ROBUST MEAN ESTIMATION AND REGRESSION WITH SUB-GAUSSIAN RATES

Zhuoyan Xu
Department of Statistics
zhuoyan.xu@wisc.edu

1 Introduction

Robust high dimensional mean estimation is a fundamental problem in learning theory. Consider we have N independent identical sample from known distribution, the goal is to estimate the mean μ with sample and computational efficiency estimators. We use ℓ_2-norm measuring loss in this project. If we have Gaussian distribution $\mathcal{N}_d(\mu, I)$, the empirical mean can bound error within $O(\sqrt{d/N})$ with high probability. We are interested in how the error bound is related to probability specifically. we define general bound:

$$\mathbb{P}\left\{\|\hat{X} - \mu\| \geq \varepsilon_\delta\right\} \leq \delta \tag{1}$$

In sub-Gaussian, we have bound

$$\varepsilon_\delta = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right)$$

known as sub-gaussian performance or sub-gaussian rates. Catoni [2012] showed this rate is optimal under sub-gaussian assumptions. We can also get above rates asymptotically when we have sample size tends to infinity due to central limit theorem.

However, for non-gaussian distribution with possibly a heavy-tail, we cannot expect empirical mean to have such sub-gaussian behavior. Moreover, consider we are given ε- corrupted sample, the empirical mean can behave worse due to outliers. A recent line of works have been working on this problem.

2 Motivation and Related Work

For non-corrupted sample, several works Catoni [2012] Joly et al. [2017] Lugosi et al. [2019] used the median-of-means on high dimension robust estimation. Lugosi et al. [2019] devised an improved estimator, based on the median-of-means framework, called the median-of-means tournament achieving the sub-gaussian rate. But the estimators in Joly et al. [2017] Lugosi et al. [2019] are not known to be computation feasible. To overcome this computation intractability, subsequent works Hopkins et al. [2020]Cherapanamjeri et al. [2019] proposed polynomial algorithm based on the sum-of-squares hierarchy.

For ε- corrupted sample in high dimension setting, some traditional approach (geometric median, coordinate median) can only attain error bound $O(\varepsilon\sqrt{d})$, which depend on dimension. Tukey [1975] Chen et al. [2018] using Tukey-median can attain $O(\varepsilon + \sqrt{d/N})$ error bound in spherical Gaussian $\mathcal{N}_d(\mu, I)$. Here we have sample efficient estimator, if we obtain sample $N = \Omega(d/\varepsilon^2)$, we obtain ℓ_2-error as $O(\varepsilon)$ independent of dimension. But Tukey median requires computation time $poly(N, 2^{O(d)})$ which is exponential in d. Diakonikolas et al. [2017] Lai et al. [2016] proposed efficient algorithm also runs in polynomial time in dimension for bounded covariance distribution. Filtering algorithm in Diakonikolas et al. [2017] can achieve error rate $O(\sqrt{\epsilon} + \sqrt{d \log d/n})$. Lugosi and Mendelson [2019] proposed trimmed mean estimator achieved error in sub-gaussian rate $O\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\frac{log(1/\delta)}{n}}\right)$ which requires exponential running time in computation.

3 Main Goal

In this project, I will investigate some papers implementing polynomial algorithms achieving sub-gaussian rates in mean estimation and regression. Diakonikolas et al. [2020] proved the robust mean estimators algorithms under *stability* condition can achieve the nearly sub-gaussian rates. The authors proved such algorithms with a preprocessing step can achieve sub-gaussian rates.

Depersin [2020] explore the problem in linear regression, the authors adapted median-of-means idea and proposed estimators robust both to heavy-tailed data and outliers in sub-gaussian rate. They proposed descent algorithm runs in nearly quadratic time to tackle the issue of inefficiency of Sum-of-Square programming in practical running.

In this project I will investigate the theorems in papers and work on the problems of how the median-of-means algorithms work under stability conditions, I will also work on the usage of stability conditions in regression setting. I'm interested in recent polynomial algorithms achieving sub-gaussian rate and how they can be generalized to regression problems.

4 Main contribution of the paper

Mean estimation for non-gaussian distribution with possibly a heavy-tail it self is a hard problem, even without outliers. In strong contamination model, the best statistical error rate we can get is $O\left(\sqrt{\epsilon}+\sqrt{d/n}+\sqrt{\frac{log(1/\delta)}{n}}\right)$, where δ is the failure probability. This is known as sub-gaussian rate.

