Finite-Sample Symmetric Mean Estimation with Fisher Information Rate

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

July 18, 2023

Asymptotic Mean Estimation

- Given n samples from a distribution, want to estimate mean μ.

Asymptotic Mean Estimation

- Given n samples from a distribution, want to estimate mean μ.

	Estimator	**Converges to**	**Notes**
Unknown Distribution	Empirical Mean	$\mathcal{N}(\mu, \frac{\sigma^2}{n})$	Central Limit Theorem
Known Distribution	MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	$\mathcal{I}$ is the *Fisher Information*
Unknown *Symmetric* Distribution	KDE + MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	[Stone; 1975]

Table: Classical Asymptotic Results

Asymptotic Mean Estimation

- Given n samples from a distribution, want to estimate mean μ.

	Estimator	Converges to	Notes
Unknown Distribution	Empirical Mean	$\mathcal{N}(\mu, \frac{\sigma^2}{n})$	Central Limit Theorem
Known Distribution	MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	$\mathcal{I}$ is the *Fisher Information*
Unknown *Symmetric* Distribution	KDE + MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	[Stone; 1975]

Table: Classical Asymptotic Results

- In ***finite-sample*** setting, when distribution is **unknown**, [Catoni; 2012], [Lee, Valiant; 2022] show estimator $\widehat{\mu}$ such that with probability $1 - \delta$,

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

Asymptotic Mean Estimation

- Given n samples from a distribution, want to estimate mean μ.

	Estimator	Converges to	Notes
Unknown Distribution	Empirical Mean	$\mathcal{N}(\mu, \frac{\sigma^2}{n})$	Central Limit Theorem
Known Distribution	MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	$\mathcal{I}$ is the *Fisher Information*
Unknown *Symmetric* Distribution	KDE + MLE	$\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$	[Stone; 1975]

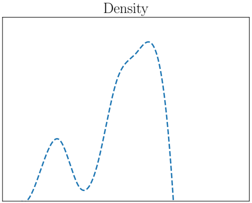
Table: Classical Asymptotic Results

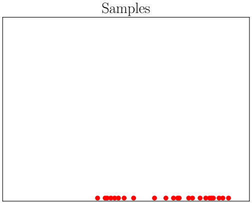
- In ***finite-sample*** setting, when distribution is **unknown**, [Catoni; 2012], [Lee, Valiant; 2022] show estimator $\widehat{\mu}$ such that with probability $1 - \delta$,

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

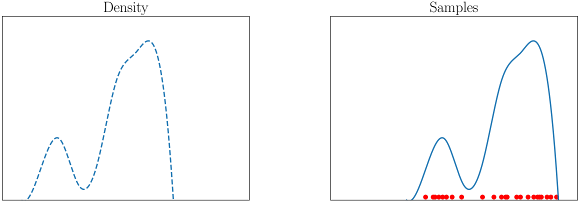
- **Natural Question:** What if distribution is **known/symmetric**?

Location Estimation (Known Distribution Case)



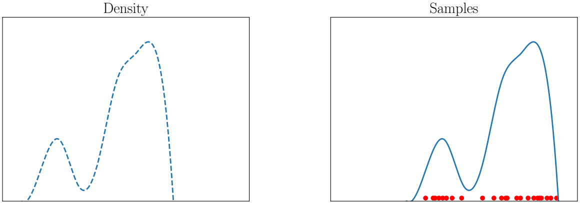


Location Estimation (Known Distribution Case)



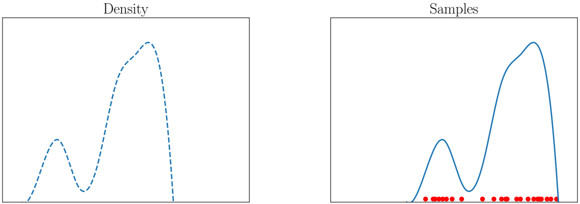
- Fit density to samples, aka **Maximum Likelihood Estimate (MLE)**

