Robust Mean Estimation Against Oblivious Adversaries

Shuchen Li

CMU-CS-23-146

December 2023

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

Thesis Committee:
Pravesh K. Kothari, Chair
David P. Woodruff

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

Keywords: robust mean estimation, sparse Fourier transform

To my family.

Abstract

We give the first algorithm that achieves *arbitrarily* high accuracy for robust mean estimation on Gaussian and Laplace distribution when the adversary corrupts the samples *before* the noise is added. For sufficiently small constant $\alpha > 0$ and accuracy $\varepsilon > 0$, our algorithm takes as input samples $Y \subseteq \mathbb{R}^d$ of size n obtained by i.i.d. samples $X \subseteq \mathbb{R}^d$ of size n from the true distribution with unknown mean μ, while an adversary corrupts an α fraction of the points before the (Gaussian or Laplace) noise is added. When the true distribution is Gaussian with unknown mean μ and covariance I, our algorithm needs sample size $n = 2^{O(d/\varepsilon^2)}$. When the true distribution is Laplace with unknown mean μ and covariance I, our algorithm needs sample size $n = \widetilde{O}(d^2/\varepsilon^4)$. Our algorithm runs in $\widetilde{O}(nd)$ time, and outputs an estimation $\hat{\mu}$ that $\|\hat{\mu} - \mu\|_2 \leq \varepsilon$ with high probability. Our method is to transform the sample to a Fourier-sparse signal that encodes the true mean μ and apply the sparse Fourier transform to decode μ.

In Huber's contamination model, where the adversary corrupts an α fraction of the sample *after* the noise is added, it is known that there is a $\Omega(\alpha)$ lower bound on the error of the estimator. In contrast, our algorithm can estimate the mean arbitrarily closely in this corruption-before-noise setting.

Our algorithm in this new setting has many possible applications. For example, the one that motivated us is max-affine regression. In max-affine regression, the model is the maximum of k linear models, where the noise is added after taking the maximum. If we can extend our algorithm to the *list-decodable* setting, then we immediately get an algorithm for the max-affine regression.

Acknowledgments

While the Master's program spanned only three semesters, my profound gratitude extends to numerous people for their support along the way.

First and foremost, I would like to express my gratitude to my amazing advisor, Professor Pravesh Kothari. Ever since I took his class in my first semester at CMU, Pravesh has been an important source of inspiration and support. He has provided many insights that helped me understand the problems, and more importantly, guided me on how to approach research in general. I am also grateful for Pravesh's generous help with my graduate school applications. I would hardly have survived these challenging months stuck on the research project and busy preparing the applications without his support and encouragement.

Next, I would like to thank Professor Manolis Zampetakis, who worked with me on this project. Although we haven't been able to meet in person throughout this project, I have learned a lot from the ideas and feedback he provided during our Zoom calls. I appreciate all the time he spent working with me, and I hope we can continue our collaboration on this project and, of course, meet in person sometime in the future.

I would like to thank Professor David Woodruff for taking the time to be on my thesis committee despite his busy schedule while at Berkeley. I also appreciate the insightful comments he provided on my thesis.

I would like to express my gratitude to Professor Anupam Gupta for helping me with my graduate school applications and for organizing the group lunch with Pravesh. I really enjoyed these interesting conversations and talks during the weekly lunch.

I would like to extend my thanks to Professor David O'Hallaron for his practical advice on giving talks and encouragement before the defense. I am also grateful to Professor Dave Eckhardt and Angy Malloy for supporting the MSCS program.

Finally, I would like to express my appreciation to my friends, my family, and my girlfriend for their unwavering support and boundless love throughout the highs and lows of this journey. I could not have come so far without them.

Contents

Chapter 1

Introduction

1.1 Background

Traditional statistical methods often assume the input samples are generated from some known distribution. However, real-world data are not necessarily from these known distributions, and may exhibit anomalies, errors, or extreme values that can significantly affect the results of statistical estimations. These outliers can arise from various sources, including measurement errors, adversarial manipulation, or genuine deviations from the assumed underlying model. Meanwhile, the field of *robust statistics* [HR09; Tuk75; HRR+11; MMY+19] has been studying this phenomenon and designing estimators that are robust in the presence of deviations of the model.

To capture the contamination in the samples, Huber introduced the following model [Hub64].

Definition 1. *In Huber's contamination model, the samples are generated from a distribution*

$$(1-\alpha)D + \alpha Z,$$

where D is the true underlying distribution, Z is a contamination distribution chosen by an adversary, and α is the fraction of corruption.

If the number of the samples is large enough, then about $(1-\alpha)$ fraction of the samples are from the true distribution D, called the *inliers*, and about α fraction of the samples are from the contamination distribution Z, called the *outliers*.

There is a stronger model, where the adversary is allowed to observe the inliers before adding outliers, defined as follows.

Definition 2. *In the strong contamination model, the samples $\{y_1, y_2, \ldots, y_n\}$ are generated as follows.*

1. *Sample $y_1, y_2, \ldots, y_{(1-\alpha)n}$ from the true distribution D.*
2. *After observing $y_1, y_2, \ldots, y_{(1-\alpha)n}$, the adversary picks the rest αn samples.*

The samples are received in random order.

For one-dimensional samples, there are many robust estimators for the mean, for example, median, trimmed mean, and winsorized mean; and robust estimators for the variance, for example, median absolute deviation and interquartile range. Moreover, they are efficient to compute.

However, for high-dimensional samples, classic estimators, such as the Tukey median, are intractable to compute or have dimension-dependent errors. Thus, there have been extensive works on *algorithmic* robust statistics, which aim at designing efficient algorithms to robustly estimate the parameters of the true distribution. Ever since the breakthrough of Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart [DKK+16] and Lai, Rao, and Vempala [LRV16] that gave the first polynomial time algorithms for robust estimation with dimension-independent error, there has been a line of works on this topic (e.g., [DKK+17; BDL+17; KKM18; HL18; DKK+18; DKK+19]) aiming at robust mean estimation, covariance estimation, regression, learning Gaussian mixtures, optimization, etc.

Another interesting setting is when there is a *majority* of the samples being corrupted. In this setting, it is impossible to produce a unique accurate estimator. Because the adversary can add a majority of the samples from a distribution with parameters far away from those of the true distribution, so that we cannot distinguish the two distributions. Therefore, our goal in this setting is to output a list of the possible parameters, such that the true parameter is close to one of the parameters in the list. This is called the *list-decodable* setting, introduced by Balcan, Blum, and Vempala [BBV08]. Begin with the influential work of Charikar, Steinhardt, and Valiant [CSV17] that gave the first list-decodable mean estimation algorithm, there has been plenty of subsequent works on list-decodable mean estimation [KSS18; DKK20; DKS18; DKK+21; CMY20], linear regression [KKK19; RY20a], subspace recovery [RY20b; BK21], covariance estimation [IK22], etc.