4.1 Ideas of Diakonikolas et al. [2020]

In this paper, the authors try to solve this question in both clean data and corrupted data settings. They proved the nearly sub-gaussian rate can be achieved in their algorithm based on stability condition, and combine median-of-means step can achieve sub-gaussian rate. The main contributions can be divided into three parts below.

4.1.1 Stability condition in bounded cavariance distribution

The $stability$ condition is widely used in recent robust estimation work (Diakonikolas et al. [2017], Diakonikolas and Kane [2019], Diakonikolas et al. [2019]). The intuitive idea of this condition is given a finite set, remove any ϵ fraction of the sample, the sample mean of the remaining set is not far away from a known vector μ and the eigen values of sample covariance is not far away from those of known covariance matrix Σ. The deviation of sample mean and sample covariance can be quantified by two parameters ϵ and δ. Several algorithms are operating on this $stability$ condition that have been developed in strong contamination model. If our original set is a (ϵ,δ)-stable set, then given any ϵ-corrupted set of the original set, we have polynomial algorithms estimating mean with error bound δ (Diakonikolas and Kane [2019]).

The authors define problem in general bounded covariance setting, then the authors proved given $i.i.d$ sample set S from bounded covariance distribution, the with probability $1-\tau$, this set contains a $(1-\epsilon')$ fraction set which is a (ϵ',δ)-stable set, where $\epsilon'=O(\log(1/\tau)/n+\epsilon)$, and $\delta=O(\sqrt{(\mathrm{r}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon}+\sqrt{\log(1/\tau)/n})$. The τ here is the failure probability. Now we can apply some known $stability$ based algorithms like filtering algorithm (Diakonikolas and Kane [2019]) to get nearly sub-gaussian rate $O(\sqrt{d\log d/n}+\sqrt{\epsilon}+\sqrt{\log(1/\tau)/n})$.

4.1.2 Stability of median-of-means sets

The authors try to remove the logarithmic factor in previous sub-gaussian rate result. Lugosi et al. [2019] devised an improved estimator, based on the median-of-means framework, called the median-of-means tournament achieving the sub-gaussian rate. But the estimators in (Joly et al. [2017], Lugosi et al. [2019]) are not known to be computation feasible. The authors in this paper combine the median-of-means technique with $stability$ condition.

The median-of-means technique first randomly partition data into K groups with equal size, then compute the means of each group $z_1,\ldots,z_k$. Then compute the multivariate median of $z_1,\ldots,z_k$. The authors use the first step to get $z_1,\ldots,z_k$, and then prove the stability condition on this intermediate dataset. Several $stability$-based algorithm can be used once this dataset has stability.

Given new set $z_1,\ldots,z_k$ computed by $x_1,\ldots,x_n$ from bounded covariance distribution with mu and Σ, Depersin and Lecué [2019] proved with probability $1-\exp(-ck)$, there are 99% of data have bounded second moment. Then authors proved with same probability, there is a 95% subset is a stable set.

In final theorem, we choose $k = \lfloor \epsilon' n \rfloor$, then after partition and compute mean steps, the $z_1, \ldots, z_k$ computed by ϵ-corrupted set is 0.01-corrupted of some set S contains 95% fraction set which is a $(0.1, \delta)$-stable set with respect to bounded covariance distribution with μ and $k\|\Sigma\|/n$, where $\epsilon' = O(\log(1/\tau)/n + \epsilon)$, and $\delta = O(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau)/n})$. The τ here is the failure probability. Now we can apply some known $stability$-based algorithms like filtering algorithm (Diakonikolas and Kane [2019]) to get nearly sub-gaussian rate $O(\sqrt{r(\Sigma)/k} + 1)$.

Then we apply stability based algorithm on above $z_1, \ldots, z_k$, we have error bound

$$\begin{aligned}\|\widehat{\mu} - \mu\| =& O\left((\sqrt{r(\Sigma)/k} + 1)(k\|\Sigma\|/n)\right) = O\left(\sqrt{\|\Sigma\| r(\Sigma)/n} + \sqrt{k\|\Sigma\|/n}\right) \\ =& O\left(\sqrt{tr(\Sigma)/n} + \sqrt{\|\Sigma\|\epsilon'}\right) \\ =& O\left(\sqrt{tr(\Sigma)/n} + \sqrt{\|\Sigma\|\epsilon} + \sqrt{\|\Sigma\| log(1/\tau)/n}\right)\end{aligned}$$

The authors proved the subgaussian-rate(both in clean data or strongly corrupted data) can be achieved in polynomial algorithm.