Location Estimation (Known Distribution Case)

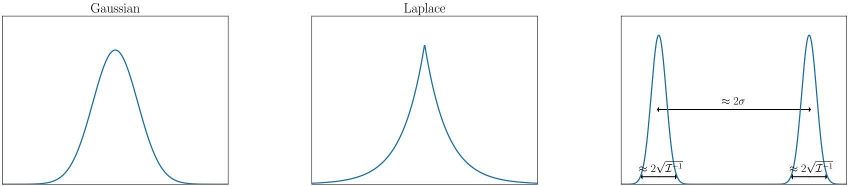


- Fit density to samples, aka **Maximum Likelihood Estimate (MLE)**
- Converges to $\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$ where $\mathcal{I}$ is the *Fisher information*

Location Estimation (Known Distribution Case)



- Fit density to samples, aka **Maximum Likelihood Estimate (MLE)**
- Converges to $\mathcal{N}(\mu, \frac{1}{n\mathcal{I}})$ where $\mathcal{I}$ is the *Fisher information*



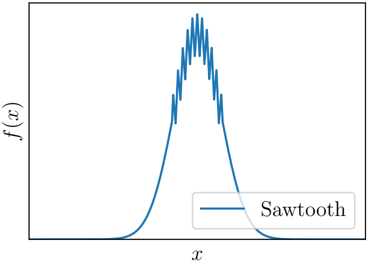
$\frac{1}{\mathcal{I}} = \sigma^2$ $\quad$ $\frac{1}{\mathcal{I}} = \frac{\sigma^2}{2}$ $\quad$ $\frac{1}{\mathcal{I}} \ll \sigma^2$

Finite-Sample Setting

- In ***finite-sample*** setting, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$

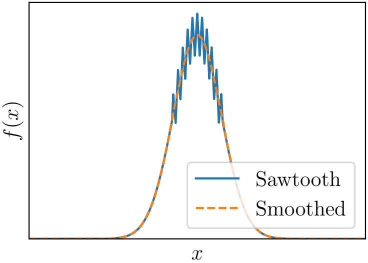
Finite-Sample Setting

- In ***finite-sample*** setting, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!



Finite-Sample Setting

- In ***finite-sample*** setting, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!

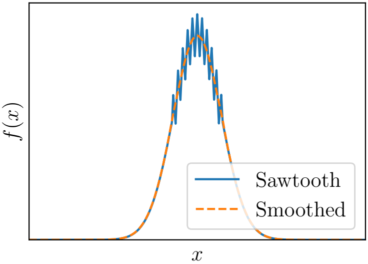


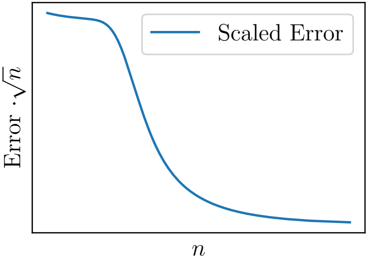
- **Solution:** smoothing **[G., Lee, Price, Valiant; NeurIPS 2022]**
Smooth samples and distribution with a radius $r \approx \sigma/n^{1/6}$ Gaussian, then run MLE. With probability $1 - \delta$,

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

Finite-Sample Setting

- In ***finite-sample*** setting, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}}}$
- Unfortunately, impossible!





