Fast Mean Estimation with Sub-Gaussian Rates

Cherapanamjeri, Flammarion, Bartlett

1 Introduction

1.1 Goal

To obtain high probability mean estimates when only the existence of the 2^{nd} moment is known. This is also called the *heavy tailed* setting, where higher order moments from the sampling distribution need not exist.

1.2 Existing results

Consider the estimator to be the sample mean $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $\{X_i\}_{i=1}^n$ are sampled from a distribution P with only finite 2^{nd} moment and mean $\theta^\star$. Markov's inequality gives:

$$\Pr(\|\widehat{\theta} - \theta^\star\|_2 > t) \leq \frac{\mathbb{E}[\|\widehat{\theta} - \theta^\star\|_2^2]}{t^2}$$

Note that $\widehat{\theta} - \theta^\star = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta^\star)$ and hence:

$$\|\widehat{\theta} - \theta^\star\|_2^2 = \frac{1}{n^2}\sum_{i=1}^{n}\|X_i - \theta^\star\|_2^2 + \frac{1}{n}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}(X_i - \theta^\star)^T(X_j - \theta^\star) \Rightarrow \mathbb{E}[\|\widehat{\theta} - \theta^\star\|_2^2] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\|X_i - \theta^\star\|_2^2\right]$$

and since $\mathbb{E}\left[\|X_i - \theta^\star\|_2^2\right] = \mathbb{E}\left[\mathrm{trace}(X_i - \theta^\star)(X_i - \theta^\star)^T\right] = \Sigma$, we get:

$$\mathbb{E}[\|\widehat{\theta} - \theta^\star\|_2^2] = \frac{\mathrm{trace}(\Sigma)}{n}$$

therefore leading to:

$$\Pr\left(\|\widehat{\theta} - \theta^\star\|_2 > \sqrt{\frac{\mathrm{trace}(\Sigma)}{n\delta}}\right) \leq \delta$$

which corresponds to: with probability at least $1 - \delta$:

$$\|\widehat{\theta} - \theta^\star\|_2 \leq \sqrt{\frac{\mathrm{trace}(\Sigma)}{n\delta}}$$

In contrast, when P is Gaussian, we get:

$$\Pr\left(\|\widehat{\theta} - \theta^\star\|_2 > O\left(\sqrt{\frac{\mathrm{trace}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}\right)\right) \leq \delta$$

We will denote $\sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}$ as $\text{OPT}_{n,\delta,\Sigma}$ as a shorthand.

To show this, consider $Z_i = X_i - \theta^\star$ for all $i \in [n]$. Then $\|\widehat{\theta} - \theta^\star\|_2 = \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2$, where Z_is are zero mean Gaussian RVs with covariance Σ. Note that $Z_i = \Sigma^{1/2} Y_i$ for all $i \in [n]$ where Y_is are standard multivariate Gaussian RVs. Now, we have that:

$$|\|Z_i\| - \|Z_i'\|_2| \leq \|Z_i - Z_i'\|_2 \leq \|\Sigma^{1/2}(Y_i - Y_i')\|_2 \leq \|\Sigma^{1/2}\|_2 \|Y_i - Y_i'\|_2$$

which shows that $\|Z_i\|$ is a $\|\Sigma^{1/2}\|_2$-Lipschitz function of Y_i. By a Lipschitz concentration lemma due to Tsirelson, Ibragimov and Sudakov, we have:

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 - \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 \right] > t \right) \leq \exp\left(-\frac{nt^2}{2\|\Sigma\|_2} \right)$$

leading to:

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 > \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 \right] + t \right) \leq \exp\left(-\frac{nt^2}{2\|\Sigma\|_2} \right)$$

and with probability at least $1 - \delta$:

$$\begin{aligned} \|\widehat{\theta} - \theta^\star\|_2 \leq \mathbb{E}[\|\widehat{\theta} - \theta^\star\|_2] + \sqrt{\frac{2\|\Sigma\|_2 \log(1/\delta)}{n}} &\leq \sqrt{\mathbb{E}[\|\widehat{\theta} - \theta^\star\|_2^2]} + \sqrt{\frac{2\|\Sigma\|_2 \log(1/\delta)}{n}} \\ &\leq \sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{2\|\Sigma\|_2 \log(1/\delta)}{n}} \end{aligned}$$

Lugosi and Mendelson showed that with only bounded 2^{nd} moment, this rate can be achieved, but the estimator proposed is intractable.

