Robust Mean Estimation 2 and Linear Regression

Instructor: Prasad Raghanvandra Scribe: Jeff Xu

Abstract

In this lecture, we continued our survey in Robust Mean Estimator and look at the generalized problem in high dimensions. Moreover, we started our discussion on Robust Linear Regression.

1 Robust Mean Estimator

1.1 Problem Set Up

Let's start with recapping from last lecture the problem and the formulation of our SoS SDP. We are given a series of observations, $x_1, x_2, ..., x_n \in R^d$, with each observation x_i being sampled from a distribution Θ w.p. $1-\epsilon$; otherwise, it is an arbitrary vector. For simplicity, we assume that $\Theta = \mathcal{N}(\hat{\mu}, Id)$ while our algorithm only requires bounded covariances of the distribution. Our goal is to recover mean $\hat{\mu}$ of Θ .

1.2 SDP formulation

With the intuition that we can recover the mena if we know the source of each sample, whether they come from Θ or from arbitrary distribution, we want to give each observation a label w_i indicating whether they come from Θ or not.

Proposition 1 The following is a SDP relaxation for the above problem.

$$\begin{split} &w_i^2 = w_i \\ &\sum_i w_i = (1-\epsilon)n \\ &\mu(w) = \frac{1}{(1-\epsilon)n} \sum_i w_i x_i \\ &Cov = \frac{1}{(1-\epsilon)n} \sum_i w_i (x_i - \mu(w))^2 \leq 2I_d \end{split}$$

Remark 2 Readers should note that the only difference between this SDP and the one before is the covariance matrix.

Observation 3 Let \hat{w} be the indicator for the sample that $\hat{w}_i = 1$ if $x \sim \Theta$ and 0 otherwise, it is not hard to observe that it is a feasible solution to the above SDP.

Remark 4 With the SDP formulation, our algorithm now simply needs to solve the above SDP, or equivalently, find a pseudomoment $\tilde{\mathbb{E}}$ that satisfies the above constraints and output $\tilde{\mathbb{E}}[\mu(w)]$.

Theorem 5 If $w \in \{0,1\}^n$ is a solution to the above SDP, then $\mu(w)$ is not too far away from the true mean $\hat{\mu}$, mathematically, $||\mu(w) - \hat{\mu}||_{2\to 2}^2 \leq O(\epsilon)$.

Our hard work in the previous lecture is mostly proving the above theorem, and similarly, we will prove a version of the theorem in high dimension and this suffices to show that our SDP gives a (reasonably) good estimator for robust mean in high dimension.

With a little abuse of notation, we call our samples in R^D to be Z_i , and we now fix a direction vector $u \in R^n$, and let $X_i := \langle u, Z_i \rangle$ be the projection of Z_i along u. Similarly, we define $\mu_u := \langle \mu, u \rangle$ to be the projection of our estimated μ along u.

The direction vector u now reduces our problem to the one-dimension case we have in the last lecture, and we can simply plug in our proof from the last lecture to show the following theorem 5 for any u, if we plug in $\mu(w) = \mu_u$ and $\hat{\mu}_u = \langle \hat{\mu}, u \rangle$. When it is clear from the context that we are in a single-dimension case, we drop the subscript u.

Claim 6 If theorem 5 holds for any direction vector $u \in \mathbb{R}^n$, we have $||\mu - \tilde{\mu}|| \leq O(\epsilon)$.

Proof: From the covariance constraint of SDP in one-dimension, $\forall u \in \mathbb{R}^n$, we have,

$$\frac{1}{(1-\epsilon)n}\sum w_i(\langle u,Z_i\rangle - \langle u,\mu\rangle)^2 \leq 2 \Leftrightarrow \frac{1}{(1-\epsilon)n}\sum w_i(\langle u,Z_i-\mu\rangle^2) \leq 2,$$

Since this holds for all $u \in \mathbb{R}^n$, we can write as an SoS proof,

$$\begin{split} &\Leftrightarrow \frac{1}{(1-\epsilon)n}u^T(\sum w_i(Z_i-\mu)(Z_i-\mu)^T)\mu) \leq 2, \quad \ \forall u \in R^n \\ &\Leftrightarrow \frac{1}{(1-\epsilon)n}\sum w_i(Z_i-\mu)(Z_i-\mu)^T \preceq 2I_d \end{split}$$

As a sanity check, we can observe that for sufficiently many true samples, $\sum w_i(Z_i-\mu)(Z_i-\mu)^T \approx I_d + \epsilon$ where $\epsilon_{i,j} \leq \frac{1}{d^4}$.

