigma_{G^0}\|_2$, we get that $\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$. Using the upper bound derived on $\|\Sigma_{G^0}\|_2$ in Lemma 41, we get that, whenever $\mathcal{E}^m$ is True,

$$\|\Sigma_{S^m}\|_2 \leq C\|\Sigma_{G^0}\|_2 \leq C(\|\Sigma\|_2 + \frac{\text{trace}\,(\Sigma)\log(p/\delta)}{n})$$

which is just our checking condition. Hence, Algorithm 10 stops whenever $\mathcal{E}^m$ is True.

□

Lemma 45. *Let* $\phi = \frac{n_{B^0}}{n}$*. Then, under the assumption that* $8\phi + 36\frac{\log(1/\delta)}{n} < \frac{1}{4}$*, we have that when* $\mathcal{E}^m$ *is True,*

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq 10\sqrt{2}(8\phi + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}}$$

Proof. Using Lemma 52, we get that,

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}}(\|\Sigma_{G^0}\|_2^{\frac{1}{2}} + \|\Sigma_{S^m}\|_2^{\frac{1}{2}}),$$

where P_1 is the equal weight discrete distribution with support on S^m, and P_2 is the equal weight discrete distribution with support on G^0. In Claim 10 we show that

$$TV(P_1, P_2) \leq 8\phi + 36\frac{\log(1/\delta)}{n}$$

When, $\mathcal{E}^m$ is True, we know by contrapositive of Lemma 51 that $\|\Sigma_{S^m}\|_2 \leq \frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma} - \psi_m}\|\Sigma_{G^m}\|_2$, where $\psi_m = (\frac{\sqrt{TV(P_1,P_3)}}{1-\sqrt{TV(P_1,P_3)}})^2$. Coupling this with Claim 11, which shows that $\|\Sigma_{G^m}\|_2 \leq 2\|\Sigma_{G^0}\|_2$, we get that $\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$.

$$\|\Sigma_{S^m}\|_2 \leq C\|\Sigma_{G^0}\|_2$$

Hence, under our assumption that $8\phi + 36\frac{\log(1/\delta)}{n} < \frac{1}{4}$, we get that,

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq C(8\phi + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}}$$

□

Lemma 46. *Given a collection of points* S *of size* n*. Let* P_1 *and* P_2 *be discrete empirical distributions on* n*. Then, we have that,*

$$\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2[x_i]}\|_2 \leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}}(\|\widehat{\Sigma}_{P_1}\|_2^{\frac{1}{2}} + \|\widehat{\Sigma}_{P_2}\|_2^{\frac{1}{2}})$$

where $\widehat{\Sigma}_{P_1}$ *is the covariance matrix when* $x_i \sim P_1$*, and* $\widehat{\Sigma}_{P_2}$ *is the empirical covariance matrix of when* $x_i \sim P_2$

Proof. Consider a joint distribution(also called coupling) $\omega^*(z, z')$ over $S \times S$ such that it's individual marginal distributions are equal to P_1 and P_2; *i.e.* $\omega(z) = P_1$ and $\omega(z') = P_2$ and

$\omega(z \neq z') = TV(P_1, P_2)$. Then, we have that

$$\begin{aligned}
&\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2[x_i]}\|_2 = \sup_{v \in \mathcal{S}^{p-1}} |\langle v, \mathbb{E}_{w^*}[z - z'] \rangle| \\
&\leq \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[|\langle v, z - z' \rangle|] \\
&\leq \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[1(z \neq z') \langle v, z - z' \rangle |] \\
&\leq (\mathbb{E}_{w^*}[(1(z \neq z'))^{1/(1-\frac{1}{2})}])^{1-\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - z' \rangle)^2]^{\frac{1}{2}} \\
&\leq TV(P_1, P_2)^{\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} (\mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] + \mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i] + \mathbb{E}_{x_i \sim P_2}[x_i] - z' \rangle)^2]^{\frac{1}{2}}) \\
&\leq TV(P_1, P_2)^{\frac{1}{2}} (\sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] \rangle)^2]^{\frac{1}{2}} + \|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i]\|_2) \\
&+ TV(P_1, P_2)^{\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_2}[x_i] \rangle)^2]^{\frac{1}{2}} \\
&\leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} \left(\|\Sigma_{P_1}\|_2^{\frac{1}{2}} + \|\Sigma_{P_2}\|_2^{\frac{1}{2}} \right)
\end{aligned}$$

□

Lemma 47. *Let S be a collection of n points. And let G be a subset of S containing n_G points. Define $\tau_i = (v^T(x_i - \widehat{\theta}_S))^2$, where v is the top unit-norm eigenvector of $\widehat{\Sigma}_S$ and $\widehat{\theta}_S$ is the sample mean of S. Let $\lambda = \|\Sigma_S\|_2$. Then, we have the following*

- *If $\lambda > \frac{1+\psi}{\frac{n}{n_G\gamma} - \psi}\|\Sigma_G\|_2$,*

$$\sum_{i: x_i \in G} \tau_i < \frac{1}{\gamma} \sum_{j=1}^{n} \tau_j,$$

  where $\psi = (\frac{1}{\sqrt{\frac{n}{n-n_G}} - 1})^2 < \frac{n}{n_G\gamma}$.

Proof. Let $\widehat{\theta}_G$ be the sample mean of points in G.

$$\begin{aligned}
\frac{1}{n_G} \sum_{i: x_i \in G} \tau_i &= \frac{1}{n_G} \sum_{i: x_i \in G} v^T(x_i - \widehat{\theta}_S)(x_i - \widehat{\theta}_S)^T v \\
&= v^T(\frac{1}{n_G} \sum_{i: x_i \in G} (x_i - \widehat{\theta}_G)(x_i - \widehat{\theta}_G)^T)v + (v^T(\widehat{\theta}_G - \widehat{\theta}_S))^2 \\
&\leq v^T \Sigma_G v + \|\widehat{\theta}_G - \widehat{\theta}_S\|_2^2 \\
&\leq v^T \Sigma_G v + \underbrace{(\frac{1}{\sqrt{\frac{n}{n-n_G}} - 1})^2}_{\psi} (\|\Sigma_S\|_2 + \|\Sigma_G\|_2) \\
&\leq \|\Sigma_G\|_2(1 + \psi) + \psi\|\Sigma_S\|_2
\end{aligned}$$

Now, if $\|\Sigma_S\|_2 \geq \frac{1+\psi}{\frac{n}{n_G\gamma}-\psi}\|\Sigma_G\|_2$, then we have that

$$\begin{aligned}
\frac{1}{n_G}\sum_{i:x_i\in G}\tau_i &\leq \frac{n}{n_G\gamma}\|\Sigma_S\|_2 \\
&= \frac{n}{n_G\gamma}\sum_{j=1}^{n}(v^T(x_j-\widehat{\theta}_S))^2 \\
\implies \sum_{i:x_i\in G}\tau_i &\leq \frac{1}{\gamma}\sum_{j=1}^{n}\tau_j
\end{aligned}$$

Claim 7. *Suppose P_1 is the equal weight discrete distribution with support on S^m, and P_2 is the equal weight discrete distribution with support on G^0. Then, when $\phi = \frac{n_{B^0}}{n}$ is such that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$,*

$$TV(P_1,P_2) \leq 8\phi + 36\frac{\log(1/\delta)}{n}$$

Proof. To bound the TV distance between P_1 and P_2, we use triangle inequality. Let P_3 be the equal weight discrete distribution with support on G^m. Let $\tau \in [T_\delta]$ be the number of "good" points thrown out in T_δ steps. For $\gamma = 3$, we have that,

$$T_\delta = 18\log(1/\delta) + 3n_{B^0}$$

$$\begin{aligned}
TV(P_1,P_2) &\leq TV(P_1,P_3) + TV(P_3,P_2) \\
&\leq \frac{n_{S^m}-n_{G^m}}{n_{S^m}} + \frac{n_{G^0}-n_{G^m}}{n_{G^0}} \\
&= \frac{n-T_\delta-(n-n_{B^0}-\tau)}{n-T_\delta} + \frac{\tau}{n-n_{B^0}} \\
&= \frac{n_{B^0}+\tau-T_\delta}{n-T_\delta} + \frac{\tau}{n-n_{B^0}} \\
&\leq \frac{n_{B^0}}{n-T_\delta} + \frac{T_\delta}{n-n_{B^0}} \\
&= \frac{\phi}{1-\frac{18\log(1/\delta)}{n}-3\phi} + \frac{\frac{18\log(1/\delta)}{n}+3\phi}{1-\phi}
\end{aligned}$$

where $\phi = \frac{n_{B^0}}{n}$. Now under the assumption that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$, the first term is less than 2ϕ. □

Claim 8. *Under the assumption that $\phi = \frac{n_{B0}}{n}$ is such that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$, and $2\phi < 0.12$, then when $\mathcal{E}^m$ is True, we have that,*

$$\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$$

Proof. Suppose P_1 is the equal weight discrete distribution with support on S^m and let P_3 be the equal weight discrete distribution with support on G^m. When $\mathcal{E}^m$ is True, we know by contrapositive of Lemma 51 that $\|\Sigma_{S^m}\|_2 \le \frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma}-\psi_m}\|\Sigma_{G^m}\|_2$, where $\psi_m = (\frac{\sqrt{TV(P_1,P_3)}}{1-\sqrt{TV(P_1,P_3)}})^2$. Note that for $TV(P_1, P_3) = \frac{n_{S_m}-n_{G^m}}{n_{S^m}}$. Hence, $\frac{n_{S^m}}{n_{G^m}\gamma} = \frac{1}{\gamma(1-TV(P_1,P_3))}$ For $\gamma = 3$, the term $\frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma}-\psi_m}$ can be rewritten solely as a function of the $TV(P_1, P_3)$. In particular, it can be written as

$$f(x) = \frac{\left(1+\left(\frac{x^{0.5}}{1\ -\ x^{0.5}}\right)^2\right)\left(3\left(1-x^{0.5}\right)^2\left(1+x^{(0.5)}\right)\right)}{1-x^{(0.5)}-3x-3x^{(1.5)}}$$

Now $TV(P_1, P_3) = \frac{n_{S^m}-n_{G^m}}{n_{S^m}} = \frac{(n-T_\delta)-(n-n_{B^0}-\tau)}{n-T_\delta} = \frac{n_{B^0}+\tau-T_\delta}{n-T_\delta} \le \frac{n_{B0}}{n-T_\delta} = \frac{\phi}{1-\frac{18\log(1/\delta)}{n}-3\phi}$. Hence, under our assumptions, $TV(P_1, P_3) < 0.12$. Some algebra shows that under $f(x)$ is monotonically increasing for $x < 0.12$, and in particular, $f(0.12) < 16$. Hence, we get that $\|\Sigma_{S^m}\|_2 \le 16\|\Sigma_{G^m}\|_2$. □

Claim 9. *Under the assumption that* $4\phi + 18\frac{\log(1/\delta)}{n} < \frac{1}{2}$, *we have that,*

$$\|\Sigma_{G^m}\|_2 \le 2\|\Sigma_{G^0}\|_2$$

Proof. We first show that $\|\Sigma_{G^m}\|_2 \le \frac{n_{G^0}}{n_{G^m}}\|\Sigma_{G^0}\|_2$.

$$\begin{aligned}
\Sigma_{G^0} &= \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i-\widehat{\theta}_{G^0})(x_i-\widehat{\theta}_{G^0})^T \\
&= \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i-\widehat{\theta}_{G^0})(x_i-\widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i\in G^m\}+\mathbb{I}\{x_i\notin G^m\}) \\
&= \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i-\widehat{\theta}_{G^0})(x_i-\widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i\in G^m\}) + \underbrace{\frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i-\widehat{\theta}_{G^0})(x_i-\widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i\notin G^m\})}_{T1} \\
&= \frac{n_{G^m}}{n_{G^0}}(\Sigma_{G^m}+(\widehat{\theta}_{G^m}-\widehat{\theta}_{G^0})(\widehat{\theta}_{G^m}-\widehat{\theta}_{G^0})^T)+T1
\end{aligned}$$

Now for v being the top eigenvector of Σ_{G^m}, we get that,

$$\frac{n_{G^m}}{n_{G^0}}v^T\Sigma_{G^m}v + \frac{n_{G^m}}{n_{G^0}}\underbrace{(v^T(\widehat{\theta}_{G^m}-\widehat{\theta}_{G^0}))^2}_{\ge 0} + \underbrace{v^T T1 v}_{\ge 0} = v^T\Sigma_{G^0}v$$

Hence, we get that,

$$\|\Sigma_{G^m}\|_2 \le \frac{n_{G^0}}{n_{G^m}}\|\Sigma_{G^0}\|_2,$$

Now,

$$\frac{n_{G^0}}{n_{G^m}} = \frac{n-n_{B^0}}{n-n_{B^0}-\tau} \le \frac{n-n_{B^0}}{n-n_{B^0}-T_\delta} = \frac{n-n_{B^0}}{n-18\log(1/\delta)-4n_{B^0}} = \frac{1-\phi}{1-18\frac{\log(1/\delta)}{n}-4\phi},$$

where $\phi = \frac{n_{B^0}}{n}$. Under our assumption, we get that, $\frac{n_{G^0}}{n_{G^m}} < 2$. □

□

Algorithm 11 Non Sample-Splitting Robust Gradient Descent

```
function RobustGD2(Initial Point $\theta^0$, Distance Estimate $R$, Gradient Estimator $g$, Data $\{z_i\}_{i=1}^n$, Cover Size $\psi$, Confidence Level $\delta$)
    Construct an $\psi$-cover of an $\ell_2$ ball of radius $R$, say $\mathcal{N}_\psi$.
    for $t = 1, \ldots, \infty$ do
        $\theta^{t+1} = \mathcal{P}_{\mathcal{N}_\psi}(\theta^t - \eta g(\theta^t, \mathcal{Z}, \delta))$
    end for
end function
```

B.17 Non sample-splitting Approach

In this section, we introduce a slight variant of our robust gradient algorithm(See Algorithm 11). which allows us to bypass sample-splitting at least theoretically. Our algorithm proceeds by constructing a covering of a certain euclidean ball, where the granularity of the covering is chosen to appropriately tradeoff estimation and approximation error.

Theorem 36. *Suppose that the gradient estimator satisfies the condition in (C.77) for the risk function $\mathcal{R} : \Theta \mapsto \mathbb{R}$. Then Algorithm 11 initialized at θ^0, with step-size $\eta = \frac{2}{\tau_\ell + \tau_u}$, upper bound on distance R, and cover-size ψ returns iterates $\{\widehat{\theta}\}_{t=1}^{\infty}$ such that with probability at least $1-\delta$, for κ in (3.13),*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{\eta\beta(n, \frac{\delta}{(R/\psi)^p}) + \psi}{1-\kappa}$$

.

Proof. Suppose we initialize at θ^0 and let θ^* be the true parameter. Let $\mathcal{N}_\psi$ be an ψ-cover of a ball of radius R centered around θ^0. Consider the following update rule,

$$\theta^{t+1} = \mathcal{P}_{\mathcal{N}_\epsilon}(\theta^t - \eta g(\theta^t, \{z_i\}_{i=1}^n)),$$

where $\mathcal{P}_{\mathcal{N}_\psi}$ is the projection operator on $\mathcal{N}_\psi$ and $g(\theta^t, \{z_i\}_{i=1}^n)$ is the output of the gradient estimator when called at θ^t using the data $\{z_i\}_{i=1}^n$. Note that the cardinality of the ψ-cover is upper bounded by $|\mathcal{N}_\psi|$. We know that at any point θ, the gradient estimator returns an estimate of the population gradient such that with probability at least $1-\delta$

$$\|g(\theta, \{z_i\}_{i=1}^n) - \nabla\mathcal{R}(\theta)\|_2 \leq \alpha(n, \delta)\|\theta - \theta^*\|_2 + \beta(n, \delta)$$

Since the cover is constructed independent of the data, hence, by a union bound, we get that with probability at least $1-\delta$

$$\|g(\theta, \{z_i\}_{i=1}^n) - \nabla\mathcal{R}(\theta)\|_2 \leq \alpha(n, \frac{\delta}{|\mathcal{N}_\psi|}) + \beta(n, \frac{\delta}{|\mathcal{N}_\psi|})\|\theta - \theta^*\|_2 \text{for all } \theta \in \mathcal{N}_\psi$$

For brevity, let $\gamma_t = \theta^t - \eta g(\theta^t, \{z_i\}_{i=1}^n)$. From proof of Theorem 1, we know that,

$$\|\gamma_t - \theta^*\|_2 \leq \kappa\|\theta^t - \theta^*\|_2 + \eta\beta(n, \frac{\delta}{|\mathcal{N}_\psi|}),$$

The projection operator returns a θ^{t+1} such that $\|\theta^{t+1} - \gamma_t\|_2 \leq \psi$. Hence, we get that,

$$\|\theta^{t+1} - \theta^*\|_2 \leq \kappa\|\theta^t - \theta^*\|_2 + \eta\beta(n, \frac{\delta}{|\mathcal{N}_\epsilon|}) + \psi,$$

Hence, we get that,

$$\|\theta^T - \theta^*\|_2 \leq \kappa^T\|\theta^0 - \theta^*\|_2 + \frac{(\eta\beta(n, \frac{\delta}{|\mathcal{N}_\epsilon|}) + \psi)}{1-\kappa},$$

Recall that $\mathcal{N}_\psi$ is an ψ-covering of an ℓ_2 ball of radius $R = \|\theta^0 - \theta^*\|_2$. Hence, we have that $\log(|\mathcal{N}_\psi|) = O(p\log(R/\psi)$. □

Observe the tradeoff between $\beta(n, \frac{\delta}{|\mathcal{N}_\psi|})$ and ψ. As the cover becomes finer, ψ decreases but $\beta(n, \frac{\delta}{|\mathcal{N}_\psi|})$ increases. Hence, for our corollaries we will set ψ to balance the two terms. To see a concrete instantiation of this algorithm, consider the following corollary for Linear Regression in the ϵ-contamination model.

B.17.1 Linear Regression

Here we observe paired samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$. We assume that the (x, y) pairs sampled from the true distribution P are linked via a linear model:

$$y = x^T\theta^* + w, \tag{B.17}$$

where w is drawn from a zero-mean distribution such as normal distribution with variance σ^2 ($\mathcal{N}(0, \sigma^2)$) or a more heavy-tailed distribution such as student-t or Pareto distribution. We suppose that under P the covariates $x \in \mathbb{R}^p$, have mean 0, and covariance Σ.

For this setting we use the squared loss as our loss function, which induces the following population risk:

$$\bar{\mathcal{L}}(\theta; (x, y)) = \frac{1}{2}\left(y - \langle x, \theta\rangle\right)^2, \quad \text{and} \quad \mathcal{R}(\theta) = \frac{1}{2}(\theta - \theta^*)^T\Sigma(\theta - \theta^*).$$

Note that the true parameter θ^* is the minimizer of the population risk $\mathcal{R}(\theta)$. The strong-convexity and smoothness assumptions in this setting require that $\tau_\ell \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \tau_u$.