While there are extensive algorithmic results on robust statistics, it is known that, even in the most fundamental setting, robust mean estimation for one-dimensional Gaussian distribution in Huber's contamination model, there is an information-theoretic lower bound stating that it is impossible to estimate the mean within additive error $\left(\sqrt{\frac{\pi}{2}} - o(1)\right)\alpha$ when there is an α fraction of corruption [DKK+18]. Thus, no algorithm can output an estimator of the mean with an error better than $O(\alpha)$, no matter how many samples are given, even in Huber's contamination model. Therefore, a natural question to ask is, whether one can further relax this adversarial model such that there is an algorithm that achieves $o(\alpha)$ additive error, given enough samples.

1.2 Our Results and Techniques

Compared to the adversary in the strong contamination model, the adversary in Huber's contamination model is oblivious to the inliers. One relaxation is to make the adversary also oblivious to the "noise". Here we view the process of generating a random variable from a certain distribution $D(\mu)$ with mean μ as adding a noise sampled from $D(0)$ to the mean μ, where $D(0)$ is the translation of $D(\mu)$ with zero mean. That is, the adversary adds αn means to the $(1-\alpha)n$ true means, then the noise is applied independently. As a concrete example, consider the true distribution to be the one-dimensional Gaussian distribution with unknown mean μ and unit variance. First, the adversary chooses $z_1, z_2, \ldots, z_{\alpha n}$ as the corrupted means, adding to $(1-\alpha)n$ true means μ. Then the standard Gaussian noise is independently added to these means. The algorithm receives these

samples in random order.

In this *noise-oblivious* model, we designed an algorithm that can achieve arbitrarily small additive error in the mean estimator, given enough samples:

Theorem 1.1. *For sufficiently small $\alpha > 0$, in the noise-oblivious model with α fraction of corruption, there is an algorithm that outputs the estimator of the mean of the true distribution D within ε additive error in ℓ_2 norm.*

1. *For D being the d-dimensional Gaussian distribution with mean μ and identity covariance, the algorithm needs $2^{O(d/\varepsilon^2)}$ samples and runs in $2^{O(d/\varepsilon^2)}$ time.*
2. *For D being the d-dimensional Laplace distribution with mean μ and identity covariance, the algorithm needs $\widetilde{O}(d^2/\varepsilon^4)$ samples and runs in $\widetilde{O}(d^3/\varepsilon^4)$ time.*

Our method is based on the characteristic function of the true distribution D. If D has its characteristic function in the form of $e^{it\mu} f(t)$, then one can divide the characteristic function by $f(t)$, and apply the Fourier transform to recover the frequency μ from $e^{it\mu}$. Note that the characteristic function is the inverse Fourier transform of the probability density function. And $f(t)$ is the characteristic function of the noise, i.e., the translation of D with zero mean. So this process is equivalent to deconvolving the noise from the distribution, producing a spike located in μ.

Another important ingredient in our algorithm is Price and Song's algorithm for sparse Fourier transform in the continuous setting [PS15]. Their algorithm enables us to recover the major frequency from the signal in the presence of noise. The noise in the signal comes from two sources: the error from estimating the characteristic function, and the corruption introduced by the adversary. We show that both of the noise can be controlled, so that we can succeed in recovering the frequency, i.e., the true mean.

Chapter 2

Preliminaries

2.1 Notations

We will use $N(\mu, \Sigma)$ to denote the d-dimensional Gaussian distribution with mean μ and covariance matrix Σ, which is positive semi-definite, with density function

$$f(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}{2}\right),$$

and Laplace(μ, Σ) to denote the d-dimensional Laplace distribution with mean μ and covariance matrix Σ, which is positive semi-definite, with density function

$$f(x) = \frac{2}{\sqrt{(2\pi)^k \det(\Sigma)}} \left(\frac{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}{2}\right)^{v/2} K_v\left(\sqrt{2(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}\right),$$

where $v = (2-d)/2$ and K_v is the modified Bessel function of the second kind. The marginal distribution of a d-dimensional Laplace distribution is a one-dimensional Laplace distribution, with density function in the form of

$$f(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|x-\mu|}{\sigma}\right).$$

2.2 Characteristic Function

For random variable $X \in \mathbb{R}^d$, the characteristic function $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ is defined as

$$\varphi_X(t) = \mathbb{E}[e^{it^{\mathrm{T}}X}].$$

We will need the following facts about the characteristic functions of certain random variables.

Fact 2.1. *If $X \sim N(\mu, \sigma^2)$, then the characteristic function*

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = e^{it\mu - \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R}.$$

Fact 2.2. *If $X \sim N(\mu, \Sigma)$, then the characteristic function*

$$\varphi_X(t) = \mathbb{E}[e^{it^{\mathrm{T}}X}] = e^{it^{\mathrm{T}}\mu - \frac{1}{2}t^{\mathrm{T}}\Sigma t}, \quad t \in \mathbb{R}^d.$$

Fact 2.3. *If $X \sim \mathrm{Laplace}(\mu, \sigma^2)$, then the characteristic function*

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \frac{e^{it\mu}}{1 + \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R}.$$

Fact 2.4. *If $X \sim \mathrm{Laplace}(\mu, \Sigma)$, then the characteristic function*

$$\varphi_X(t) = \mathbb{E}[e^{it^{\mathrm{T}}X}] = \frac{e^{it^{\mathrm{T}}\mu}}{1 + \frac{1}{2}t^{\mathrm{T}}\Sigma t}, \quad t \in \mathbb{R}^d.$$

Fact 2.5. *Suppose X is distributed according to the noncentral chi-squared distribution, that is,*

$$X = \sum_{i=1}^{k} X_i^2,$$

where X_i are i.i.d. Gaussian random variables with means μ_i and unit variances, for $i = 1, 2, \ldots, k$. Then the characteristic function

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \frac{\exp\left(\frac{i\lambda t}{1-2it}\right)}{(1-2it)^{k/2}}, \quad t \in \mathbb{C}, s.t.\ |\Im t| \le 1/2,$$

where $\lambda = \sum_{i=1}^{k} \mu_i^2$.

2.3 Sparse Fourier Transform

Price and Song gave an algorithm for sparse Fourier transform in the continuous setting [PS15]. Their result is as follows.

Theorem 2.6 (Theorem 1.1 in [PS15])**.** *Consider any signal $x(t) = x^*(t) + g(t) \in \mathbb{C}$, $t \in [0, T]$, for arbitrary noise $g(t)$ and exactly k-sparse $x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi i f_j t}$ with $f_i \in [-F, F]$ and frequency separation $\eta \le \min_{i \neq j} |f_i - f_j|$. For some $\delta > 0$, define the "noise level"*

$$\mathcal{N}^2 := \frac{1}{T}\int_0^T |g(t)|^2 \mathrm{d}t + \delta \sum_{j=1}^{k} |v_j|^2.$$

Then there's an algorithm that takes $O(k \log(FT) \log(k/\delta) \log k)$ samples from $x(t)$ over any duration $T > O(\log(k/\delta)/\eta)$, returns $\{(v_i', f_i')\}$ in $O(k \log(FT) \log(FT/\delta) \log k)$ time, such that for any v_i with $|v_i| = \Omega(\mathcal{N})$, $|f_i' - f_i| \le O(\frac{\mathcal{N}}{T|v_i|})$, with constant probability.