Remark 7 Alternatively, we can turn the last observation into a constraint.

To extract the solution, we can simply output $\tilde{\mathbb{E}}[\mu] \in \mathbb{R}^d$, and by Cauchy-Schwarz, we have

$$\left|\left|\tilde{\mathbb{E}}[\mu] - \hat{\mu}\right|\right|^2 \leq \tilde{\mathbb{E}}[||\mu - \hat{\mu}||^2] \leq O(\epsilon)$$

2 Robust Linear Regression

2.1 Problem Set Up

Let $\hat{l}: R^d \to \mathbb{R}$ be the unknown function and we assume $||l||_2 \le 1$ for simplicity. We are given a series of samples (x_i, y_i) and we know that $y_i = \hat{l}(x) + \gamma$ w.p. $1 - \epsilon$; otherwise, y_i is arbitrary. For simplicity, we also assume x_i to be $\{0,1\}^n$ and γ is bounded noise, and our goal is to learn \hat{l} , i.e., we want to minimize $\frac{1}{n} \sum (y_i - \langle x_i, l \rangle)^2$ over all possible functions l.

2.2 SDP Formulation

This is similar to our Robust Mean Estimator, that we use w_i to indicate whether the samples comes from the true distribution or function.

$$\begin{split} \min \frac{1}{(1-\epsilon)n} \sum w_i (y_i - \langle l, x_i \rangle)^2 \\ w_i^2 &= w_i \\ \sum w_i &= (1-\epsilon)n \\ ||l||_2^2 &= \sum l_i^2 \leq 1 \end{split}$$

Now, our algorithm simply needs to solve for the above SDP and output $l(x) = \langle \tilde{\mathbb{E}}[l], x \rangle$. It remains to show that it gives a good estimation.

Definition 8 For any function l, we define the error on $(1-\epsilon)n$ true samples to be

$$err(l) = \frac{1}{(1-\epsilon)n} \sum \hat{w_i} (y_i - \langle l, x_i \rangle)^2$$

Theorem 9

$$err(\tilde{\mathbb{E}}[l]) \leq err(\hat{l}) + O(\sqrt{\epsilon})$$

Proof:

$$LHS = \frac{1}{(1-\epsilon)n} \sum \hat{w_i} (y_i - \langle \tilde{\mathbb{E}}[l], x_i \rangle)^2 \tag{1}$$

$$\leq \frac{1}{(1-\epsilon)n} \tilde{\mathbb{E}}[\sum \hat{w_i}(y_i - \langle l, x_i \rangle)^2 \tag{2}$$

$$=\frac{1}{(1-\epsilon)n}\left[\tilde{\mathbb{E}}\left[\sum \hat{w_i}w_i(y_i-\langle l,x_i\rangle)^2\right]+\tilde{\mathbb{E}}\left[\sum \hat{w_i}(1-w_i)(y_i-\langle l,x_i\rangle)^2\right]\right] \tag{3}$$

where (2) follows from Cauchy-Schwarz, and we separate the sum over all true samples into two parts in (3).

Observation 10

$$\widetilde{\mathbb{E}}\left[\sum \hat{w_i}w_i(y_i-\langle l,x_i\rangle)^2\right] \leq err(\hat{l})$$

since it is the objective value of the SDP, which is a relaxation of the problem.

To bound the second part of (3), we apply Cauchy-Schwarz again, and we have

$$\begin{split} \frac{1}{(1-\epsilon)n} \tilde{\mathbb{E}} \left[\sum \hat{w_i} (1-w_i) (y_i - \langle l, x_i \rangle)^2 \right] &\leq \frac{1}{(1-\epsilon)n} \left(\tilde{\mathbb{E}} [\sum (1-w_i)^2] \right)^{1/2} \left(\tilde{\mathbb{E}} [\sum \hat{w_i}^2 (y_i - , x_i \rangle)^4] \right)^{1/2} \\ &\leq \frac{1}{(1-\epsilon)n} (\epsilon n)^{1/2} (\tilde{\mathbb{E}} [\sum \hat{w_i} (y_i - \langle l, x_i \rangle)^4]^{1/2}) \\ &\leq \frac{1}{(1-\epsilon)n} (\epsilon n)^{1/2} (\tilde{\mathbb{E}} [\sum \hat{w_i} (\gamma_i + \langle \hat{l} - l, x_i \rangle)^{1/2}) \end{split}$$

By our last SDP constraint, we can conclude that $\langle \hat{l} - l, x_i \rangle \leq O(1)$; along with assumption that γ is some bounded noise, we have the desired conclusion.