Corollary 37 (Robust Linear Regression). *Consider the statistical model in* (B.17)*, and suppose that the number of samples n is large enough such that $\widetilde{\gamma}(n, p, \delta, \epsilon) < \frac{C_1\tau_\ell}{\|\Sigma\|_2\sqrt{\log p}}$ and the contamination level is such that*

$$\epsilon < \left(\frac{C_2\tau_\ell}{\|\Sigma\|_2\sqrt{\log p}} - \widetilde{\gamma}(n, p, \epsilon, \delta)\right)^2,$$

for some constants C_1 and C_2. Then, there are universal constants C_3, C_4, such that if Algorithm 11 is initialized at 0 with stepsize $\eta = 2/(\tau_u + \tau_\ell)$, cover size $\psi = \sigma\sqrt{\epsilon\|\Sigma\|_2\log p}$, *distance estimate $R = \|\theta^*\|_2$ and the dimension halving estimator of [28] as gradient estimator, then it returns iterates $\{\widehat{\theta}^t\}_{t=1}^{\infty}$ such that for a contraction parameter $\kappa < 1$, with probability at least $1 - \delta$,*

$$\|\theta^t - \theta^*\|_2 \le \kappa^t\|\theta^*\|_2 + \frac{1}{1-\kappa}(2\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon} + \sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\widetilde{\gamma}(n,p,\delta,\epsilon)), \tag{B.18}$$

where

$$\begin{aligned}\widetilde{\gamma}(n,p,\delta,\epsilon) = &\Big(\frac{p^2\log p\log(\frac{\|\theta^*\|_2}{\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}})}{n} + \frac{p\log p\log\big(n/(p\delta)\big)}{n}\Big)^{3/8}\\ &+ \Big(\frac{\epsilon p^3\log p\log(\frac{\|\theta^*\|_2}{\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}})}{n} + \frac{\epsilon p^2\log p\log\big(\frac{p\log(p)}{\delta}\big)}{n}\Big)^{1/4}.\end{aligned}$$

Before giving a detailed proof, we remark on the main differences between this result and the result in Theorem 2 of our paper. This result applies to an algorithm which does not use sample-splitting, and provides a similar guarantee as in Theorem 2, i.e. that the error of robust GD decreases linearly up to an error floor roughly determined by ϵ and $\widetilde{\gamma}$. However, $\widetilde{\gamma}$ in this result is worse by a factor of (at most) $p^{3/8}$ from the corresponding term in Theorem 2, indicating the statistical price for requiring uniform control of the distance between the sample and population gradients over the entire ψ-cover.

Proof. We initialize the algorithm θ^0 at 0, and suppose we know the signal strength $\|\theta^*\|_2$. Recall that the gradient estimator of [28] satisfies that with probability $1 - \delta$,

$$\|g(\theta, \mathcal{Z}_n, \delta) - \nabla\mathcal{R}(\theta)\|_2 \le C_1(\sqrt{\epsilon} + \gamma(n,p,\delta,\epsilon))\sqrt{\log p}\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta,z))\|_2^{\frac{1}{2}},$$

where $\gamma(n,p,\delta,\epsilon) := \Big(\frac{p\log p\log\big(n/(p\delta)\big)}{n}\Big)^{3/8} + \Big(\frac{\epsilon p^2\log p\log\big(\frac{p\log(p)}{\delta}\big)}{n}\Big)^{1/4}$. Moreover, recall that for linear regression we have that,

$$\begin{aligned}\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta,z))\|_2 &\le \sigma^2\|\Sigma\|_2 + C_4\|\Delta\|_2^2\|\Sigma\|_2^2\\ \alpha(n,\delta) &\le (\sqrt{\epsilon} + \gamma(n,p,\delta))\sqrt{\log p}\|\Sigma\|_2\\ \beta(n,\delta) &\le (\sqrt{\epsilon} + \gamma(n,p,\delta))\sqrt{\log p}\sigma\sqrt{\|\Sigma\|_2}\end{aligned}$$

Setting $\psi = \sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}$, we get that $\log(|\mathcal{N}_\psi|) \le Cp\log(\frac{\|\theta^*\|_2}{\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}})$. Note that γ has only a logarithmic dependence on the confidence level. Hence, we only get hit by a logarithmic term on the size of the covering. In particular, let $\widetilde{\gamma}(n,p,\delta,\epsilon) := \gamma(n,p,\frac{\delta}{|\mathcal{N}_\psi|},\epsilon)$, then we have that

$$\begin{aligned}\widetilde{\gamma}(n,p,\delta,\epsilon) = &\Big(\frac{p^2\log p\log(\frac{\|\theta^*\|_2}{\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}})}{n} + \frac{p\log p\log\big(n/(p\delta)\big)}{n}\Big)^{3/8}\\ &+ \Big(\frac{\epsilon p^3\log p\log(\frac{\|\theta^*\|_2}{\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon}})}{n} + \frac{\epsilon p^2\log p\log\big(\frac{p\log(p)}{\delta}\big)}{n}\Big)^{1/4}\end{aligned}$$

As stated above that, we get hit by only an additional multiplicative factor of p, and a logarithmic factor of $\|\theta^*\|_2$.

$$\beta(n, \frac{\delta}{|\mathcal{N}_\psi|}) \leq \sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon} + \sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\widetilde{\gamma}(n, p, \delta, \epsilon).$$

Pluging the values of ψ and $\beta(n, \frac{\delta}{|\mathcal{N}_\psi|})$ into Theorem 36, we get that with probability at least $1 - \delta$,

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t\|\theta^*\|_2 + \frac{1}{1-\kappa}(2\sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\sqrt{\epsilon} + \sigma\sqrt{\|\Sigma\|_2}\sqrt{\log p}\widetilde{\gamma}(n, p, \delta, \epsilon))$$

□

Appendix C

Supplementary Material for Chapter 5

C.1 Proof of Theorems and Lemmas in Section 5.2

C.1.1 Proof of Lemma 48

For completeness, we first restate Lemma 48 for sake of completeness.

Lemma 48. *Let S be any arbitrary collection of points, and let $G^0 \subset S$ be an unknown subset of size n_G^0 such that $8\frac{n-n_{G^0}}{n}+36\frac{\log(1/\delta)}{n} < \frac{1}{4}$. Then, when Algorithm 7 is run for $T^* = \lceil 3(n-n_{G^0}) + 18\log(1/\delta)\rceil$ steps on S, it returns an estimate $\widehat{\theta}_\delta$ such that with probability at least $1-\delta$,*

$$\Big\|\widehat{\theta}_\delta - \frac{1}{n_{G^0}}\sum_{x_i \in G^0} x_i\Big\|_2 \lesssim \|\Sigma_{G^0}\|_2^{\frac{1}{2}}\Big(\frac{n-n_{G^0}}{n}+\frac{\log(1/\delta)}{n}\Big)^{\frac{1}{2}},$$

where Σ_{G^0} is the covariance of the unknown subset of points.

Proof. Our proof is split into two keys Lemmas. Firstly, in Lemma 49, we show that the with probability at least $1-\delta$, when the algorithm terminates after $T^*_\delta = \lceil 18\log(1/\delta) + 3(n-n_{G^0})\rceil$, then the covariance of the remaining samples is well-behaved. Finally, in Lemma 50 we show that under our assumptions that $8\frac{n-n_{G^0}}{n}+36\frac{\log(1/\delta)}{n} < \frac{1}{4}$, when the algorithm stops after T^*_δ steps, the sample mean of points, $\widehat{\theta}_{S^{T^*_\delta}}$ is close to the mean of G^0. In particular, we show that

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^{T^*_\delta}}\|_2 \le C_1\big(8\frac{n-n_{G^0}}{n}+36\frac{\log(1/\delta)}{n}\big)^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}}, \tag{C.1}$$

which recovers the statement of the Lemma.

Lemma 49. *When Algorithm 7 is instantiated on S^0 for $T^*_\delta = \lceil 18\log(1/\delta) + 3(n-n_{G^0}\rceil$ steps, then with probability $1-\delta$,*

$$\|\Sigma_{S^{T^*}})\|_2 \le C_2\|\Sigma_{G^0}\|_2$$

Proof. At each step of Algorithm 7, we remove one sample based on the probability distribution of the scores. Let $l = 1, 2, \ldots, n$ be the steps of the algorithm. Note that the steps of the Algorithm are dependent, hence to obtain a high probability statement, we will have to use martingale style analysis. The martingale analysis in the proof mostly follows from [119, 120].

Let $\mathcal{F}^l$ be the filtration generated by the sets of events until step l. At step l, let S^l be the set of samples, G^l be the subset of G^0 stil in S^l, *i.e.* $\{x_i \in S^l \cap G^0\}$. Let $B^l = S^l \backslash G^l$ be the remaining samples. Note that $|S^l| = n_l = n - l$, and $S^l, G^l, B^l \in \mathcal{F}^l$.

Let τ_i be some score for each point. Define $\mathcal{E}^l$ be an event variable at step l which is True if

$$\sum_{i \in G^l} \tau_i \geq \frac{1}{(\gamma - 1)} \sum_{j \in B^l} \tau_j, \equiv \sum_{i \in G^l} \tau_i \geq \frac{1}{\gamma} \sum_{j \in S^l} \tau_j$$

for say $\gamma = 3$. Intuitively, this means the event is true when the sum of the scores of the good points is larger compared to the bad points. Now, when $\mathcal{E}^l$ is false, we sample a point j according τ_j and remove it. Some algebra shows, that when $\mathcal{E}^l$ is false, then with constant probability of 2/3, we throw a point from B^l.

$$\Pr(\text{sample removed at Step } l \in B^l | \mathcal{F}^l) = \frac{\sum\limits_{i \in B^l} \tau_i}{\sum_{j \in S^l} \tau_j} \geq \frac{\gamma - 1}{\gamma} = 2/3$$

Essentially, our argument shows that whenever $\mathcal{E}^l$ is false, then we are more likely to throw a point from the bad set. This means, that in the next iteration the fraction of bad points will reduce. To argue more formally, let $T \overset{\text{def}}{=} \min\{l : \mathcal{E}^l \text{ is true}\}$ be the first time that $\mathcal{E}^l$ is True. Then, our goal is to show that T is small.

To show this, based on T, define Y^l, as

$$Y^l = \begin{cases} |B^{T-1}| + \frac{\gamma-1}{\gamma}(T - 1), \text{ if } l \geq T \\ |B^l| + \frac{\gamma-1}{\gamma} l, \text{ if } l < T \end{cases}$$

Now, we show that $\{Y^l, \mathcal{F}^l\}$ is a supermartingale, *i.e.* $\mathbb{E}[Y^l | \mathcal{F}^{l-1}] \leq Y^{l-1}$. To see this, we split it into three cases:

- **Case 1.** $l < T$. This means that $\mathcal{E}^l$ is false.

$$Y^l - Y^{l-1} = |B^l| - |B^{l-1}| + \frac{\gamma - 1}{\gamma}, \tag{C.2}$$

Now, $|B^l| = |B^{l-1}|$ if no bad point is thrown, and $|B^l| = |B^{l-1}| - 1$ if the point thrown is bad. Since, $\mathcal{E}^{l-1}$ is false, hence, we have that,

$$\mathbb{E}[Y^l - Y^{l-1} | \mathcal{F}^{l-1}] = -1(\Pr(\text{sample removed at Step } l - 1 \in B^{l-1})) + \frac{\gamma - 1}{\gamma} \overset{(i)}{\leq} 0$$

where (i) is true because $\mathcal{E}^{l-1}$ is false.

- **Case 2.** $l = T$, This follows by construction, because at $l = T$, $Y^l = Y^{l-1}$.
- **Case 3.** $l > T$, This also follows by construction.

So, we have that $Y^l, \mathcal{F}^l$ is a supermartingale. Now, we need to bound the steps T_δ such that the probability that the algorithm doesn't stop in T_δ steps is less than δ, *i.e.*

$$\Pr(\bigcap_{l=1}^{T_\delta}(\mathcal{E}^l)^c) \leq \delta$$

Note, that,

$$\Pr(\bigcap_{l=1}^{T_\delta}(\mathcal{E}^l)^c) = \Pr(T \geq T_\delta) \overset{(ii)}{\leq} \Pr(Y^{T_\delta} \geq \frac{\gamma-1}{\gamma}T_\delta) \tag{C.3}$$

where (ii) follows because, if $T > T_\delta \implies Y^{T_\delta} = |B^{T_\delta}| + \frac{\gamma-1}{\gamma}T_\delta \geq \frac{\gamma-1}{\gamma}T_\delta$. Now,

$$\Pr(Y^{T_\delta} \geq \frac{\gamma-1}{\gamma}T_\delta) = \Pr(Y^{T_\delta} - Y^0 \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0)$$

Now, defining $D^l = Y^l - Y^{l-1}$, and let $Z^l = D^l - \mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}]$. Then,

$$Y^{T_\delta} - Y^0 = \sum_{l=1}^{T_\delta} D^l = \sum_{l=1}^{T_\delta} Z^l + \sum_{l=1}^{T_\delta} \mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}]$$

Since, we know that $\{Y^l, \mathcal{F}^l\}$ is a supermartingale, hence the difference process is such that

$$\mathbb{E}[D^l|D^1, D^2, \ldots, D^{l-1}] \leq 0$$

This implies that

$$Y^{T_\delta} - Y^0 \leq \sum_{l=1}^{T_\delta} Z^l \implies \Pr(Y^{T_\delta} - Y^0 \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0) \leq \Pr(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0)$$

Since, $|D^l| \leq 1$, and $Z^l \leq 2$ are bounded, hence we can use Azuma-Hoeffding to bound the above probability. In particular,

$$\Pr(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma-1}{\gamma}T_\delta - Y_0) \leq \exp(-\frac{(\frac{\gamma-1}{\gamma}T_\delta - Y_0)^2}{8T_\delta})$$

Now, we want a T_δ such that, $\exp(-\frac{(\frac{\gamma-1}{\gamma}T_\delta - Y_0)^2}{8T_\delta}) \leq \delta$. Solving the quadratic, we need a T_δ such that,

$$(\frac{\gamma-1}{\gamma})^2 T_\delta^2 - (8\log(1/\delta) + 2Y^0\frac{\gamma-1}{\gamma})T_\delta + Y_0^2 \geq 0$$

Some algebra shows that $T^*_\delta = \left\lceil 8\log(1/\delta)\frac{\gamma^2}{(\gamma-1)^2} + 2Y^0\frac{\gamma}{\gamma-1}\right\rceil$ satisfies the above equation. Hence, we know that with probability at least $1-\delta$, there exists at least one good event in 1 to T^*_δ iterations. Note than $Y^0 = n_{B^0} = n - n_{G^0}$.

While we have established that there is at least one good event in 1 to T^*_δ iterations, suppose $m \in [1, T_{\delta^*}]$ is the first index such that $\mathcal{E}^m$ is true. Next, we establish a series of deterministic results.

- When $\mathcal{E}^m$ is True, then $\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$(See Claim 12).
- Coupling this with Claim 11, which shows that $\|\Sigma_{G^m}\|_2 \leq 2\|\Sigma_{G^0}\|_2$, we get that $\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$.
- Hence, we have that with probability $1-\delta$, there exists a point in time $m \in [1, T_\delta]$ such that,
$$\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$$
- Now, observe that $S^{T^*} \subseteq S^m$, *i.e.* the final returned set of points is a subset of the points at m. Claim 13 shows that the covariance at S^{T^*} is such that $\|\Sigma_{S^{T^*}}\|_2 \leq \frac{n-m}{n-T^*}\|\Sigma_{S^m}\|_2 \leq C_1\|\Sigma_{S^m}\|_2$.

Chaining the above arguments shows that $\|\Sigma_{T^*}\|_2 \leq C\|\Sigma_{G^0}\|_2$. □

Next, we state and prove Lemma 50. Recall that $\mathcal{E}^l$ is defined to be an event variable at step l which is True if

$$\sum_{i\in G^l} \tau_i \geq \frac{1}{(\gamma-1)}\sum_{j\in B^l}\tau_j, \equiv \sum_{i\in G^l}\tau_i \geq \frac{1}{\gamma}\sum_{j\in S^l}\tau_j,$$

where S^l is set of samples at step l, and $G^l = \{x_i \in S^l \cap G^0\}$ is the subset of samples from G^0 which are still in S^l. Also, recall that for Algorithm 7, the sampling weights τ_i at any step ℓ are defined as $\tau_i = (v^T(x_i - \widehat{\theta}_{S^l}))^2$, where v is the top unit-norm eigenvector of $\widehat{\Sigma}_{S^l}$ and $\widehat{\theta}_{S^l}$ is the sample mean of S^l. Then, in Lemma 49 we showed that with probability $1-\delta$,

$$\|\Sigma_{S^{T^*_\delta}}\|_2 \leq C_2\|\Sigma_{G^0}\|_2.$$

Lemma 50. *Let $\phi = \frac{n-n_{G^0}}{n}$. Then, under the assumption that $8\phi + 36\frac{\log(1/\delta)}{n} < \frac{1}{4}$, we have that for $m = T^*_\delta$*

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq 10\sqrt{2}(8\phi + 36\frac{\log(1/\delta)}{n})^{\frac{1}{2}}\|\Sigma_{G^0}\|_2^{\frac{1}{2}},$$

Proof. Using Lemma 52, we get that,

$$\|\widehat{\theta}_{G^0} - \widehat{\theta}_{S^m}\|_2 \leq \frac{\sqrt{TV(P_1,P_2)}}{1-\sqrt{TV(P_1,P_2)}}(\|\Sigma_{G^0}\|_2^{\frac{1}{2}} + \|\Sigma_{S^m}\|_2^{\frac{1}{2}}),$$

where P_1 is the equal weight discrete distribution with support on S^m, and P_2 is the equal weight discrete distribution with support on G^0. Lemma 49 already controls tell us that for $m = T^*_\delta$, $\|\Sigma_{S^m}\|_2 \leq C_2\|\Sigma_{G^0}\|_2$. We show next that

$$TV(P_1,P_2) \leq 8\phi + 36\frac{\log(1/\delta)}{n},$$

which finishes the proof of the Lemma.