2.4 Lower Bounds for Huber's Contamination Model

Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart proved an information-theoretic lower bound on the error of the mean estimator for one-dimensional unit-variance Gaussian distribution in Huber's contamination model [DKK+18].

Lemma 2.7 (Lemma 17 in [DKK+18])**.** *It is impossible to estimate the mean of a one-dimensional unit-variance Gaussian distribution within additive error* $\left(\sqrt{\frac{\pi}{2}} - o(1)\right)\alpha$ *in Huber's contamination model with* α *fraction of corruption.*

Their proof also extends to Laplace distributions. Here we present their proof first.

Proof. It suffices to show that $p_1 = N(-\varepsilon, 1)$ and $p_2 = N(\varepsilon, 1)$ can be corrupted into the same distribution p with density function $f(x) = \max\{p_1(x), p_2(x)\}/\eta$, where η is the normalizing constant. Compute

$$\begin{aligned}
\eta &= \int_{-\infty}^{\infty} \max\{p_1(x), p_2(x)\}\,\mathrm{d}x \\
&= 2\int_{0}^{\infty} p_2(x)\mathrm{d}x \\
&= 2\left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\varepsilon}{\sqrt{2}}\right)\right) \\
&= 1 + \operatorname{erf}\left(\frac{\varepsilon}{\sqrt{2}}\right) \\
&= 1 + \sqrt{\frac{2}{\pi}}\varepsilon - O\left(\varepsilon^3\right).
\end{aligned}$$

Thus, $f(x) = \max\{p_1(x), p_2(x)\}/\eta \geq p_1(x)/\eta \geq \left(1 - \sqrt{\frac{2}{\pi}}\varepsilon + O(\varepsilon^3)\right) p_1(x)$. Therefore, when $\varepsilon \leq \left(\sqrt{\frac{\pi}{2}} - o(1)\right)\alpha$, we have that $f(x) \geq (1-\alpha)p_1(x)$. Let $q_1(x) = \frac{f(x)-(1-\alpha)p_1(x)}{\alpha}$ be the contamination distribution chosen by the adversary, since $q_1(x) \geq 0$ and $\int_{-\infty}^{\infty} q_1(x) = 1$. Then $p_1(x)$ can be corrupted into $f(x)$, and similarly $p_2(x)$ can also be corrupted into $f(x)$. So the algorithm cannot distinguish whether the true distribution is p_1 or p_2 given the corrupted distribution f, and the best mean estimator in this scenario is 0, with error ε. □

Lemma 2.8. *It is impossible to estimate the mean of a one-dimensional unit-variance Laplace distribution within additive error* $\left(\frac{1}{2\sqrt{2}} - o(1)\right)\alpha$ *in Huber's contamination model with* α *fraction of corruption.*

Proof. The proof is similar to that of the Gaussian case. For $p_1 = \text{Laplace}(-\varepsilon, 1)$ and $p_2 =$

Laplace$(\varepsilon, 1)$, $f(x) = \max\{p_1(x), p_2(x)\}/\eta$. Compute

$$\begin{aligned}\eta &= \int_{-\infty}^{\infty} \max\{p_1(x), p_2(x)\}\,\mathrm{d}x \\ &= 2\int_{0}^{\infty} p_2(x)\mathrm{d}x \\ &= e^{-\sqrt{2}\varepsilon}(2e^{\sqrt{2}\varepsilon} - 1) \\ &\le 2(1 + \sqrt{2}\varepsilon + O(\varepsilon^2)) - 1 \\ &= 1 + 2\sqrt{2}\varepsilon + O(\varepsilon^2).\end{aligned}$$

So similarly, when $\varepsilon \le \left(\frac{1}{2\sqrt{2}} - o(1)\right)\alpha$, p_1 and p_2 can be corrupted into the same distribution with density f so that the algorithm cannot distinguish within error ε. □

2.5 Noise-Oblivious Contamination Model

Let $D(\mu)$ denote the translation of distribution D with mean μ. For example, if D is the Gaussian distribution with identity covariance, then $D(\mu)$ is the Gaussian distribution with mean μ and identity covariance.

Our *noise-oblivious contamination model* is defined as follows.

Definition 3. *Let α be the fraction of corruption. The noise-oblivious contamination model generates a set of samples $X_1, X_2, \ldots, X_n$ on a distribution $D(\mu)$ with unknown mean μ by the following process:*

1. *The model produces $(1-\alpha)n$ true means μ.*
2. *An adversary adds αn points $z_1, z_2, \ldots, z_{\alpha n}$ arbitrarily as the means of contamination.*
3. *The model adds i.i.d. noise from the distribution $D(0)$ to the n means.*

Then the algorithm is given the set of samples $X_1, X_2, \ldots, X_n$ in random order, where $(1-\alpha)$ fraction of the samples are generated from $D(\mu)$, and the rest are generated from $D(z_i)$ for $i = 1, 2, \ldots, \alpha n$, independently. We say the set of samples $X_1, X_2, \ldots, X_n$ is α-corrupted.

This model is a relaxation of Huber's contamination model, in the sense that we required the contamination distribution chosen by the adversary to be *structured*, that is, a mixture of the translations of the underlying distribution.

Meanwhile, recovering the mean μ from this model can be also viewed as an easier problem than the problem of learning mixtures, when α is small. For example, when the underlying distribution D is the Gaussian distribution with identity covariance, the output distribution of this model can be viewed as a mixture of spherical Gaussians. However, we only care about the location of the component with the largest weight $1-\alpha$, which is the true mean μ.

2.6 Median

For some one-dimensional symmetric distribution D, the median can estimate the mean up to $O(\alpha)$ additive error under our noise-oblivious model. Let $F(x) = \Pr_{X\sim D(0)}[X \le x]$ denote the cumulative distribution function of $D(0)$. Then we have the following bound.

