High-dimensional Location Estimation via Norm Concentration for Subgamma Vectors

Shivam Gupta [1] Jasper C.H. Lee [2] Eric Price [1]

[1]The University of Texas at Austin [2]University of Wisconsin–Madison

Asymptotic Mean Estimation

- Given n samples from a distribution on $\mathbb{R}^d$, want to estimate mean μ.

	Estimator	Converges to	Notes
Unknown Distribution	Empirical Mean	$\mathcal{N}(\mu, \frac{\Sigma}{n})$	Central Limit Theorem
Known Distribution	MLE	$\mathcal{N}(\mu, \frac{\mathcal{I}^{-1}}{n})$	$\mathcal{I}$ is the *Fisher Information*

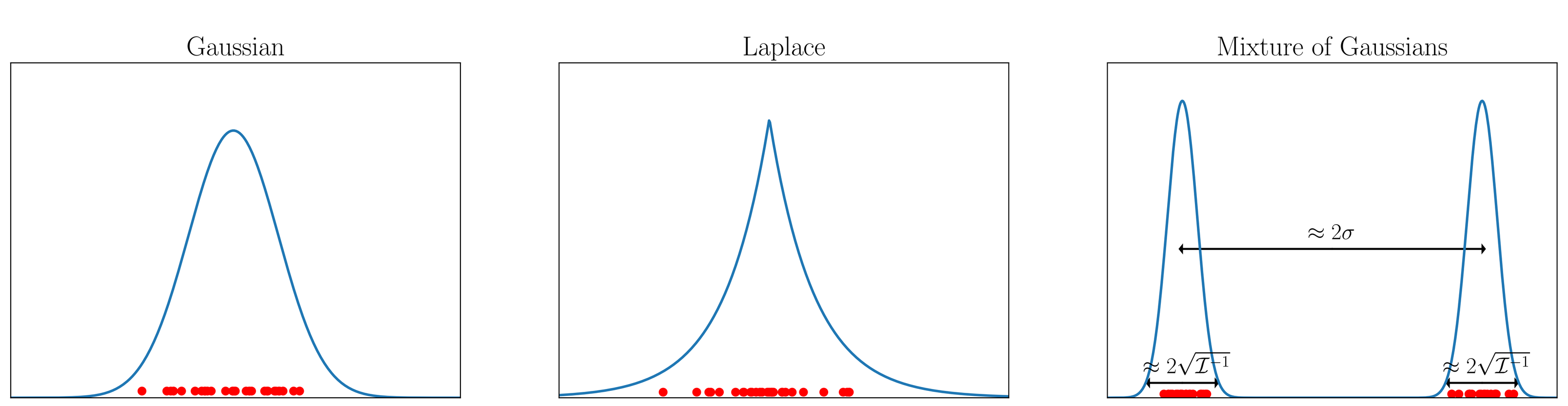
Table 1. Classical Asymptotic Results

- In ***finite-sample*** setting, when $d = 1$ and distribution is **unknown**, [Catoni '12], [Lee, Valiant '21] show estimator $\widehat{\mu}$ such that with probability $1 - \delta$,

$$|\widehat{\mu} - \mu| \le \sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}(1 + o(1))$$

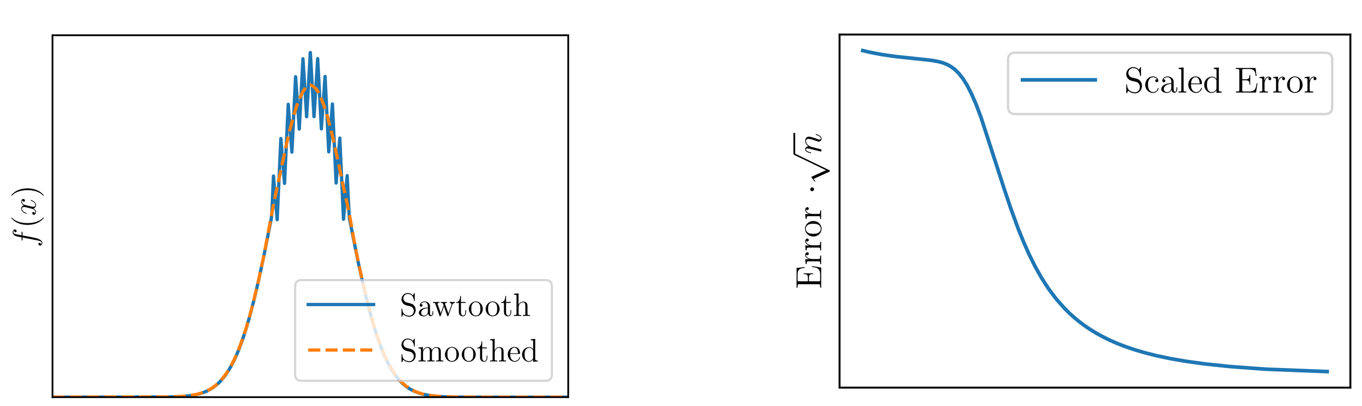
- **Natural Question:** What if distribution is **known**?