To bound the TV distance between P_1 and P_2, we use triangle inequality. Let P_3 be the equal weight discrete distribution with support on G^m. Let $\tau \in [1, m] \leq T_\delta$ be the number of "good" points thrown out in $m \leq T_\delta$ steps. For $\gamma = 3$, we have that,

$$T_\delta = 18 \log(1/\delta) + 3n_{B^0}$$

$$\begin{aligned}
TV(P_1, P_2) &\leq TV(P_1, P_3) + TV(P_3, P_2) && \text{(C.4)}\\
&\leq \frac{n_{S^m} - n_{G^m}}{n_{S^m}} + \frac{n_{G^0} - n_{G^m}}{n_{G^0}} && \text{(C.5)}\\
&= \frac{n - T_\delta - (n - n_{B^0} - \tau)}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} && \text{(C.6)}\\
&= \frac{n_{B^0} + \tau - T_\delta}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} && \text{(C.7)}\\
&\leq \frac{n_{B^0}}{n - T_\delta} + \frac{T_\delta}{n - n_{B^0}} && \text{(C.8)}\\
&= \frac{\phi}{1 - \frac{18\log(1/\delta)}{n} - 3\phi} + \frac{\frac{18\log(1/\delta)}{n} + 3\phi}{1 - \phi} && \text{(C.9)}
\end{aligned}$$

where $\phi = \frac{n_{B^0}}{n}$. Now under the assumption that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$, the first term is less than 2ϕ.

□

Auxillary Results for Proof of Theorem 48

Lemma 51. *Let S be a collection of n points. And let G be a subset of S containing n_G points. Define $\tau_i = (v^T(x_i - \widehat{\theta}_S))^2$, where v is the top unit-norm eigenvector of $\widehat{\Sigma}_S$ and $\widehat{\theta}_S$ is the sample mean of S. Let $\lambda = \|\Sigma_S\|_2$. Then, we have the following*

- *If $\lambda > \frac{1+\psi}{\frac{n}{n_G\gamma} - \psi} \|\Sigma_G\|_2$,*

$$\sum_{i:x_i \in G} \tau_i < \frac{1}{\gamma} \sum_{j=1}^{n} \tau_j,$$

where $\psi = (\frac{1}{\sqrt{\frac{n}{n-n_G}} - 1})^2 < \frac{n}{n_G\gamma}$.

Proof. Let $\widehat{\theta}_G$ be the sample mean of points in G.

$$\begin{aligned}
\frac{1}{n_G}\sum_{i:x_i\in G}\tau_i &= \frac{1}{n_G}\sum_{i:x_i\in G} v^T(x_i-\widehat{\theta}_S)(x_i-\widehat{\theta}_S)^T v && \text{(C.10)}\\
&= v^T(\frac{1}{n_G}\sum_{i:x_i\in G}(x_i-\widehat{\theta}_G)(x_i-\widehat{\theta}_G)^T)v + (v^T(\widehat{\theta}_G-\widehat{\theta}_S))^2 && \text{(C.11)}\\
&\le v^T\Sigma_G v + \|\widehat{\theta}_G-\widehat{\theta}_S\|_2^2 && \text{(C.12)}\\
&\le v^T\Sigma_G v + \underbrace{(\frac{1}{\sqrt{\frac{n}{n-n_G}}-1})}_{\psi}{}^2(\|\Sigma_S\|_2+\|\Sigma_G\|_2) && \text{(C.13)}\\
&\le \|\Sigma_G\|_2(1+\psi)+\psi\|\Sigma_S\|_2 && \text{(C.14)}
\end{aligned}$$

Now, if $\|\Sigma_S\|_2 \ge \frac{1+\psi}{\frac{n}{n_G\gamma}-\psi}\|\Sigma_G\|_2$, then we have that

$$\begin{aligned}
\frac{1}{n_G}\sum_{i:x_i\in G}\tau_i &\le \frac{n}{n_G\gamma}\|\Sigma_S\|_2 && \text{(C.15)}\\
&= \frac{n}{n_G\gamma}\sum_{j=1}^{n}(v^T(x_j-\widehat{\theta}_S))^2 && \text{(C.16)}\\
\implies \sum_{i:x_i\in G}\tau_i &\le \frac{1}{\gamma}\sum_{j=1}^{n}\tau_j && \text{(C.17)}
\end{aligned}$$

□

Claim 10. *Suppose P_1 is the equal weight discrete distribution with support on S^m, and P_2 is the equal weight discrete distribution with support on G^0. Then, when $\phi = \frac{n_{B^0}}{n}$ is such that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$,*

$$TV(P_1,P_2) \le 8\phi + 36\frac{\log(1/\delta)}{n}$$

Proof. To bound the TV distance between P_1 and P_2, we use triangle inequality. Let P_3 be the equal weight discrete distribution with support on G^m. Let $\tau \in [T_\delta]$ be the number of "good" points thrown out in T_δ steps. For $\gamma = 3$, we have that,

$$T_\delta = 18\log(1/\delta) + 3n_{B^0}$$

$$
\begin{aligned}
TV(P_1, P_2) &\leq TV(P_1, P_3) + TV(P_3, P_2) && \text{(C.18)}\\
&\leq \frac{n_{S^m} - n_{G^m}}{n_{S^m}} + \frac{n_{G^0} - n_{G^m}}{n_{G^0}} && \text{(C.19)}\\
&= \frac{n - T_\delta - (n - n_{B^0} - \tau)}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} && \text{(C.20)}\\
&= \frac{n_{B^0} + \tau - T_\delta}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} && \text{(C.21)}\\
&\leq \frac{n_{B^0}}{n - T_\delta} + \frac{T_\delta}{n - n_{B^0}} && \text{(C.22)}\\
&= \frac{\phi}{1 - \frac{18\log(1/\delta)}{n} - 3\phi} + \frac{\frac{18\log(1/\delta)}{n} + 3\phi}{1 - \phi} && \text{(C.23)}
\end{aligned}
$$

where $\phi = \frac{n_{B^0}}{n}$. Now under the assumption that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$, the first term is less than 2ϕ. □

Lemma 52. *[29] Given a collection of points S of size n. Let P_1 and P_2 be discrete empirical distributions on n. Then, we have that,*

$$
\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2[x_i]}\|_2 \leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} (\|\widehat{\Sigma}_{P_1}\|_2^{\frac{1}{2}} + \|\widehat{\Sigma}_{P_2}\|_2^{\frac{1}{2}}) \tag{C.24}
$$

where $\widehat{\Sigma}_{P_1}$ is the covariance matrix when $x_i \sim P_1$, and $\widehat{\Sigma}_{P_2}$ is the empirical covariance matrix of when $x_i \sim P_2$

Proof. Consider a joint distribution(also called coupling) $\omega^*(z, z')$ over $S \times S$ such that it's individual marginal distributions are equal to P_1 and P_2; *i.e.* $\omega(z) = P_1$ and $\omega(z') = P_2$ and $\omega(z \neq z') = TV(P_1, P_2)$. Then, we have that

$$
\begin{aligned}
&\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2[x_i]}\|_2 = \sup_{v \in \mathcal{S}^{p-1}} |\langle v, \mathbb{E}_{w^*}[z - z'] \rangle| && \text{(C.25)}\\
&\leq \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[|\langle v, z - z' \rangle|] && \text{(C.26)}\\
&\leq \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[1(z \neq z') \langle v, z - z' \rangle|] && \text{(C.27)}\\
&\leq (\mathbb{E}_{w^*}[(1(z \neq z'))^{1/(1-\frac{1}{2})}])^{1-\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - z' \rangle)^2]^{\frac{1}{2}} && \text{(C.28)}\\
&\leq TV(P_1, P_2)^{\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} (\mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] + \mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i] + \mathbb{E}_{x_i \sim P_2}[x_i] - z' \rangle)^2]^{\frac{1}{2}}) && \text{(C.29)}\\
&\leq TV(P_1, P_2)^{\frac{1}{2}} (\sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] \rangle)^2]^{\frac{1}{2}} + \|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i]\|_2)\\
&+ TV(P_1, P_2)^{\frac{1}{2}} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*}[(\langle v, z - \mathbb{E}_{x_i \sim P_2}[x_i] \rangle)^2]^{\frac{1}{2}} && \text{(C.30)}\\
&\leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} \left(\|\Sigma_{P_1}\|_2^{\frac{1}{2}} + \|\Sigma_{P_2}\|_2^{\frac{1}{2}}\right) && \text{(C.31)}
\end{aligned}
$$

□

Claim 11. *Under the assumption that $4\phi + 18\frac{\log(1/\delta)}{n} < \frac{1}{2}$, we have that,*

$$\|\Sigma_{G^m}\|_2 \le 2\|\Sigma_{G^0}\|_2$$

Proof. We first show that $\|\Sigma_{G^m}\|_2 \le \frac{n_{G^0}}{n_{G^m}}\|\Sigma_{G^0}\|_2$.

$$\Sigma_{G^0} = \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T \tag{C.32}$$

$$= \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i \in G^m\} + \mathbb{I}\{x_i \notin G^m\}) \tag{C.33}$$

$$= \frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i \in G^m\}) + \underbrace{\frac{1}{n_{G^0}}\sum_{i\in G^0}(x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T(\mathbb{I}\{x_i \notin G^m\})}_{T1} \tag{C.34}$$

$$= \frac{n_{G^m}}{n_{G^0}}(\Sigma_{G^m} + (\widehat{\theta}_{G^m} - \widehat{\theta}_{G^0})(\widehat{\theta}_{G^m} - \widehat{\theta}_{G^0})^T) + T1 \tag{C.35}$$

Now for v being the top eigenvector of Σ_{G^m}, we get that,

$$\frac{n_{G^m}}{n_{G^0}}v^T\Sigma_{G^m}v + \frac{n_{G^m}}{n_{G^0}}\underbrace{(v^T(\widehat{\theta}_{G^m} - \widehat{\theta}_{G^0}))^2}_{\ge 0} + \underbrace{v^T T1 v}_{\ge 0} = v^T\Sigma_{G^0}v$$

Hence, we get that,

$$\|\Sigma_{G^m}\|_2 \le \frac{n_{G^0}}{n_{G^m}}\|\Sigma_{G^0}\|_2,$$

Now,

$$\frac{n_{G^0}}{n_{G^m}} = \frac{n - n_{B^0}}{n - n_{B^0} - \tau} \le \frac{n - n_{B^0}}{n - n_{B^0} - T_\delta} = \frac{n - n_{B^0}}{n - 18\log(1/\delta) - 4n_{B^0}} = \frac{1-\phi}{1 - 18\frac{\log(1/\delta)}{n} - 4\phi},$$

where $\phi = \frac{n_{B^0}}{n}$. Under our assumption, we get that, $\frac{n_{G^0}}{n_{G^m}} < 2$. □

Claim 12. *Under the assumption that $\phi = \frac{n_{B^0}}{n}$ is such that $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$, and $2\phi < 0.12$, then when $\mathcal{E}^m$ is True, we have that,*

$$\|\Sigma_{S^m}\|_2 \le 16\|\Sigma_{G^m}\|_2$$

Proof. Suppose P_1 is the equal weight discrete distribution with support on S^m and let P_3 be the equal weight discrete distribution with support on G^m. When $\mathcal{E}^m$ is True, we know by contrapositive of Lemma 51 that $\|\Sigma_{S^m}\|_2 \le \frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma} - \psi_m}\|\Sigma_{G^m}\|_2$, where $\psi_m = (\frac{\sqrt{TV(P_1,P_3)}}{1-\sqrt{TV(P_1,P_3)}})^2$. Note that for $TV(P_1, P_3) = \frac{n_{S_m} - n_{G^m}}{n_{S^m}}$. Hence, $\frac{n_{S^m}}{n_{G^m}\gamma} = \frac{1}{\gamma(1-TV(P_1,P_3))}$ For $\gamma = 3$, the term

$\frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma}-\psi_m}$ can be rewritten solely as a function of the $TV(P_1, P_3)$. In particular, it can be written as

$$f(x) = \frac{\left(1 + \left(\frac{x^{0.5}}{1\ -\ x^{0.5}}\right)^2\right)\left(3\left(1 - x^{0.5}\right)^2\left(1 + x^{(0.5)}\right)\right)}{1 - x^{(0.5)} - 3x - 3x^{(1.5)}}$$

Now $TV(P_1, P_3) = \frac{n_{S^m} - n_{G^m}}{n_{S^m}} = \frac{(n-T_\delta)-(n-n_{B^0}-\tau)}{n-T_\delta} = \frac{n_{B^0}+\tau-T_\delta}{n-T_\delta} \leq \frac{n_{B^0}}{n-T_\delta} = \frac{\phi}{1-\frac{18\log(1/\delta)}{n}-3\phi}$. Hence, under our assumptions, $TV(P_1, P_3) < 0.12$. Some algebra shows that under $f(x)$ is monotonically increasing for $x < 0.12$, and in particular, $f(0.12) < 16$. Hence, we get that $\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$. □

Claim 13. *Let S_1 be any collection of points of size n_1. Let $S_2 \subseteq S_1$ be a subset of size $n_2 \leq n_1$. Then, we have that*

$$\|\Sigma_{S_2}\|_2 \leq \frac{n_1}{n_2}\|\Sigma_{S_1}\|_2$$

Proof. Let $\widehat{\theta}_{S_2}$ be the mean of points in S_2. Similarly, let $\widehat{\theta}_{S_1}$ be mean of points in S_1.

$$\begin{aligned}
\Sigma_{S_1} &= \frac{1}{n_1}\sum_{i\in S_1}(x_i - \widehat{\theta}_{S_1})(x_i - \widehat{\theta}_{G^0})^T && \text{(C.36)}\\
&= \frac{1}{n_1}\sum_{i\in S_1}(x_i - \widehat{\theta}_{S_1})(x_i - \widehat{\theta}_{S_1})^T(\mathbb{I}\{x_i \in S_2\} + \mathbb{I}\{x_i \notin S_2\}) && \text{(C.37)}\\
&= \frac{1}{n_1}\sum_{i\in S_1}(x_i - \widehat{\theta}_{S_1})(x_i - \widehat{\theta}_{S_1})^T(\mathbb{I}\{x_i \in S_2\}) + \underbrace{\frac{1}{n_1}\sum_{i\in S_1}(x_i - \widehat{\theta}_{S_1})(x_i - \widehat{\theta}_{S_1})^T(\mathbb{I}\{x_i \notin S_2\})}_{T1} && \text{(C.38)}\\
&= \frac{n_2}{n_1}(\Sigma_{S_2} + (\widehat{\theta}_{S_2} - \widehat{\theta}_{S_1})(\widehat{\theta}_{S_2} - \widehat{\theta}_{S_1})^T) + T1 && \text{(C.39)}
\end{aligned}$$

Now for v being the top eigenvector of Σ_{S_2}, we get that,

$$\frac{n_2}{n_1}v^T\Sigma_{S_2}v + \frac{n_{S_2}}{n_{S_1}}\underbrace{(v^T(\widehat{\theta}_{S_2} - \widehat{\theta}_{S_1}))^2}_{\geq 0} + \underbrace{v^T T1 v}_{\geq 0} = v^T\Sigma_{S_1}v$$

□

□

C.1.2 Proof of Lemma 15

For sake of completeness, we restate the complete Lemma. In particular, given n-samples from a distribution P, we define a *good point selector* $\mathcal{O} : \mathbb{R}^p \mapsto \{0, 1\}$ by

$$\mathcal{O}(x) = \mathbb{I}\left\{\|x - \mu(P)\|_2 \leq R\right\}, \tag{C.40}$$

and let $G = \{x_i | \mathcal{O}(x_i) = 1\}$ to be the set of points chosen by $\mathcal{O}$. Note that this (unknown) subset of points chosen by the ℓ_2-radius based point selector, is precisely our unknown subset from the previous subsection. Let

$$\widehat{\mu}_n = \Big(\sum_{i=1}^{n} \mathcal{O}(x_i)\Big)^{-1} \sum_{i=1}^{n} x_i \mathcal{O}(x_i),$$

be the sample mean of the points within the subsets, and let

$$\widehat{\Sigma}_n^{\mathcal{O}} = (\sum_{i=1}^{n} \mathcal{O}(x_i))^{-1} \sum_{i=1}^{n} (x_i - \widehat{\mu}_n)(x_i - \widehat{\mu}_n)^T \mathcal{O}(x_i).$$

Then, we have that

Lemma 53. *Let P be any distribution with mean μ and covariance Σ and bounded $2k$-moments for $k \in \{1, 2\}$. For any $\delta \in (0, 0.5)$ such that $(\frac{\sqrt{trace(\Sigma)}}{R})^{2k} + \frac{\log(1/\delta)}{n} < c$ with probability at least $1 - 3\delta$,*

$$\frac{n - |G|}{n} \leq C_1 \frac{\log(1/\delta)}{n} + \frac{(\sqrt{trace\,(\Sigma)})^{2k}}{R^{2k}}$$

$$\begin{aligned}\|\widehat{\mu}_n - \mu\|_2 \lesssim & OPT_{n,\Sigma,\delta} + \frac{R\log(1/\delta)}{n} \\ & + \|\Sigma\|_2^{\frac{1}{2}} \left(\frac{\sqrt{trace\,(\Sigma)}}{R}\right)^{2k-1}.\end{aligned}$$

$$\|\widehat{\Sigma}_n^{\mathcal{O}}\|_2 \lesssim \|\Sigma\|_2 + R\|\Sigma\|_2^{\frac{1}{2}} \sqrt{\frac{\log(p/\delta)}{n}} + \frac{R^2 \log(p/\delta)}{n}.$$

Proof. **Controlling size of Set $|G|$.** Using Chebyshev's inequality, we have that,

$$\Pr(\|x - \mu\|_2 \geq R) \leq \frac{\mathbb{E}[\|x - \mu\|_2^{2k}]}{R^{2k}}$$

Now, to see that $\mathbb{E}[\|x - \mu\|_2^{2k}] \leq C(\sqrt{\text{trace}\,(\Sigma)})^{2k}$. The case for $k = 1$ is clear. We now show it for $k = 2$.