Lemma 2.9. *For $X_1, X_2, \ldots, X_n$ generated from the α-corrupted noise-oblivious model on distribution $D(\mu)$, where $\alpha < 1/2$. Let $m = \text{median}(X_1, X_2, \ldots, X_n)$, then*

$$\Pr[|m-\mu| > t] \le 2\exp\left(-2(1-\alpha)n\left(F(t) - \frac{1}{2} - \alpha\right)^2\right).$$

Proof. It suffices to prove $\Pr[m-\mu > t] \le \exp\left(-2(1-\alpha)n\left(F(t) - 1/2 - \alpha\right)^2\right)$. Since there are αn outliers in the samples, the median is at most the $\frac{1/2}{1-\alpha}$-quantile of the inliers (all the outliers are larger than all the inliers), which is at most the $(\frac{1}{2}+\alpha)$-quantile of the inliers, denoted by m'. Let I denote the set of the indices of the inliers, and for $i \in I$, let $Z_i = \mathbb{1}[X_i - \mu > t]$. Then $m' - \mu > t$ is equivalent to $\frac{1}{(1-\alpha)n}\sum_{i\in I} Z_i \ge \frac{1}{2} - \alpha$. Note that $\mathbb{E}[Z_i] = 1 - F(t)$. By the Chernoff bound, for $s > 0$,

$$\Pr\left[\frac{1}{(1-\alpha)n}\sum_{i\in I} Z_i \ge 1 - F(t) + s\right] \le e^{-2(1-\alpha)ns^2}.$$

Let $s = F(t) - \frac{1}{2} - \alpha$. Then

$$\begin{aligned}\Pr[m-\mu > t] &\le \Pr[m'-\mu > t] \\ &= \Pr\left[\frac{1}{(1-\alpha)n}\sum_{i\in I} Z_i \ge \frac{1}{2} - \alpha\right] \\ &\le \exp\left(-2(1-\alpha)n\left(F(t) - \frac{1}{2} - \alpha\right)^2\right).\end{aligned}$$

□

Gaussian If D is the one-dimensional unit-variance Gaussian distribution, then $F(t) = \frac{1}{2} + \frac{t}{\sqrt{2\pi}} + O(t^2)$, and we have the following from Lemma 2.9.

Corollary 2.10. *For $X_1, X_2, \ldots, X_n$ generated from the α-corrupted noise-oblivious model on distribution $N(\mu, 1)$, where $\alpha < 1/2$. Let $m = \text{median}(X_1, X_2, \ldots, X_n)$, then for $t > 0$,*

$$\Pr[|m-\mu| > \sqrt{2\pi}\alpha(1+t)] \le 2\exp\left(-\Omega(t^2(1-\alpha)\alpha^2 n)\right).$$

That is, for $\alpha < 1/2$, given $n = O(1/\alpha^2)$ samples, the median can robustly estimate the mean up to $O(\alpha)$ additive error for Gaussian distributions, with constant probability.

Laplace If D is the one-dimensional unit-variance Laplace distribution, then $F(t) = \frac{1}{2} + \sqrt{2}t + O(t^2)$, and similarly we have the following.

Corollary 2.11. *For $X_1, X_2, \ldots, X_n$ generated from the α-corrupted noise-oblivious model on distribution* Laplace$(\mu, 1)$*, where $\alpha < 1/2$. Let $m = \text{median}(X_1, X_2, \ldots, X_n)$, then for $t > 0$,*

$$\Pr[|m - \mu| > \alpha(1 + t)/\sqrt{2}] \leq 2 \exp\left(-\Omega(t^2(1 - \alpha)\alpha^2 n)\right).$$

That is, for $\alpha < 1/2$, given $n = O(1/\alpha^2)$ samples, the median can robustly estimate the mean up to $O(\alpha)$ additive error for Laplace distributions, with constant probability.

Chapter 3

Algorithm for Robust Mean Estimation

3.1 One-Dimensional Gaussian Case

As a starting point, we can consider the case where $d = 1$, and D is the Gaussian distribution with unit variance. The input of the algorithm can be view as n independent random variables, with a $(1 - \alpha)$ fraction being sampled from $N(\mu, 1)$, and the rest α fraction being sampled from $N(z_k, 1)$, for $k = 1, 2, \ldots, \alpha n$, where z_k is chosen by the adversary.

3.1.1 Preprocessing

From Corollary 2.10, we can robustly estimate the mean up to $O(\alpha)$ by the median, and translate the samples so that the true mean has magnitude $O(\alpha)$.

3.1.2 Transform into Sparse Signal

For a sample y_j generated by one of the Gaussian distributions, say $N(z, 1)$, from Fact 2.1, we have for $t \in \mathbb{R}$

$$\mathbb{E}[e^{ity_j}] = e^{itz - \frac{1}{2}t^2}.$$

Averaging for all $j = 1, 2, \ldots, n$, we have

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[e^{ity_j}] = (1-\alpha)e^{it\mu - \frac{1}{2}t^2} + \frac{1}{n}\sum_{k=1}^{\alpha n} e^{itz_k - \frac{1}{2}t^2}.$$

Multiplying by $e^{\frac{1}{2}t^2}$ on both sides gives

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[e^{ity_j}]e^{\frac{1}{2}t^2} = (1-\alpha)e^{it\mu} + \frac{1}{n}\sum_{k=1}^{\alpha n} e^{itz_k}.$$

This is a noisy 1-sparse signal if we treat the contributions from the corrupted points as noise, and $e^{it\mu}$ as the exactly 1-sparse signal. Replacing the expectation with empirical value, we will

apply the sparse Fourier transform in Theorem 2.6 on the signal

$$x(t) = \frac{1}{n}\sum_{j=1}^{n} e^{ity_j} e^{\frac{1}{2}t^2}.$$

Proof of Theorem 1.1, part 1. Let $x^* = (1-\alpha)e^{it\mu}$, and the noise

$$\begin{aligned} g(t) &= x(t) - x^*(t) \\ &= \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right)e^{\frac{1}{2}t^2}}_{g_1(t)} + \underbrace{\frac{1}{n}\sum_{k=1}^{\alpha n} e^{itz_k}}_{g_2(t)}. \end{aligned}$$

For $g_1(t)$, we can use concentration inequalities to bound the difference between the empirical average and the expectation. By Hoeffding's inequality, we can bound the real part and the imaginary part of the difference separately:

$$\begin{aligned} \Pr\left[\left|\frac{1}{n}\sum_{j=1}^{n}\cos(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\cos(ty_j)]\right| \geq \sqrt{\frac{C\log n}{n}}\right] &\leq \frac{2}{n^{C/2}}, \\ \Pr\left[\left|\frac{1}{n}\sum_{j=1}^{n}\sin(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\sin(ty_j)]\right| \geq \sqrt{\frac{C\log n}{n}}\right] &\leq \frac{2}{n^{C/2}}, \end{aligned}$$

