Learning Entangled Single-Sample Distributions via Iterative Trimming

Hui Yuan
Department of Statistics and Finance
University of Science and Technology of China

Yingyu Liang
Department of Computer Sciences
University of Wisconsin-Madison

Abstract

In the setting of entangled single-sample distributions, the goal is to estimate some common parameter shared by a family of distributions, given one *single* sample from each distribution. We study mean estimation and linear regression under general conditions, and analyze a simple and computationally efficient method based on iteratively trimming samples and re-estimating the parameter on the trimmed sample set. We show that the method in logarithmic iterations outputs an estimation whose error only depends on the noise level of the $\lceil \alpha n \rceil$-th noisiest data point where α is a constant and n is the sample size. This means it can tolerate a constant fraction of high-noise points. These are the first such results under our general conditions with computationally efficient estimators. It also justifies the wide application and empirical success of iterative trimming in practice. Our theoretical results are complemented by experiments on synthetic data.

1 INTRODUCTION

This work considers the novel parameter estimation setting called entangled single-sample distributions. Different from the typical i.i.d. setting, here we have n data points that are independent, but each is from a different distribution. These distributions are entangled in the sense that they share some common parameter, and our goal is to estimate the common parameter. For example, in the problem of mean estimation for entangled single-sample distributions, we have n data points from n different distributions with a common mean but different variances (the mean and all the variances are unknown), and our goal is to estimate the mean.

This setting is motivated for both theoretical and practical reasons. From the theoretical perspective, it goes beyond the typical i.i.d. setting and raises many interesting open questions, even on basic topics like mean estimation for Gaussians. It can also be viewed as a generalization of the traditional mixture modeling, since the number of distinct mixture components can grow with the number of samples. From the practical perspective, many modern applications have various forms of heterogeneity, for which the i.i.d. assumption can lead to bad modeling of their data. The entangled single-sample setting provides potentially better modeling. This is particularly the case for applications where we have no control over the noise levels of the samples. For example, the images taken by self-driving cars can have varying degrees of noise due to changing weather or lighting conditions. Similarly, signals collected from sensors on the Internet of Things can come with interferences from a changing environment.

Though theoretically interesting and practically important, few studies exist in this setting. Chierichetti et al. (2014) considered the mean estimation for entangled Gaussians and showed the existence of a gap between estimation error rates of the best possible estimator in this setting and the maximum likelihood estimator when the variances are known. Pensia et al. (2019) considered means estimation for symmetric, unimodal distributions including the symmetric multivariate case (i.e., the distributions are radially symmetric) with sharpened bounds, and provided extensive discussion on the performance of their estimators in different configurations of the variances. These existing results focus on specific family of distributions or focus on the case where most samples are "high-noise" points.

On the contrary, we focus on the case with a constant fraction of "high-noise" points, which is more inter-

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

esting in practice. We study multivariate mean estimation and linear regression under more general conditions and analyze a simple and efficient estimator based on iterative trimming. The iterative trimming idea is simple: the algorithm keeps an iterate and repeatedly refines it; each time it trims a fraction of bad points based on the current iterate and then uses the trimmed sample set to compute the next iterate. It is computationally very efficient and widely used in practice as a heuristic for handling noisy data. It can also be viewed as an alternating-update version of the classic trimmed estimator (e.g., Huber (2011)) which typically takes exponential time:

$$\hat{\theta} = \mathop{\arg\min}_{\theta \in \Theta; S \subseteq [n], |S| = \lceil \alpha n \rceil} \sum_{i \in S} \mathrm{Loss}_i(\theta)$$

where Θ is the feasible set for the parameter θ to be estimated, $\lceil \alpha n \rceil$ is the size of the trimmed sample set S, and $\mathrm{Loss}_i(\theta)$ is the loss of θ on the i-th data point (e.g., ℓ_2 error for linear regression).

For mean estimation, only assuming the distributions have a common mean and bounded covariances, we show that the iterative trimming method in logarithmic iterations outputs a solution whose error only depends on the noise level of the $\lceil \alpha n \rceil$-th noisiest point for $\alpha \geq 4/5$. More precisely, the error only depends on the $\lceil \alpha n \rceil$-th largest value among all the norms of the n covariance matrices. This means the method can tolerate a 1/5 fraction of "high-noise" points. We also provide a similar result for linear regression, under a regularity condition that the explanatory variables are sufficiently spread out in different directions (satisfied by typical distributions like Gaussians). As far as we know, these are the first such results of iterative trimming under our general conditions in the entangled single-sample distributions setting. These results also theoretically justify the wide application and empirical success of the simple iterative trimming method in practice. Experiments on synthetic data provide positive support for our analysis.

2 RELATED WORK

Entangled distributions. This setting is first studied by Chierichetti et al. (2014), which considered mean estimation for entangled Gaussians and presented a algorithm combining the k-median and the k-shortest gap algorithms. It also showed the existence of a gap between the error rates of the best possible estimator in this setting and the maximum likelihood estimator when the variances are known. Pensia et al. (2019) considered a more general class of distributions (unimodal and symmetric) and provided analysis on both individual estimator (r-modal interval, k-shortest gap, k-median estimators) and hybrid estimator, which combines Median estimator with Shortest Gap or Modal Interval estimator. They also discussed slight relaxation of the symmetry assumption and provided extensions to linear regression. Our work considers mean estimation and linear regression under more general conditions and analyzes a simpler estimator. However, our results are not directly comparable to the existing ones above, since those focus on the case where most of the points have high noise or have extra constraints on distributions are assumed. For the constrained distributions, our results are weaker than the existing ones. See the detailed discussion in the remarks after our theorems.

This setting is also closely related to robust estimation, which have been extensively studied in the literature of both classic statistics and machine learning theory.

Robust mean estimation. There are several classes of data distribution models for robust mean estimators. The most commonly addressed is adversarial contamination model, whose origin can be traced back to the malicious noise model by Valiant (1985) and the contamination model by Huber (2011). Under contamination, mean estimation has been investigated in Diakonikolas et al. (2017, 2019a); Cheng et al. (2019). Another related model is the mixture of distributions. There has been steady progress in algorithms for leaning mixtures, in particular, leaning Gaussian mixtures. Starting from Dasgupta (1999), a rich collection of results are provided in many studies, such as Sanjeev and Kannan (2001); Achlioptas and McSherry (2005); Kannan et al. (2005); Belkin and Sinha (2010a,b); Kalai et al. (2010); Moitra and Valiant (2010); Diakonikolas et al. (2018a).

