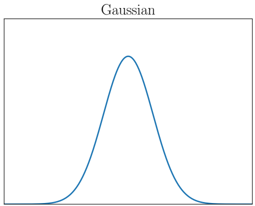
High-dimensional Location Estimation via Norm Concentration for Subgamma Vectors

Shivam Gupta (UT Austin),
Jasper C.H. Lee (UW Madison), Eric Price (UT Austin)

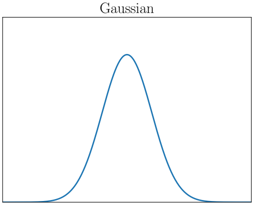
August 11, 2023

Motivating Examples



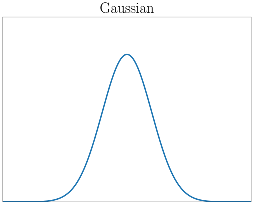
- Given n samples from a Gaussian with variance σ^2, would like to estimate mean.

Motivating Examples

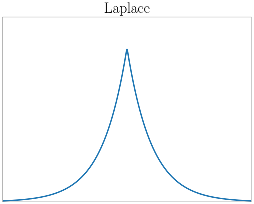


- Given n samples from a Gaussian with variance σ^2, would like to estimate mean.
- The optimal estimator is the empirical mean, which has $1-\delta$ confidence radius $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$

Motivating Examples

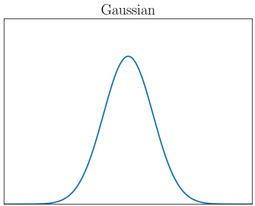


- Given n samples from a Gaussian with variance σ^2, would like to estimate mean.
- The optimal estimator is the empirical mean, which has $1-\delta$ confidence radius $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$

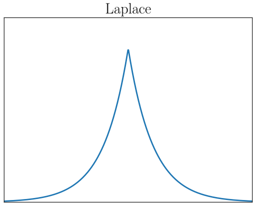


- For the Laplace distribution, the median achieves error $\sigma\sqrt{\frac{\log\frac{1}{\delta}}{n}}$, a factor $\sqrt{2}$ savings over the above

Motivating Examples



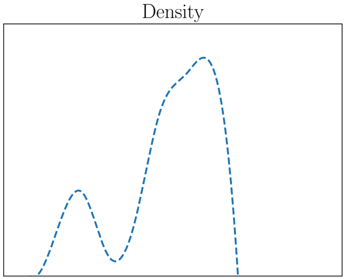
- Given n samples from a Gaussian with variance σ^2, would like to estimate mean.
- The optimal estimator is the empirical mean, which has $1 - \delta$ confidence radius $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$

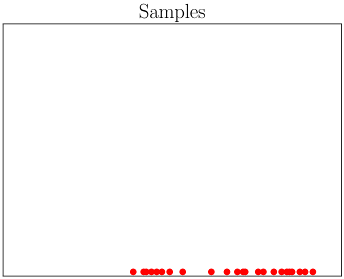


- For the Laplace distribution, the median achieves error $\sigma\sqrt{\frac{\log\frac{1}{\delta}}{n}}$, a factor $\sqrt{2}$ savings over the above

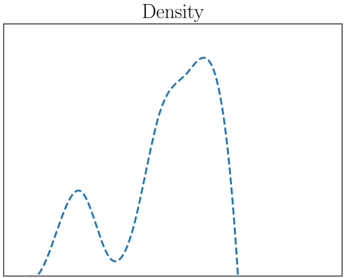
Given a density f (up to shift) on $\mathbb{R}^d$, and n samples $X_1, \ldots, X_n$, what is the best estimator of the mean? **Mean Estimation with known density**.

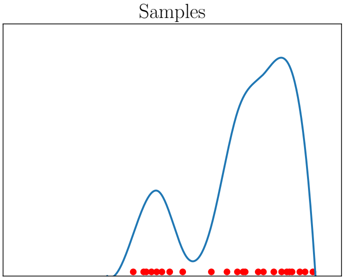
First attempt





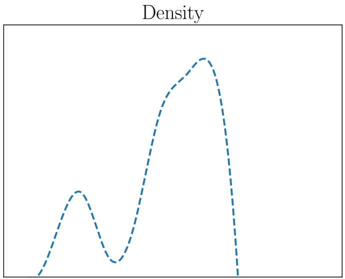
First attempt

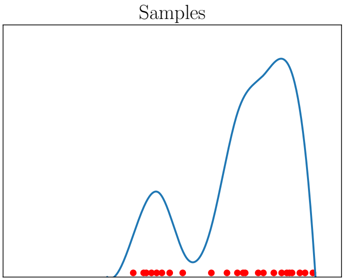




- Fit samples to density: find mean that is most likely to have generated samples – aka Maximum Likelihood Estimation (MLE)