4.1.3 Stability condition in finite higher moments

In Section 4.1.1 we talked about the stability on bounded covariance distribution. The authors add one more condition bounding the $k-$th central moment σ_k, usually $k \geq 4$. The authors also imposed the identity covariance assumption to bound have the sharper lower bound of covariance matrix. Combining these two got a sharper error rate. Specifically, the error rate improves the $\sqrt{\epsilon}$ to $\epsilon^{1-\frac{1}{k}}$.

The steps are similar in section 4.1.1. The authors starts with the stability condition. The authors proved the upper and lower bound of covariance Σ. With probability $1-\tau$, this set contains a $(1-\epsilon')$ fraction set which is a (ϵ', δ)-stable set, where $\epsilon' = O(\log(1/\tau)/n + \epsilon)$, and $\delta = O(\sigma_k \epsilon^{1-\frac{1}{k}} + \sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sigma_k\sqrt{\log(1/\tau)/n})$. The τ here is the failure probability. Now we can apply some known $stability$ based algorithms like filtering algorithm (Diakonikolas and Kane [2019]) to get nearly sub-gaussian rate $O(\sigma_k \epsilon^{1-\frac{1}{k}} + \sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sigma_k\sqrt{\log(1/\tau)/n})$.

4.2 Depersin [2020]

In this paper, the author tackle the problem in linear regression setting. Consider the predict X and response Y, the author tries to find the β minimizes risk:

$$\beta^* = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \ell(\beta) = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}\left(Y_1 - \langle \beta, X_1 \rangle\right)^2$$

Cherapanamjeri et al. [2020] proposed first polynomial algorithm achieved the risk of β has sub-gaussian rate, but the power of polynomial time can be large since they need to solve large semi-definite programming. The author in Depersin [2020] proposed a practical algorithm achieving nearly quadratic time reaching sub-gaussian rate. The author runs simulations and shows the practical rate is nearly linear in most cases.

In clean data setting, the author assumes the data have bounded covariance and adds one more condition L2/L4 norm equivalence. In the outliers contamination setting, the author uses the strong contamination model. The adversary can observe the data and only keep the data with a certain property (like 9/10 proportion with the largest euclidean norm), then add the outliers arbitrarily(which can be correlated with data). The author bound the size of outliers by $|\mathcal{O}|$, which is equivalent to the setting in Diakonikolas et al. [2020].

The main result in paper is the error rate can be bounded by $O(\frac{1}{N}(\log(1/\delta) \vee d \vee |\mathcal{O}|))$, which attained the sub-gaussian rate in contaminated setting.

4.2.1 Median-of-means framework

The author partition the data into K groups, where K is larger than the VC dimension of given binary classification function set $\mathcal{F}$ and the $|\mathcal{O}|$. The author introduces rate $r = \sqrt{K/N}$, then he proved with failure probability exponential

in K, at least $19/20K$ groups B_k will have the following property:

$$\begin{aligned}
\frac{1}{m}\left|\sum_{i\in B_k}\left(\tilde{Y}_i-\left\langle\beta^*,\tilde{X}_i\right\rangle\right)\left\langle u,\tilde{X}_i\right\rangle\right| \leq & r \\
\left|\frac{1}{m}\sum_{i\in B_k}\left\langle u,\tilde{X}_i\right\rangle\left\langle v,\tilde{X}_i\right\rangle-\langle u,\Sigma v\rangle\right| \leq & 6\gamma\sqrt{\frac{1}{m}}\|u\|_\Sigma\|v\|_\Sigma \\
\frac{1}{m}\sum_{i\in B_k}\left(\tilde{Y}_i-\beta_c\tilde{X}_i\right)\Sigma^{-1/2}\tilde{X}_i \leq & \sqrt{d}\left(r+\|\beta_c-\beta^*\|_\Sigma\right)
\end{aligned}$$

for all $u, v, \beta_C \in \mathbb{R}^d$.

4.2.2 Descent algorithm

The author introduces the basic descent algorithm. Given data, K, and iteration steps, the algorithm outputs a robust estimator β^*. The algorithm will approximate the descent and step-size by some criteria in each iteration, and implement gradient descent. The author proved the algorithm will decrease the risk by a valid step until bounded by $O(r) = O(\sqrt{K/N})$.