- **Solution:** smoothing **[G., Lee, Price, Valiant; NeurIPS 2022]** Smooth samples and distribution with a radius $r \approx \sigma/n^{1/6}$ 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 Mean Estimation

	Error	**Notes**
Unknown Distribution	$\sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}$	[Cat12, LV22]
Known Distribution	$\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$	[GLPV22], $\mathcal{I}_r$ is the *smoothed* Fisher Information
Unknown *Symmetric* Distribution	$\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$	**This paper**

Table: Finite-Sample Results

Finite-Sample Mean Estimation

	Error	Notes
Unknown Distribution	$\sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}$	[Cat12, LV22]
Known Distribution	$\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$	[GLPV22], $\mathcal{I}_r$ is the *smoothed* Fisher Information
Unknown *Symmetric* Distribution	$\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$	**This paper**

Table: Finite-Sample Results

Main Theorem (Informal)

For $r \approx \sigma / n^{1/13}$, our estimator $\widehat{\mu}$ given n samples from a *symmetric* distribution satisfies, with probability $1 - \delta$,

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

The $o(1)$ depends on δ, n, but is independent of the distribution.

Kernel Density Estimate

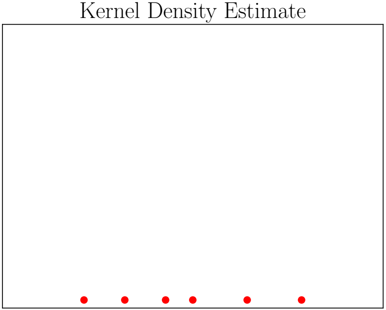
- If we knew the density, we could run (smoothed) MLE

Kernel Density Estimate

- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples

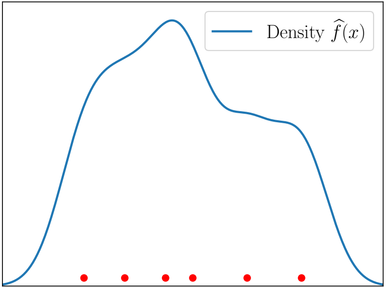
Kernel Density Estimate

- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples



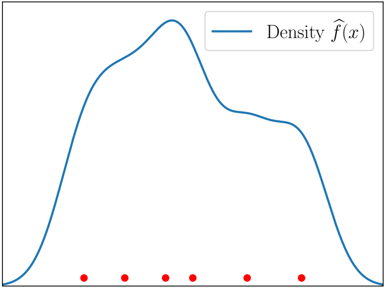
Kernel Density Estimate

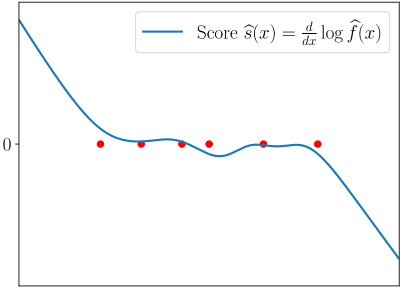
- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples



Kernel Density Estimate

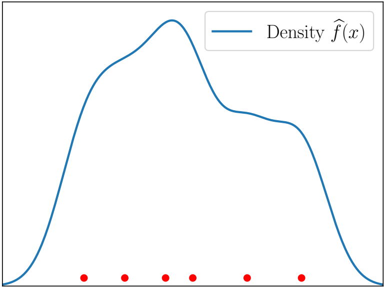
- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples

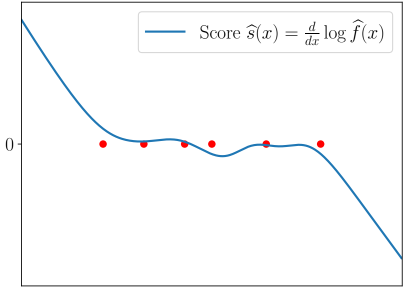




Kernel Density Estimate

- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples

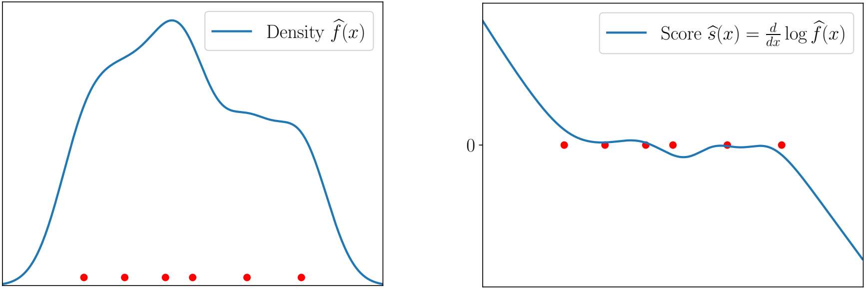




- Naive algorithm: Run (smoothed) MLE on the KDE/Find zero of KDE score using remaining samples.