Since

$$\begin{aligned} &\left|\frac{1}{n}\sum_{j=1}^{n} e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right| \\ &= \left|\frac{1}{n}\sum_{j=1}^{n}\cos(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\cos(ty_j)] + i\left(\frac{1}{n}\sum_{j=1}^{n}\sin(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\sin(ty_j)]\right)\right| \\ &\leq \left|\frac{1}{n}\sum_{j=1}^{n}\cos(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\cos(ty_j)]\right| + \left|\left(\frac{1}{n}\sum_{j=1}^{n}\sin(ty_j) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\sin(ty_j)]\right)\right|, \end{aligned}$$

by the union bound, we have

$$\left|\frac{1}{n}\sum_{j=1}^{n} e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right| \leq O\left(\sqrt{\frac{\log n}{n}}\right)$$

with probability at least $1 - \frac{1}{\mathsf{poly}(n)}$. Thus, w.h.p.,

$$|g_1(t)| = \left| \left(\frac{1}{n} \sum_{j=1}^{n} e^{ity_j} - \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[e^{ity_j}] \right) e^{\frac{1}{2}t^2} \right|$$
$$\leq O\left(\sqrt{\frac{\log n}{n}} \cdot e^{\frac{1}{2}t^2} \right).$$

Then we can apply Theorem 2.6, by setting $k = 1$, $\delta = O(1)$, $\eta = O(1)$, $F = O(\alpha)$, and $T = \frac{1}{2}\sqrt{\log n}$. Then we will need $M := O(\log \log n)$ samples from $x(t)$, say $x(t_1), \ldots, x(t_M)$. By the union bound, $|g_1(t_j)| \leq O\left(\sqrt{\frac{\log n}{n}} \cdot e^{\frac{1}{2}T^2} \right) \leq O(n^{-1/4})$ w.p. at least $1 - \frac{1}{\mathsf{poly}(n)}$. As we are only accessing $x(t)$ via these points, we can pretend that $|g_1(t)| \leq O(n^{-1/4})$ for all $t \in [0, T]$. Then compute

$$\begin{aligned}
\frac{1}{T} \int_0^T |g(t)|^2 \mathrm{d}t &= \frac{1}{T} \int_0^T |g_1(t) + g_2(t)|^2 \, \mathrm{d}t \\
&\leq \frac{2}{T} \int_0^T |g_1(t)|^2 \, \mathrm{d}t + \frac{2}{T} \int_0^T |g_2(t)|^2 \, \mathrm{d}t \\
&\leq O(n^{-1/2}) + \frac{2}{T} \int_0^T \left| \frac{1}{n} \sum_{k=1}^{\alpha n} e^{itz_k} \right| \mathrm{d}t \\
&\leq O(n^{-1/2}) + \frac{2}{T} \int_0^T \frac{1}{n^2} \cdot \alpha n \cdot \sum_{k=1}^{\alpha n} |e^{itz_k}|^2 \mathrm{d}t \\
&= O(n^{-1/2}) + 2\alpha^2.
\end{aligned}$$

So the noise level $\mathcal{N}^2 \leq O(n^{-1/2}) + 2\alpha^2 + \delta(1-\alpha)^2 = O(1)$. So for small enough constant α, we have $|v_1| = 1 - \alpha = \Omega(\mathcal{N})$, and we can recover the frequency (i.e. μ^2) up to $O\left(\frac{1}{T}\right) = O\left(\frac{1}{\sqrt{\log n}}\right)$ additive error, in $O((\log \log n)^2)$ time with constant probability.

Overall, note that each sample of the signal takes $O(n)$ time to compute, and thus our algorithm for the 1-dimensional case runs in time $O(nM) = \widetilde{O}(n)$ and estimates the true mean μ up to additive error $O\left(\frac{1}{\sqrt{\log n}}\right)$ with constant probability. That is, for our algorithm to estimate the true mean up to additive error $\varepsilon > 0$, the sample complexity is $2^{O(1/\varepsilon^2)}$, and the time complexity is $2^{O(1/\varepsilon^2)}$. □

3.2 One-Dimensional Laplace Case

The analysis for the Laplace case follows the same line as that in the Gaussian case. For a sample y_j generated by one of the Laplace distributions, say Laplace$(z, 1)$, from Fact 2.3, we have for $t \in \mathbb{R}$

$$\mathbb{E}[e^{ity_j}] = \frac{e^{itz}}{1 + \frac{1}{2}t^2}$$

Averaging for all $j = 1, 2, \ldots, n$, we have

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}] = (1-\alpha)\frac{e^{it\mu}}{1+\frac{1}{2}t^2} + \frac{1}{n}\sum_{k=1}^{\alpha n}\frac{e^{itz_k}}{1+\frac{1}{2}t^2}.$$

Multiplying by $(1+\frac{1}{2}t^2)$ on both side gives

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\left(1+\frac{t^2}{2}\right) = (1-\alpha)e^{it\mu} + \frac{1}{n}\sum_{k=1}^{\alpha n}e^{itz_k}.$$

Proof of Theorem 1.1, part 2. We will apply the sparse Fourier transform on the signal

$$x(t) = \frac{1}{n}\sum_{j=1}^{n}e^{ity_j}\left(1+\frac{t^2}{2}\right).$$

And similarly noise

$$g_1(t) = \left(\frac{1}{n}\sum_{j=1}^{n}e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right)\left(1+\frac{t^2}{2}\right),$$

and $g_2(t)$ is the same as that in the Gaussian case. We know from the previous section that

$$\left|\frac{1}{n}\sum_{j=1}^{n}e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right| \le O\left(\sqrt{\frac{\log n}{n}}\right)$$

with probability at least $1 - \frac{1}{\mathsf{poly}(n)}$. Thus, w.h.p.,

$$\begin{aligned}|g_1(t)| &= \left|\left(\frac{1}{n}\sum_{j=1}^{n}e^{ity_j} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{ity_j}]\right)\left(1+\frac{t^2}{2}\right)\right| \\ &\le O\left(\sqrt{\frac{\log n}{n}}\cdot t^2\right).\end{aligned}$$

Then we can apply Theorem 2.6, by setting $k = 1$, $\delta = O(1)$, $\eta = O(1)$, $F = O(\alpha)$, and $T = n^{1/4}/\log n$. Then we will need $M := O(\log n)$ samples from $x(t)$, say $x(t_1), \ldots, x(t_M)$. By the union bound, $|g_1(t_j)| \le O\left(\sqrt{\frac{\log n}{n}}\cdot T^2\right) \le O(1/\log n)$ w.p. at least $1 - \frac{1}{\mathsf{poly}(n)}$. As we are only accessing $x(t)$ via these points, we can pretend that $|g_1(t)| \le O(1/\log n)$ for all $t \in [0, T]$. Then similarly, the noise level $\mathcal{N}^2 \le O((\log n)^{-2}) + 2\alpha^2 + \delta(1-\alpha)^2 = O(1)$. So for small enough constant α, we have $|v_1| = 1 - \alpha = \Omega(\mathcal{N})$, and we can recover the frequency (i.e. μ^2) up to $O\left(\frac{1}{T}\right) = O\left(\frac{\log n}{n^{1/4}}\right)$ additive error, in $O((\log n)^2)$ time with constant probability.