Observation 11 This is a degree-4 SoS.

CS 294 Sum of Squares

Fall 2018

Lecture 9: Gaussian Mixture Models

Lecturer: Prasad Raghavendra Scribe: Morris Yau

9.1 Background

Today we discuss how to cluster a mixture of gaussians via sum of squares. Let $\mu_1, ..., \mu_k \in \mathbb{R}^d$ be k means. The input consists of samples drawn i.i.d from the mixture

$$X_1, ..., X_N \sim \frac{1}{k} \sum_{i=1}^n N(\mu_i, I_d)$$
 (9.1)

Our task is to assign each point X_i to their respective cluster between 1 and k. The goal will be to recover a clustering that is correct on $1 - \frac{1}{poly(k)}$ fraction of the points. A similar question is to recover the means $\mu_1, ..., \mu_k$ which we can be obtain from the good clustering with some robust postprocessing procedure.

Intuitively this problem would be impossible if the means were allowed to be arbitrarily close. Indeed, such a condition is necessary. A result by [Kalai-Moitra-Valiant] states that there exists two mixtures of k-gaussians on \mathbb{R} such that the statistical distance between them is $O(2^{-k})$ and exp(k) samples and runtime are required to distinguish the two.

Thus for our purposes we assume that for all $i, j \in [k]$ we have separation $||\mu_i - \mu_j||_2 \ge \Delta$. For simplicity we will also assume k = d the number of clusters is the same as the dimension of the space. The precise separation required for information theoretic recovery is as follows.

Theorem 9.1. Recovery of means $\mu_1, ..., \mu_k$ is information theoretically feasible if and only if $\Delta >> \sqrt{\log k}$

The SOS algorithm can perform recover up to the information theoretic threshold

Theorem 9.2. SOS SDP can be used to recover means $\mu_1, ..., \mu_k$ as long as $\Delta >> \sqrt{\log k}$

9.2 Algorithm

We now write down the SOS SDP for clustering gaussian mixtures in dimension 1. More specifically, we will solve the following constraints for identifying a single cluster. For variables $w_1, ..., w_n$ indicators the constraints are as follows.

$$\forall i \in [b] \ w_i^2 = w_i$$

$$\sum_{i=1}^n w_i = \frac{n}{k}$$

$$\mu(w) = \frac{k}{n} \sum_{i=1}^n w_i X_i$$

$$\frac{k}{n} \sum_{i=1}^n w_i (X_i - \mu(w))^t \le 1.1 t^{t/2}$$
(9.2)

The first constraint is the standard indicator for cluster membership. The second is the constraint for cluster size. The third equation is not a constraint in the SDP, but a definition for the empirical mean. The final constraint corresponds to the empirical moment boundedness of a single cluster. We now work towards formally stating our theorem guarantee.

Definition 9.3. For a cluster S_i for $i \in [k]$, let $w(S_i) = \frac{k}{n} \sum_{i=1}^n w_i$

We observe that $\sum_{i \in [k]} w(S_i) = \frac{k}{n} \sum_{i=1}^n w_i = 1$ and conclude that

$$\sum_{i \in [k]} w(S_i)^2 \le (\max_{i \in [k]} w(S_i)) \sum_{i \in [k]} w(S_i) \le \max_{i \in [k]} w(S_i)$$
(9.3)

We are now ready to state our main theorem

Theorem 9.4. If $w_1, ..., w_n$ satisfy the SDP then the corresponding $w(S_i)$ satisfy

$$\sum_{i=1}^{k} w(S_i)^2 \ge 1 - \frac{2^{3t}}{\Delta^t} k t^{t/2}$$

Which implies

$$\max_{i \in [k]} w(S_i) \ge 1 - \frac{2^{3t}}{\Delta^t} k t^{t/2}$$

Proof.