- Let $\Sigma = Q\Lambda Q^T$ and $\{q_i\}_{i=1}^p$ be the eigenvectors of Σ and let $\lambda_i = q_i^T\Sigma q_i$ be the associated eigenvalue. Then,

$$(x-\mu)^T(x-\mu) = \sum_i (q_i^T(x-\mu))^2 = \sum_i \nu_i^2, \tag{C.41}$$

where $\nu_i = q_i^T(x-\mu)$. Now, $\|x-\mu\|_2^4 = (\sum_i \nu_i^2)^2 = \sum_i \nu_i^4 + 2\sum_{i\neq j}\nu_i^2\nu_j^2$. Now, since we assume bounded fourth moments, we get that, $\mathbb{E}[\nu_i^4] \leq C(q_i^T\Sigma q_i)^2 = C\lambda_i^2$, Using Cauchy-Schwartz inequality, we get that $\mathbb{E}[\nu_i^2\nu_j^2] \leq \sqrt{\mathbb{E}[\nu_i^4]}\sqrt{\mathbb{E}[\nu_j^4]} = C\lambda_i\lambda_j$. Hence, we have that,

$$\mathbb{E}[\|x-\mu\|_2^4] \leq C(\sum_i \lambda_i^2 + 2\sum_{i\neq j}\lambda_i\lambda_j) = C_4 \text{trace}\,(\Sigma)^2$$

$$\Pr(\|x-\mu\|_2 \geq R) \leq \frac{\mathbb{E}[\|x-\mu\|_2^4]}{R^4} = C_4\frac{\text{trace}\,(\Sigma)^2}{R^4}$$

Hence, for $k = 1, 2$, we have that,

$$\Pr(\|x-\mu\|_2 \geq R) \leq \frac{(\sqrt{\text{trace}\,(\Sigma)})^{2k}}{R^{2k}} \tag{C.42}$$

Hence, know that for $x_i \sim P$, $\Pr(\mathcal{O}(x_i) = 1) \geq 1-\alpha$, where $\alpha = \frac{(\sqrt{\text{trace}(\Sigma)})^{2k}}{R^{2k}}$. Now, let $G^0 \stackrel{\text{def}}{=} \{x_i \text{ s.t. } \mathcal{O}(x_i) = 1\}$. Then, using Bernstein's inequality, we know that with probability at least $1-\delta$.

$$n_{G^0} = |G^0| \geq n(1 - \alpha - C_1\sqrt{\alpha\frac{\log(1/\delta)}{n}} - C_2\frac{\log(1/\delta}{n}) \tag{C.43}$$

Hence, we control the size of the good subset $|G|$.

Controlling the mean of G. Recall from our assumption that

$$\alpha + C_2\frac{\log(1/\delta)}{n} < \frac{1}{2},$$

hence we have that $|G| = |n_G| > n/2$. Let $\widehat{\theta}_{G^0} = \widehat{\mu}_n$ be the mean of the points in G.

1. Controlling $\|\mu - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. This is a deterministic statement and essentially quantifies the amount the mean can shift, when the random variable is conditioned on an event. We show this in Claim 14 which was shown in [28, 126]. We also provide a proof of the statement for completeness in Section **??**.

   Claim 14. *[General Mean shift,[28, 126]] Suppose that a distribution P has mean μ and covariance Σ and bounded $2k$ moments. Then, for any event $\mathcal{A}$ which occurs with probability at least $1-\epsilon \geq \frac{1}{2}$,*

$$\|\mu - E[x|\mathcal{A}]\|_2 \leq 2\|\Sigma\|_2^{\frac{1}{2}}\epsilon^{1-\frac{1}{2k}} \tag{C.44}$$

Now using this Claim 14 with $\mathcal{A}$ being the event that $\mathcal{O}(x) = 1$, we get that

$$\|\mu - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq 2\|\Sigma\|_2^{\frac{1}{2}} \alpha^{1-1/(2k)} \tag{C.45}$$

2. **Controlling** $\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. This term measures how quickly the samples within G^0 converge to their true mean. To show this we use vector version of Bernstein's inequality. Let $z_i \overset{\text{def}}{=} x_i - \mathbb{E}[\widehat{\theta}_{G^0}]$ be the centered random variables. Then, we have that

$$\begin{aligned}\|z_i\|_2 &\leq \|\theta^* - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 + \|x_i - \theta^*\|_2 \\ &\leq 2\|\Sigma\|_2^{\frac{1}{2}} \alpha^{1-1/(2k)} + R \\ &\leq 2R\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[\|z_i\|_2^2] &= \mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 | x \in \mathcal{A}] && \text{(C.46)} \\ &= \frac{\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 \mathbb{I}\left\{x \in \mathcal{A}\right\}]}{P(\mathcal{A})} && \text{(C.47)} \\ &\leq 2\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2] && \text{(C.48)} \\ &\leq 2\mathbb{E}[\|x - E[x]\|_2^2] + 2\|\theta^* - E[x|\mathcal{A}]\|_2^2 && \text{(C.49)} \\ &\leq 2\text{trace}\left(\Sigma\right) + 4\|\Sigma\|_2 \alpha^{2-1/(k)} && \text{(C.50)} \\ &\leq 4\text{trace}\left(\Sigma\right) && \text{(C.51)}\end{aligned}$$

Now, we first state the vector version of Bernstein's inequality.

Lemma 54. *(Vector Bernstein, Corollary 8.45 [134]) Let $Y_1, \ldots, Y_M$ be independent copies of a random vector $Y \in \mathcal{C}^p$ satisfying $\mathbb{E}Y = 0$. Assume $\|Y\|_2 \leq K$ for some $K > 0$. Let,*

$$Z = \|\sum_{l=1}^{M} Y_l\|_2, \mathbb{E}[Z^2] = M\mathbb{E}[\|Y\|_2^2], \sigma^2 = \sup_{\|v\|_2 \leq 1} \mathbb{E}[| \langle v, Y \rangle |^2]$$

Then for $t > 0$,

$$\Pr(Z \geq \sqrt{\mathbb{E}Z^2} + t) \leq \exp(-\frac{t^2/2}{M\sigma^2 + 2K\sqrt{\mathbb{E}Z^2} + tK/3}) \tag{C.52}$$

We use the above lemma, with $Y_i = \frac{z_i}{n_{G^0}}$. Hence, we have that, $K = \frac{2R}{n_{G^0}}$. Hence, we have that $Z = \|\sum_{k=1}^{n_{G^0}} Y_k\|_2 = \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2$. Hence, we have the following,

- $\mathbb{E}[Z^2] \leq n\frac{4\text{trace}(\Sigma)}{n^2} = 4\frac{\text{trace}(\Sigma)}{n}$.

- $\sigma^2 \leq 4\frac{\|\Sigma\|_2}{n^2}$. To see this, for any $v \in \mathcal{S}^{p-1}$,

$$\mathbb{E}[(v^T Y)^2] = \frac{1}{n^2}\mathbb{E}[(v^T(x-\mu_A))^2 | x \in \mathcal{A}]$$

where μ_A is the conditional mean, and $\mathcal{A}$ is the event that x s.t. $\|x-\mu\|_2 \leq R$. We know that $P(A) \geq 1/2$. Hence, we get that,

$$\begin{aligned}
\mathbb{E}[(v^T Y)^2] &= \frac{1}{n^2}\frac{\mathbb{E}[(v^T(x-\mu_A))^2 \mathbb{I}\{x \in \mathcal{A}\}]}{P(A)} \\
&\leq \frac{2}{n^2}\mathbb{E}[(v^T(x-\mu_A))^2] \\
&= \frac{2}{n^2}(\mathbb{E}[(v^T(x-\mu))^2] + \|\mu-\mu_A\|_2^2) \\
\implies \sigma^2 &\leq \frac{2}{n^2}(\|\Sigma\|_2 + \|\Sigma\|_2\alpha) \\
&\leq \frac{4\|\Sigma\|_2}{n^2}
\end{aligned}$$

Hence, we get that, with probability at least $1-\delta$,

$$\begin{aligned}
\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq & C_1\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\log(1/\delta)}{n_{G^0}}} + C_3 R^{\frac{1}{2}}(\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}})^{\frac{1}{2}}\sqrt{\frac{\log(1/\delta)}{n_G^0}} \\
& + C_4 R\frac{\log(1/\delta)}{n_{G^0}}
\end{aligned}$$

Now, we use that $\sqrt{ab} \leq a+b \;\forall\; a,b \geq 0$. Hence, we get that with probability at least $1-\delta$

$$\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq \underbrace{C_5\sqrt{\frac{\text{trace}\,(\Sigma)}{n_{G^0}}} + C_2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\log(1/\delta)}{n_{G^0}}}}_{T1} + C_3 R\frac{\log(1/\delta)}{n_{G^0}}$$

Using the bound on $\|\mathbb{E}[\widehat{\theta}_{G^0}] - \mu\|_2$ from (C.45), we get that,

$$\begin{aligned}
\|\widehat{\theta}_{G^0} - \mu\|_2 &\leq \|\mathbb{E}[\widehat{\theta}_{G^0}] - \mu\|_2 + \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 && \text{(C.53)} \\
&\leq T1 + C_3 R\frac{\log(1/\delta)}{n_{G^0}} + 2\|\Sigma\|_2^{\frac{1}{2}}((\frac{\sqrt{\text{trace}\,(\Sigma)}}{R})^{2k})^{1-1/(2k)} && \text{(C.54)} \\
&= T1 + C_3 R\frac{\log(1/\delta)}{n_{G^0}} + 2\|\Sigma\|_2^{\frac{1}{2}}(\frac{(\sqrt{\text{trace}\,(\Sigma)})^{2k-1}}{R^{2k-1}}) && \text{(C.55)}
\end{aligned}$$

Under our assumption that $(\frac{\sqrt{\text{trace}(\Sigma)}}{R})^{2k} + \frac{\log(1/\delta)}{n} < c$, we know that $n_{G^0} \geq n/2$. Hence, we get get that $T1 \precsim \text{OPT}_{n,\Sigma,\delta}$.

Controlling the covariance of points in $|G|$. Let $G^0 = \{x_i | \mathcal{O}(x_i) = 1\}$ be the empirical collection of points chosen by the oracle. Let $n_{G^0} = |G^0|$. Then, we study and bound the operator norm of Σ_{G^0}. Recall that all oracles have the form $\mathbb{I}\left\{\|x_i - \mu\|_2 \leq R\right\}$, *i.e.*, $\forall x_i$ s.t. $\mathcal{O}(x_i) = 1$, we have that $\|x_i - \mu\|_2 \leq R$.

Note that from Proof of Theorem **??**, we know that $\Pr(x \in G^0) \geq 1 - \alpha$, where $\alpha = (\frac{\sqrt{\text{trace}(\Sigma)}}{R})^{2k}$. Let Σ_{G^0} be the empirical covariance matrix. Then,

$$\Sigma_{G^0} = \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \widehat{\theta}_{G^0})(x_i - \widehat{\theta}_{G^0})^T,$$

where $\widehat{\theta}_{G^0}$ is the empirical mean of the points in G^0. Recentering it around the true mean θ^* of P, we get that,

$$\Sigma_{G^0} = \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \theta^*)(x_i - \theta^*)^T - (\widehat{\theta}_{G^0} - \theta^*)(\widehat{\theta}_{G^0} - \theta^*)^T$$

Hence, we have that $\|\Sigma_{G^0}\|_2 \leq \| \underbrace{\frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \theta^*)(x_i - \theta^*)^T}_{A} \|_2$. To control, $\|A\|_2$, we use triangle inequality,

$$\|A\|_2 \leq \underbrace{\|A - \mathbb{E}[A]\|_2}_{T1} + \underbrace{\|\mathbb{E}[A]\|_2}_{T2} \tag{C.56}$$

1. **Controlling T2.** Note that $\mathbb{E}[A] = \mathbb{E}[(x - \theta^*)(x - \theta^*)^T | x \in G]$.

   $$\mathbb{E}[A] = \frac{\mathbb{E}[(x - \theta^*)(x - \theta^*)^T \mathbb{I}\left\{x \in G^0\right\}]}{P(x \in G^0)} \tag{C.57}$$

   Let $\Pr(x \in G^0) \geq 1 - \alpha$. Hence, for any $v \in \mathcal{S}^{p-1}$,

   $$v^T \mathbb{E}[A] v = \frac{\mathbb{E}[(v^T(x - \theta^*))^2 \mathbb{I}\left\{x \in G^0\right\}]}{P(x \in G^0)} \leq \frac{\|\Sigma\|_2}{1 - \alpha}$$

   Under the assumption that $\alpha < \frac{1}{2}$, we get that,

   $$\|\mathbb{E}[A]\|_2 \leq 2\|\Sigma\|_2$$

2. **Controlling T1.** Note that T1 can be controlled using a concentration of measure argument, and in particular exploits concentration of covariance for bounded random vectors.

Lemma 55. *[Theorem 5.44 [135]] Let $\{y_i\}_{i=1}^n$ samples such that $y_i \in \mathbb{R}^p$ and $\|y_i\|_2 \leq \sqrt{m}$ and $\mathbb{E}[yy^T] = \Sigma$. Then, with probability at least $1-\delta$,*

$$\|\frac{1}{n}\sum_{i=1}^{n} y_i y_i^T - \Sigma\|_2 \leq \max(\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\log(p/\delta)}\sqrt{\frac{m}{n}}, \log(p/\delta)\frac{m}{n})$$

$$T1 = \|\frac{1}{n_{G^0}}\sum_{i=1}^{n_{G^0}}(x_i-\theta^*)(x_i-\theta^*)^T - \mathbb{E}[A]\|_2 \tag{C.58}$$

We use Lemma 55 with $y_i = x_i - \theta^*$. Note that $\sqrt{m} = R$. This means that with probability $1-\delta$,

$$T1 \leq C_1 R\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2\frac{\log(p/\delta)}{n_{G^0}}$$

Hence, we get that under the assumption that $\alpha + \sqrt{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{2}$, we recover statement of the result.

□

C.1.3 Proof of Theorem 25

We consider the ℓ_2 oracle of $\mathcal{O}$ radius $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/4}}$. Using chebychevs inequality, we know that $\Pr(\mathcal{O}(x) = 1) \geq 1 - \alpha$, where $\alpha = \frac{\log(1/\delta)}{n}$.

Suppose we are given n-samples from P. Let G^0 be the set of points such that $\mathcal{O}(x_i) = 1$. Using bernstein's inequality we know that with probability $1 - \delta$,

$$|n_{G^0}| \geq n(1 - C\frac{\log(1/\delta)}{n}) \tag{C.59}$$

Hence, we have that,

$$\frac{n - n_{G^0}}{n} \lesssim \frac{\log(1/\delta)}{n} \tag{C.60}$$

Let $\widehat{\mu}_n$ and Σ_{G^0} be the empirical mean and covariance of the points in G^0.

Let $\widehat{\theta}_\delta$ be the output of Algorithm 7. Then, we know that with probability at least $1-\delta$,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma_{G^0}\|_2^{\frac{1}{2}}(\frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n})^{\frac{1}{2}} \tag{C.61}$$

Using Lemma 15, we bound $\|\Sigma_{G^0}\|_2^{\frac{1}{2}}$.

$$\|\Sigma_{n,\mathcal{O}}\|_2 \leq C_1\|\Sigma\|_2 + C_2 R\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2\frac{\log(p/\delta)}{n_{G^0}}$$

$$\|\Sigma_{n,\mathcal{O}}\|_2^{\frac{1}{2}} \leq C_1\|\Sigma\|_2^{\frac{1}{2}} + C_2 R^{\frac{1}{2}}\|\Sigma\|_2^{1/4}(\frac{\log(p/\delta)}{n_{G^0}})^{1/4} + R\sqrt{\frac{\log(p/\delta)}{n_{G^0}}} \tag{C.62}$$

Plugging $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/4}}$, we get,

$$\|\Sigma_{n,\mathcal{O}}\|_2^{\frac{1}{2}} \leq C_1\|\Sigma\|_2^{\frac{1}{2}} + \underbrace{C_2\text{trace}\,(\Sigma)^{1/4}\,\|\Sigma\|_2^{1/4}\frac{(\frac{\log(p/\delta)}{n_{G^0}})^{1/4}}{(\frac{\log(1/\delta)}{n})^{1/8}}}_{T1} + \underbrace{\sqrt{\text{trace}\,(\Sigma)}\frac{\sqrt{\frac{\log(p/\delta)}{n_{G^0}}}}{(\frac{\log(1/\delta)}{n})^{1/4}}}_{T2} \tag{C.63}$$

Plugging (C.60) and (C.63) into (C.61), we get that,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} + T1\sqrt{\frac{\log(1/\delta)}{n}} + T2\sqrt{\frac{\log(1/\delta)}{n}} \tag{C.64}$$

When $T1$ and $T2$ are less than $C\sqrt{\|\Sigma\|_2}$, then we have that,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} \tag{C.65}$$

Some algebra shows that when $\frac{r(\Sigma)^2 \log^2(p/\delta)}{n \log(1/\delta)} \leq C$, the both $T1$ and $T2$ are $O(\sqrt{\|\Sigma\|_2})$. Hence, we get that,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n}} \tag{C.66}$$

Using Lemma 15, and plugging $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/4}}$, we get that with probability at least $1 - \delta$,

$$\|\mu(P) - \widehat{\mu}_n\|_2 \lesssim \text{OPT}_{n,\Sigma,\delta} + \underbrace{\sqrt{\text{trace}\,(\Sigma)}(\frac{\log(1/\delta)}{n})^{3/4}}_{T3} \tag{C.67}$$

Under our assumption that $r^2(\Sigma)\frac{\log(1/\delta)}{n} \leq C$, $T3 \lesssim \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n}}$. Combining the above equation and C.66, we recover the theorem statement.

C.1.4 Proof of Corollary 26

We consider the ℓ_2 oracle of $\mathcal{O}$ radius $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/2}}$. Using chebychevs inequality, we know that $\Pr(\mathcal{O}(x) = 1) \geq 1 - \alpha$, where $\alpha = \frac{\log(1/\delta)}{n}$.

Suppose we are given n-samples from P. Let G^0 be the set of points such that $\mathcal{O}(x_i) = 1$. Using bernstein's inequality we know that with probability $1 - \delta$,

$$|n_{G^0}| \geq n(1 - C\frac{\log(1/\delta)}{n}) \tag{C.68}$$

Hence, we have that,

$$\frac{n - n_{G^0}}{n} \lesssim \frac{\log(1/\delta)}{n} \tag{C.69}$$

Let $\widehat{\mu}_n$ and Σ_{G^0} be the empirical mean and covariance of the points in G^0.

Let $\widehat{\theta}_\delta$ be the output of Algorithm 7. Then, we know that with probability at least $1 - \delta$,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma_{G^0}\|_2^{\frac{1}{2}}(\frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n})^{\frac{1}{2}} \tag{C.70}$$

Using Lemma 15, we bound $\|\Sigma_{G^0}\|_2^{\frac{1}{2}}$.

$$\|\Sigma_{n,\mathcal{O}}\|_2 \leq C_1\|\Sigma\|_2 + C_2 R\|\Sigma\|_2^{\frac{1}{2}}\sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2\frac{\log(p/\delta)}{n_{G^0}}$$

$$\|\Sigma_{n,\mathcal{O}}\|_2^{\frac{1}{2}} \leq C_1\|\Sigma\|_2^{\frac{1}{2}} + C_2 R^{\frac{1}{2}}\|\Sigma\|_2^{1/4}(\frac{\log(p/\delta)}{n_{G^0}})^{1/4} + R\sqrt{\frac{\log(p/\delta)}{n_{G^0}}} \tag{C.71}$$

Plugging $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/2}}$, we get,

$$\|\Sigma_{n,\mathcal{O}}\|_2^{\frac{1}{2}} \leq C_1\|\Sigma\|_2^{\frac{1}{2}} + C_2\text{trace}\,(\Sigma)^{1/4}\,\|\Sigma\|_2^{1/4}(\frac{\log(p/\delta)}{\log(1/\delta)})^{1/4} + \frac{\sqrt{\text{trace}\,(\Sigma)}}{\sqrt{\frac{\log(1/\delta)}{n}}}\sqrt{\frac{\log(p/\delta)}{n}} \tag{C.72}$$

Plugging (C.69) and (C.72) into (C.70), we get that,

$$\|\widehat{\theta}_\delta - \widehat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\text{trace}\,(\Sigma)\log(p/\delta)}{n}} \tag{C.73}$$

Using Lemma 15, and plugging $R = \frac{\sqrt{\text{trace}(\Sigma)}}{(\frac{\log(1/\delta)}{n})^{1/2}}$, we get that with probability at least $1 - \delta$,

$$\|\mu(P) - \widehat{\mu}_n\|_2 \lesssim \text{OPT}_{n,\Sigma,\delta} + \sqrt{\frac{\text{trace}\,(\Sigma)\log(p/\delta)}{n}} \tag{C.74}$$

Combining the above equation and C.73, we recover the corollary statement.