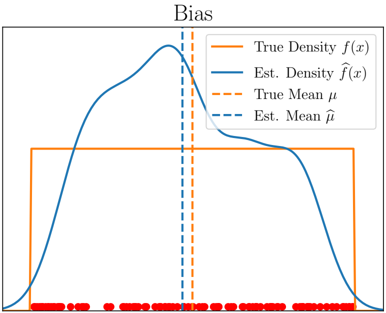
Kernel Density Estimate

- If we knew the density, we could run (smoothed) MLE
- Since we don't know the density, let's try to estimate it using the Kernel Density Estimate (KDE) on the first (say) $n^{1/100}$ samples

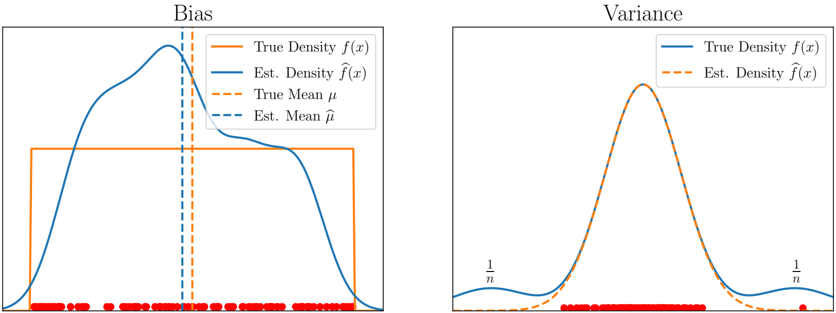


- Naive algorithm: Run (smoothed) MLE on the KDE/Find zero of KDE score using remaining samples. Two issues:
  1. Bias
  2. Variance

Bias and Variance from KDE



Bias and Variance from KDE



Correcting the KDE – Bias

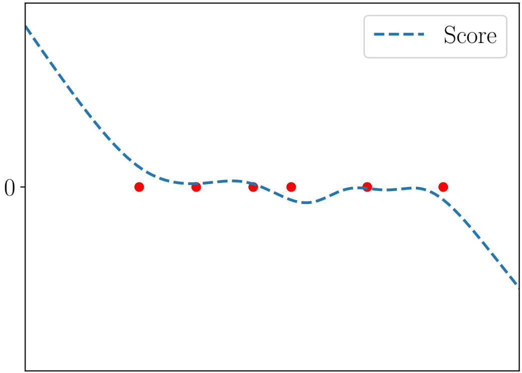
- For a *symmetric* distribution, MLE with respect to *any* (possibly different) symmetric distribution is an unbiased estimator

Correcting the KDE – Bias

- For a *symmetric* distribution, MLE with respect to *any* (possibly different) symmetric distribution is an unbiased estimator
- **Idea:** (Anti)-symmetrize the KDE score.

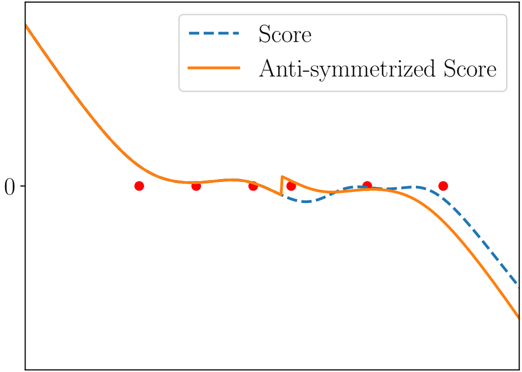
Correcting the KDE – Bias

- For a *symmetric* distribution, MLE with respect to *any* (possibly different) symmetric distribution is an unbiased estimator
- **Idea:** (Anti)-symmetrize the KDE score.