Location Estimation, Known Distribution, $d = 1$ case



$\frac{1}{\mathcal{I}} = \sigma^2$ $\frac{1}{\mathcal{I}} = \frac{\sigma^2}{2}$ $\frac{1}{\mathcal{I}} \ll \sigma^2$

- **Finite-Sample Setting:** Might expect $|\widehat{\mu} - \mu| \le \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$. Unfortunately, impossible!
- **Solution:** Smoothing [Gupta, Lee, Price, Valiant; NeurIPS 2022]



Smooth with radius $r \approx \sigma/n^{1/8}$ Gaussian, then run MLE. With prob. $1 - \delta$,

$$|\widehat{\mu} - \mu| \le \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}(1 + o(1))$$

where $\mathcal{I}_r$ is the Fisher information of the smoothed distribution.

Finite-Sample Mean Estimation

	Error	Notes
Unknown Distribution, $d = 1$	$\sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}$	[Catoni '12; Lee-Valiant '21]
Known Distribution, $d = 1$	$\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$	[GLPV22], $\mathcal{I}_r$ is the *smoothed* Fisher Information
Unknown Distribution, any d	$O\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\| \log \frac{1}{\delta}}{n}}\right)$	[Hopkins '20; Lee-Valiant '22]
Known Distribution, any d	$\approx \sqrt{\frac{\mathrm{Tr}(\mathcal{I}_R^{-1})}{n}} + 5\sqrt{\frac{\|\mathcal{I}_R^{-1}\| \log \frac{4}{\delta}}{n}}$	This paper

Table 2. Finite-Sample Results

- *Empirical mean* does not benefit from "spiky" distributions with high Fisher information.
- *MLE* is asymptotically optimal (Cramer-Rao), but number of samples needed *depends on distribution*. For any fixed number of samples, some distribution makes MLE arbitrarily bad!
- *Smoothed MLE* [GLPV '22] has Fisher information guarantees for all distributions + number of samples, but *only in one dimension* and *needs* $\delta \to 0$.

The MLE maximizes the log likelihood,

$$L(\theta) := \sum \log p(x_i - \theta)$$

so it is a zero of the average score $\sum \nabla \log p(x_i - \theta)$.

Contributions

Main Result

After smoothing, *one step of Newton's method* to approximate the MLE gives fast, accurate results for any distribution in 1 or more dimensions.

- In one dimension matches [GLPV '22] but without requiring $\delta \to 0$.
- In high dimensions, fast algorithm that is $1 + o(1)$ of smoothed optimal for $n, d_{\mathrm{eff}}(\mathcal{I}_R^{-1}) \gg \log \frac{1}{\delta}$

Main Theorem: high dimensions

For $n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2 \left(\log \frac{2}{\delta} + d_{\mathrm{eff}}(\mathcal{I}_R^{-1}) + \frac{d_{\mathrm{eff}}(\Sigma)^2}{d_{\mathrm{eff}}(\mathcal{I}_R^{-1})}\right)\right)$, and $R = r^2 I_d$, with probability $1 - \delta$,

$$\|\widehat{\mu} - \mu\| \le (1 + \eta)\sqrt{\frac{\mathrm{Tr}(\mathcal{I}_R^{-1})}{n}} + 5\sqrt{\frac{\|\mathcal{I}_R^{-1}\| \log \frac{4}{\delta}}{n}}$$

Norm Concentration from Subgamma Projections

Norm Concentration Lemma

Let $x \in \mathbb{R}^d$ be *subgamma* in every projection: for every vector v,

$$\mathbb{E}[e^{\lambda\langle x, v\rangle}] \le e^{\lambda^2 v^T \Sigma v/2}$$

for all $|\lambda| \le \frac{1}{C\|v\|}$. Then x has concentrated norm: with probability $1 - \delta$,

$$\|x\| \le \sqrt{\mathrm{Tr}(\Sigma)} + 4\sqrt{\|\Sigma\| \log \frac{2}{\delta}} + 16\|C\| \log \frac{2}{\delta}$$
$$+ \min\left(4\|C\|_F\sqrt{\log \frac{2}{\delta}}, 8\frac{\|C\|_F^2}{\sqrt{\mathrm{Tr}(\Sigma)}} \log \frac{1}{\delta}\right)$$

The first term is tight, and the next two are tight up to constants (from the Gaussian and 1d subgamma case, respectively).

Experiments

- We perform experiments on a mixture of three Gaussians. Here, $d = 20$, $x \sim \mathcal{N}(-e_1, I) + \mathcal{N}(e_1, 9I) + 10^{-4}\mathcal{N}(10^4 e_2, 10^{-6} I)$.

N	10^1	10^2	10^3	10^4	10^5	10^6
Empirical Mean	10.15	11.18	18.76	51.09	34.58	51.82
Newton w/out smoothing	9.45	10.34	17.00	45.94	0.66	0.63
Newton w/ $R = 0.01I$	6.21	6.09	5.95	5.87	5.87	5.70

Table 3. Median Error

- We observe better finite-sample performance by smoothing: it makes the problem better conditioned, so easier to find a good solution (until enough samples that the initial estimate is sufficiently precise).

Summary

1. Gaussian Smoothing + MLE → Finite sample bound for mean estimation with **known** density in **one dimension**
2. Gaussian Smoothing + single step of Newton's method on gradient of log-likelihood
   - Faster and more accurate
   - Finite sample bound in **high dimensions**