Overall, our algorithm for the 1-dimensional case runs in time $O(nM) = \widetilde{O}(n)$ and estimates the true mean μ up to additive error $O\left(\frac{\log n}{n^{1/4}}\right)$ with constant probability. That is, for our algorithm to estimate the true mean up to additive error $\varepsilon > 0$, the sample complexity is $\widetilde{O}((1/\varepsilon)^4)$, and the time complexity is $\widetilde{O}((1/\varepsilon)^4)$. □

3.3 High-Dimensional Case

With this one-dimensional algorithm that can estimate the mean arbitrarily closely, it is easy to design the algorithm for the high-dimensional case: just project the samples onto each axis, and apply the one-dimensional algorithm on the projected samples to estimate the corresponding coordinate of the true mean.

Take the Laplace case for example. For a direction along the axis e_i, project the input sample $Y = \{y_j\}_{j=1}^n$ along e_i to get $Y_i = \{\langle e_i, y_j\rangle\}_{j=1}^n$, which consists of $(1-\alpha)n$ points being distributed as Laplace($\langle e_i, \mu\rangle$, 1), and the rest αn points being distributed as Laplace($\langle e_i, z_k\rangle$, 1), for $k = 1, 2, \ldots, \alpha n$. Then our one-dimensional algorithm, taking Y' as the input, can estimate $\langle e_i, \mu\rangle$ up to additive error $\varepsilon/\sqrt{d}$ with constant probability with sample and time complexity $\widetilde{O}(d^2/\varepsilon^4)$. Let $i = 1, 2, \ldots, d$, we can estimate each coordinate of μ up to $\varepsilon/\sqrt{d}$ additive error, and thus we can estimate μ up to ε additive error in ℓ_2 norm. Also, we need to repeat each one-dimensional algorithm $O(\log d)$ times to boost the success probability from constant to $1 - \frac{1}{\mathsf{poly}(d)}$ to apply the union bound over all axis-aligned directions. So the total sample complexity is $\widetilde{O}(d^2/\varepsilon^4)$, and time complexity is $\widetilde{O}(d^3/\varepsilon^4)$.

Similarly, for the d-dimensional Gaussian case, the total samples and time complexity are $2^{O(d/\varepsilon^2)}$.

Chapter 4

Applications

In this chapter, we will give two possible applications of our noise-oblivious model that motivated us to study this model. While currently our results only solve the robust mean estimation problem when the fraction of corruption is small in our model, these two applications require an algorithm that solves the robust linear regression problem in the list-decodable setting.

4.1 Max-Affine Regression

Max-affine regression refers to the following model

$$y = \max_{1 \le j \le k} \{\langle \ell_j, x \rangle\} + \eta,$$

where x is a d-dimensional covariate, ℓ_j are unknown linear models for $j = 1, 2, \ldots, k$, and η is a zero-mean noise, independent of x. Max-affine regression is useful in a broad range of applications. For $k = 2$, and $\ell_1 = -\ell_2 = \theta$, the model becomes $y = |\langle \theta, x \rangle| + \eta$. This is called real phase retrieval, which has extensive applications in science and engineering [Tay03; SEC+15; MCK+99; FWd13]. For general k, the model can be viewed as a tractable approximation of the convex regression problem [MB09; Bal16], as the maximum of k linear functions is convex. While the convex regression problem suffers from the curse of dimensionality [GS13], it is natural to constrain the number of affine pieces of the function and hope to solve this more structured problem.

There have been a lot of algorithmic results for max-affine regression based on alternating minimization and stochastic gradient descent [GPG+19; GPG+20; GPG+22; KL23]. As list-decodable linear regression algorithms in the strong contamination model can be used to solve the problem of learning the maximum of k linear models, we want to solve the max-affine regression in the robust statistics perspective, i.e., using a list-decodable linear regression algorithm in our noise-oblivious model. Before that, let us first introduce how the problem of learning the maximum of k linear models reduces to the problem of list-decodable linear regression in the strong contamination model. Here, the problem of learning the maximum of k linear models is defined as, given independent samples $\{(x^{(i)}, y^{(i)})\}_i$, where $x^{(i)}$ are i.i.d. standard Gaussian random variables, and $y^{(i)} = \max_{1 \le j \le k} \{\langle \ell_j, x^{(i)} \rangle + \eta_j^{(i)}\}$, for i.i.d. standard Gaussian noise $\eta_j^{(i)}$,

the algorithm needs to estimate the true linear models ℓ_j for $j = 1, 2, \ldots, k$. One can directly apply the list-decodable linear regression algorithm to this problem: to recover ℓ_j, we can view the samples where $(\langle \ell_j, x^{(i)} \rangle + \eta_j^{(i)})$ attains the maximum as the inliers, and the rest as the outliers. That is, the adversary corrupts all the $y^{(i)}$ to the maximum of the k linear models. As long as the probability p_j of each model being selected as the maximum is bounded from below, there is an about p_j fraction of the samples are uncorrupted, and we can recover all ℓ_j from the list output by the algorithm.

The reduction above hints at the correspondence between taking the maximum and corruption. In the max-affine regression model, the difference is that the noise η is added *after* taking the maximum. This is our motivation to consider the contamination model where the noise is added after corruption.

For the max-affine regression problem, the samples can be viewed as generated from the noise-oblivious model as follows: suppose the true linear function is ℓ_j, then

1. the model generate the covariates $x^{(i)} \sim N(0, I)$ independently;
2. the model compute $y^{(i)} = \langle \ell_j, x^{(i)} \rangle$;
3. an adversary changes each $y^{(i)} = \max_{1 \le j' \le k} \{\langle \ell_{j'}, x^{(i)} \rangle\}$ (the fraction of uncorrupted samples is equal to the fraction of samples where $(\langle \ell_j, x^{(i)} \rangle)$ attains the maximum);
4. the model adds i.i.d. noise from some distribution $D(0)$ to all the $y^{(i)}$.

Similarly, if the probability p_j of each linear function being selected as the maximum is bounded from below, we can recover all ℓ_j from the list output by the list-decodable linear regression algorithm in the noise-oblivious model.

4.2 Mixed Linear Regression

A list-decodable linear regression algorithm in the noise-oblivious model can also solve the mixed linear regression problem, in a way the same as that in the strong contamination model. Mixed linear regression is the problem of given k clusters of sample pairs $(x^{(i)}, y^{(i)})$, each generated by one of the k unknown linear models, estimating the k linear functions. The problem of mixed linear regression has been well-studied before robust linear regression [De 89; JJ93; FS10; SJA16; BWY17; LL18], but these techniques work under some assumptions, such as pairwise separation or bounded condition number. Meanwhile, a list-decodable linear regression algorithm in the strong contamination model immediately gives an algorithm for mixed linear regression by taking one cluster as inliers and the rest as outliers [KKK19], similarly to the reduction for max-affine regression. Note that when reducing the mixed linear regression problem to the list-decodable linear regression problem, the criterion of adversarial corruption is independent of the value of the samples (in contrast to the case in max-affine regression, where the adversary corrupts a sample if it is not the maximum), especially, the noise. Therefore, there is no difference between adding the noise after and before corruption. Thus, a list-decodable linear regression algorithm in the noise-oblivious model also works for mixed linear regression.