Robust regression. Robust Least Squares Regression (RLSR) addresses the problem of learning regression coefficients in the presence of corruptions in the response vector. A class of robust regression estimator solving RLSR is Least Trimmed Square (LTS) estimator, which is first introduced by Rousseeuw (1984) and has high breakdown point. The algorithm solutions of LTS are investigated in Hössjer (1995); Rousseeuw and Van Driessen (2006); Shen et al. (2013) for the linear regression setting. Recently, for robust linear regression in the adversarial setting (i.e., a small fraction of responses are replaced by adversarial values), there is a line of work providing algorithms with theoretical guarantees following the idea of LTS, e.g., Bhatia et al. (2015); Vainsencher et al. (2017); Yang et al. (2018) for example. For robust linear regression in the adversary setting where both explanatory and response variables can be replaced by adversarial values, a line of work provided algorithms and guarantees, e.g., Diakonikolas et al. (2018b); Prasad et al. (2018); Klivans et al.

(2018); Shen and Sanghavi (2019), while some others like Chen et al. (2013); Balakrishnan et al. (2017); Liu et al. (2018) considered the high-dimensional scenario.

3 MEAN ESTIMATION

Suppose we have n independent samples $\boldsymbol{x}_i \sim F_i \in \mathbb{R}^d$, $d \in \mathbb{N}^\star$, where the mean vector and the covariance matrix of each distribution F_i exist. Assume F_i's have a common mean $\boldsymbol{\mu}^\star$ and denote their covariance matrices as Σ_i. When $d = 1$, each $\boldsymbol{x}_i$ degenerate to an univariate random variable x_i, and we also write $\boldsymbol{\mu}^\star$ as $\mu^\star$ and write Σ_i as σ_i^2. Our goal is to estimate the common mean $\boldsymbol{\mu}^\star$.

Notations For an integer m, $[m]$ denotes the set $\{1, \cdots, m\}$. $|S|$ is the cardinality of a set S. For two sets S_1, S_2, $S_1 \backslash S_2$ is the set of elements in S_1 but not in S_2. $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues. Denote the order statistics of $\{\lambda_{\max}(\Sigma_i)\}_{i=1}^n$ as $\{\lambda_{(i)}\}_{i=1}^n$. c or C denote constants whose values can vary from line to line.

3.1 Iterative Trimmed Mean Algorithm

First, recall the general version of the least trimmed loss estimator. Let $f_{\boldsymbol{\mu}}(\cdot)$ be the loss function, given currently learned parameter $\boldsymbol{\mu}$. In contrast to minimizing the total loss of all samples, the least trimmed loss estimator of $\boldsymbol{\mu}^\star$ is given by

$$\hat{\boldsymbol{\mu}}^{(\mathrm{TL})} = \mathop{\arg\min}_{\boldsymbol{\mu}, S:|S|=\lceil \alpha n \rceil} \sum_{i \in S} f_{\boldsymbol{\mu}}(\boldsymbol{x}_i), \tag{1}$$

where $S \subseteq [n]$ and $\alpha \in (0, 1]$ is the fraction of samples we want to fit. Finding $\hat{\boldsymbol{\mu}}^{(\mathrm{TL})}$ requires minimizing the trimmed loss over both the set of all subsets S with size $\lceil \alpha n \rceil$ and the set of all available values of the parameter $\boldsymbol{\mu}$. However, solving the minimization above is hard in general. Therefore, we attempt to minimize the trimmed loss in an iterative way by alternating between minimizing over S and $\boldsymbol{\mu}$. That is, it follows a natural iterative strategy: in the t-th iteration, first select a subset S_t of samples with the least loss on the current parameter $\boldsymbol{\mu}_t$, and then update to $\boldsymbol{\mu}_{t+1}$ by minimizing the total loss over S_t.

By taking $f_{\boldsymbol{\mu}}(\boldsymbol{x})$ in (1) as $\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2$, in each iteration, $\lceil \alpha n \rceil$ samples are selected due to their least distance to the current $\boldsymbol{\mu}$ in l_2 norm. Besides, note that for a given sample set S,

$$\mathop{\arg\min}_{\boldsymbol{\mu}} \sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2 = \frac{1}{|S|} \sum_{i \in S} \boldsymbol{x}_i,$$

that is, the parameter $\boldsymbol{\mu}$ minimizing the total loss over sample set S is the empirical mean over S. This leads to our method described in Algorithm 1, which we referred to as Iterative Trimmed Mean.

Algorithm 1 Iterative Trimmed MEAN (ITM)

```
Input: Samples $\{\boldsymbol{x}_i\}_{i=1}^n$, number of rounds $T$, fraction of samples $\alpha$
$\boldsymbol{\mu}_0 \leftarrow \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$
for $t = 0, \cdots, T-1$ do
    Choose samples with smallest current loss:
        $S_t \leftarrow \mathop{\arg\min}_{S:|S|=\lceil \alpha n \rceil} \sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{\mu}_t\|_2^2$
    $\boldsymbol{\mu}_{t+1} = \frac{1}{|S_t|}\sum_{i \in S_t} \boldsymbol{x}_i$
end for
Output: $\boldsymbol{\mu}_T$
```

Each update in our algorithm is similar to the trimmed-mean (or truncated-mean) estimator, which is, in univariate case, defined by removing a fraction of the sample consisting of the largest and smallest points and averaging over the rest; see Tukey and McLaughlin (1963); Huber (2011); Bickel et al. (1965). A natural generalization to multivariate case is to remove a fraction of the sample consisting of points which have the largest distances to $\boldsymbol{\mu}^\star$. The difference between our algorithm and the generalized trimmed-mean estimator is that ours select points based on the estimated $\boldsymbol{\mu}_t$ while trimmed-mean is based on the ground-truth $\boldsymbol{\mu}^\star$.

3.2 Theoretical Guarantees

Firstly, we introduce a lemma giving an upper bound of the sum of ℓ_2 distances between points $\boldsymbol{x}_i$ and mean vector $\boldsymbol{\mu}^\star$ over a sample set S.

Lemma 1. *Define* $\lambda_S = \max_{i \in S}\{\lambda_{\max}(\Sigma_i)\}$. *Then we have*

$$\sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{\mu}^\star\|_2 \leq 2|S|\sqrt{\lambda_S d} \tag{2}$$

with probability at least $1 - \frac{1}{|S|}$.

