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 $\widehat{\theta} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ 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 $\widehat{\theta} - \theta^\star = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \theta^\star)$ 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 2.1. *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* $\mathsf{Descent-Mean-Estimate}$ *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 $1 - \delta$:*

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

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