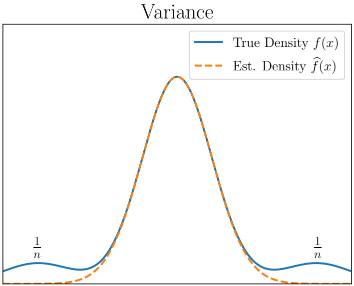


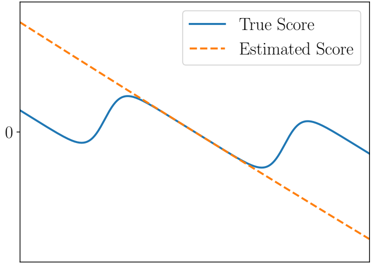
Correcting the KDE – Bias

- For a *symmetric* distribution, MLE with respect to *any* (possibly different) symmetric distribution is an unbiased estimator
- **Idea:** (Anti)-symmetrize the KDE score.

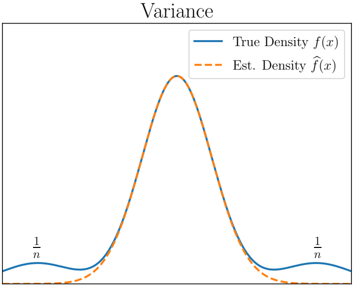


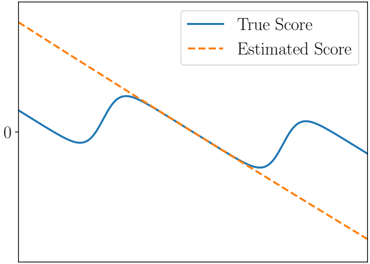
Correcting the KDE – Variance





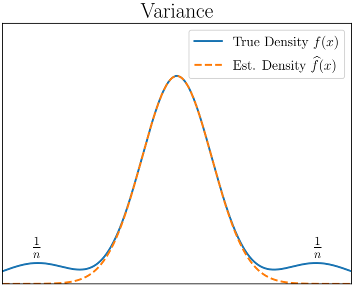
Correcting the KDE – Variance

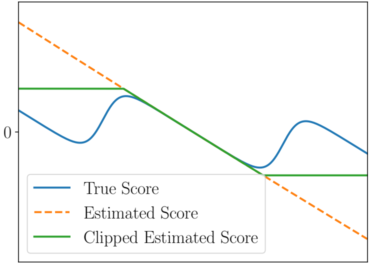




- The true score is close to 0 near the small bumps

Correcting the KDE – Variance





- The true score is close to 0 near the small bumps
- **Solution:** Clip the score

High-Level Summary

- Use the first (say) $n^{1/100}$ samples to compute the KDE

High-Level Summary

- Use the first (say) $n^{1/100}$ samples to compute the KDE
- Symmetrize and clip the KDE score appropriately

High-Level Summary

- Use the first (say) $n^{1/100}$ samples to compute the KDE
- Symmetrize and clip the KDE score appropriately
- Run (variant of) smoothed MLE using the symmetrized and clipped KDE score

High-Level Summary

- Use the first (say) $n^{1/100}$ samples to compute the KDE
- Symmetrize and clip the KDE score appropriately
- Run (variant of) smoothed MLE using the symmetrized and clipped KDE score

Main Theorem (Informal)

For $r \approx \sigma / n^{1/13}$, our estimator $\widehat{\mu}$ given n samples from a *symmetric* distribution satisfies, with probability $1 - \delta$,

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

The $o(1)$ depends on δ, n, but is independent of the distribution.