Proof. For each $i \in S$, denote the transformed random vector $\Sigma_i^{-\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu}^\star)$ as $\tilde{\boldsymbol{x}}_i$. Then $\tilde{\boldsymbol{x}}_i$ has mean $\mathbf{0}$ and identical covariance matrix. Since $\|\boldsymbol{x}_i - \boldsymbol{\mu}^\star\|_2 = \sqrt{\tilde{\boldsymbol{x}}_i^T \Sigma_i \tilde{\boldsymbol{x}}_i} \leq \sqrt{\lambda_S}\,\|\tilde{\boldsymbol{x}}_i\|_2$, we can write

$$\sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{\mu}^\star\|_2 \leq \sqrt{\lambda_S} \sum_{i \in S} \|\tilde{\boldsymbol{x}}_i\|_2 .$$

Let $\xi_i = \|\tilde{\boldsymbol{x}}_i\|_2$, thus $\{\xi_i\}_{i \in S}$ are independent and for any $i \in S$, it holds that

$$\mathbb{E}\xi_i \leq \sqrt{\mathbb{E}\xi_i^2} = \sqrt{d}, \quad \mathrm{Var}(\xi_i) \leq \mathbb{E}\xi_i^2 = d.$$

By Chebyshev's inequality, we have

$$\mathbb{P}\left(\sum_{i\in S}\xi_i \ge 2|S|\sqrt{d}\right) \le \frac{1}{|S|},$$

which completes the proof. ■

We are now ready to prove our key lemma, which shows the progress of each iteration of the algorithm. The key idea is that the selected subset S_t of samples have a large overlap with the subset $S^\star$ of the $\lceil \alpha n\rceil$ "good" points with smallest covariance. Furthermore, $S_t \backslash S^\star$ is not that "bad" since by the selection criterion, they have less loss on $\boldsymbol{\mu}_t$ than the points in $S^\star \backslash S_t$. This thus allows to show the progress.

Lemma 2. *Given ITM with fraction $\alpha \ge \frac{4}{5}$, we have*

$$\left\|\boldsymbol{\mu}_{t+1}-\boldsymbol{\mu}^\star\right\|_2 \le \frac{1}{2}\left\|\boldsymbol{\mu}_t-\boldsymbol{\mu}^\star\right\|_2 + 2\sqrt{d\lambda_{(\lceil \alpha n\rceil)}} \qquad (3)$$

with probability at least $1-\frac{5}{4n}$.

Proof. Define $S^\star = \left\{i : \lambda_{\max}(\Sigma_i) \le \lambda_{(\lceil \alpha n\rceil)}\right\}$. Without loss of generality, assume αn is an integer. Then by the algorithm,

$$\boldsymbol{\mu}_{t+1} = \frac{1}{|S_t|}\sum_{i\in S_t}\boldsymbol{x}_i = \boldsymbol{\mu}^\star + \frac{1}{\alpha n}\sum_{i\in S_t}\left(\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right).$$

Therefore, the ℓ_2 distance between the learned parameter $\boldsymbol{\mu}_{t+1}$ and the ground truth parameter $\boldsymbol{\mu}^\star$ can be bound by:

$$\begin{aligned}
&\left\|\boldsymbol{\mu}_{t+1}-\boldsymbol{\mu}^\star\right\|_2 = \frac{1}{\alpha n}\left\|\sum_{i\in S_t}\left(\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right)\right\|_2 \\
&\le \frac{1}{\alpha n}\left(\sum_{i\in S_t\cap S^\star}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2 + \sum_{i\in S_t\backslash S^\star}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2\right) \\
&\le \frac{|S_t\backslash S^\star|}{\alpha n}\cdot\left\|\boldsymbol{\mu}_t-\boldsymbol{\mu}^\star\right\|_2 \\
&\quad + \frac{1}{\alpha n}\left(\sum_{i\in S^\star\cap S_t}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2 + \sum_{i\in S^\star\backslash S_t}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}_t\right\|_2\right) \\
&\le \underbrace{\frac{2|S_t\backslash S^\star|}{\alpha n}}_{\kappa}\cdot\left\|\boldsymbol{\mu}_t-\boldsymbol{\mu}^\star\right\|_2 + \frac{1}{\alpha n}\sum_{i\in S^\star}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2,
\end{aligned}$$

where the second inequality is guaranteed by line 3 in Algorithm 1. Note that $|S^\star| = \alpha n = |S_t|$, thus $|S^\star\backslash S_t| = |S_t\backslash S^\star|$. Due to the choice of $|S_t|$, we have the distance $\left\|\boldsymbol{x}_i-\boldsymbol{\mu}_t\right\|_2$ of each sample in $S_t\backslash S^\star$ is less than that of samples in $S^\star\backslash S_t$.

By Lemma 1, we can bound $\frac{1}{\alpha n}\sum_{i\in S^\star}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2$ by $2\sqrt{d\lambda_{(\lceil \alpha n\rceil)}}$ with probability at least $1-\frac{1}{\alpha n}$. Meanwhile, by $|S^\star| = |S_t| = \alpha n$, it guarantees $|S_t\backslash S^\star| \le (1-\alpha)n$. Thus, when $\alpha \ge \frac{4}{5}$, $\kappa \le \frac{2(1-\alpha)}{\alpha} \le \frac{1}{2}$. Combining the inequalities completes the proof. ■

Based the error bound per-round in Lemma 2, it is easy to show that $\left\|\boldsymbol{\mu}_t-\boldsymbol{\mu}^\star\right\|_2$ can be upper bounded by $\Theta(\sqrt{d\lambda_{(\lceil \alpha n\rceil)}})$ after sufficient iterations. This leads to our final guarantee.

Theorem 1. *Given ITM with $\alpha \ge \frac{4}{5}$ within $T = \Theta\left(\log_2\frac{\|\boldsymbol{\mu}_0-\boldsymbol{\mu}^\star\|_2}{\sqrt{d\lambda_{(\lceil \alpha n\rceil)}}}\right)$ iterations, it holds that*

$$\left\|\boldsymbol{\mu}_T-\boldsymbol{\mu}^\star\right\|_2 \le c\sqrt{d\lambda_{(\lceil \alpha n\rceil)}} \qquad (4)$$

with probability at least $1-\frac{5T}{4n}$.