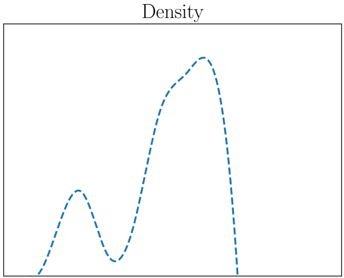
First attempt

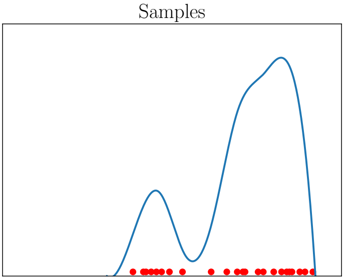




- Fit samples to density: find mean that is most likely to have generated samples – aka Maximum Likelihood Estimation (MLE)
- Enjoys great properties asymptotically – converges to $\mathcal{N}(\mu, \mathcal{I}^{-1}/n)$, where $\mathcal{I}$ is the *Fisher Information*

First attempt





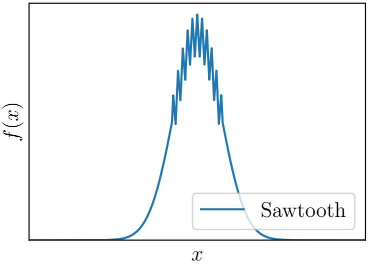
- Fit samples to density: find mean that is most likely to have generated samples – aka Maximum Likelihood Estimation (MLE)
- Enjoys great properties asymptotically – converges to $\mathcal{N}(\mu, \mathcal{I}^{-1}/n)$, where $\mathcal{I}$ is the *Fisher Information*
- Basically tight: Cramér-Rao bound says any unbiased estimator must have variance at least $\mathcal{I}^{-1}/n$

Finite-Sample Setting

- In 1-d, when density known, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$

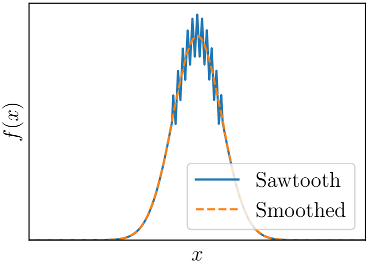
Finite-Sample Setting

- In 1-d, when density known, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!



Finite-Sample Setting

- In 1-d, when density known, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!

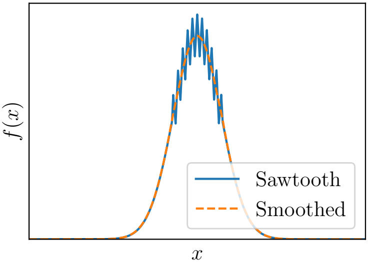


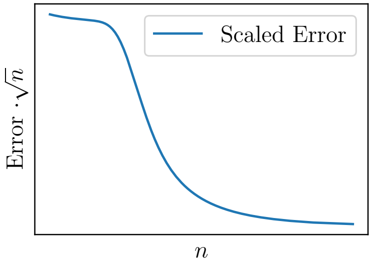
- **Solution:** smoothing [Gupta, Lee, Price, Valiant, NeurIPS 2022] Smooth with a radius $r = \sigma/n^{1/6}$ Gaussian, then run MLE. With probability $1 - \delta$,

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

Finite-Sample Setting

- In 1-d, when density known, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!





- **Solution:** smoothing [Gupta, Lee, Price, Valiant, NeurIPS 2022] Smooth with a radius $r = \sigma / n^{1/6}$ Gaussian, then run MLE. With probability $1 - \delta$,

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

This paper

1. A **new algorithm** for 1d location estimation.
   - **Faster:** one step of Newton's method rather than full MLE
   - **More accurate:** Smaller $o(1)$ term, particularly for constant δ.

This paper

1. A **new algorithm** for 1d location estimation.
   - **Faster:** one step of Newton's method rather than full MLE
   - **More accurate:** Smaller $o(1)$ term, particularly for constant δ.
2. An extension to **high dimensions**.
   - Possible because of simplified algorithm
   - Bound matches Gaussian tail bound for large effective dimension

Our results

Theorem (Informal)

Let $R = r^2 I_d$ and let $\mathcal{I}_R$ be the R-smoothed Fisher information. For large enough r decaying polynomially in n, and any constant $0 < \eta < 1$

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

Our results

Theorem (Informal)

Let $R = r^2 I_d$ and let $\mathcal{I}_R$ be the R-smoothed Fisher information. For large enough r decaying polynomially in n, and any constant $0 < \eta < 1$

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

- Based on new theorem for concentration of norm of vectors with subgamma projections

Summary

1. Gaussian Smoothing + MLE $\rightarrow$ 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**

Contact: shivamgupta@utexas.edu