C.2 Proofs for Section 5.3

C.2.1 Common Proof Template for Corollaries 27, 28 and 29.

. In this section, we provide the proofs of the technical results in Section 5.3. We follow the template provided by [1] to prove the corollaries appearing in this section.

- In particular, given a distribution $z \sim P$, and a loss function $\bar{\mathcal{L}}(\theta, z)$, we look at the distribution of the gradients $\nabla\bar{\mathcal{L}}(\theta^t, z)$ for any θ^t, and in particular calculate the trace and operator norm of the covariance of gradient distribution $\Sigma(\bar{\mathcal{L}}(\theta^t, z))$. We show that for GLMs, they are of the form,

$$\text{trace}\left(\Sigma(\bar{\mathcal{L}}(\theta^t, z))\right) \le A\|\theta^t - \theta^*\|_2^2 + B \tag{C.75}$$

$$\|\Sigma(\bar{\mathcal{L}}(\theta^t, z))\|_2 \le C\|\theta^t - \theta^*\|_2^2 + D \tag{C.76}$$

- From Theorem 25, we know that given n samples the output of the mean estimator satisfies the guarantee that with probability at least $1-\delta$,

$$\|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta^t, z)] - \text{FilterpD}(\{\nabla\bar{\mathcal{L}}(\theta^t, z_i)\}_{i=1}^n)\|_2 \le \sqrt{\frac{\text{trace}\left(\Sigma(\bar{\mathcal{L}}(\theta^t, z))\right)}{n}} + \sqrt{\frac{\|\Sigma(\bar{\mathcal{L}}(\theta^t, z))\|_2 \log(1/\delta)}{n}},$$

or equivalently,

$$\|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta^t, z)] - \text{FilterpD}(\{\nabla\bar{\mathcal{L}}(\theta^t, z_i)\}_{i=1}^n)\|_2 \le (\sqrt{\frac{A}{n}} + \sqrt{\frac{C}{n}})\|\theta^t - \theta^*\|_2 + (\sqrt{\frac{B}{n}} + \sqrt{\frac{D\log 1/\delta}{n}}).$$

- The last step is to use the following result from [1] on the stability of gradient descent with inexact gradients.

Lemma 56. *[Prasad et al. [1]] For a given sample-size n and confidence parameter $\delta \in (0,1)$, suppose we have a gradient estimator $g(\theta; \{\nabla\bar{\mathcal{L}}(\theta, z_i)\}_{i=1}^n, \delta)$ such that for any fixed $\theta \in \Theta$, the estimator satisfies the following inequality:*

$$\|g(\theta; \{\nabla\bar{\mathcal{L}}(\theta, z_i)\}_{i=1}^n, \delta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta^t, z)]\|_2 \le \alpha(n,\delta)\|\theta - \theta^*\|_2 + \beta(n,\delta). \tag{C.77}$$

Then Algorithm 8 initialized at θ^0 with step-size $\eta = 2/(\tau_\ell + \tau_u)$, returns iterates $\{\widehat{\theta}^t\}_{t=1}^T$ such that with probability at least $1-\delta$

$$\|\widehat{\theta}^t - \theta^*\|_2 \le \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\beta(\widetilde{n}, \widetilde{\delta}), \tag{C.78}$$

where $\widetilde{n} = n/T, \widetilde{\delta} = \delta/T$, $\kappa = \sqrt{1 - \frac{2\eta\tau_\ell\tau_u}{\tau_\ell+\tau_u}} + \eta\alpha(\widetilde{n}, \widetilde{\delta}) < 1$ is a contraction and $\theta^* = \text{argmin}_{\theta\in\Theta}\mathbb{E}[\bar{\mathcal{L}}(\theta, z)]$ is the minimizer of the population loss.

- Using the above we get that

$$\|\hat{\theta}^t - \theta^*\|_2 \lesssim \kappa^t\|\theta^0 - \theta^*\|_2 + \sqrt{\frac{B}{(n/T)}} + \sqrt{\frac{D\log(T/\delta)}{(Tn)}}, \tag{C.79}$$

as long as $\alpha(\widetilde{n}, \widetilde{\delta}) < \tau_\ell$.

- Hence, all that remains is to calculate (A, B, C, D) for linear regression and GLMs.

C.2.2 Proof of Corollary 27.

We refer the reviewer to Section C.2.1 for a common proof template. In this section we simply focus on deriving upper bounds for the gradient distribution for Linear Regression. This result can also be found in [1], but we provide it for the sake of completeness. Recall that for linear regression we have that, $\bar{\mathcal{L}}(\theta, (x, y)) = \frac{1}{2}(y - x^T\theta)^2$.

Lemma 57 (Prasad et al. [1])**.** *Consider the model in* (B.17)*. Suppose the covariates $x \in \mathbb{R}^p$ have bounded 8^{th}-moments and the noise w has bounded 4^{th} moments. Then there exist universal constants C_1, C_2 such that*

$$\begin{aligned}
\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] &= \Sigma\Delta \\
trace\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &\leq \underbrace{C_4 trace\,(\Sigma)\,\|\Sigma\|_2}_{A}\|\Delta\|_2^2 + \underbrace{\sigma^2 trace\,(\Sigma)}_{B}, \\
\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 &\leq \|\Delta\|_2^2\underbrace{C_4\|\Sigma\|_2^2}_{C} + \underbrace{\sigma^2\|\Sigma\|_2}_{D} \\
\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] &\leq C_2(\mathrm{Var}[\nabla\bar{\mathcal{L}}(\theta)^T v])^2
\end{aligned}$$

where $\Delta = \theta - \theta^*$ and $E[xx^T] = \Sigma$.

From the above lemma, we recover the values of (A, B, C, D) for linear regression which we simply plug into (C.79) to recover the statement of the corollary.

Proof of Lemma 57

Proof. We start by deriving the results for $\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]$.

$$\begin{aligned}
\bar{\mathcal{L}}(\theta) = \frac{1}{2}(y - x^T\theta)^2 &= \frac{1}{2}(x^T(\Delta) - w)^2 \\
\nabla\bar{\mathcal{L}}(\theta) &= xx^T\Delta - x.w \\
\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] &= \Sigma\Delta.
\end{aligned}$$

Next, we bound the operator norm of the covariance of the gradients $\nabla\bar{\mathcal{L}}(\theta)$ at any point θ.

Covariance.

$$\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) = \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T$$

For any unit vector $z \in \mathcal{S}^{p-1}$, we have that,

$$\begin{aligned}
z^T\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))z &= z^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]z - (\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T z)^2 \\
&\leq z^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]z \\
\implies \sup_{z\in\mathcal{S}^{p-1}} z^T\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))z &\leq \sup_{z\in\mathcal{S}^{p-1}} z^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]z
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
\lambda_{\max}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))) &\leq \sup_{z\in\mathcal{S}^{p-1}} z^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]z \\
&= \sup_{z\in\mathcal{S}^{p-1}} z^T\mathbb{E}[(xx^T\Delta - x.w)(xx^T\Delta - x.w)^T]z \\
&= \sup_{z\in\mathcal{S}^{p-1}} z^T(\mathbb{E}[xx^T\Delta\Delta^T xx^T] + \sigma^2\mathbb{E}[xx^T])z \\
&\leq \sup_{z\in\mathcal{S}^{p-1}} z^T(\mathbb{E}[xx^T\Delta\Delta^T xx^T])z + \sigma^2\|\Sigma\|_2 \\
&\leq \sigma^2\|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y,z\in\mathcal{S}^{p-1}} \mathbb{E}[(z^Tx)^2(y^Tz)^2] \\
&\leq \sigma^2\|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y,z\in\mathbb{S}^{p-1}} \sqrt{\mathbb{E}\left[(y^Tx)^4\right]}\sqrt{\mathbb{E}\left[(z^Tx)^4\right]} \\
&\leq \sigma^2\|\Sigma\|_2 + \|\Delta\|_2^2 C_4\|\Sigma\|_2^2
\end{aligned}$$

where the second last step follows from Cauchy-Schwartz and the last step follows from our assumption of bounded 4^{th} moments. Now to bound the trace of the covariance matrix,

$$
\begin{aligned}
\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) &= \mathbb{E}[(xx^T-\Sigma)\Delta - xw)(xx^T-\Sigma)\Delta - xw)^T] \\
\mathrm{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &= \mathbb{E}[\|(xx^T-\Sigma)\Delta - xw)\|_2^2] \\
&= \underbrace{\mathbb{E}[\|(xx^T-\Sigma)\Delta\|_2^2]}_{T1} + \underbrace{\mathbb{E}[\|x\|_2^2w^2]}_{\sigma^2\mathrm{trace}(\Sigma)} \\
T1 &= \mathbb{E}[\|(xx^T-\Sigma)\Delta\|_2^2] = \Delta^T\mathbb{E}[(xx^T-\Sigma)^2]\Delta \\
&= \Delta^T\mathbb{E}[(x^Tx)xx^T + \Sigma^2 - \Sigma xx^T - xx^T\Sigma]\Delta \\
&= \Delta^T\mathbb{E}[(x^Tx)xx^T]\Delta - \Delta^T\Sigma^2\Delta \\
&\leq \Delta^T\mathbb{E}[(x^Tx)xx^T]\Delta \\
&\leq \|\Delta\|_2^2\mathbb{E}[(x^Tx)(x^Tu)^2], \quad \text{where } u = \frac{\Delta}{\|\Delta\|_2} \in \mathcal{S}^{p-1} \\
&\leq \|\Delta\|_2^2\mathbb{E}[(x^Tx)^2]^{\frac{1}{2}} \underbrace{\mathbb{E}[(x^Tu)^4]^{\frac{1}{2}}}_{\leq\sqrt{C_4}\|\Sigma\|_2} \\
x &\overset{\mathrm{def}}{=} \sum_{i=1}^{p} \underbrace{(x^Tq_i)}_{\nu_i} q_i, \quad \text{where } \{q_i\}_{i=1}^p \text{ are eigenvectors of } \Sigma \\
\mathbb{E}[(x^Tx)(x^Tx)] &= \mathbb{E}[(\sum_i \nu_i^2)(\sum_i \nu_i^2)] \\
&= \mathbb{E}[\sum_i \nu_i^4 + 2\sum_{i<j}\nu_i^2\nu_j^2] \\
\mathbb{E}[\nu_i^4] &= \mathbb{E}[(x^Tq_i)^4] \leq C_4\mathbb{E}[(x^Tq_i)^2]^2 = C_4\lambda_i^2 \\
\mathbb{E}[\nu_i^2\nu_j^2] &\leq \sqrt{\mathbb{E}[\nu_i^4]}\sqrt{\mathbb{E}[\nu_j^4]} = C_4\lambda_i\lambda_j \\
\mathbb{E}[(x^Tx)(x^Tx)] &\leq C_4(\sum_i \lambda_i^2 + 2\sum_{i<j}\lambda_i\lambda_j) = C_4\mathrm{trace}\,(\Sigma)^2 \\
\mathrm{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &\leq \sigma^2\mathrm{trace}\,(\Sigma) + C_4\mathrm{trace}\,(\Sigma)\,\|\Sigma\|_2\|\Delta\|_2^2
\end{aligned}
$$

Bounded Fourth Moment. We start from the LHS

$$
\begin{aligned}
\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] &\le \mathbb{E}\left[\left[\left|(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right|\right]^4\right] && \text{(C.80)}\\
&= \mathbb{E}\left[\left|((xx^T - \Sigma)\Delta - wx)^T v\right|^4\right] && \text{(C.81)}\\
&= \mathbb{E}\left[\left|(\Delta^T x)(x^T v) - (\Sigma\Delta)^T v - wv^T x\right|^4\right] && \text{(C.82)}\\
&\le 8\left[8\left[\underbrace{\mathbb{E}\left|(\Delta^T x)(x^T v)\right|^4}_{A} + \underbrace{\mathbb{E}\left|(\Sigma\Delta)^T v\right|^4}_{B}\right] + \underbrace{\mathbb{E}\left|w(x^T v)\right|^4}_{C}\right]. && \text{(C.83)}
\end{aligned}
$$

The last line follows from two applications of the following inequality:

C_r **inequality**. If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^4 < \infty$ where $r \ge 1$ then:

$$\mathbb{E}|X+Y|^r \le 2^{r-1}\left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r\right).$$

Now to control each term:

- **Control of** A. Using Cauchy Schwartz, that C_8 is bounded for x

$$
\begin{aligned}
A &\le \sqrt{\mathbb{E}[|\Delta^T x|^8]}\sqrt{\mathbb{E}[|x^T v|^8]} && \text{(C.84)}\\
&\lesssim \|\Delta\|_2^4 C_8 \|\Sigma\|_2^4. && \text{(C.85)}
\end{aligned}
$$

- **Control of** B, $B \lesssim \|\Delta\|_2^4\|\Sigma\|_2^4$.
- **Control of** C, $C \lesssim C_4\|\Sigma\|_2^2$, using independence of w and bounded moments of x.

Therefore the $\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] \lesssim c + \|\Sigma\|_2^4\|\Delta\|_2^4$.

For the RHS:

$$\mathrm{Var}(\nabla\bar{\mathcal{L}}(\theta)^T v)^2 = (v^T \mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))v)^2 \le \|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2^2$$

We saw that the $\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2 \lesssim c + \|\Sigma\|_2^2\|\Delta\|_2^2$, so both the LHS and RHS scale with $\|\Sigma\|_2^4\|\Delta\|_2^4$.

□

C.2.3 Proof of Corollary 28

We refer the reviewer to Section C.2.1 for a common proof template. In this section we simply focus on deriving upper bounds for the gradient distribution for Generalized Linear Models. This result can also be found in [1], but we provide it for the sake of completeness. Recall that for generalized linear models we have that,

$$\bar{\mathcal{L}}(\theta;(x,y)) = -y\langle x,\theta\rangle + \Phi(\langle x,\theta\rangle). \tag{C.86}$$

.

Lemma 58 ([1]). *Consider the model in* (5.10)*, then there exist universal constants* $C_1, C_2 > 0$ *such that*

$$\textit{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\right) \leq \underbrace{\sqrt{C_4}\textit{trace}\,(\Sigma)\,\sqrt{L_{\Phi,4}}}_{A}\|\Delta\|_2^2 + \underbrace{\sqrt{C_4}\textit{trace}\,(\Sigma)\,(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}})}_{B}$$

$$\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)\|_2 \leq \underbrace{\sqrt{C}\sqrt{C_4}\|\Sigma\|_2(\sqrt{L_{\Phi,4}})_C\|\Delta\|_2^2} + \underbrace{\sqrt{C}\sqrt{C_4}\|\Sigma\|_2\left(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,}}\right.}_{D}$$

and

$$\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] \leq C_2(\mathrm{Var}[\nabla\bar{\mathcal{L}}(\theta)^T v])^2.$$

From the above lemma, we recover the values of (A, B, C, D) for GLMs which we simply plug into (C.79) to recover the statement of the corollary.

Proof of Lemma 58

.

Proof. The gradient $\nabla\bar{\mathcal{L}}(\theta)$ and it's expectation can be written as:

$$\nabla\bar{\mathcal{L}}(\theta) = -y.x + u(\langle x, \theta\rangle).x$$
$$\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] = \mathbb{E}[x\left(u(x^T\theta) - u(x^T\theta^*)\right)],$$

where $u(t) = \Phi'(t)$.

$$\begin{aligned}\|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\|_2 &= \sup_{y\in\mathbb{S}^{p-1}} y^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)] \\ &\leq \sup_{y\in\mathbb{S}^{p-1}} \mathbb{E}[(y^T x)\left(u(x^T\theta) - u(x^T\theta^*)\right)] \\ &\leq \sup_{y\in\mathbb{S}^{p-1}} \sqrt{\mathbb{E}[(y^T x)^2]}\sqrt{\mathbb{E}[(u(x^T\theta) - u(x^T\theta^*))^2]} \\ &\leq C_1\|\Sigma\|_2^{\frac{1}{2}}\sqrt{L_{\Phi,2}\|\Delta\|_2^2 + B_{\Phi,2}}\end{aligned}$$

where the last line follows from our assumption of smoothness.

Now, to bound the maximum eigenvalue of the $\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))$,

$$\begin{aligned}\lambda_{\max}(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))) &\leq \sup_{z\in\mathcal{S}^{p-1}} z^T\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]z \\ &= \sup_{z\in\mathbb{S}^{p-1}} z^T\left(\mathbb{E}\left[xx^T\left(u(x^T\theta) - y)\right)^2\right]\right)z \\ &\leq \sup_{z\in\mathbb{S}^{p-1}} \mathbb{E}\left[z^T\left(xx^T\left(u(x^T\theta) - y\right)^2\right)z\right] \\ &\leq \sup_{z\in\mathbb{S}^{p-1}} \sqrt{\mathbb{E}\left[(z^T x)^4\right]}\sqrt{\mathbb{E}\left[(u(x^T\theta) - y)^4\right]}\end{aligned}$$

To bound $\mathbb{E}\left[\left(u(x^T\theta) - y\right)^4\right]$, we make use of the C_r inequality.