Proof. By Lemma 2, we derive

$$\begin{aligned}
\left\|\boldsymbol{\mu}_T-\boldsymbol{\mu}^\star\right\|_2 &\le \frac{1}{2}\left\|\boldsymbol{\mu}_{T-1}-\boldsymbol{\mu}^\star\right\|_2 + 2\sqrt{d\lambda_{(\lceil \alpha n\rceil)}} \\
&\le \frac{1}{2^T}\left\|\boldsymbol{\mu}_0-\boldsymbol{\mu}^\star\right\|_2 + 2\sum_{i=0}^{T-1}\frac{1}{2^i}\sqrt{d\lambda_{(\lceil \alpha n\rceil)}} \\
&\le \frac{1}{2^T}\left\|\boldsymbol{\mu}_0-\boldsymbol{\mu}^\star\right\|_2 + 4\sqrt{d\lambda_{(\lceil \alpha n\rceil)}} \\
&\le c\sqrt{d\lambda_{(\lceil \alpha n\rceil)}}
\end{aligned}$$

with probability at least $1-\frac{5T}{4n}$ when $T = \Theta\left(\log_2\frac{\|\boldsymbol{\mu}_0-\boldsymbol{\mu}^\star\|_2}{\sqrt{d\lambda_{(\lceil \alpha n\rceil)}}}\right)$. ■

The theorem shows that the error bound only depends on the order statistics $\lambda_{(\lceil \alpha n\rceil)}$, irrelevant of the larger covariances for $\alpha \ge 4/5$. Therefore, using the iterative trimming method, the magnitudes of a $1/5$ fraction of highest noises will not affect the quality of the final solution. However, it is still an open question whether the bound can be made $c\sqrt{d\lambda_{(\lceil \alpha n\rceil)}/n}$, i.e., the optimal rate given an oracle that knows the covariances.

The method is also computationally very simple and efficient. Since a single iteration of ITM takes time $\tilde{\mathcal{O}}(nd)$, $\mathcal{O}(\sqrt{d\lambda_{(\lceil \alpha n\rceil)}})$ error can be computed in $\tilde{\mathcal{O}}(nd)$ time, which is nearly linear in the input size.

Remark 1. With a further assumption that $\boldsymbol{x}_i$ are independently sampled from the Gaussians $\mathcal{N}(\boldsymbol{\mu}^\star, \Sigma_i)$, the same bound in the theorem holds with probability converging to 1 with exponential rate of n. This is because (2) can be guaranteed with a higher probability. The sketch of the proof follows: $\left(\sum_{i\in S}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\right\|_2\right)^2 \le |S|\lambda_S\sum_{i\in S}\left\|\tilde{\boldsymbol{x}}_i\right\|_2^2$, by Cauchy-Schwarz inequality. Here $\tilde{\boldsymbol{x}}_i \sim \mathcal{N}(0, I_d)$ due to the Gaussian distribution of $\boldsymbol{x}_i$. By the Lemma 3 in Fan and Lv (2008), $\sum_{i\in S}\left\|\tilde{\boldsymbol{x}}_i\right\|_2^2 \le cd|S|$ with probability at least $1-e^{-c'd|S|}$. Hence,

$\mathbb{P}\left(\left(\sum_{i\in S}\|\boldsymbol{x}_i-\boldsymbol{\mu}^\star\|_2\right)^2 \le cd|S|^2\lambda_S\right) \ge 1-e^{-c'd|S|}$, where c can be any constant greater than 1 and $c'=[c-1-\log(c)]/2$.

Remark 2. For the univariate case with $d=1$, it is sufficient to regard Σ_i as one dimensional matrix σ_i^2. Then we obtain that when $\alpha \ge \frac{4}{5}$ and $T=\Theta\left(\log_2\frac{\|\mu_0-\mu^\star\|}{\sigma_{(\lceil\alpha n\rceil)}}\right)$ iterations: $\|\mu_T-\mu^\star\|_2 \le c\sigma_{(\lceil\alpha n\rceil)}$, with probability at least $1-\frac{5T}{4n}$.

Remark 3. It is worth mentioning that there is a trade off between the accuracy and running time in ITM. In particular, the constant α can be any constant greater than $\frac{2}{3}$, since it suffices to guarantee κ in the proof of Lemma 2 is less than 1. On the other hand, a smaller α will slow down the speed for $\|\boldsymbol{\mu}_t-\boldsymbol{\mu}^\star\|_2$ to shrink, although the computational complexity is still in the same order of $\tilde{\mathcal{O}}(nd)$.

Remark 4. We now discuss existing results in detail.

In the univariate case, Chierichetti et al. (2014) achieved $\min_{2\le k\le\log n}\tilde{\mathcal{O}}\left(n^{1/2(1+1/(k-1))}\sigma_k\right)$ error in time $\mathcal{O}(n\log^2 n)$. Among all estimators studied in Pensia et al. (2019), the superior performance is obtained by the hybrid estimators, which includes version (1): combining k_1-median with k_2-shorth and version (2): combining k_1-median with modal interval estimator. These two versions achieve similar guarantees while version 1 has lower run time $\mathcal{O}(n\log n)$. Version 1 of the hybrid estimator outputs $\hat{\mu}_{k_1,k_2}$ such that $|\hat{\mu}_{k_1,k_2}-\mu| \le \frac{4\sqrt{n}\log n}{k_2}r_{2k_2}$ with probability $1-2\exp(-c'k_2)-2\exp(-c\log^2 n)$, where $k_1=\sqrt{n}\log n$ and $k_2\ge C\log n$. Since here r_k is defined as $\inf\left\{r:\frac{1}{n}\sum_{i=1}^n\mathbb{P}(|x_i-\mu^\star|\le r)\ge\frac{k}{n}\right\}$, the error bound giving above varies with specific $\{F_i\}_{i=1}^n$, while the worst-case error guarantee is $\mathcal{O}(\sqrt{n}\sigma_{(C\log n)})$. When take $k_2=\frac{\lceil\alpha n\rceil}{2}$, the error can be $\mathcal{O}(\frac{\log n}{\sqrt{n}}\sigma_{(\lceil\alpha n\rceil)})$. However, this result for symmetric and unimodal distributions F_i's.

In the multivariate case, Chierichetti et al. (2014) studied the special case where $\boldsymbol{x}_i\sim\mathcal{N}(\boldsymbol{\mu},\sigma_i^2 I_d)$ and provided an algorithm with the error bound $\min_{2\le k\le\log n}\tilde{\mathcal{O}}\left(n^{(1+1/(k-1))/d}\sigma_k\right)$ in time $\tilde{\mathcal{O}}(n^2)$. Pensia et al. (2019) mainly considered the special case where the overall mixture distribution is radically symmetric, and sharpens the bound above, resulting in the worst-case error bound $\mathcal{O}(\sqrt{d}\sqrt{n}^{1/d}\sigma_{(Cd\log n)})$. Pensia et al. (2019) also provided computationally efficient estimator with running time $\mathcal{O}(n^2 d)$.