The author approximate the descent direction by $Z_k(\beta_c) = \frac{1}{m}\sum_{i\in B_k}(Y_i - \beta_c X_i)\Sigma^{-1/2}X_i$ and proved the efficiency of algorithm using the three properties in previous section.

4.2.3 Experiments

The author showed his algorithm in an experimental aspect with actual code. This section proved the efficiency of the algorithm in practical settings and it's replicable. The author generated predictors X follows t distribution and corresponding response Y :

$$Y = \langle\beta^*, X\rangle + \sigma\xi$$

where noise xi is t distribution and σ is signal to noise ratio. Then the author added ϵ-fraction by adding or multiplying predictors of ϵN of the data. The author first compared his method with Ordinary Least Square, the Huber-loss M-estimator, RANdom SAmple Consensus (RANSAC) and the MOM-estimator from Hsu and Sabato [2016]. The algorithm in paper achieved the best error rate in different dimension d and SNR rate σ.

Then the author showed the trend of error rate under different K, the number of median-of-means groups. The author generates clean data. The result showed the best error rate is achieved when K is nearby the dimension of data. The author interpreted it as "bias-variance trade-off": when $K \ll d$, our algorithm can not seize the complexity of the regression task, and that when $K \gg d$, there are not enough data per block and thus the block are "not informative enough".

5 Main work

In this section, I will state the main contribution of Diakonikolas et al. [2020], I provide the proof using slightly different notation from claim 2.1(D1) in the paper. I prove the simpler case of lemma B.5 focusing on the 2-nd moment covariance and upper bound. I elaborate the details of proofs in Main theorem 2.9(theorem 1.4) for stability of bound covariance sample. The details includes the stability in bounded support case with slightly different notation and the Chernoff bound for sample size in general case. I got the results in different notation but equivalent to the conclusions in the paper. For completeness, I will show the main steps below.

5.1 Stability Characterization

In this section, I will state state the alternate definition of stability with different notation.

Let S be a set such that $\|\mu_S - \mu\| \leq \sigma\delta$, and $\left\|\bar{\Sigma}_S - \sigma^2 I\right\| \leq \sigma^2\delta^2/\epsilon$ for some $0 \leq \epsilon \leq \delta$. Then S is (ϵ, δ')-stable with respect to μ and σ, where $\delta' = 4\delta + 2\sqrt{\epsilon}$.

Proof. With out loss of generality, we assume $\sigma = 1$. Consider $S' \subseteq S : |S'| \geq (1-\epsilon)|S|$.

First we bound second moment.

$$\begin{aligned}\frac{1}{|S'|}\sum_{i\in S'}\left((x_i-\mu)^T v\right)^2-1 \leq& \frac{|S|}{|S'|}\frac{1}{|S|}\sum_{i\in S}\left((x_i-\mu)^T v\right)^2-1\\ \leq& \frac{1}{1-\epsilon}\left(1+\frac{\delta^2}{\epsilon}\right)-1\\ \leq& \frac{1}{\epsilon}(2\delta^2+2\epsilon^2)\leq\frac{(\delta')^2}{\epsilon}\end{aligned}$$

where $\delta'=4\delta+2\sqrt{\epsilon}>2\delta+2\epsilon$. We also have $\|\Sigma_{S'}\|-1\geq-\frac{(\delta')^2}{\epsilon}$ since $\delta'>\sqrt{\epsilon}$ and $\|\Sigma_{S'}\|$ is PSD matrix.

Then we bound mean. Let event E denote x comes from S' when selecting x from S where $S'\subseteq S:|S'|\geq(1-\epsilon)|S|$. We denote this event in emirical sense.

$$\begin{aligned}|\frac{1}{|S'|}\sum_{i\in S'}\left((x_i-\mu)^T v\right)|=&|\frac{|S|}{|S'|}\frac{1}{|S|}\sum_{i\in S}\left((x_i-\mu)^T v*\mathbb{1}(x_i\in S')\right)|\\ \leq&|\frac{1}{1-\epsilon}(\delta-\frac{1}{|S|}\sum_{i\in S}(x_i-\mu)^T v*\mathbb{1}(x_i\notin S'))|\\ \leq&2\delta+2\frac{1}{|S|}|\sum_{i\in S}(x_i-\mu)^T v*\mathbb{1}(x_i\notin S'))|\\ \leq&2\delta+2\frac{1}{|S|}\sqrt{\sum_{i\in S}(v^T(x_i-\mu))^2}\sqrt{\sum_{i\in S}\mathbb{1}(x_i\notin S')}\\ \leq&2\delta+2\sqrt{1+\frac{\delta^2}{\epsilon}}\sqrt{\epsilon}\\ \leq&4\delta+2\sqrt{\epsilon}\\ \leq&\delta'\end{aligned}$$

□

In this claim, I used the ϵ and δ' where ϵ can be the corrupted coefficient in strong contamination model.