Chapter 5

Future Work

One important question is whether there exists a polynomial time algorithm for the Gaussian case. Intuitively, Laplace distributions have an exponential tail, which is heavier than the Gaussian tail, which indicates that the Gaussian case should be easier than the Laplace case, as the samples will be more concentrated. Note that the exponential sample and time complexity of our algorithm for the Gaussian case completely come from the characteristic function of Gaussian distributions. One natural direction is to operate on the samples, that is, to transform the Gaussian distribution into some distribution with a tamed characteristic function. Here we will present a failed but interesting attempt in this direction.

For a Gaussian sample $y_j \sim N(\mu, 1)$, the square y_j^2 is distributed according to the *noncentral chi-squared distribution* with one degree of freedom. By Fact 2.5, for $t \in \mathbb{C}$ with $|\Im t| \leq 1/2$,

$$\mathbb{E}[e^{ity_j^2}] = \frac{\exp(\frac{itz^2}{1-2it})}{(1-2it)^{-1/2}}.$$

Then the analysis is similar to that in Section 3.1.2. Summing up for all y_j in the sample set, we have for $t \in \mathbb{C}$ with $|\Im t_1| \leq 1/2$,

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[e^{it_1y_j^2}] = (1-\alpha)\frac{\exp\left(\frac{it_1\mu^2}{1-2it_1}\right)}{(1-2it_1)^{1/2}} + \frac{1}{n}\sum_{k=1}^{\alpha n}\frac{\exp\left(\frac{it_1z_k^2}{1-2it_1}\right)}{(1-2it_1)^{1/2}}.$$

Substituting $t = \frac{t_1}{1-2it_1}$ (and thus $t_1 = \frac{t}{1+2it}$), then for $t \in \mathbb{R}$ we have $|\Im t_1| = \frac{2t^2}{1+4t^2} \leq 1/2$, and

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\exp\left(i\frac{t}{1+2it}y_j^2\right)\right] = \left((1-\alpha)e^{it\mu^2} + \frac{1}{n}\sum_{k=1}^{\alpha n}e^{itz_k^2}\right)(1+2it)^{1/2}.$$

That is,

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\exp\left(i\frac{t}{1+2it}y_j^2\right)\right](1+2it)^{-1/2} = (1-\alpha)e^{it\mu^2} + \frac{1}{n}\sum_{k=1}^{\alpha n}e^{itz_k^2}.$$

Although we can get a noisy 1-sparse signal from the left-hand side, the problem is that the real part of the exponent $i\frac{t}{1+2it}y_j^2$ is $\frac{2t^2}{1+4t^2}y_j^2$. Thus the norm of $\exp\left(i\frac{t}{1+2it}y_j^2\right)$ is $\exp\left(\frac{2t^2}{1+4t^2}y_j^2\right)$,

which is too large, almost canceling out the Gaussian tail $e^{-y^2/2}$. This makes it hard to apply concentration inequalities to bound the difference between the expectation and the empirical value of $\exp\left(i\frac{t}{1+2it}y_j^2\right)$. Therefore, one possible future direction is to consider another way of transforming the Gaussian distribution to get a polynomial time algorithm for the Gaussian case.

Here we only consider the mean estimation problem. So another question is, are there polynomial time algorithms for other statistical estimation problems in our noise-oblivious model, such as covariance estimation and linear regression? The latter is especially important to us, because our motivation is to solve the max-affine regression problem. Thus, another question is whether our method extends to the list-decodable setting, where a majority of the samples are corrupted. Note that similar to the reduction for the mixed linear regression problem, a list-decodable mean estimation algorithm in the noise-oblivious model immediately yields an algorithm for learning mixtures of Gaussians. Since there is an $\exp(k)$ lower bound in learning mixtures of k Gaussians [MV10], there should also be an $\exp(1/c)$ lower bound in list-decodable mean estimation in the noise-oblivious model, where $c = 1 - \alpha$ is the fraction of uncorrupted samples.

References

[Bal16] Gábor Balázs. "Convex Regression: Theory, Practice, and Applications". PhD thesis. Sept. 2016. DOI: 10.7939/R3T43J98B.

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. "A discriminative framework for clustering via similarity functions". In: *STOC'08*. ACM, New York, 2008, pp. 671–680. ISBN: 978-1-60558-047-0. DOI: 10.1145/1374376.1374474. URL: https://doi.org/10.1145/1374376.1374474.

[BDL+17] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. "Computationally Efficient Robust Sparse Estimation in High Dimensions". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, 2017, pp. 169–212. URL: https://proceedings.mlr.press/v65/balakrishnan17a.html.

[BK21] Ainesh Bakshi and Pravesh K. Kothari. "List-Decodable Subspace Recovery: Dimension Independent Error in Polynomial Time". In: *Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '21. Virtual Event, Virginia: Society for Industrial and Applied Mathematics, 2021, 1279–1297. ISBN: 9781611976465.

[BWY17] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *The Annals of Statistics* 45.1 (2017), pp. 77 –120. DOI: 10.1214/16-AOS1435. URL: https://doi.org/10.1214/16-AOS1435.

[CMY20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. "List decodable mean estimation in nearly linear time". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*. IEEE Computer Soc., Los Alamitos, CA, 2020, pp. 141–148. ISBN: 978-1-7281-9621-3. DOI: 10.1109/FOCS46700.2020.00022. URL: https://doi.org/10.1109/FOCS46700.2020.00022.

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. "Learning from untrusted data". In: *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2017, pp. 47–60. ISBN: 978-1-4503-4528-6.

[De 89] Richard D. De Veaux. "Mixtures of linear regressions". In: *Computational Statistics Data Analysis* 8.3 (1989), pp. 227–245. ISSN: 0167-9473. DOI: https://doi.org/10.1016/0167-9473(89)90043-1. URL: https://www.sciencedirect.com/science/article/pii/0167947389900431.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robust estimators in high dimensions without the computational intractability". In: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, 2016, pp. 655–664. ISBN: 978-1-5090-3933-3. DOI: 10.1109/FOCS.2016.85. URL: https://doi.org/10.1109/FOCS.2016.85.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Being Robust (in High Dimensions) Can Be Practical". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, 999–1008.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robustly Learning a Gaussian: Getting Optimal Error, Efficiently". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '18. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2018, 2683–2702. ISBN: 9781611975031.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. "Sever: A Robust Meta-Algorithm for Stochastic Optimization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1596–1606. URL: https://proceedings.mlr.press/v97/diakonikolas19a.html.

[DKK20] Ilias Diakonikolas, Daniel M. Kane, and Daniel Kongsgaard. "List-Decodable Mean Estimation via Iterative Multi-Filtering". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[DKK+21] Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. "List-Decodable Mean Estimation in Nearly-PCA Time". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 10195–10208. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/547b85f3fafdf30856386753dc21c4e1-Paper.pdf.

