Section: Module 6

Asymptotics of MLE

Kentaro Nakamura

GOV 2003

March 6th, 2026

Today's Agenda

- Basic tools for calculus
  - Mean Value Theorem
  - Taylor Series Expansion
  - Continuous Mapping Theorem
  - Covariance Inequality

- Asymptotics for MLE
  - Consistency
  - Asymptotic Normality
  - Efficiency (Cramer-Rao inequality)
  - Delta Method

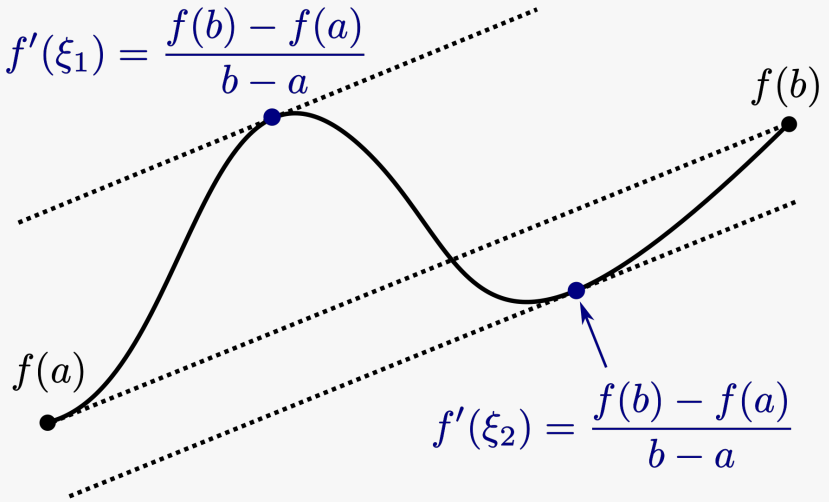
- Logistics
  - HW3 is released on next Monday, due 3/23 (Monday after spring break)
  - Review Session on 3/23 (Monday)
  - Midterm 3/25 (Wednesday)

Mean Value Theorem

Theorem (Mean Value Theorem)

Let $f : [a, b] \mapsto \mathbb{R}$ be a continuous function that is differentiable on (a, b). Then, there exists at least one point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



Taylor Expansion with Mean Value Theorem

- From mean value theorem, we get

$$f(x) = f(x_0) + f'(x^*)(x - x_0) \quad x^* \in (x_0, x)$$

  - This is called **first-order Taylor expansion**

- Similarly, we can have the **second-order Taylor expansion**

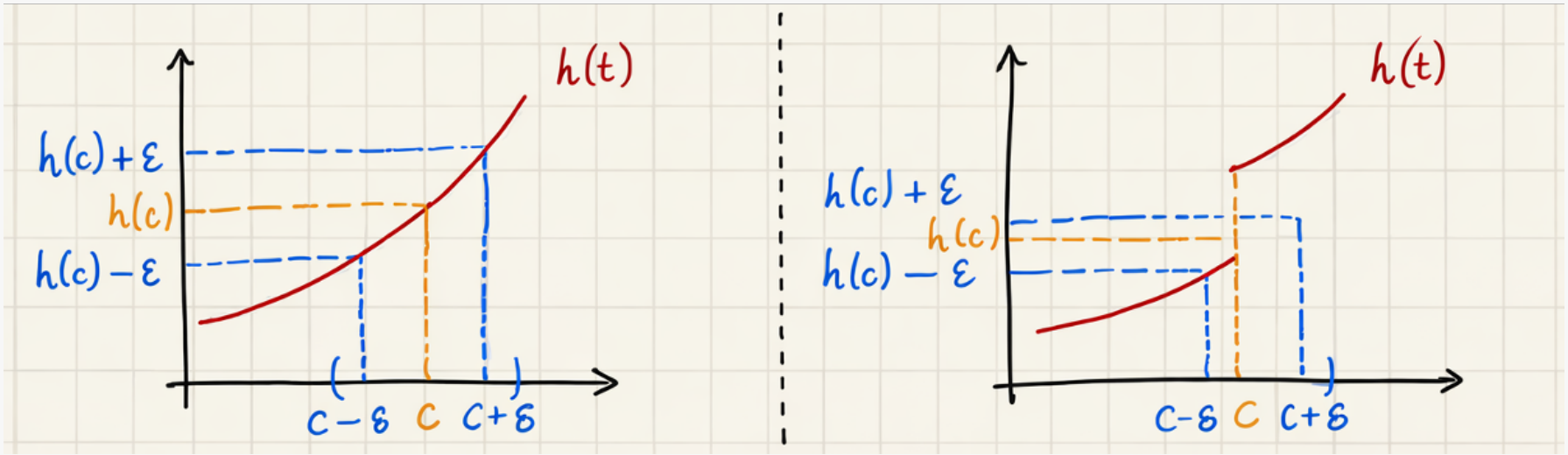
$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x^*)(x - x_0)^2 \quad x^* \in (x_0, x)$$

Continuous Mapping Theorem

Theorem (Continuous Mapping Theorem)

Let $X_n \xrightarrow{*} X$ with $* = p$ or $* = d$. Then, for continuous function g, we have

$$g(X_n) \xrightarrow{*} g(X)$$



Covariance Inequality

Lemma (Covariance Inequality)

For any two random variables X and Y with finite second moments ($\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$),

$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$$

- Recall that pearson's corr. coef. $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$, which is upper bounded by 1.