These existing results depend on $\sigma_{(C\log n)}$ (or alike) at the expense of an additional factor of roughly $\sqrt{n}^{1/d}$, which is most relevant when $C\log n$ of the points have small noises, i.e., when the samples are dominated by high noises. Our results depend on $\sigma_{(\alpha n)}$ (or $\sqrt{\lambda_{(\lceil\alpha n\rceil)}}$ for multivariate) for $\alpha\ge 4/5$, which is more relevant when 1/5 fraction of the points have high noises. We believe our bounds are more applicable for many practical scenarios. Finally, for the multivariante case, our result holds under more general assumptions and the method is significantly simpler and more efficient.

4 LINEAR REGRESSION

Given observations $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ from the linear model

$$y_i=\boldsymbol{x}_i^T\boldsymbol{\beta}^\star+\epsilon_i,\quad \forall 1\le i\le n, \tag{5}$$

our goal is to estimate $\boldsymbol{\beta}^\star$. Here $\{\epsilon_i\}_{i=1}^n$ are independently distributed with expectation 0 and variances $\{\sigma_i^2\}$, and $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent with $\{\epsilon_i\}_{i=1}^n$ and satisfy some regularity conditions described below. Denote the stack matrix $(\boldsymbol{x}_1^T,\cdots,\boldsymbol{x}_n^T)^T$ as X, the noise vector $(\epsilon_1,\cdots,\epsilon_n)^T$ as $\boldsymbol{\epsilon}$ and the response vector $(y_1,\cdots,y_n)^T$ as $\boldsymbol{y}$.

Assumption 1. *Assume $\|\boldsymbol{x}_i\|_2=1$ for all i. Define*

$$\psi^-(k)=\min_{S:|S|=k}\lambda_{\min}\left(X_S^T X_S\right),$$

where X_S is the submatrix of X consisting of rows indexed by $S\subseteq[n]$. Assume that for $k=\Omega(n)$, $\psi^-(k)\ge k/c_1$ for a constant $c_1>0$.

Remark 5. $\|\boldsymbol{x}_i\|_2=1$ is assumed without loss of generality, as we can always normalize $(\boldsymbol{x}_i,y_i)$, without affecting the assumption on ϵ_i's. The assumption on $\psi^-(k)$ states that every large enough subset of X is well conditioned, and has been used in previous work, e.g., Bhatia et al. (2015); Shen and Sanghavi (2019). It is worth mentioning that the uniformity over S assumed is not for convenience but for necessity since the covariances of samples are unknown. Still, the assumption holds under several common settings in practice. For example, by Theorem 17 in Bhatia et al. (2015), this regularity is guaranteed for c_1 close to 1 w.h.p. when the rows of X are i.i.d. spherical Gaussian vectors and n is sufficiently large.

Remark 6. In our setting, the noise terms are independent but not necessarily identically distributed. It is also referred to as heteroscedasticity in linear regression by Rao (1970) and Horn et al. (1975).

4.1 Iterative Trimmed Squares Minimization

We now apply iterative trimming to linear regression. The first step is to use the square loss, i.e. let $f_{\boldsymbol{\beta}}(\boldsymbol{x}_i,y_i)=\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2$, then the form of the trimmed loss estimator turns to

$$\hat{\boldsymbol{\beta}}^{(\mathrm{TL})}=\operatorname*{arg\,min}_{\boldsymbol{\beta},S:|S|=\lceil\alpha n\rceil}\sum_{i\in S}\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2. \tag{6}$$

Algorithm 2 Iterative Trimmed Squares Minimization (ITSM)

```
Input: Samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, number of rounds $T$,
    fraction of samples $\alpha$
$\boldsymbol{\beta}_0 \leftarrow \hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$
for $t = 0, \cdots, T-1$ do
    Choose samples with smallest current loss:
        $S_t \leftarrow \underset{S:|S|=\lceil \alpha n \rceil}{\arg\min} \sum_{i \in S} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_t\right)^2$
    $\boldsymbol{\beta}_{t+1} = \hat{\boldsymbol{\beta}}_{S_t}^{(\mathrm{LS})}$
end for
Output: $\boldsymbol{\beta}_T$
```

Such $\hat{\boldsymbol{\beta}}^{(\mathrm{TL})}$ is first introduced by Rousseeuw (1984) as least trimmed squares estimator and its statistical efficiency has been studied in previous literature. However, the principal shortcoming is also its high computational complexity; see, e.g., Mount et al. (2014). Hence, we again use iterative trimming. Let

$$\hat{\boldsymbol{\beta}}_S^{(\mathrm{LS})} = \arg\min_{\boldsymbol{\beta}} \sum_{i \in S} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2$$

which is the least square estimator obtained by the sample set S. When $S = [n]$, we omit the subscript S and write it as $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$. The resulting algorithm is called Iterative Trimmed Squares Minimization and described in Algorithm 2.

4.2 Theoretical Guarantees

The key idea for the analysis is similar to that for mean estimation: the selected set S has sufficiently large overlap with the set $S^\star$ of $\lceil \alpha n \rceil$ "good" points with smallest noises, while the points in $S \backslash S^\star$ are not that "bad" by the selection criterion. Therefore, the algorithm makes progress in each iteration.

Lemma 3. *Under Assumption 1, given ITSM $\alpha \geq \frac{4c_1}{1+4c_1}$, with probability at least $1 - \frac{1+4c_1}{4c_1 n}$, we have*

$$\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star\right\|_2 \leq \frac{1}{2}\left\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\right\|_2 + 2c_1 \sigma_{(\lceil \alpha n \rceil)}. \tag{7}$$

Proof. First we introduce some notations. Define

$$\psi^+(k) = \max_{S:|S|=k} \lambda_{\max}\left(X_S^T X_S\right)$$

where X_S is the submatrix of X consisting of rows indexed by S. Note that $\psi^+(k) \leq k$, since for any S of size k, by $\|\boldsymbol{x}_i\|_2 = 1$, $\lambda_{\max}\left(X_S^T X_S\right)$ is bounded by

$$\operatorname{Tr}\left(X_S^T X_S\right) = \operatorname{Tr}\left(X_S X_S^T\right) = \sum_{i \in S} \|\boldsymbol{x}_i\|_2^2 = k.$$

Denote W_t as the diagonal matrix where $W_{t,ii} = 1$ if the i-th sample is in set S_t, otherwise $W_{t,ii} = 0$. Let $S^\star$ be a subset of $\left\{i : \sigma_i \leq \sigma_{(\lceil \alpha n \rceil)}\right\}$ with size $\lceil \alpha n \rceil$ and denote $W^\star$ as the diagonal matrix w.r.t. $S^\star$.