[DKS18] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "List-decodable robust mean estimation and learning mixtures of spherical Gaussians". In: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1047–1060. ISBN: 978-1-4503-5559-9. DOI: 10.1145/3188745.3188758. URL: https://doi.org/10.1145/3188745.3188758.

[FS10] Susana Faria and Gilda Soromenho. “Fitting mixtures of linear regressions”. In: *Journal of Statistical Computation and Simulation* 80.2 (2010), pp. 201–225. DOI: 10.1080/00949650802590261. eprint: https://doi.org/10.1080/00949650802590261. URL: https://doi.org/10.1080/00949650802590261.

[FWd13] Fajwel Fogel, Irène Waldspurger, and Alexandre d’Aspremont. “Phase retrieval for imaging problems”. In: *arXiv e-prints*, arXiv:1304.7735 (Apr. 2013), arXiv:1304.7735. DOI: 10.48550/arXiv.1304.7735. arXiv: 1304.7735 [math.OC].

[GPG+19] Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. *Max-Affine Regression: Provable, Tractable, and Near-Optimal Statistical Estimation*. 2019. arXiv: 1906.09255 [stat.ML].

[GPG+20] Avishek Ghosh, Ashwin Pananjady, Aditya Guntuboyina, and Kannan Ramchandran. “Max-Affine Regression with Universal Parameter Estimation for Small-Ball Designs”. In: *2020 IEEE International Symposium on Information Theory (ISIT)*. Los Angeles, CA, USA: IEEE Press, 2020, 2706–2710. DOI: 10.1109/ISIT44484.2020.9174116. URL: https://doi.org/10.1109/ISIT44484.2020.9174116.

[GPG+22] Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. “Max-Affine Regression: Parameter Estimation for Gaussian Designs”. In: *IEEE Transactions on Information Theory* 68.3 (2022), pp. 1851–1885. DOI: 10.1109/TIT.2021.3130717.

[GS13] Adityanand Guntuboyina and Bodhisattva Sen. “Covering Numbers for Convex Functions”. In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 1957–1965. DOI: 10.1109/TIT.2012.2235172.

[HL18] Samuel B. Hopkins and Jerry Li. “Mixture models, robustness, and sum of squares proofs”. In: *STOC’18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1021–1034. ISBN: 978-1-4503-5559-9. DOI: 10.1145/3188745.3188748. URL: https://doi.org/10.1145/3188745.3188748.

[HR09] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2009, xvi+354 pp. + loose erratum. ISBN: 978-0-470-12990-6. DOI: 10.1002/9780470434697. URL: https://doi.org/10.1002/9780470434697.

[HRR+11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

[Hub64] Peter J. Huber. “Robust estimation of a location parameter”. In: *Ann. Math. Statist.* 35 (1964), pp. 73–101. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732.

[IK22] Misha Ivkov and Pravesh K. Kothari. "List-Decodable Covariance Estimation". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2022. Rome, Italy: Association for Computing Machinery, 2022, 1276–1283. ISBN: 9781450392648. DOI: 10.1145/3519935.3520006. URL: https://doi.org/10.1145/3519935.3520006.

[JJ93] M.I. Jordan and R.A. Jacobs. "Hierarchical mixtures of experts and the EM algorithm". In: *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*. Vol. 2. 1993, 1339–1344 vol.2. DOI: 10.1109/IJCNN.1993.716791.

[KKK19] Sushrut Karmalkar, Adam R. Klivans, and Pravesh K. Kothari. "List-Decodeable Linear Regression". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[KKM18] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. "Efficient Algorithms for Outlier-Robust Regression". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1420–1430. URL: https://proceedings.mlr.press/v75/klivans18a.html.

[KL23] Seonho Kim and Kiryung Lee. "Fast max-affine regression via stochastic gradient descent". In: *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2023, pp. 1–5. DOI: 10.1109/Allerton58177.2023.10313409.

[KSS18] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1035–1046. ISBN: 978-1-4503-5559-9.

[LL18] Yuanzhi Li and Yingyu Liang. "Learning Mixtures of Linear Regressions with Nearly Optimal Complexity". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1125–1144. URL: https://proceedings.mlr.press/v75/li18b.html.

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. "Agnostic estimation of mean and covariance". In: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, 2016, pp. 665–674. ISBN: 978-1-5090-3933-3.

[MB09] Alessandro Magnani and Stephen P. Boyd. "Convex piecewise-linear fitting". In: *Optimization and Engineering* 10.1 (2009), pp. 1–17. ISSN: 1573-2924. DOI: 10.1007/s11081-008-9045-3. URL: https://doi.org/10.1007/s11081-008-9045-3.

[MCK+99] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens". In: 400.6742 (July 1999), pp. 342–344. DOI: 10.1038/22498.

[MMY+19] Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. *Robust statistics*. Second. Wiley Series in Probability and Statistics. Theory and methods (with R). John Wiley & Sons, Inc., Hoboken, NJ, 2019, pp. xxvii+430. ISBN: 978-1-119-21468-7.

[MV10] Ankur Moitra and Gregory Valiant. "Settling the Polynomial Learnability of Mixtures of Gaussians". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 2010, pp. 93–102. DOI: 10.1109/FOCS.2010.15.

[PS15] Eric Price and Zhao Song. "A robust sparse Fourier transform in the continuous setting". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*. IEEE Computer Soc., Los Alamitos, CA, 2015, pp. 583–600. ISBN: 978-1-4673-8191-8. DOI: 10.1109/FOCS.2015.42. URL: https://doi.org/10.1109/FOCS.2015.42.

[RY20a] Prasad Raghavendra and Morris Yau. "List Decodable Learning via Sum of Squares". In: *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '20. Salt Lake City, Utah: Society for Industrial and Applied Mathematics, 2020, 161–180.

[RY20b] Prasad Raghavendra and Morris Yau. "List Decodable Subspace Recovery". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3206–3226. URL: https://proceedings.mlr.press/v125/raghavendra20a.html.

[SEC+15] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. "Phase Retrieval with Application to Optical Imaging: A contemporary overview". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109. DOI: 10.1109/MSP.2014.2352673.

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. "Provable Tensor Methods for Learning Mixtures of Generalized Linear Models". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016, pp. 1223–1231. URL: https://proceedings.mlr.press/v51/sedghi16.html.

[Tay03] Garry Taylor. "The phase problem". In: *Acta Crystallographica Section D* 59.11 (2003), pp. 1881–1890. DOI: 10.1107/S0907444903017815. URL: https://doi.org/10.1107/S0907444903017815.

[Tuk75] John W. Tukey. "Mathematics and the picturing of data". In: *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*. Canad. Math. Congr., Montreal, QC, 1975, pp. 523–531.