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 $\mu$. 

Table: Classical Asymptotic Results

- In finite-sample setting, when distribution is unknown, [Catoni; 2012], [Lee, Valiant; 2022] show estimator $\hat{\mu}$ such that with probability $1 - \delta$,
  \[ |\hat{\mu} - \mu| \leq \sqrt{\frac{2}{\sigma^2} \log \frac{2}{\delta}} \cdot \frac{1}{n} (1 + o(1)) \]

- Natural Question: What if distribution is known/symmetric?
Asymptotic Mean Estimation

- Given $n$ samples from a distribution, want to estimate mean $\mu$.

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

Table: Classical Asymptotic Results
Asymptotic Mean Estimation

- Given $n$ samples from a distribution, want to estimate mean $\mu$.

<table>
<thead>
<tr>
<th></th>
<th>Estimator</th>
<th>Converges to</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unknown Distribution</strong></td>
<td>Empirical Mean</td>
<td>$\mathcal{N}(\mu, \frac{\sigma^2}{n})$</td>
<td>Central Limit Theorem</td>
</tr>
<tr>
<td><strong>Known Distribution</strong></td>
<td>MLE</td>
<td>$\mathcal{N}(\mu, \frac{1}{nI})$</td>
<td>$I$ is the Fisher Information</td>
</tr>
<tr>
<td><strong>Unknown Symmetric Distribution</strong></td>
<td>KDE + MLE</td>
<td>$\mathcal{N}(\mu, \frac{1}{nI})$</td>
<td>[Stone; 1975]</td>
</tr>
</tbody>
</table>

**Table: Classical Asymptotic Results**

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

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

- Given \( n \) samples from a distribution, want to estimate mean \( \mu \).

<table>
<thead>
<tr>
<th></th>
<th>Estimator</th>
<th>Converges to</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown Distribution</td>
<td>Empirical Mean</td>
<td>( \mathcal{N}(\mu, \frac{\sigma^2}{n}) )</td>
<td>Central Limit Theorem</td>
</tr>
<tr>
<td>Known Distribution</td>
<td>MLE</td>
<td>( \mathcal{N}(\mu, \frac{1}{n\mathcal{I}}) )</td>
<td>( \mathcal{I} ) is the Fisher Information</td>
</tr>
<tr>
<td>Unknown Symmetric</td>
<td>KDE + MLE</td>
<td>( \mathcal{N}(\mu, \frac{1}{n\mathcal{I}}) )</td>
<td>[Stone; 1975]</td>
</tr>
</tbody>
</table>

Table: Classical Asymptotic Results

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

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

- Natural Question: What if distribution is known/symmetric?
Location Estimation (Known Distribution Case)

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

- Gaussian
  - $\approx 2\sigma$

- Laplace
  - $\approx 2\sqrt{I}$

- $I \approx \frac{3}{10}$
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 \( \mathcal{N}(\mu, \frac{1}{nI}) \) where \( I \) is the **Fisher information**

\[
\frac{1}{\mathcal{I}} = \sigma^2 \\
\frac{1}{\mathcal{I}} = \frac{\sigma^2}{2} \\
\frac{1}{\mathcal{I}} \ll \sigma^2
\]
Finite-Sample Setting

- In finite-sample setting, might expect $|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{nI}}$
Finite-Sample Setting

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

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 $1 - \delta$,

$$|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{nI_r}} (1 + o(1))$$
Finite-Sample Setting

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

\[ f(x) \]

\[ x \]

\[ f(x) \]

\[ Error \cdot \sqrt{n} \]

\[ n \]

\[ Error \cdot \sqrt{n} \]

- **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 $1 - \delta$,

\[
|\hat{\mu} - \mu| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{n I r}} (1 + o(1))
\]
## Finite-Sample Mean Estimation

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Error</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unknown Distribution</strong></td>
<td>$\sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}$</td>
<td>[Cat12, LV22]</td>
</tr>
<tr>
<td><strong>Known Distribution</strong></td>
<td>$\sqrt{\frac{2 \log \frac{2}{\delta}}{nI_r}}$</td>
<td>[GLPV22], $I_r$ is the <em>smoothed</em> Fisher Information</td>
</tr>
<tr>
<td><strong>Unknown Symmetric Distribution</strong></td>
<td>$\sqrt{\frac{2 \log \frac{2}{\delta}}{nI_r}}$</td>
<td>This paper</td>
</tr>
</tbody>
</table>

**Table:** Finite-Sample Results
### Finite-Sample Mean Estimation

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Error</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown Distribution</td>
<td>$\sqrt{2\sigma^2 \log \frac{2^\frac{1}{2}}{\delta}}$</td>
<td>[Cat12, LV22]</td>
</tr>
<tr>
<td>Known Distribution</td>
<td>$\sqrt{\frac{2 \log 2^\frac{1}{2}}{nI_r}}$</td>
<td>[GLPV22], $I_r$ is the <em>smoothed</em> Fisher Information</td>
</tr>
<tr>
<td>Unknown Symmetric Distribution</td>
<td>$\sqrt{\frac{2 \log 2^\frac{1}{2}}{nI_r}}$</td>
<td>This paper</td>
</tr>
</tbody>
</table>

**Table:** Finite-Sample Results

### Main Theorem (Informal)

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

$$|\hat{\mu} - \mu| \leq (1 + o(1)) \sqrt{\frac{2 \log 2^\frac{1}{2}}{nI_r}}$$

The $o(1)$ depends on $\delta, n$, but is independent of the distribution.
Kernel Density Estimate

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

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left( \frac{x - x_i}{h} \right)
\]

\[
\hat{s}(x) = \frac{d}{dx} \log \hat{f}(x)
\]

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

Two issues:
1. Bias
2. Variance
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.

\[
\hat{s}(x) = \frac{d}{dx} \log \hat{f}(x)
\]
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

\[
\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)
\]

Kernel Density Estimate

\[
\hat{s}(x) = \frac{d}{dx} \log \hat{f}(x)
\]
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

\[
\text{Density } \hat{f}(x) = \frac{d}{dx} \log \hat{f}(x)
\]
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

![Graph showing density estimate and score](image)

- 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

$$\hat{f}(x)$$

Score $\hat{s}(x) = \frac{d}{dx} \log \hat{f}(x)$

- 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

True Density $f(x)$
Est. Density $\hat{f}(x)$
True Mean $\mu$
Est. Mean $\hat{\mu}$
Bias and Variance from KDE

**Bias**
- True Density $f(x)$
- Est. Density $\hat{f}(x)$
- True Mean $\mu$
- Est. Mean $\hat{\mu}$

**Variance**
- True Density $f(x)$
- Est. Density $\hat{f}(x)$

\[
\text{Bias} = \frac{1}{n}
\]

\[
\text{Variance} = \frac{1}{n}
\]
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.

![Graph showing Score and Anti-symmetrized Score](image-url)
Correcting the KDE – Variance

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

Solution: Clip the score.
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

Main Theorem (Informal)

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

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

The $o(1)$ depends on $\delta, n$, but is independent of the distribution.
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 $\hat{\mu}$ given $n$ samples from a symmetric distribution satisfies, with probability $1 - \delta$,

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

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