Under Assumption 1, $X_{S_t}^T X_{S_t} = X^T W_t X$ is nonsingular, so we have $\boldsymbol{\beta}_{t+1} = \left(X^T W_t X\right)^{-1} X^T W_t \boldsymbol{y}$, where we have used $W_t^2 = W_t$. Then

$$\begin{aligned}\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}^\star + \left(X^T W_t X\right)^{-1} X^T W_t \boldsymbol{\epsilon} \\ &= \boldsymbol{\beta}^\star + \left(X^T W_t X\right)^{-1} X^T W_t \left((I - W^\star)\boldsymbol{\epsilon} + W^\star \boldsymbol{\epsilon}\right).\end{aligned}$$

Therefore, the error can be bounded by:

$$\begin{aligned}&\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star\right\|_2 \\ =&\left\|\left(X^T W_t X\right)^{-1} X^T \left(W_t (I - W^\star)\boldsymbol{\epsilon} + W_t W^\star \boldsymbol{\epsilon}\right)\right\|_2 \\ \leq& \frac{1}{\psi^-(\lceil \alpha n \rceil)}\left(\underbrace{\left\|X^T W_t (I - W^\star)\boldsymbol{\epsilon}\right\|_2}_{\mathcal{T}_1} + \underbrace{\left\|X^T W_t W^\star \boldsymbol{\epsilon}\right\|_2}_{\mathcal{T}_2}\right),\end{aligned}$$

where we use the spectral norm inequality and triangle inequality and the fact that $\operatorname{Tr}(W_t) = \lceil \alpha n \rceil$. In the following, we bound the two terms $\mathcal{T}_1$ and $\mathcal{T}_2$.

First, $\mathcal{T}_1$ can be bounded as:

$$\begin{aligned}\mathcal{T}_1 &= \left\|X^T W_t (I - W^\star)\boldsymbol{\epsilon}\right\|_2 \\ &\leq \left\|X^T W_t (I - W^\star) X (\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star))\right\|_2 \\ &\quad + \left\|X^T W_t (I - W^\star)(\boldsymbol{y} - X\boldsymbol{\beta}_t)\right\|_2 \\ &\leq \psi^+(|S_t \backslash S^\star|)\left\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\right\|_2 \\ &\quad + \left\|X^T W_t (I - W^\star)(\boldsymbol{y} - X\boldsymbol{\beta}_t)\right\|_2,\end{aligned}$$

since $\operatorname{Tr}((I - W^\star)W_t) = |S_t \backslash S^\star|$.

By the fact that $|S_t \backslash S^\star| = |S^\star \backslash S_t|$, there exists a bijection between $S_t \backslash S^\star$ and $S^\star \backslash S_t$. Denote the image of $i \in S_t \backslash S^\star$ as k_i. Since the loss $(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_t)^2$ of sample in $S_t \backslash S^\star$ is less than that of sample in $S^\star \backslash S_t$, we have $\left|y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_t\right| \leq \left|y_{k_i} - \boldsymbol{x}_{k_i}^T \boldsymbol{\beta}_t\right|$ for any $i \in S_t \backslash S^\star$. Hence we can write

$$\begin{aligned}&\left\|X^T W_t (I - W^\star)(\boldsymbol{y} - X\boldsymbol{\beta}_t)\right\|_2 \\ =&\left\|\sum_{i \in S_t \backslash S^\star} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_t)\boldsymbol{x}_i\right\|_2 \\ \leq&\left\|\sum_{i \in S_t \backslash S^\star} n_i (y_{k_i} - \boldsymbol{x}_{k_i}^T \boldsymbol{\beta}_t)\boldsymbol{x}_i\right\|_2 \\ \leq&\left\|\sum_{i \in S_t \backslash S^\star} n_i \epsilon_{k_i} \boldsymbol{x}_i\right\|_2 + \left\|\sum_{i \in S_t \backslash S^\star} n_i \boldsymbol{x}_i {\boldsymbol{x}_{k_i}}^T (\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star)\right\|_2\end{aligned}$$

where n_i is either 1 or -1.

By the assumption that $\|\boldsymbol{x}_i\|_2 = 1$,

$$\left\| \sum_{i \in S_t \backslash S^\star} n_i \epsilon_{k_i} \boldsymbol{x}_i \right\|_2 \leq \sum_{i \in S_t \backslash S^\star} |\epsilon_{k_i}| = \sum_{i \in S^\star \backslash S_t} |\epsilon_i|.$$

Meanwhile,

$$\left\| \sum_{i \in S_t \backslash S^\star} n_i \boldsymbol{x}_i {\boldsymbol{x}_{k_i}}^T (\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star) \right\|_2 \leq s_{\max} \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2 .$$

where $s_{\max} := s_{\max}\left(\sum_{i \in S_t \backslash S^\star} n_i \boldsymbol{x}_i {\boldsymbol{x}_{k_i}}^T\right)$ denotes the maximum singular value and is bounded as follows.

Claim 1. $s_{\max} \leq \psi^+ (|S_t \backslash S^\star|)$.