C_r **inequality**. If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^4 < \infty$ where $r \geq 1$ then:

$$\mathbb{E}|X+Y|^r \leq 2^{r-1}\left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r\right)$$

Using the C_r inequality, we have that

$$\begin{aligned}\mathbb{E}\left[\left(u(x^T\theta) - y\right)^4\right] &\leq 8\left(\mathbb{E}\left[\left(u(x^T\theta) - u(x^T\theta^*)\right)^4\right] + \mathbb{E}\left[\left(u(x^T\theta^*) - y\right)^4\right]\right)\\ &\leq C\left(L_{\Phi,4}\|\Delta\|_2^4 + B_{\Phi,4} + c(\sigma)^3 M_{\Phi,4,1} + 3c(\sigma)^2 M_{\Phi,2,2}\right)\end{aligned}$$

where the last line follows from our assumption that $P_{\theta^*}(y|x)$ is in the exponential family, hence, the cumulants are higher order derivatives of the log-normalization function.

$$\|\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\|_2 \leq \sqrt{C}\sqrt{C_4}\|\Sigma\|_2\left(\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}\right)$$

Now, to control the trace. We have that,

$$\begin{aligned}\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta)) &= \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T\\ \mathrm{trace}\left(\mathrm{Cov}(\nabla\bar{\mathcal{L}}(\theta))\right) &= \mathrm{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]\right) - \mathrm{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T\right)\\ &\leq \mathrm{trace}\left(\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\nabla\bar{\mathcal{L}}(\theta)^T]\right)\\ &\leq \mathrm{trace}\left(\mathbb{E}\left[xx^T\left(u(x^T\theta) - y)\right)^2\right]\right)\\ &= \mathbb{E}\left[\mathrm{trace}\left(xx^T\left(u(x^T\theta) - y)\right)^2\right)\right]\\ &= \mathbb{E}[\mathrm{trace}\left((xx^T)\right)u(x^T\theta) - y)^2] \;\; \text{Because } (u(x^T\theta) - y)^2 \in \mathbb{R}\\ &\leq \sqrt{\mathbb{E}[\mathrm{trace}\left((xx^T)\right)^2]}\sqrt{\mathbb{E}[(u(x^T\theta) - y)^4]}\\ &\leq \sqrt{C_4}\mathrm{trace}\left(\Sigma\right)\left(\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}\right)\\ &= \sqrt{C_4}\mathrm{trace}\left(\Sigma\right)\sqrt{L_{\Phi,4}}\|\Delta\|_2^2 + \sqrt{C_4}\mathrm{trace}\left(\Sigma\right)(\sqrt{B_{\Phi,4}} + c(\sigma)\sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}}\end{aligned}$$

Bounded Fourth Moment. To show that the fourth moment of the gradient distribution is bounded, we have

$$\begin{aligned}\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right]^4\right] &\leq \mathbb{E}\left[\left[\left|(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)])^T v\right|\right]^4\right]\\ &\leq 8\left[\underbrace{\mathbb{E}[|\nabla\bar{\mathcal{L}}(\theta)])^T v|^4]}_{A} + \underbrace{\mathbb{E}[|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T v|^4]}_{B}\right].\end{aligned}$$

Control of A.

$$\begin{aligned}
\mathbb{E}[|\nabla\bar{\mathcal{L}}(\theta])^T v|^4] &= \mathbb{E}[(x^T v)^4 (u(x^T\theta) - y)^4] \\
&\leq \sqrt{\mathbb{E}[(x^T v)^8]}\sqrt{\mathbb{E}[(u(x^T\theta) - y)^8]} \\
&\leq \sqrt{C_8}\|\Sigma\|_2^2 \sqrt{\mathbb{E}[(u(x^T\theta) - u(x^T\theta^*))^8] + \mathbb{E}[(u(x^T\theta^*) - y)^8]} \\
&\leq \sqrt{C_8}\|\Sigma\|_2^2 \sqrt{L_{\Phi,8}\|\Delta\|_2^8 + B_{\Phi,8} + \sum_{t,k=2}^{8} g_{t,k} M_{\Phi,t,k}} \\
&\leq \sqrt{C}\|\Sigma\|_2^2 \sqrt{L_{\Phi,8}}\|\Delta\|_2^4 + \sqrt{B_{\Phi,8}} + \sqrt{\sum_{t,k=2}^{8} g_{t,k} M_{\Phi,t,k}}
\end{aligned}$$

where the last step follows from the fact that the 8th central moment can be written as a polynomial involving the lower cumulants, which in turn are the derivatives of the log-normalization function.

Control of B.

$$\mathbb{E}[|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)]^T v|^4] \leq \|\mathbb{E}[\nabla\bar{\mathcal{L}}(\theta)\|_2^4 \leq C_1\|\Sigma\|_2^2 \left(L_{\Phi,2}^2\|\Delta\|_2^2 + B_{\Phi,2}^2\right)$$

By assumption $L_{\Phi,k}, B_{\Phi,k}, M_{\Phi,t,k}$ are all bounded for $k, t \leq 8$, which implies that there exist constants $c_1, c_2 > 0$ such that

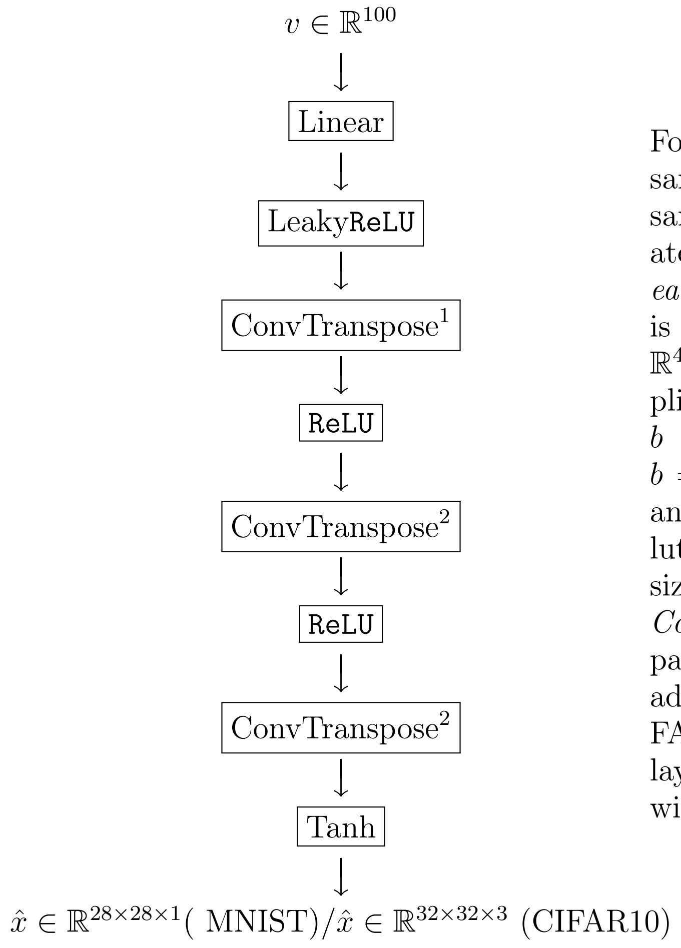
$$\mathbb{E}\left[\left[(\nabla\bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla\bar{\mathcal{L}}(\theta])^T v\right]^4\right] \leq c_1\|\Sigma\|_2^2\|\Delta\|_2^4 + c_2$$

Previously, we say that $\|\mathrm{Cov}\nabla\bar{\mathcal{L}}(\theta)\|_2 \leq c_3\|\Sigma\|_2\|\Delta\|_2^2 + c_4$, for some universal constants $c_3, c_4 > 0$, hence the gradient $\nabla\bar{\mathcal{L}}(\theta)$ has bounded fourth moments. □

C.3 Architectures used for the GAN experiment

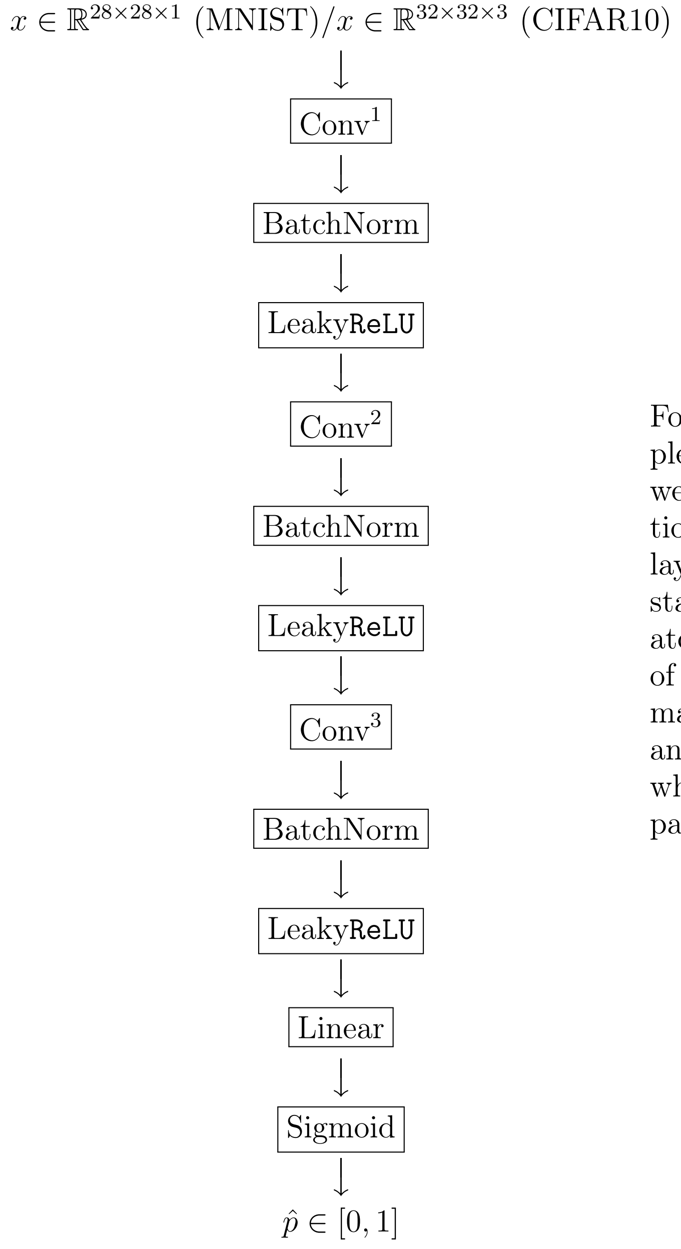
We implement the experiments and the networks using PyTorch, and use the default initialization used.

C.3.1 Generator for MNIST and CIFAR10


$v \in \mathbb{R}^{100}$
Linear
LeakyReLU
ConvTranspose^1
ReLU
ConvTranspose^2
ReLU
ConvTranspose^2
Tanh
$\hat{x} \in \mathbb{R}^{28\times28\times1}$(MNIST)/$\hat{x} \in \mathbb{R}^{32\times32\times3}$ (CIFAR10)

For MNIST and CIFAR10, to generate a sample, we use a 100-dimensional vector sampled from a 100-dimensional Multivariate Normal, with identity covariance. *Linear* denotes a fully connected layer which is an affine transformation from $\mathbb{R}^{100}$ to $\mathbb{R}^{4096}$. *LeakyReLU* is a layer that applies a function: $f(x) = \max\{bx, x\}$ for $b \in (0, 1)$ elementwise, and we choose $b = 0.2$. *ConvTranspose^1*, *ConvTranspose^2* and *ConvTranspose^3* are transposed convolution layers. All of these have a kernel of size 4 × 4; however for the MNIST case, *ConvTranspose^1* uses a stride of 1, and no padding and the other two use stride 2 and add padding with width 1. For the CIFAR10 case, all the transposed convolution layers have a stride of 2 and pad the input with width 1.

C.3.2 Discriminator for MNIST and CIFAR10



For MNIST and CIFAR10, to classify a sample as belonging to the distribution or not, we pass it through a sequence of convolution, batch normalization and LeakyReLU layers in succession. The LeakyReLU constant is set to be 0.2, as done for the generator. All the convolution layers use a kernel of size 4 × 4. For the CIFAR10 dataset, we maintain the stride and padding width as 2 and 1 respectively for all convolution layers, whereas for the MNIST dataset, we use no padding and a stride of 1 for *Conv*^{3}.

Appendix D

Supplementary Material for Chapter 4

D.1 Beyond Dobrushin's Conditions.

All of our previous results are under the high temperature condition (4.3), where we rely of special properties of Ising models namely sub-Gaussianity of Ising models random variables. Following this effort, we attempt to analyze classes of Ising models where this condition doesn't hold to present an even more general analysis. Towards this end, we present moduli of continuity bounds as presented in Theorem 20. Here, we look out for dependence in the model width parameter in addition to the effective dimensionality of the problem (d in the case of $\mathcal{G}_{p,d}$ and k in the case of $\mathcal{G}_{p,k}$, and the tolerance parameter ϵ.

Theorem 38. *Consider two Ising models defined over two graphs $G^{(1)}$ and $G^{(2)}$ over p vertices with parameters $\theta^{(1)}$ and $\theta^{(2)}$ respectively, satisfying $\omega(\theta^{(1)}), \omega(\theta^{(2)}) \leq \omega$. If $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$, then we have the following results for all $i \in [p]$:*

(a) If $G^{(1)}, G^{(2)} \in \mathcal{G}_{p,d}$, then

$$\|\theta^{(2)}(i) - \theta^{(1)}(i)\|_2 \lesssim \min\{\sqrt{\epsilon}, \epsilon\sqrt{d}\}\ \omega \exp(O(\omega)). \tag{D.1a}$$

(b) If $G_1, G_2 \in \mathcal{G}_{p,k}$, then

$$\|\theta^{(2)}(i) - \theta^{(1)}(i)\|_2 \lesssim \min\{\sqrt{\epsilon}, \epsilon\sqrt{k}\}\ \omega \exp(O(\omega)). \tag{D.1b}$$

Similar to Theorem 38, we get a modulus of continuity bound for the loss function defined by the $(2, \infty)$-norm. Note that as ϵ tends to 0, the bounds also tend to 0. However, it is worth noting that our primitive analysis contains an additional factor in d/k based on the graph class considered. The sub-optimality is clear when we set $\omega = O(1)$, and the bounds while retaining a optimal dependence on ϵ have an additional dependence with d/k when compared to the result in Theorem 20. Our analysis of the Yatracos estimator (4.7) does not depend of any specific bounds on the model width, and hence with the derived modulus of continuity bound, we arrive at the following corollary for the estimation error of the Yatracos estimate:

Corollary 39. *Given n samples from the distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^*} + \epsilon Q$, where $\mathbb{P}_{\theta^*} \in \mathcal{G}_{p,k}(\lambda, \omega)$ and Q is an arbitrary distribution supported over $\{-1,+1\}^p$, the parameter of Yatracos estimate (4.7) satisfies:*

$$\|\widehat{\theta}(i) - \theta^*(i)\|_2 \lesssim \sqrt{k}\omega e^{\mathcal{O}(\omega)}\epsilon + \mathcal{O}\left(k\omega e^{\mathcal{O}(\omega)}\sqrt{\frac{\log(p^2 e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \quad \text{for all } i \in [p].$$

Note that as $n \to \infty$, the bias of the estimator has optimal dependence on ϵ, but incurs an additional dependence of $\sqrt{k}$. For $\epsilon = 0$ *i.e.* no contamination, the rate we achieve is approximately $\omega e^{\mathcal{O}(\omega)} k \sqrt{\frac{\log(p)}{n}}$, which leads to the number of samples $n \geq \mathcal{O}\left(\frac{k^2\omega^2 e^{\mathcal{O}(\omega)}\log(p)}{\lambda^2}\right)$ required to recover the true edge set $E(\theta^*)$, and this is comparable to existing sample complexity results for learning Ising models belonging to $\mathcal{G}_{p,k}(\lambda, \omega)$ [90]. We present the proof of Theorem 38 in Section D.6.

D.2 Useful Properties of Ising models

In this section, we summarize some useful properties of Ising models which we use judiciously in our proofs. These results have appeared in previous work, but we state them for the sake of completeness.

D.2.1 Sub-Gaussianity of Ising model distributions in the high temperature regime

First, we present a result from [136], which states that a random variable distributed according to an Ising model in the high temperature regime is sub-Gaussian.

Proposition 59 ([136, Theorem 1.4])**.** *Let $z \sim \mathbb{P}$ be a random variable whose distribution $\mathbb{P}$ is an Ising model over p nodes in the high temperature regime (4.3) with constant α. Then for $v \sim \mathbb{R}^p$:*

$$\Pr_{z\sim\mathbb{P}}\left(|\langle v, z\rangle| > t\right) \leq 2\exp\left(-\frac{t^2}{C(\alpha)||v||_2^2}\right), \tag{D.2}$$

where $C(\alpha)$ is a constant depending on α.

D.2.2 Strong convexity of the negative conditional log-likelihood

Here we present a proposition that states that the population negative conditional log-likelihood is strongly convex. This proposition is obtained using a result by Dagan et al. [97]. We first state the result by Dagan et al. [97] below, and then use it to show that the population negative condition log-likelihood is strongly convex.

Proposition 60 ([97, Lemma 10])**.** *Let z be a random variable distributed w.r.t. an Ising model over p nodes whose parameter θ satisfies $\max_{i\in[p]}\|\theta(i)\|_\infty \leq \omega$ and $\min_{i\in[p]}\mathbb{P}_\theta(X_i = 1|X_{-i} = x_{-i})(1-\mathbb{P}_\theta(X_i = 1|X_{-i} = x_{-i})) \geq \gamma$. Then for any $v \in \mathbb{R}^p$, we have that:*

$$\mathrm{Var}[\langle v, z\rangle] \geq \frac{C_1\gamma^2||v||_2^2}{\omega},$$

where C_1 is a universal constant.

Now, let $\mathcal{L}_{\theta,i}(w)$ be the population negative conditional log-likelihood for node X_i, where X is sampled from the Ising model distribution $\mathbb{P}_\theta$. Formally, $\mathcal{L}_{\theta,i}(w) = -\mathbb{E}_{z\sim\mathbb{P}_\theta}[\ell_i(w;z)]$, where $\ell_i(w;z)$ is the conditional log-likelihood of z under $\mathbb{P}_\theta$ with respect to the i^{th} node. As stated earlier, by the maximum likelihood principle, $\nabla\mathcal{L}_{\theta,i}(2\theta(i)) = \mathbf{0}$. With this definition, we have the Hessian of the population negative conditional log-likelihood as $\nabla^2\mathcal{L}_{\theta,i}(w) = \mathbb{E}_{z\sim\mathbb{P}_\theta}[\nabla^2\ell_i(w;z)]$. Then, we have the following result.