5.2 Proof of stability of bounded covariance

In this section, I will follow the steps in proving the main theorem 2.9(theorem 1.4) of stability in bound covariance sample with fewer notations. I will elaborate the omitted steps in the paper and prove the simpler version of lemma B.5. I will invoke lemma 2.3 and lemma 2.6 to prove the main theorem.

Lemma 5.1 (lemma 2.3 in paper)**.** *Let $x_1,\ldots,x_n$ be n i.i.d sample in $\mathbb{R}^d$ from a distribution with mean μ and covariance $\Sigma\preceq I$. Assume $\|x_i-u\|=O(\sqrt{\mathrm{tr}(\Sigma)/\epsilon})$ holds almost surely. Then there exists $c,c'>0$ such that $\epsilon\in(0,c')$ with probability $1-2\exp(-cn\epsilon)$. We have*

$$\min_{w\in\Delta_{n,\epsilon}}\left\|\bar{\Sigma}_w-I\right\|\leq\delta^2/\epsilon$$

where $\delta=O(\sqrt{(\mathrm{tr}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon})$ and

$$\Delta_{n,\epsilon}=\left\{w\in\mathbb{R}^n:0\leq w_i\leq 1/((1-\epsilon)n);\sum_{i=1}^n w_i=1\right\}$$

Lemma 5.2 (lemma 2.6 in paper)**.** *Let $x_1,\ldots,x_n$ be n i.i.d sample in $\mathbb{R}^d$ from a distribution with mean μ and covariance $\Sigma\preceq I$. Let $\epsilon<1/2$ and $u\in\Delta_{n,\epsilon}$. Then there exists $c>0$ such that with probability $1-\exp(-cn\epsilon)$. We have*

$$\min_{w\in\Delta_{n,4\epsilon,u}}\|\mu_w-\mu\|\leq\delta$$

where $\delta=O(\sqrt{\epsilon}+\sqrt{\mathrm{tr}(\Sigma)/n})$ and

$$\Delta_{n,\epsilon,u}=\left\{w:\sum_{i=1}^n w_i=1,w_i\leq u_i/(1-\epsilon)\right\}$$

Then we state the main theorem for stability in bound covariance distribution.

Theorem 5.3. *Let $x_1, \ldots, x_n$ be n i.i.d sample in $\mathbb{R}^d$ from a distribution with mean μ and covariance Σ. Consider S is the sample containing $x_1, \ldots, x_n$ with sample size n. Then, with probability at least $1-\tau$, there exist a subset $S' \subseteq S : |S'| \geq (1-\epsilon)|S|$ and S' is (ϵ', δ')-stable with respect to μ and $\|\Sigma\|$, where*

$$\delta' = O\left(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon'}\right)$$

and $\epsilon' = O(\log(1/\tau)/n + \epsilon)$ smaller than a small constant c.

Combine the previous notation we get the $\delta' = O\left(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{log(1/\tau)}/n\right)$

Then we can use filtering algorithm in Diakonikolas and Kane [2019] to compute the mean estimator $\hat{\mu}$ with polynomial time algorithm. We have error bound in $\|\hat{\mu} - \mu\| = O(\delta') = O\left(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{log(1/\tau)}/n\right) = O\left(\sqrt{(d\log d)/n} + \sqrt{\epsilon} + \sqrt{log(1/\tau)}/n\right)$. This rate is nearly sub-gaussian rate and can be computed efficiently in polynomial time.

Now we prove the theorem.

Proof. Without loss of generality, we assume $\|\Sigma\| = 1$.

Since the failure probability τ is independent of corruption rate ϵ, we choose $\epsilon' = \max(log(1/\tau)/n, \epsilon) \leq log(1/\tau)/n + \epsilon$.