2 Main Result

Theorem 1. *Let $\{X_i\}_{i=1}^n$ be a set of n i.i.d. random vectors i.e. $X_i \in \mathbb{R}^p$, sampled from a distribution with mean $\theta^\star$ and covariance Σ. Then* $\mathsf{Descent} - \mathsf{Mean} - \mathsf{Estimate}$ *with stepsize $\gamma = \frac{1}{20}$ and number of iterations $T = 1000 \frac{\log(\|\theta^\star\|_2)}{\epsilon}$ returns a mean estimate $\widehat{\theta}_{n,\delta}$ that satisfies with probability at least $1 - \delta$:*

$$\|\widehat{\theta}_{n,\delta} - \theta^\star\|_2 \leq \max\left(\epsilon, 480000 \cdot \text{OPT}_{n,\delta,\Sigma}\right)$$

in $O\left(\log(1/\delta)^{3.5} + \log(1/\delta)^2 d + nd\right)$ time.

$\mathsf{Descent} - \mathsf{Mean} - \mathsf{Estimate}$ is Algorithm 1 in the main text.

Remark. Note that this result achieves the sub-Gaussian rate upto constants. However there is a direct dependence of $\log(\|\theta^\star\|_2)$ in the number of steps. The author comment that this can be brought down to $\log(d)$ with an appropriate initialization.

2.1 Idea behind the algorithm

Lugosi and Mendelson show that when performing bucketing, most of the bucket means are close to the true mean in any direction. That is:

$$|\{i : |\langle v, Z_i - \theta^\star \rangle| \leq r\}| \geq 0.9k \tag{1}$$

where k is the number of buckets, and $\{Z_i\}_{i=1}^k$ are the bucket means, for an appropriately chosen r.

This provides a method for outlier detection. If a point x is far away from the mean, then x could be far away from most of the bucket means themselves. Define the optimization problem below:

$$\begin{aligned} &\max \sum_{i=1}^{k} b_i \\ &b_i \in \{0,1\}, v \in \mathcal{S}^{d-1} \\ &b_i \langle v, Z_i - x \rangle \geq b_i r, \;\; i \in [k] \qquad (*) \end{aligned}$$

Let's parse this optimization problem. Given an x, we would like to find out the maximum number of buckets number and the direction along which these bucket means are far away from x.

How is this related to outlier-detection? Let the solution for the optimization problem be $v^\star, b^\star$.

- If $\|b^\star\|_1$ is small, then this indicates that even along the direction $v^\star$, which is the direction of maximum deviation, the number of "activations" i.e., $|\{i : b_i = 1\}|$ in $(*)$ is small. This means that the supremum over all directions, which is the distance, is also going to be small.
- On the otherhand, if $\|b^\star\|_1$ is large, $(*)$ is activated for many bucket means along the direction of maximum deviation, which is the distance.

As seen above, the direction $v^\star$ is vital in giving the direction of maximum deviation. An important insight of this paper is that the direction $v^\star$ is well-aligned with the vector joining x and μ. That is:

$$\left\langle v^\star, \frac{x-\mu}{\|x-\mu\|_2} \right\rangle \approx 1$$

for specific choices of r. Therefore, moving in small steps along $v^\star$ takes us closer to the mean.

Remark. I believe that the choice of r is pertinent to the property of the point we are trying to estimate. In other words, this seems to me as a result can be applied for any point $x^\star$ for an appropriately chosen r.

2.2 The (sub-optimal) algorithm

Bucketing: Given n datapoints, we bucket them into k groups of size $\lfloor \frac{n}{k} \rfloor$ each. $\{Z_i\}_{i=1}^k$ are the means of these buckets.

Iterations: At each iteration of the algorithm, we solve the optimization problem described above. Moreover, our update is: $x_{t+1} = x_t + \gamma d_t g_t$. d_t is the maximum r for which we satisfy Eqn 1. g_t is the direction corresponding to this choice of r.

For this algorithm, we have the following result:

Theorem 2. *Let $\{X_i\}_{i=1}^n$ be a set of n i.i.d. random vectors i.e. $X_i \in \mathbb{R}^p$, sampled from a distribution with mean $\theta^\star$ and covariance Σ. Then the* $\mathsf{sub-optimal\ algorithm}$ *with stepsize $\gamma = \frac{1}{4}$ and number of iterations $T = 50\frac{\log(\|\theta^\star\|_2)}{\epsilon}$ returns a mean estimate $\widehat{\theta}_{n,\delta}$ that satisfies with probability at least $1-\delta$:*

$$\|\widehat{\theta}_{n,\delta} - \theta^\star\|_2 \leq \max\left(\epsilon, 108000 \cdot \mathrm{OPT}_{n,\delta,\Sigma}\right)$$

One key assumption to be made is as follows:

Assumption 1. *For the bucket means $\{Z_i\}_{i=1}^k$, we have:*

$$\forall v \in \mathcal{S}^{d-1}, \;\; \left| \{i : \langle Z_i - \mu, v \rangle \geq 300 \left(\sqrt{\frac{\mathrm{trace}(\Sigma)}{n}} + \sqrt{\frac{k\|\Sigma\|_2}{n}} \right) \right| \leq 0.05k$$

The above assumption is telling that the proportion of buckets which are "far away" in all directions is small. Now we shift our focus on proving the correctness of the distance and gradient estimates d_t and g_t respectively.