Proposition 61. *Let $\mathbb{P}_\theta$ be an Ising model over p nodes whose parameter satisfies $\max_{i\in[p]}\|\theta(i)\|_\infty \leq \omega$, and let $w \in \mathbb{R}^{p-1}$ be such that $\|w\|_1 \leq 2\omega$. Then, for any vector $v \in \mathbb{R}^{p-1}$, there exists a universal constant $C > 0$ such that:*

$$v^T\nabla^2\mathcal{L}_{\theta,i}(w)v \geq C\frac{\exp(-O(\omega))}{\omega}\|v\|_2^2.$$

Proof. First, observe that

$$\nabla^2 \mathcal{L}_{\theta,i}(w) = \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle))z_{-i}z_{-i}^T\right]$$
$$\Rightarrow v^T\nabla^2 \mathcal{L}_{\theta,i}(w)v = \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle))\langle z_{-i}, v\rangle^2\right].$$

In Lemma 62, we show that for any $\|w\|_1 \leq 2\omega$, we have that

$$\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle)) \geq \frac{\exp(-2\omega)}{4}. \tag{D.3}$$

We now lower bound $\mathbb{E}[\langle z_{-i}, v\rangle^2]$. Since Ising model has zero mean field, we have that $\mathbb{E}[\langle z_{-i}, v\rangle^2] = \mathrm{Var}[\langle z_{-i}, v\rangle]$. Furthermore, due the assumptions placed on the parameter of the Ising model, we obtain that for any $x \in \{-1,+1\}^{p-1}$, $\mathbb{P}_\theta(X_i = 1|X_{-i} = x)(1-\mathbb{P}_\theta(X_i = 1|X_{-i} = x)) \geq \frac{1}{4}\exp(-4\omega)$. This can be shown as follows. For any $z \in \{-1,+1\}$ and $x \in \{-1,+1\}^{p-1}$, we have that:

$$\begin{aligned}
\mathbb{P}_\theta(X_i = z|X_{-i} = x) &= \frac{1}{1+\exp(-z\,\langle 2\theta(i,-i), x\rangle)} \\
&\overset{(i)}{\geq} \frac{1}{1+\exp(2\omega)} \\
&\geq \frac{1}{2\exp(2\omega)} = \frac{\exp(-2\omega)}{2} \\
\Rightarrow \mathbb{P}_\theta(X_i = 1|X_{-i} = x)\mathbb{P}_\theta(X_i = 0|X_{-i} = x) &\geq \frac{\exp(-2\omega)}{2}\frac{\exp(-2\omega)}{2} \\
&= \frac{\exp(-4\omega)}{4}
\end{aligned}$$

where Step (i) uses Hölder's inequality as: $|\,\langle 2\theta(i,-i), x\rangle\,| \leq 2\omega \Rightarrow -z\,\langle 2\theta(i,-i), x\rangle \leq 2\omega$.

Using this in Proposition 60, we have that:

$$\mathrm{Var}[\langle v, z_{-i}\rangle] \geq C\frac{\exp(-8\omega)\|v\|_2^2}{\omega} \tag{D.4}$$

where C is a universal constant.

Combining (D.3) and (D.4), we obtain the statement of the lemma. □

Auxiliary Lemmata

Lemma 62. *If $w \in \mathbb{R}^{p-1}$ such that $||w||_1 \leq 2\omega$, then for $x, y \in \{-1,+1\}^{p-1} \times \{-1,+1\}$:*

$$\sigma(y\langle w, x\rangle)(1-\sigma(y\langle w, x\rangle)) = \frac{\exp(-y\langle w, x\rangle)}{(1+\exp(-y\langle w, x\rangle))^2} \geq \frac{\exp(-|y\langle w, x\rangle|)}{4} \geq \frac{\exp(-2\omega)}{4} \tag{D.5}$$

Proof. Consider $f(a) = \sigma(a)(1-\sigma(a)) = \frac{\exp(-a)}{(1+\exp(-a))^2} = \frac{\exp(a)}{(1+\exp(a))^2}$. Now for $a > 0$:

$$e^{-a} < 1 \Leftrightarrow e^{-a} + 1 < 2 \Leftrightarrow (e^{-a}+1)^2 < 4 \Leftrightarrow \frac{\exp(-a)}{(1+\exp(-a))^2} \geq \frac{\exp(-a)}{4}$$

For $a < 0$:

$$e^a < 1 \Leftrightarrow e^a + 1 < 2 \Leftrightarrow (e^a + 1)^2 < 4 \Leftrightarrow \frac{\exp(a)}{(1+\exp(a))^2} \geq \frac{\exp(a)}{4}$$

Therefore:

$$f(a) \geq \frac{\exp(-|a|)}{4}$$

By Hölder's inequality, $|y\langle w, x\rangle| \leq ||w||_1||x||_\infty \leq 2\omega$. This implies that

$$\sigma(y\langle w, x\rangle)(1 - \sigma(y\langle w, x\rangle)) = f(y\langle w, x\rangle) \geq \frac{\exp(-|y\langle w, x\rangle|)}{4} \geq \frac{\exp(-2\omega)}{4}$$

□

D.3 Proofs of Propositions in Section 4.2

In this section, we present the proofs for Theorem 20 and Lemma 12.

D.3.1 Proof of Theorem 20

Here, we derive bounds on the modulus of continuity defined in (4.4) with the loss function given by the $\ell_{2,\infty}$ norm of the parameters.

Proof Sketch. We begin by giving a brief proof outline. $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$ are two Ising models in the high temperature regime (4.3) with constant α, and additionally satisfy $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$. Consider $\mathcal{L}_{\theta^{(1)},i}$ to be the population negative conditional log-likelihood for the i^{th} node with respect to $\mathbb{P}_{\theta^{(1)}}$ defined earlier. We earlier noted that $\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) = 0$ by the maximum likelihood principle.

In Lemma 63, we show that under these conditions, the gradient $\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))$ satisfies $\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}$, where $C(\alpha)$ is a universal constant only depending on α. With this intermediate result, we complete the proof of the theorem as follows. Considering the Taylor series expansion of $\mathcal{L}_{\theta^{(1)},i}$ around $2\theta^{(2)}(i)$, we get

$$\begin{aligned}
\mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) &= \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle + \frac{1}{2}\Delta_i^T \nabla^2 \mathcal{L}_{\theta^{(1)},i}(\widetilde{w})\Delta_i \\
&\overset{(i)}{\geq} \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle + \frac{C}{2}\frac{\exp(-O(\omega))}{\omega}\|\Delta_i\|_2^2 \\
&\overset{(ii)}{\geq} \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle + C'\frac{\exp(-c(1-\alpha))}{1-\alpha}\|\Delta_i\|_2^2,
\end{aligned}$$

where $\widetilde{w}$ lies between $2\theta^{(2)}(i)$ and $2\theta^{(1)}(i)$, and $\Delta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$. In step (i), we have used the result in Proposition 61 and in step (ii) we use the fact that $\omega \leq 1 - \alpha$.

We also know by the maximum likelihood principle that $\mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) \leq \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))$, and substituting this in the inequality above yields

$$C'\frac{\exp(-c(1-\alpha))}{1-\alpha}\|\Delta_i\|_2^2 \leq -\left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle \leq \left| \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle \right| .$$

Finally, we bound the right hand side using the Cauchy-Schwarz inequality and the result from Lemma 63 to get

$$\left| \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \Delta_i \right\rangle \right| \leq \|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \|\Delta_i\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}\|\Delta_i\|_2,$$

and substituting this in the quadratic bound above gives

$$\|\Delta_i\|_2 \leq C_1(\alpha)\epsilon\sqrt{\log(1/\epsilon)}, \qquad C_1(\alpha) = \frac{1}{C'}(1-\alpha)\exp(c(1-\alpha))\sqrt{C(\alpha)}.$$

□

We now state Lemma 63 and prove it below.

Lemma 63. *Let $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$ be two Ising models in the high temperature regime (4.3) with constant α that satisfies $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$. Then, there exists a universal constant $C(\alpha)$ that only depends on α such that*

$$\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log\left(1/\epsilon\right)} \qquad \textit{for all } i \in [p]$$

Proof. Recall that $\mathcal{L}_{\theta^{(1)},i}(w) = \mathbb{E}_{z\sim\mathbb{P}_{\theta^{(1)}}}[\ell_i(w;z)]$. By the maximum likelihood principle, we know that

$$\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) = \mathbf{0} \qquad \nabla \mathcal{L}_{\theta^{(2)},i}(2\theta^{(2)}(i)) = \mathbf{0}$$

Since $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$, there exists an ϵ-coupling $\mathcal{C}$ between $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$. In particular, $\mathcal{C}$ is a joint distribution over z_1, z_2 such that the respective marginals are $z_1 \sim \mathbb{P}_{\theta^{(1)}}$ and $z_2 \sim \mathbb{P}_{\theta^{(2)}}$, and $\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\mathbb{I}\{z_1 \neq z_2\}] \leq \epsilon$.

The rest of the proof begins by making the observation that $\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) = \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_1)]$. By introducing indicator random variables for the cases when z_1 and z_2 are equal or not, we have

$$\begin{aligned}
\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) &= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 = z_2\}] \\
&= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 = z_2\}] \\
&\overset{(a)}{=} \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] - \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 \neq z_2\}],
\end{aligned}$$

where step (a) follows from the stationarity of $2\theta^{(2)}(i)$ under $\mathbb{P}_{\theta^{(2)}}$ like so.

$$\begin{aligned}
\mathbf{0} &= \nabla \mathcal{L}_{\theta^{(2)},i}(2\theta^{(2)}(i)) \\
&= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_2)] \\
&= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 = z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla \ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 \neq z_2\}].
\end{aligned}$$

Therefore, for any vector $v \in \mathcal{S}^{p-2}$, we have that

$$\begin{aligned}
\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| &= \big|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\{z_1 \neq z_2\}] \\
&\quad - \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\big| \\
&\leq \underbrace{\left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\right|}_{T_1} \\
&\quad + \underbrace{\left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\right|}_{T_2}.
\end{aligned}$$

Bounding T_2: Note that $\nabla \ell_i(w;z_1) = (\sigma(\langle w, z_1(-i)\rangle z_1(i)) - 1)z_1(-i)z_1(i)$. Since $z_1 \sim \{-1,+1\}^p$, we have that $|(\sigma(\langle w, z_1(-i)\rangle z_1(i)) - 1)z_1(i)| < 1$, and hence we get $|\langle v, \nabla \ell_i(w;z_1)| < |\langle v, z_1(-i)|$.

This in turn implies

$$\Pr(|\langle v, \nabla \ell_i(w;z_1)\rangle| > t) \leq \Pr(|\langle v, z_1(-i)\rangle| > t) \overset{(b)}{\leq} 2\exp\left(-\frac{t^2}{C(\alpha)}\right)$$

where step (b) follows from the sub-Gaussianity of random variables distributed with respect to an Ising model in the high temperature regime (Proposition 59). Using standard tail bounds (see [127, Chapter 2]), we obtain that $\mathbb{E}[\exp(\lambda(\langle v, \nabla \ell_i(w; z_1)\rangle))] \leq \exp\left(\frac{C\lambda^2 C(\alpha)}{2}\right)$. To finally bound T_2, we use the following result from [137].

Proposition 64 ([137, Lemma 2.3])**.** *Let Z be a random variable such that $\mathbb{E}[\exp(\lambda Z)] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. For any measurable event A, we have*

$$|\mathbb{E}[Z \cdot \mathbb{I}\{A\}]| \leq \sigma P(A)\sqrt{\log(1/P(A))}.$$

In T_2, the event A is $z_1 \neq z_2$ and this occurs with probability less than ϵ. Hence, we get $T_2 \leq C\sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}$.

Bounding T_1: This can be bounded in an analogous manner as T_2, thus yielding $T_1 \leq C\sqrt{C(\alpha)}\epsilon\sqrt{\log(2/\epsilon)}$.

Plugging these bounds above, we get

$$\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq C\sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon},$$

which proves the statement of the lemma. □

D.3.2 Proof of Lemma 12

Proof. Consider two Ising models with p vertices. For the first Ising model, consider one edge with parameter 2ϵ. The second Ising model has no edges.

Via a simple calculation, the TV distance between these Ising models can be computed to be $\frac{1}{2}\tanh(2\epsilon) \leq \epsilon$. Consequently, the $\ell_{2,\infty}$ norm of the difference in parameters is ϵ, and this proves the lower bound. □

D.4 Proofs of Propositions in Section 4.3

D.4.1 A general result for estimators based on Yatracos classes

Here, we present a result for estimators of the form

$$\mathbb{P}_{\text{est}} = \operatorname*{argmin}_{\mathbb{P}\in\mathcal{P}} \sup_{A\in\mathcal{A}} \left|\mathbb{P}(A) - \widehat{\mathbb{P}}_{n,\epsilon}(A)\right|, \tag{D.6}$$

where $\widehat{\mathbb{P}}_{n,\epsilon}$ the empirical distribution of n samples from the mixture model $\mathbb{P}_\epsilon$ defined in (4.1) and $\mathcal{P}$ is the class of all distributions. Recall that $\mathcal{A}$ is defined as

$$\mathcal{A} = \{A(\mathbb{P}_1, \mathbb{P}_2) : \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}\}, \text{ and } A(\mathbb{P}_1, \mathbb{P}_2) = \{x : \mathbb{P}_1(x) > \mathbb{P}_2(x)\}$$

The result in formally stated in Proposition 13.

Proposition 65. *Given n samples from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q$, the estimator $\mathbb{P}_{est}$ defined in (D.6) satisfies*

$$d_{\text{TV}}(\mathbb{P}_{\text{est}}, \mathbb{P}^\star) \leq 2\epsilon + 2\sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x)\right|$$

Proof. We begin by using $2d_{\text{TV}}(\mathbb{P}_{\text{est}}, \mathbb{P}^\star) = \sum_{x\in\mathcal{X}} |\mathbb{P}_{\text{est}}(x) - \mathbb{P}^\star(x)|$. Consider the sets $B = \{x : \mathbb{P}_{\text{est}}(x) > \mathbb{P}^\star(x)\}$ and $C = \{x : \mathbb{P}_{\text{est}}(x) \leq \mathbb{P}^\star(x)\}$.

This gives us:

$$\begin{aligned}
\sum_{x\in\mathcal{X}} |\mathbb{P}_{\text{est}}(x) - \mathbb{P}^{\star}(x)| &= 2 \max_{A\in\{B,C\}} \left| \sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \mathbb{P}^{\star}(x) \right| \\
&\leq 2 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&= 2 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) + \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&\leq 2 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) \right| + 2 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&\overset{(i)}{\leq} 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&= 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_{\epsilon}(x) + \sum_{x\in A} \mathbb{P}_{\epsilon}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&\leq 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_{\epsilon}(x) \right| + 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \mathbb{P}_{\epsilon}(x) - \sum_{x\in A} \mathbb{P}^{\star}(x) \right| \\
&= 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_{\epsilon}(x) \right| + 4 d_{\text{TV}}(\mathbb{P}_{\epsilon}, \mathbb{P}^{\star}) \\
&\overset{(ii)}{\leq} 4 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_{\epsilon}(x) \right| + 4\epsilon,
\end{aligned}$$

where in step (i) we have used the optimality of $\mathbb{P}_{\text{est}}$ and in step (ii) we have used the fact that $d_{\text{TV}}(\mathbb{P}_{\epsilon}, \mathbb{P}^{\star}) \leq \epsilon$ and this completes the proof. □

D.4.2 Proof of Lemma 13

With the general result for estimators based on Yatracos classes, we state the proof of Lemma 13.

Proof. For the estimator in (4.7), the class of distributions is $\mathcal{G}_{p,k}$. Via Proposition 65, we have that:

$$d_{\text{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^*}) \leq 2\epsilon + \underbrace{2 \sup_{A\in\mathcal{A}} \left| \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_{\epsilon}(x) \right|}_{T_1}$$

Note that distributions in $\mathcal{G}_{p,k}$ are Ising model distributions and are parameterized. Thus, we can alternatively identify the sets $A(\mathbb{P}_1, \mathbb{P}_2)$ via the parameters of Ising model distributions as $A(\theta^{(1)}, \theta^{(2)})$.

Bounding T_1: The set $A(\theta^{(1)}, \theta^{(2)})$ is equivalent to

$$A(\theta^{(1)}, \theta^{(2)}) = \{x : \log \mathbb{P}_{\theta^{(1)}}(x) > \log \mathbb{P}_{\theta^{(2)}}(x)\}$$

Recalling the definitions of $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$, and flattening the parameters to $\mathbb{R}^{\binom{p}{2}}$, we have:

$$A(\theta^{(1)}, \theta^{(2)}) = \left\{y : \left\langle \theta^{(1)}_{\text{flat}} - \theta^{(2)}_{\text{flat}}, y \right\rangle + \log(Z(\theta^{(2)})) - \log(Z(\theta^{(1)})) > 0 \right\} = \{y : \langle w, \widetilde{y} \rangle > 0\}$$

where $w = [\theta^{(1)}_{\text{flat}} - \theta^{(2)}_{\text{flat}}, \log(Z(\theta^{(2)})) - \log(Z(\theta^{(1)}))]$ and $\tilde{y} = [y, 1]$. $Z(\theta)$ is the normalization constant of the probability mass function of an Ising model $\mathbb{P}_\theta$ and $y \in \mathbb{R}^{\binom{p}{2}}$ is a vector of sufficient statistics. Since $\theta^{(1)}, \theta^{(2)} \in \mathcal{G}_{p,k}$, both $\theta^{(1)}_{\text{flat}}$ and $\theta^{(2)}_{\text{flat}}$ can have at most k entries. Consequently, the vector w can have at most $2k + 1$ non-zero entries. Hence, $\mathcal{A}$ can be viewed as a collection of sets:

$$\mathcal{A} = \{\mathbb{I}\{\langle w, y \rangle > 0\} : w \in \mathbb{R}^{\binom{p}{2}}, ||w||_0 \leq 2k + 1\}$$

The following proposition bounds the VC-dimension of sparse linear classifiers:

Proposition 66 ([138, Corollary 1])**.** *Consider the class of linear predictors, defined by the set $S_s = \{v : ||v||_0 \leq s, v \in \mathbb{R}^m\}$* i.e. *the set of s-sparse vectors. The VC-dimension of this class is upper bounded as: $O(s \log(em/s))$.*

Therefore, from the above proposition, we have that the VC-dimension of $\mathcal{A}$ is bounded from above by $\mathcal{O}\,(2k + 1) \log(ep^2/4k+2)$ which is $\mathcal{O}(k \log(ep/k))$. Hence, by a concentration of measure argument, we have that with probability at least $1 - \delta$:

$$T_1 \lesssim \sqrt{\frac{k \log(ep/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

Finally, we obtain

$$d_{\text{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^*}) \leq 2\epsilon + \mathcal{O}\left(\sqrt{\frac{k \log(ep/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

and this recovers the statement of the lemma. □

D.5 Proof of Propositions in Section 4.4

D.5.1 Proof of Theorem 23

Proof Sketch. We give an outline of the proof of the theorem. $\mathbb{P}_{\theta^*}$ is an Ising model in the high temperature regime with constant α. Recall the proposed estimator:

$$\widehat{\theta}(i) = \underset{w \in \mathcal{N}_d^{\gamma}(\mathcal{S}^{p-2})}{\operatorname{argmin}} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(w; x^{(j)})\}_{j=1}^n \right) \right|. \tag{D.7}$$

Proposition 61 states that the negative conditional log-likelihood $\mathcal{L}_{\theta^*,i}$ is $C_2(\alpha)$-strongly convex, where $C_2(\alpha)$ is a universal constant only depending on α. Therefore, by the monotonicity of the gradient of strongly-convex function, we bound the parameter error $\|\widehat{\theta}(i) - \theta^*(i)\|_2$ as

$$\|\widehat{\theta}(i) - \theta^*(i)\|_2^2 \le \frac{1}{C_2(\alpha)} \left\langle \nabla \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) - \nabla \mathcal{L}_{\theta^*,i}(\theta^*(i)), \widehat{\theta}(i) - \theta^*(i) \right\rangle.$$

Next, note that

$$\begin{aligned}
\|\widehat{\theta}(i) - \theta^*(i)\|_2 &\le \frac{1}{C_2(\alpha)} \frac{\left\langle \nabla \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) - \nabla \mathcal{L}_{\theta^*,i}(\theta^*(i)), \widehat{\theta}(i) - \theta^*(i) \right\rangle}{\|\widehat{\theta}(i) - \theta^*(i)\|_2} \\
&\overset{(i)}{\le} \frac{1}{C_2(\alpha)} \sup_{u \in \mathcal{N}_{2d}(\mathcal{S}^{p-2})} \left| \left\langle u, \nabla \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) \right\rangle \right| \\
&\overset{(ii)}{\le} \frac{2}{C_2(\alpha)} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \nabla \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) \right\rangle \right|,
\end{aligned}$$

where in step (i) we have used the facts that 1) $\frac{\widehat{\theta}(i) - \theta^*(i)}{\|\widehat{\theta}(i) - \theta^*(i)\|_2}$ is a unit vector with at most $2d$ non-zero elements, and 2) $\nabla \mathcal{L}_{\theta^*,i}(\theta^*(i)) = \mathbf{0}$ by the maximum likelihood principle, and in step (ii) we have constructed a $1/2$-cover of the set $\mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})$.