From Markov's inequality, we have

$$P(\|x_i - \mu\| > t) \leq \frac{tr(\Sigma)}{t^2} \leq \epsilon'$$

we have $\|x_i - \mu\| = O(\sqrt{\frac{tr(\Sigma)}{\epsilon'}})$ with probability $1-\epsilon'$.

We first consider simpler case where this bounded covariance distribution, where $\|x_i - \mu\| = O(\sqrt{\frac{tr(\Sigma)}{\epsilon}})$ almost surely.

Now we bound second moment. Consider $u^* \in \Delta_{n,\epsilon'}$ achieves minimum in lemma 2.3, and $w^* \in \Delta_{n,4\epsilon',u^*}$ achieves minimum in lemma 2.6. Then by lemma 2.3, we have with probability $1 - 2\exp(-\Omega(cn\epsilon')) \geq 1-\tau$, we have $\left\|\bar{\Sigma}_{u^*} - I\right\| \leq \delta^2/\epsilon'$, where $\delta = O(\sqrt{(\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon'})$. Then

$$\begin{aligned}
\left\|\bar{\Sigma}_{w*} - I\right\| \leq & \sup_v \sum_{i\in S'} w_i^* \left(v^T(x_i - \mu)\right)^2 - 1 \leq \frac{1}{1-4\epsilon'} \sup_v \sum_{i\in S'} u_i^* \left(v^T(x_i-\mu)\right)^2 - 1 \\
= & \frac{1}{1-4\epsilon'}\left[\sup_v \sum_{i\in S'} u_i^* \left(v^T(x_i-\mu)\right)^2 - 1 + 4\epsilon'\right] \\
\leq & 4\frac{\delta^2}{\epsilon'}
\end{aligned}$$

for $\delta > \sqrt{\epsilon'}$, where $\delta = O\left(\sqrt{(\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon'}\right)$.

Now we bound mean. By lemma 2.6, we have with probability $1-\exp(-\Omega(cn\epsilon')) \geq 1-\tau$, we have $\|\mu_{w^*} - \mu\| \leq \delta'$, where $\delta' = O(\sqrt{(\mathrm{tr}(\Sigma)/n} + \sqrt{\epsilon'}) \leq \delta$, where $\delta = O\left(\sqrt{(\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon'}\right)$.

Then there exist $u^* \in \Delta_{n,\epsilon'}$ and $w^* \in \Delta_{n,4\epsilon',u^*}$ and ϵ' such that $S' \subseteq S : |S'| \geq (1-\epsilon)|S|$ satisfies stability.

Now we consider the general case, where $\|x_i - \mu\| = O\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{\epsilon'}}\right)$ with probability $1-\epsilon'$.

Define

$$E = \left\{X \in \mu, \Sigma : \|X - \mu\| \leq C'\sqrt{\frac{\mathrm{tr}(\Sigma)}{\epsilon'}}\right\}$$

in probabilistic sense. Then we have $P(E^c) \leq \epsilon'$.

Now we prove second moment version of lemma B.5. First we prove the deviation between $\mathbb{E}Z$ and μ can be bounded by $\sqrt{\epsilon'}$. For any v with norm 1:

$$\begin{aligned}|v^T(\mathbb{E}Z-\mu)| =& |\frac{1}{p(E)}v^T\mathbb{E}\left[(X-\mu)\mathbb{1}(X\in E)\right]| = |\frac{1}{p(E)}v^T\mathbb{E}\left[(X-\mu)\mathbb{1}(X\notin E)\right]| \\ \leq& \frac{1}{1-\epsilon'}\sqrt{\mathbb{E}\left[v^T(X-\mu)\right]^2}\sqrt{P(X\in E^c)} \leq 2\sqrt{\|\Sigma\|}\sqrt{\epsilon'} \\ \leq& 2\sqrt{\epsilon'}\end{aligned}$$

Then we have $\|\mathbb{E}Z-\mu\| = O(\sqrt{\epsilon'})$.