$$\sum_{i \in [k]} w(S_i)^2 = \left(\sum_{i \in [k]} w(S_i)\right)^2 - \sum_{i,j \in [k]} w(S_i)w(S_j)$$
(9.4)

$$=1 - \sum_{i,j \in [k] i \neq j} w(S_i) w(S_j)$$
(9.5)

Now it suffices to upper bound

$$\sum_{i,j \in [k] i \neq j} w(S_i) w(S_j) \le \sum_{i,j \in [k] i \neq j} w(S_i) w(S_j) \frac{|\mu_{S_i} - \mu_{S_j}|^t}{\Delta^t}$$
(9.6)

$$2^{t-1} \sum_{i,j \in [k] i \neq j} w(S_i) w(S_j) (|\mu_S - \mu|^t + |\mu_{S_j} - \mu|^t)$$
(9.7)

Nos using SOS version of triangle inequality on $|\mu_{S_i} - \mu_{S_j}|^t \le 2^{t-1}(|\mu_{S_i} - \mu|^t + |\mu - \mu_{S_j}|^t)$ we obtain

$$=2^{t} \sum_{i \in [k]} w(S_{i}) |\mu_{S} - \mu|^{t}$$
(9.8)

Where we also used the fact that $\sum_{i \in [k]} w(S_i) = 1$. This is equal to

$$2^{t} \frac{k}{n} \sum_{i=1}^{n} w_{i} |\mu_{S_{i}} - \mu|^{t}$$
(9.9)

Where μ_{S_i} is the mean associated with the i'th sample X_i . Using SOS triangle inequality we obtain

$$=2^{t}\frac{k}{n}2^{t-1}\sum_{i=1}^{n}w_{i}(|\mu_{S_{i}}-X_{i}|^{t}+|X_{i}-\mu|^{t})$$
(9.10)

$$=2^{2t-1}\left[\frac{k}{n}\left(\sum_{i}i\in[n]w_{i}|\mu_{S_{i}}-X_{i}|^{t}\right)\right]+2^{2t-1}\left[\frac{k}{n}\sum_{i\in[n]}w_{i}|X_{i}-\mu|^{t}\right]$$
(9.11)

We will upper bound both terms in the above sum separately. Consider the first term. Using the fact that $w_i \leq 1$

$$2^{2t-1} \left[\frac{k}{n} \left(\sum_{i \in [n]} w_i |\mu_{S_i} - X_i|^t \right) \right] \le 2^{2t-1} \frac{k}{n} \sum_{i \in [n]} |\mu_{S_i} - X_i|^t \le 2^{2t-1} k t^{t/2}$$

$$(9.12)$$

Where the last equality follows from the fact that X_i are drawn from a gaussian. As for the second term we have by the moment boundedness of the SDP constraint

$$2^{2t-1} \left[\frac{k}{n} \sum_{i \in [n]} w_i |X_i - \mu|^t \right] \le 2^{2t-1} t^{t/2}$$
(9.13)

Putting the upper bounds on both terms together we obtain

$$\sum_{i \in [k]} w(S_i)^2 \ge 1 - \frac{2^{3t} k t^{t/2}}{\Delta^t}$$
(9.14)

9.2.1 Higher Dimensions

Given samples $X_1, ..., X_n \in \mathbb{R}^d$ cluster the points. The proof sketch is that to set

$$1 \le \max_{||u|| \le 1} \frac{\langle u, \mu_{S_i} - \mu_{S_j} \rangle^t}{\Delta^t} \tag{9.15}$$

instead of $1 \leq (\mu_{S_i} - \mu_{S_j})^t / \Delta^t$. The rest of the proof remains the same except for replacing X_i with $\langle i, Z_i \rangle$ We will also need the high dimensional version of moment boundedness. For all $u \in \mathbb{R}^d$

$$\frac{k}{n} \sum_{i \in [n]} w_i(\langle u, X_i \rangle - \frac{k}{n} \sum_{i \in [n]} w_i \langle u, X_i \rangle)^t \le t^{t/2}$$
(9.16)

In the next class we will discuss how to add this constraint for every $u \in \mathbb{R}^d$