Lemma 1. *Under Assumption 1, we have that:*

$$|d_t - \|x_t - \theta^\star\|_2| \leq 300\left(\sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{k\|\Sigma\|_2}{n}}\right)$$

Proof. For convenience, let $r_\delta = 300\left(\sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{k\|\Sigma\|_2}{n}}\right)$. The equation can be hashed out into two pieces:

$$-r_\delta \leq d_t - \|x_t - \theta^\star\|_2 \leq r_\delta$$

Lower Bound: If $\|x_t - \theta^\star\|_2 \leq r_\delta$, then $d_t - \|x_t - \theta^\star\| \geq -r_\delta$. If $\|x_t - \theta^\star\|_2 > r_\delta$. Note that d_t is that value of parameter r such that:

$$\forall v \in \mathcal{S}^{d-1},\ \ |\{i : \langle Z_i - x, v\rangle \geq r\}| \leq 0.1k$$

From Assumption 1, we know that:

$$\forall v \in \mathcal{S}^{d-1},\ \ |\{i : |\langle Z_i - \theta^\star, v\rangle| \geq r\}| \leq 0.05k$$

Furthermore, for any $v \in \mathcal{S}^{d-1}$:

$$\langle v, Z_i - x_t\rangle = \langle v, Z_i - \theta^\star\rangle + \langle v, \theta^\star - x_t\rangle \overset{(i)}{\geq} \|x_t - \theta^\star\|_2 - r_\delta \geq 0$$

where in Step (i) we have chosen $v = \frac{\theta^\star - x_t}{\|\theta^\star - x_t\|_2}$ and the assumption. Therefore, by definition, the lower bound holds.

Upper Bound: The upper bound can be proven by contradiction. If there existed $r > \|x_t - \theta^\star\|_2 + r_\delta$ such that the solutions of the optimization problem $b^\star, v^\star$, with $\|b^\star\|_1 \geq 0.9k$, then:

$$\langle Z_i - \theta^\star, v^\star\rangle = \langle Z_i - x_t, v^\star\rangle + \langle x_t - \theta^\star, v^\star\rangle \geq r - \|x_t - \theta^\star\|_2 \geq r_\delta$$

for $0.9k$ indices $i \in [k]$. This is contradiction to the assumption made, and hence the upper bound holds. □

Lemma 2. *Under Assumption 1 and the iterate satisfying*

$$\|x_t - \theta^\star\| \geq 1200\left(\sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{k\|\Sigma\|_2}{n}}\right)$$

we have that:

$$\langle g_t, \Delta\rangle \geq \frac{1}{2}$$

where $\Delta = \frac{\theta^\star - x_t}{\|\theta^\star - x_t\|_2}$.

Proof. For convenience, let $r_\delta = 300\left(\sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{k\|\Sigma\|_2}{n}}\right)$. As seen earlier, d_t is such that:

$$|\{i : \langle Z_i - x_t, g_t\rangle > d_t\}| \geq 0.9k$$

From Assumption 1, we have that:

$$|\{i : \langle Z_i - x_t, g_t\rangle > r_\delta\}| \leq 0.95k$$

Hence there must exist points in common to these groups, and the intersection is at least $0.5k$ in size. Let Z_j belong to the intersection. Then:

$$\|\theta^\star - x_t\| - r_\delta \leq d_t \leq \langle Z_i - x_t, g_t \rangle = \langle Z_i - \theta^\star, g_t \rangle + \langle \theta^\star - x_t, g_t \rangle \leq r_\delta + \|\mu - x_t\|_2 \langle \Delta, g_t \rangle$$

Some rearrangement followed by using the condition of the lower bound on the distance between the estimate and the mean finishes the proof. □

Remark. Note that in all its generality, we only use the properties of the optimization problem and the property of the mean in Assumption 1. This means that we can loosely search for points that satisfy an assumed property, by merely tweaking the relevant portions of the optimization. Also note that we have a k term sitting in the form for r_δ. Choosing this to be $O(\log(1/\delta)$ would give the sub-Gaussian rate.