Now we prove Z has bounded covariance:

$$\mathbb{E}\left[v^T(Z-\mu)\right]^2 = \frac{1}{P(E)}\mathbb{E}\left[v^T(X-\mu)\right]^2\mathbb{1}(X\in E) \leq 2\mathbb{E}\left[v^T(X-\mu)\right]^2 \leq 2$$

We have random variable Z has bound support with mean $\mathbb{E}Z$ and bounded covariance $2I$. We apply simpler case on Z. We have:

- Let $x_1,\ldots,x_n$ be n i.i.d sample in $\mathbb{R}^d$ from a distribution Z with mean $\mathbb{E}Z$ and covariance Σ. Consider S is the sample containing $x_1,\ldots,x_n$ with sample size n. Then, with probability at least $1-\tau$, there exist a subset $S'\subseteq S : |S'| \geq (1-\epsilon)|S|$ and S' is (ϵ',δ')-stable with respect to μ and $\|\Sigma\|$, where

$$\delta' = O\left(\sqrt{(\mathrm{r}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon'}\right)$$

Now we prove for $i.i.d$ sample $x_1,\ldots,x_n$, with probability $1-\exp(-n\epsilon')$, let $\tilde{S} = \{x_i : x_i \in E\}$, then $|\tilde{S}| \geq (1-\epsilon'/2)n$. We elaborate the Chernoff-bound in paper to prove this conclusion. Consider $Y_i = \mathbb{1}(\|X-\mu\| \leq C\sqrt{\frac{\mathrm{tr}(\Sigma)}{\epsilon'}})$, then Y_i is Bernoulli variable with $P(Y_i=1) = 1-\epsilon'$. Then $|\tilde{S}| \geq (1-\epsilon'/2)n \iff \sum_i Y_i > (1-\epsilon'/2)n$. Consider $Z_i = 1-Y_i$, we have $\mathbb{E}\sum_i Z_i = \epsilon' n$, apply Chernoff-lower bound on Z_i, we have:

$$P(\sum_i Z_i \leq \frac{\epsilon' n}{2}) \leq e^{-\epsilon' n}\left(\frac{e\epsilon' n}{\epsilon' n/2}\right)^{\epsilon' n/2} \leq exp(-c\epsilon' n)$$

Then $P(\sum_i Y_i > (1-\epsilon'/2)n) \leq exp(-c\epsilon' n) \leq \tau$ since $\epsilon' \leq log(1/\tau)/n + \epsilon$.

To sum up, we have for $x_1,\ldots,x_n$ be n i.i.d sample in $\mathbb{R}^d$ from a distribution with mean μ and covariance Σ, $S = \{x_1,\ldots,x_n\}$.

- with probability $1-\tau$, $|\tilde{S}| \geq (1-\epsilon'/2)n$ such that $x_1,\ldots,x_n \sim Z$.
- with probability $1-\tau$,$|S'| \geq |\tilde{S}| \geq (1-2\epsilon')n$ such that S' is (ϵ',δ) stable with $\mathbb{E}Z$ and Σ, where $\delta = O\left(\sqrt{(\mathrm{r}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon'}\right)$.

Combine above with $\|\mathbb{E}Z-\mu\| = O(\sqrt{\epsilon'})$, we have with probability $1-2\tau$, $\exists |S'| \geq (1-2\epsilon')n$ such that S' is (ϵ',δ) stable with μ and Σ, where

$$\delta = O\left(\sqrt{(\mathrm{r}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon'}+\sqrt{\epsilon'}\right) = O\left(\sqrt{(\mathrm{r}(\Sigma)\log\mathrm{r}(\Sigma))/n}+\sqrt{\epsilon}+\sqrt{log(1/\tau)/n}\right)$$

□

Now we prove the stability condition with bound sample covariance.

6 Conclusion

In this project I state the main contribution of Diakonikolas et al. [2020]. I proved some claims with different notations. I proved one of the theorems with different notations and elaborated the omitted steps in the proof.

I also state the main contribution of Depersin [2020], this paper proposed the practical algorithm and efficient code. I think the experiment is not general enough to cover the worst case. The author generated data in t-distribution, which is unimodal and symmetric. The shape of t distribution will approximate gaussian as sample size grows larger. The author can use other general bounded covariance distribution to feed the algorithm with bad-behaved data. The way author added outliers is not representative. The author increased the magnitude of coordinates of predictors X or setting response Y to 0 or some extremely large number. These outliers are far from the good sample and can be observed easily. The author can add noise to different directions while maintaining the same magnitude, and provide more convincing results.

References

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

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

Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019.

Samuel B Hopkins et al. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2): 1193–1213, 2020.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under huber's contamination model. *Annals of Statistics*, 46(5):1932–1960, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

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

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391*, 2019.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *arXiv preprint arXiv:2007.15618*, 2020.

Jules Depersin. A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*, 2020.

Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 601–609, 2020.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.