Proof. The proof is implicit in the proof for Theorem 7 in Shen and Sanghavi (2019). With matrix notations, $\sum_{i \in S_t \backslash S^\star} n_i \boldsymbol{x}_i {\boldsymbol{x}_{k_i}}^T$ can be then written as $X^T W_t (I - W^\star) NPX$, where N is diagonal with diagonal entries in $\{1, -1\}$ and P is some permutation matrix. Then

$$s_{\max} = \max_{\|u\|_2 = 1, \|v\| = 1} u^T X^T W_t (I - W^\star) NPXv.$$

Let $\tilde{u} = Xu$ and $\tilde{v} = Xv$, $s_{\max}$ is bounded by

$$\sum_{i \in S_t \backslash S^\star} |\tilde{u}_{r_i} \tilde{v}_{t_i}| \leq \max\{ \sum_{i \in S_t \backslash S^\star} \tilde{u}_{r_i}^2, \sum_{i \in S_t \backslash S^\star} \tilde{v}_{t_i}^2 \}$$

for some sequences $\{r_i\}$ and $\{t_i\}$. So $s_{\max}$ is bounded by the larger of the maximum singular values of $X^T W_t (I - W^\star) X$ and $X^T P^T N^T W_t (I - W^\star) NPX$, which is bounded by $\psi^+ (|S_t \backslash S^\star|)$. ■

With this claim, we can derive

$$\mathcal{T}_1 \leq \sum_{i \in S^\star \backslash S_t} |\epsilon_i| + 2\psi^+ (|S_t \backslash S^\star|) \|\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t\|_2 .$$

Next, $\mathcal{T}_2$ can be bounded as:

$$\mathcal{T}_2 = \left\| \sum_{i \in S_t \cap S^\star} \epsilon_i \boldsymbol{x}_i \right\|_2 \leq \sum_{i \in S_t \cap S^\star} |\epsilon_i| .$$

Combining the inequalities above, we have

$$\begin{aligned} \|\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t\|_2 &= \frac{1}{\psi^-(\lceil \alpha n \rceil)} (\mathcal{T}_1 + \mathcal{T}_2) \\ &\leq \frac{1}{\psi^-(\lceil \alpha n \rceil)} \left(\sum_{i \in S^\star} |\epsilon_i| + 2\psi^+ (|S_t \backslash S^\star|) \|\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t\|_2 \right) \\ &= \underbrace{\frac{2\psi^+ (|S_t \backslash S^\star|)}{\psi^-(\lceil \alpha n \rceil)}}_{\kappa} \|\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t\|_2 + \frac{\sum_{i \in S^\star} |\epsilon_i|}{\psi^-(\lceil \alpha n \rceil)}. \end{aligned}$$

By assumption, there exist a constant c_1 such that $\frac{\alpha n}{\psi^-(\lceil \alpha n \rceil)} \leq c_1$ and $\frac{\psi^+(|S_t \backslash S^\star|)}{\psi^-(\lceil \alpha n \rceil)} \leq c_1 \cdot \frac{|S_t \backslash S^\star|}{\lceil \alpha n \rceil}$. Without loss of generality, assume αn is an integer. Since $|S_t \backslash S^\star| \leq (1 - \alpha) n$, when $\alpha \geq \frac{4c_1}{1+4c_1}$, $\kappa \leq 2c_1 \cdot \frac{1-\alpha}{\alpha} \leq \frac{1}{2}$. And $\sum_{i \in S^\star} |\epsilon_i|$ is bounded by $2|S^\star| \sigma_{(\lceil \alpha n \rceil)}$ with probability at least $1 - \frac{1}{|S^\star|}$, using Markov's inequality. ■

Theorem 2. *Under Assumption 1, given ITSM with $\alpha \geq \frac{4c_1}{1+4c_1}$ and $T = \Theta\left(\log_2 \frac{\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^\star\|_2}{\sigma_{(\lceil \alpha n \rceil)}}\right)$, it holds that*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^\star\|_2 \leq c c_1 \sigma_{(\lceil \alpha n \rceil)} \tag{8}$$

with probability at least $1 - T \frac{1+4c_1}{4c_1 n}$.

Proof. It follows from Lemma 3 by an argument similar to that for Theorem 1. ■

Remark 7. When $\boldsymbol{x_i}$'s are i.i.d. spherical Gaussians, Assumption 1 can be satisfied with c_1 close to 1. Then we require $\alpha \geq 4/5$ and the error bound holds with probability $\geq 1 - 5T/(4n)$, similar to that for mean estimation. Also, (8) is obtained without extra assumption on noise ϵ_i except for assuming its second order moment exists. If $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, the previous bound holds with a higher probability $1 - e^{-c'n}$.

Remark 8. We now discuss existing results. Pensia et al. (2019) also adapted its mean estimation methodology to linear regression. When $\boldsymbol{x}_i$'s are from a multivariate Gaussian with covariance matrix Σ, w.h.p. it obtained error bound $\frac{c' n \sigma_{(cd \log n)}}{\sqrt{\lambda_{\min}(\Sigma)}}$. Again, the result depends on $\sigma_{(cd \log n)}$ with an additional factor n, while ours depends on $\sigma_{(\alpha \log n)}$ for a constant α (our bound will also have a $\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}$ factor for such Gaussians). However, their run time is $\mathcal{O}(n^d)$, exponential in the dimension d, while ours is polynomial.

There also exist studies for robust linear regression in the adversary setting, where a small fraction of points are being corrupted by an adversary (e.g., Bhatia et al. (2015); Liu et al. (2018); Shen and Sanghavi (2019); Diakonikolas et al. (2019b)). It is unclear if their results directly apply to our setting, since they have additional assumptions on the data.

5 Experiments

5.1 Mean Estimation

To validate Theorem 1, we repeat the procedure that first generating samples $\{\boldsymbol{x}_i\}_{i=1}^n$ under designed distribution $\{F_i\}_{i=1}^n$, then run Algorithm 1 with fraction $\alpha = \frac{4}{5}$ and iteration $T = 20$ and finally output error $|\boldsymbol{\mu}_T - \boldsymbol{\mu}^\star|$. We report the average error over R repetitions; we set $R = 200$ for the univariate case and

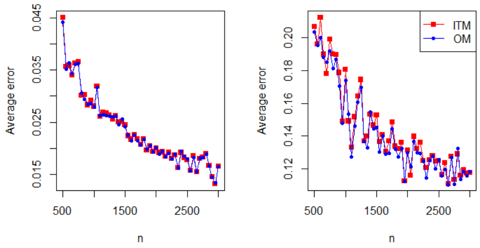

Figure 1: Univariate mean estimation. Left: setting 1; Right: setting 2. x-axis: sample size; y-axis: error after $T = 20$ iterations.

$R = 20$ for the multivariate case. For comparison, we also report the error of the empirical mean over the first αn samples with smallest variance (or norm of covariance matrix), which is the estimator given an oracle knowing all the covariances, and thus referred to as Oracle Mean (OM). It is easy to be seen that the latter average is only relevant to $\{\sigma_{(i)}\}_{i=1}^{\alpha n}$ (or $\{\lambda_{(i)}\}_{i=1}^{\alpha n}$) and regardless of the rest $(1-\alpha)n$ samples.