We further analyze the right hand side by splitting it into two different terms as follows.

$$\begin{aligned}
&\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) \right\rangle \right| \le \\
&\qquad \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \mathcal{L}_{\theta^*,i}(\widehat{\theta}(i)) \right\rangle - \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i), x^{(j)})\}_{j=1}^n \right) \right|}_{T_1} \\
&\qquad\qquad + \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i), x^{(j)})\}_{j=1}^n \right) \right|}_{T_2}.
\end{aligned}$$

In Lemmas 67 and 68, considering $\gamma = \max\left\{ \frac{\epsilon}{p}, \frac{\log(1/\delta)}{np} \right\}$, and for sufficiently large n (4.10), we bound T_1 and T_2 as $T_1 \le \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d \log(p)}{n}} + \sqrt{\frac{d}{n} \log\left(\frac{3ep}{d\gamma}\right)} \right\}$, and in Lemma

68, we bound T_2 as $T_2 \leq \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)$ respectively.

Plugging these bound into the previous right hand side, we obtain

$$\|\widehat{\theta}(i) - \theta^*(i)\|_2 \lesssim \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right),$$

and this recovers the statement of the theorem. □

We state Lemmas 67 and 68 and prove them below.

Lemma 67. *Consider samples $\{x^{(j)}\}_{j=1}^n$ from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^*} + \epsilon Q$, where $\mathbb{P}_{\theta^*}$ is an Ising model over p nodes in the high temperature regime (4.3) with constant α and with maximum vertex degree d. Suppose n, confidence δ and contamination level ϵ satisfy (4.10). Then,* 1DMean *satisfies*

$$\sup_{w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})} \sup_{u \in \mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})} \left|\langle u, \nabla\mathcal{L}_{\theta^*,i}(w)\rangle - \mathsf{1DMean}\left(\{u^T\nabla\ell_i(w; x^{(j)})\}_{j=1}^n\right)\right|$$
$$\leq \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\}.$$

Proof. Let $z \sim \mathbb{P}_{\theta^*}$. In the proof of Lemma 63, we showed that

$$\Pr(|\langle u, \nabla\ell_i(w; z)\rangle|) \leq 2\exp\left(-\frac{t^2}{C(\alpha)}\right)$$

holds due to the form of the gradient and the sub-Gaussianity of the Ising model distribution. This implies that the gradients of ℓ_i due to non-outlier samples are sub-Gaussian. This allows us to leverage techniques from [103] to produce a guarantee for the 1DMean algorithm when the true distribution is sub-Gaussian in Lemma 69. This states that

$$\left|\langle u, \nabla\mathcal{L}_{\theta^*,i}(w)\rangle - \mathsf{1DMean}\left(\{u^T\nabla\ell_i(w; x^{(j)})\}_{j=1}^n\right)\right| \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{C(\alpha)}{n}\log\left(\frac{1}{\delta}\right)},$$

where $w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $u \in \mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})$.

Finally, to convert the point-wise bound to a uniform bound, we perform a union bound over all the elements in $\mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $\mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})$, and use the fact that the number of elements in the cover can be bounded as $|\mathcal{N}_k^\gamma(\mathcal{S}^{p-2})| \leq \left(\frac{3ep}{k\gamma}\right)^k$ to recover the statement of the result. □

Lemma 68. *Given samples $\{x^{(j)}\}_{j=1}^n$ from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^*} + \epsilon\mathbb{Q}$, where $\mathbb{P}_{\theta^*}$ is an Ising model over p nodes in the high temperature regime (4.3) with constant α,* *there exists a constant $C(\alpha)$ that only depends on α such that:*

$$\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i); x^{(j)}\}_{j=1}^n\right)\right| \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d \log(p)}{n}} + \sqrt{\frac{d}{n} \log\left(\frac{3ep}{d\gamma}\right)} \right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)$$

where $\widehat{\theta}(i)$ is as defined in (4.9) with $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{p}\right\}$.

Proof. First, define $C_\gamma(\theta^*(i))$ as the element closest to $\theta^*(i)$ in the set $\mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$.

We begin the proof by recognizing that

$$\begin{aligned} &\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i); x^{(j)})\}_{j=1}^n\right)\right| \\ &\quad \overset{(i)}{\leq} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^*(i)); x^{(j)})\}_{j=1}^n\right)\right| \\ &\overset{(ii)}{\leq} \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^*(i)); x^{(j)})\}_{j=1}^n\right) - \langle u, \nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i)))\rangle\right|}_{T_{2,1}} \\ &\qquad + \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} |\langle u, \nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i)))\rangle|}_{T_{2,2}} \end{aligned}$$

where Step (i) uses the optimality of $\widehat{\theta}(i)$ and Step (ii) performs splitting by addition and subtraction as mentioned earlier.

Bounding $T_{2,1}$: $T_{2,1}$ can be bounded using Lemma 67, since it holds for any $w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $C_\gamma(\theta^*(i)) \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2}$ by definition. Therefore, we get

$$\begin{aligned} &\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^*(i)); x^{(j)})\}_{j=1}^n\right) - \langle u, \nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i)))\rangle\right| \\ &\qquad \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d \log(p)}{n}} + \sqrt{\frac{d}{n} \log\left(\frac{3ep}{d\gamma}\right)} \right\}. \end{aligned}$$

Bounding $T_{2,2}$: $T_{2,2}$ can be bounded as follows:

$$\begin{aligned} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} |\langle u, \nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i)))\rangle| &\leq \|\nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i)))\|_2 \\ &= \|\nabla \mathcal{L}_{\theta^*,i}(C_\gamma(\theta^*(i))) - \nabla \mathcal{L}_{\theta^*,i}(\theta^*(i))\|_2 \\ &\leq L\|C_\gamma(\theta^*(i)) - \theta^*(i)\|_2 \leq L\gamma, \end{aligned}$$

where L is the Lipschitz constant of $\mathcal{L}_{\theta^*,i}$. A simple calculation reveals that:

$$\begin{aligned}\nabla^2 \mathcal{L}_{\theta^*,i}(w) &= \mathbb{E}_{x\sim P_{\theta^*}}[\sigma(\langle w, x(-i)\rangle\, x_i)(1-\sigma(\langle w, x(-i)\rangle\, x_i))x(-i)x(-i)^T] \\ \Rightarrow v^T\nabla^2 \mathcal{L}_{\theta^*,i}(w)v &= \mathbb{E}_{x\sim P_{\theta^*}}[\sigma(\langle w, x(-i)\rangle\, x_i)(1-\sigma(\langle w, x(-i)\rangle\, x_i))(\langle v, x(-i)\rangle)^2] \\ &\overset{(i)}{\leq} \frac{1}{4}\mathbb{E}_{x\sim P_{\theta^*}}[(v^T x_i)^2] \overset{(ii)}{\leq} \frac{p}{4}\|v\|_2^2\end{aligned}$$

where in Step (i) we have used the fact that $\sigma(z)(1-\sigma(z)) \leq \frac{1}{4}$ and in Step (ii) we have used the Cauchy-Schwarz inequality, leading to $L = p$.

With the choice of $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{n}\right\}$, we have the final result

$$\begin{aligned}&\sup_{u\in\mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left|\mathsf{1DMean}\left(\{u^T\nabla\ell_i(\widehat{\theta}(i); x^{(j)})\}_{j=1}^n\right)\right| \\ &\qquad \leq \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)\end{aligned}$$

and this completes the proof. □

Auxiliary Results

Here we state and prove Lemma 69, which we use in the proof of Lemma 67.

Lemma 69 ([103, Lemma 3])**.** *Suppose $\mathbb{P}^\star$ is a sub-Gaussian distribution with variance proxy σ^2 and mean $\mu = \mathbb{E}_{x\sim\mathbb{P}^\star}[x]$. Given n samples from the mixture distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q$, Algorithm 6 returns an estimate $\widehat{\theta}_\delta$ that satisfies*

$$|\widehat{\theta}_\delta - \mu| \lesssim \epsilon\sqrt{\sigma^2\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\sigma^2\log\left(\frac{1/\delta}{n}\right)}$$

with probability at least $1-\delta$.

Proof. The proof mostly follows the proof in [103].

Let $I^\star$ be the interval $\mu \pm \sqrt{\sigma^2\log\left(\frac{1}{\delta_1}\right)}$. For notational convenience, let $f_n(u,v) = \sqrt{u(1-u)}\sqrt{\frac{\log(1/v)}{n}} + \frac{2}{3}\frac{\log(1/v)}{n}$. Let $\widehat{I} = [a,b]$ be the interval obtained using the first split of the sample set $\mathcal{Z}_1$ *i.e.* the shortest interval containing $n(1-(\delta_1+\epsilon+f_n(\epsilon+\delta_1,\delta_3)))$ points of $\mathcal{Z}_1$. In Algorithm 6, we have $\delta_1 = \epsilon$ and $\delta_3 = \delta/4$.

From [103, Claim 5], we have that

$$\text{length}(\widehat{I}) \leq \text{length}(I^\star) \leq 2\sqrt{\sigma^2\log\left(\frac{1}{\delta_1}\right)}.$$

To bound the error of the estimator, we analyze the quantity

$$\left| \frac{1}{|\widehat{I}|} \sum_{z_i \in \mathcal{Z}_2} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} - \mu \right|,$$

where $|\widehat{I}| = \sum_{z_i \in \mathcal{Z}_2} \mathbb{I}\left\{ z_i \in \widehat{I} \right\}$.

We do so by casing on whether a sample z_i was sampled from $\mathbb{P}^\star$ or from Q, like so.

$$\begin{aligned}
\left| \frac{1}{|\widehat{I}|} \sum_{z_i \in \mathcal{Z}_2} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} - \mu \right| &= \left| \frac{1}{|\widehat{I}|} \left(\sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim \mathbb{P}^\star}} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} + \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim Q}} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} \right) - \mu \right| \\
&\leq \underbrace{\left| \frac{1}{|\widehat{I}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim \mathbb{P}^\star}} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} - \mu \right|}_{T_1} + \underbrace{\left| \frac{1}{|\widehat{I}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim Q}} z_i \mathbb{I}\left\{ z_i \in \widehat{I} \right\} - \mu \right|}_{T_2}.
\end{aligned}$$

Bounding T_1: From [103, Claim 6], we bound T_1 with probability at least $1 - \delta_3 - \delta_5$ as

$$T_1 \leq \frac{\epsilon + f_n(\epsilon, \delta_5)}{1 - \delta_4} \cdot 4\sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)},$$

where $\delta_4 = (\delta_1 + \epsilon) + f_n(\delta_1 + \epsilon, \delta_3)$.

Bounding T_2: To bound T_2, we split the terms further.

$$\begin{aligned}
T_2 = \left| \frac{1}{|\widehat{I}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} (z_i - \mu) \right| &= \frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \frac{1}{|\widehat{I}_{\mathbb{P}^\star}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} (z_i - \mu) \right| \\
&\leq \underbrace{\frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \left(\frac{1}{|\widehat{I}_{\mathbb{P}^\star}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} z_i \right) - \mathbb{E}[x | x \in \widehat{I}, x \sim \mathbb{P}^\star] \right|}_{T_{2,1}} \\
&\quad + \underbrace{\frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \mathbb{E}[x | x \in \widehat{I}, x \sim \mathbb{P}^\star] - \mu \right|}_{T_{2,2}},
\end{aligned}$$

where $|\widehat{I}_{\mathbb{P}^\star}| = \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim \mathbb{P}^\star}} \mathbb{I}\left\{z_i \in \widehat{I}\right\}$ is the number of elements in $\mathcal{Z}_2$ that were originally sampled from $\mathbb{P}^\star$.

$T_{2,1}$ is the deviation of the mean of the samples originally sampled from Q and remain in $\widehat{I}$ from the mean of $\mathbb{P}^\star$ conditioned on the event that they belong to $\widehat{I}$ as well. $T_{2,2}$ measures the deviation of the mean of $\mathbb{P}^\star$ from the mean of the same distribution conditioned on $\widehat{I}$.

Bounding $T_{2,1}$: We bound $T_{2,1}$ using [103, Lemma 15]. With this result, we get that with probability at least $1-\delta_7$,

$$T_{2,1} \leq \sqrt{\frac{2\sigma^2 \log(3/\delta_7)}{\mathbb{P}^\star(\widehat{I})}} + 2\sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)\frac{\log(3/\delta_7)}{|\widehat{I}_{\mathbb{P}^\star}|}}.$$

Bounding $T_{2,2}$: To control $T_{2,2}$ we make use of Proposition 64 in conjuction with [103, Lemma 14] to get

$$T_{2,2} \leq 2\mathbb{P}^\star(\widehat{I}^c)\sqrt{\sigma^2 \log\left(\frac{1}{\mathbb{P}^\star(\widehat{I}^c)}\right)},$$

where $\mathbb{P}^\star(A)$ is the probability that $z \sim \mathbb{P}^\star$ lies in A. Finally, we bound $\mathbb{P}^\star(\widehat{I}^c$ using [103, Claim 7] to obtain with probability at least $1-\delta_6$ that

$$\mathbb{P}^\star(\widehat{I}^c) \leq C_1\epsilon + C_2\delta_1 + C_3\frac{\log(n)}{n} + C_4\frac{\log(1/\delta_6)}{n} + C_5\frac{\log(1/\delta_3)}{n},$$

where $\{C_i\}_{i=1}^6$ are universal constants.

Therefore, combining the bounds for T_1, $T_{2,1}$ and $T_{2,2}$, and setting $\delta_1 = \epsilon$, $\delta_3 = \delta_5 = \delta_6 = \delta_7 = \delta/4$ and noting that for the choice of n $|\widehat{I}_{\mathbb{P}^\star}| \geq \frac{n}{2}$, we get the final deviation bound:

$$T_1 + T_{2,1} + T_{2,2} \lesssim \epsilon\sqrt{\sigma^2 \log\left(\frac{1}{\epsilon}\right)} + \sqrt{\sigma^2 \log\left(\frac{1/\delta}{n}\right)},$$

and this completes the proof of the lemma. □

D.6 Proof of Theorem 38

In this section, we present the proof of Theorem D.6. The proof mostly follows the analysis in the proofs of Lemma 63 and Theorem 20. The only difference is that we will not be able to use the sub-Gaussianity of Ising model distributions anymore, as it is no longer applicable.

Proof. Following the proof of Lemma 63, we have for any v such that $\|v\|_1 = 1$ that

$$\begin{aligned}\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| &\leq \left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\left\{z_1 \neq z_2\right\}\right]\right| \\ &\quad + \left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\left\{z_1 \neq z_2\right\}\right]\right| \\ &\overset{(i)}{\leq} \underbrace{\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right]}_{T_1} \\ &\quad + \underbrace{\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right]}_{T_2},\end{aligned}$$

where in step (i), we have used Jensen's inequality for $f(x) = |x|$.

Bounding T_1: By Hölder's inequality $\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq \|v\|_1 \left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty$. Again by Jensen's inequality, and the explicit form of $\nabla \ell_i$, we have $\left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty = \left\|\mathbb{E}\left[\nabla \ell_i(2\theta^{(2)}(i), z_1)\right]\right\|_\infty \leq \mathbb{E}\left[\|\nabla \ell_i(2\theta^{(2)}(i), z_1)\|_\infty\right] \leq 1$. Therefore,

$$\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right] \leq \mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\mathbb{I}\left\{z_1 \neq z_2\right\}\right] \leq \epsilon.$$

Bounding T_2: T_2 can be bounded in the exact same way as T_1.

Plugging these bounds, we get that

$$\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq 2\epsilon.$$

Now, following the first part of the proof of Theorem 20, we have using Hölder's inequality and the bound above that

$$\frac{C}{2}\frac{\exp(-O(\omega))}{\omega}\|\Delta_i\|_2^2 \leq \left|\left\langle \Delta_i, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq \|\Delta_i\|_1 \left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty \leq 2\epsilon\|\Delta_i\|_1$$

where $\Delta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$. Now, since $\theta^{(1)}$ and $\theta^{(2)}$ are parameters of Ising models with maximum vertex degree d, $\Delta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$ has atmost $2d$ non-zero elements. Consequently, we get $\|\Delta_i\|_1 \leq \sqrt{d}\|\Delta_i\|_2$.

Finally, plugging the above norm inequality in the previous bound, we have:

$$\|\Delta_i\|_2 \lesssim \epsilon\sqrt{d}\omega \exp(O(\omega)).$$

Analogously, since $d \leq k$ when $G^{(1)}, G^{(2)} \in \mathcal{G}_{p,k}$, we have

$$\|\Delta_i\|_2 \lesssim \epsilon\sqrt{k}\omega \exp(O(\omega)),$$

Alternatively, note that by the triangle inequality: $\|\Delta_i\|_1 \leq \|2\theta^{(1)}(i)\|_1 + \|2\theta^{(2)}(i)\|_1 \leq 4\omega$. This gives us:

$$\|\Delta_i\|_2 \lesssim \sqrt{\epsilon}\omega \exp(O(\omega))$$

Since both types of inequalities holds simultaneously, we recover the statements of the theorem for $\mathcal{G}_{p,d}$ and $\mathcal{G}_{p,k}$. $\square$