Univariate Case We first present experiments when $d = 1$ under two designed settings of the entangled distributions $\{F_i\}_{i=1}^{n}$.

Setting 1 For $i \leq \alpha n$, $F_i = \mathcal{N}(0, 1)$ and $F_i = \mathcal{N}(0, i^2)$ for $i > \alpha n$.

Setting 2 For $i \leq \alpha n$, $F_i = \mathcal{N}(0, (\log i)^2)$ and $F_i = \mathcal{N}(0, i^2)$ for $i > \alpha n$.

Figure 1 shows the performances of ITM and OM. In both settings, ITM obtains roughly the same average error as OM, showing that the error bound of ITM is regardless of all $\sigma_i \geq \sigma_{(\lceil \alpha n \rceil)}$. Besides, ITM actually converges to the truth $\boldsymbol{\mu}^\star$, in the same rate as OM. It suggests ITM can achieve a vanishing error, at least in some special cases. This is left as a future direction.

Multivariate Case For the multivariate case, we set $d = 10$ and provide two simulation settings of $\{F_i\}_{i=1}^{n}$.

Setting 3 (radically symmetric) $F_i = \mathcal{N}(0, I_{10})$ for $i \leq \alpha n$, and $F_i = \mathcal{N}(0, 100 \cdot I_{10})$ for $i > \alpha n$.

Setting 4 (radically asymmetric) Let Σ_0 be a stochastic positive definite matrix generated as follows. (1) Set the diagonal entries to 1. (2) Off-diagonal entries are set to zero or nonzero with equal probability. For each nonzero off-diagonal entry, it is sampled from the uniform distribution on interval $(-0.5, 0.5)$. Then we make it symmetric by forcing the lower triangular matrix equal to the upper triangular, and then add a diagonal matrix $c\mathbf{I}_{10}$ to make sure it is positive definite, where c is chosen to make the smallest eigenvalue of the matrix equal to 0.2. Finally, let $F_i = \mathcal{N}(0, \Sigma_0)$ for $i \leq \alpha n$ and $F_i = \mathcal{N}(0, 100 \cdot \Sigma_0)$ for $i > \alpha n$.

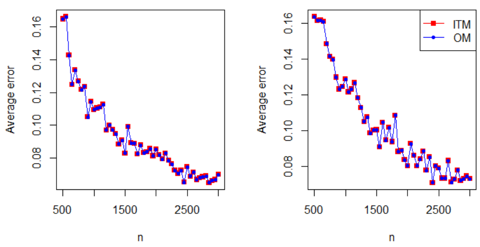

Figure 2: Multivariate mean estimation. Left: setting 3; Right: setting 4. x-axis: sample size; y-axis: error after $T = 20$ iterations.

Figure 2 shows the results. Again, in both settings, ITM also obtain the same average error with OM. The results for setting 4 suggest that the radical symmetry assumption may not be needed.

5.2 Linear Regression

We now consider the linear regression problem with $d = 100$. We generate data according to the model (5), where each row of matrix X is independently sampled from $\mathcal{N}(\mathbf{0}, I_d)$ and we choose $\boldsymbol{\beta}^\star$ to be a random vector with l_2 norm 1. The noise vector is generated s.t. for $i \leq \alpha n$, $\epsilon_i \sim \mathcal{N}(0, 1)$ and $\epsilon_i \sim \mathcal{N}(0, 100)$ for $i > \alpha n$. We repeatedly generate data and run ITSM for $R = 20$ times, then report the average errors of both ITSM and the least square estimator over the first αn samples with smallest noise variances, which we refer to as Oracle Least Square (OLS). We also set $\alpha = \frac{4}{5}$ and $T = 20$ for simplicity though the smallest possible value of α in Lemma 3 is dependent on X.

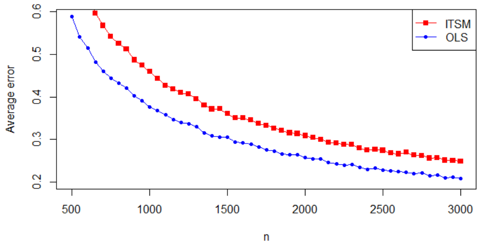

Figure 3: Linear Regression. x-axis: sample size; y-axis: error after $T = 20$ iterations.

Figure 3 shows the results. Similar to mean estimation, the iterative trimming method has error closely tracking those of the oracle method. This again verifies the effectiveness of iterative trimming and provide positive support for our analysis.

Acknowledgement

This work was supported in part by FA9550-18-1-0166. The authors would also like to acknowledge the support provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

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.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010a.

Mikhail Belkin and Kaushik Sinha. Toward learning gaussian mixtures with arbitrary separation. In *COLT*, pages 407–419. Citeseer, 2010b.

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

Peter J Bickel et al. On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3): 847–858, 1965.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. Learning entangled single-sample gaussians. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 511–522. Society for Industrial and Applied Mathematics, 2014.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

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*, pages 999–1008. JMLR. org, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018a.

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*, 2018b.

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, 2019a.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019b.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Susan D Horn, Roger A Horn, and David B Duncan. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70(350):380–385, 1975.

Ola Hössjer. Exact computation of the least trimmed squares estimate in simple linear regression. *Computational Statistics & Data Analysis*, 19(3):265–282, 1995.

Peter J Huber. *Robust statistics*. Springer, 2011.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

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

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. On the least trimmed squares estimator. *Algorithmica*, 69 (1):148–183, 2014.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Estimating location parameters in entangled single-sample distributions. *arXiv preprint arXiv:1907.03087*, 2019.

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

C Radhakrishna Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172, 1970.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Peter J Rousseeuw and Katrien Van Driessen. Computing lts regression for large data sets. *Data mining and knowledge discovery*, 12(1):29–45, 2006.

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

Fumin Shen, Chunhua Shen, Anton van den Hengel, and Zhenmin Tang. Approximate least trimmed sum of squares fitting and applications in image analysis. *IEEE Transactions on Image Processing*, 22(5):1836–1847, 2013.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748, 2019.

John W Tukey and Donald H McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352, 1963.

Daniel Vainsencher, Shie Mannor, and Huan Xu. Ignoring is a bliss: Learning with large noise through reweighting-minimization. In *Conference on Learning Theory*, pages 1849–1881, 2017.

Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

Eunho Yang, Aurélie C Lozano, Aleksandr Aravkin, et al. A general family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12(2):3519–3553, 2018.