- **Proof**: We use **Caughy-Schwartz** inequality:

$$|\mathbb{E}[UV]| \leq \sqrt{\mathbb{E}[U^2]\mathbb{E}[V^2]}$$

  - Applying $U = X - \mathbb{E}[X]$ and $V = Y - E[Y]$ gives you the inequality

Consistency of MLE: Proof (1/4)

- **Setup**: You have i.i.d. data $Y_1, \cdots, Y_n \sim f_{\theta_0}(Y_i)$ with true parameter θ_0
- The MLE maximizes the log-likelihood $\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(Y_i)$, which is equivalent to maximizing the average log-likelihood:

$$\bar{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(Y_i)$$

- Also define the expected log-likelihood as

$$\tilde{\ell}(\theta, \theta_0) = \mathbb{E}_{\theta_0}[\log f_\theta(Y_1)]$$

Proof strategy: We will show two facts:

1. $\bar{\ell}_n(\theta) \xrightarrow{P} \tilde{\ell}(\theta, \theta_0)$ for each θ (by WLLN)
2. $\tilde{\ell}(\theta, \theta_0)$ is uniquely maximized at $\theta = \theta_0$ (by Jensen's inequality)

Consistency of MLE: Proof (2/4)

- For each fixed θ, the random variables $\log f_\theta(Y_1), \ldots, \log f_\theta(Y_n)$ are iis. Thus, by the LLN,

$$\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(Y_i) \xrightarrow{P} \tilde{\ell}(\theta, \theta_0)$$

- Now, by the simple algebra, we have

$$\begin{aligned}\tilde{\ell}(\theta, \theta_0) - \tilde{\ell}(\theta_0, \theta_0) &= \mathbb{E}_{\theta_0}\left[\log f_\theta(Y_1)\right] - \mathbb{E}_{\theta_0}\left[\log f_{\theta_0}(Y_1)\right] \\ &= \mathbb{E}_{\theta_0}\left[\log \frac{f_\theta(Y_1)}{f_{\theta_0}(Y_1)}\right]\end{aligned}$$

Consistency of MLE: Proof (3/4)

- Applying Jensen's inequality (as log is concave):

$$\tilde{\ell}(\theta, \theta_0) - \tilde{\ell}(\theta_0, \theta_0) = \mathbb{E}_{\theta_0}\left[\log \frac{f_\theta(Y_1)}{f_{\theta_0}(Y_1)}\right] \leq \log \mathbb{E}_{\theta_0}\left[\frac{f_\theta(Y_1)}{f_{\theta_0}(Y_1)}\right]$$

$$= \log \int \frac{f_\theta(y)}{f_{\theta_0}(y)} f_{\theta_0}(y)\, dy = \log 1 = 0$$

with equality iff $f_\theta(y) = f_{\theta_0}(y)$ for all y.

  - This implies $\theta = \theta_0$[1]
  - So $\tilde{\ell}(\theta, \theta_0)$ is **uniquely maximized** at θ_0.

- Thus, for each $\theta \neq \theta_0$:

$$\mathbb{P}_{\theta_0}(\bar{\ell}_n(\theta) \geq \bar{\ell}_n(\theta_0)) \to 0$$

[1]We implicitly assume the **identifiablity** here: i.e.., $f_\theta(y) = f_{\theta_0}(y)$ implies $\theta = \theta_0$

Consistency of MLE: Proof (4/4)

- Finally, since θ_0 maximizes $\bar{\ell}_n$ means no other θ beats it, we have

$$\begin{aligned}
\mathbb{P}_{\theta_0}(\theta_0 \text{ maximizes } \bar{\ell}_n(\theta)) &= 1 - \mathbb{P}_{\theta_0}\left(\bigcup_{\theta \neq \theta_0} \{\bar{\ell}_n(\theta) \geq \bar{\ell}_n(\theta_0)\}\right) \\
&\leq 1 - \sum_{\theta \neq \theta_0} \mathbb{P}(\bar{\ell}_n(\theta) \geq \bar{\ell}_n(\theta_0)) \\
&\to 1
\end{aligned}$$

  where the second line inequality is by union bound (i.e., $\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i)$)

- Therefore, $\hat{\theta}_n$, which is the maximizer of the average log-likelihood $\bar{\ell}_n(\theta)$, is asymptotically equivalent to θ_0, which proves the convergence in probability

Fisher Information

Definition (Fisher Information)

The Fisher information for data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is defined as

$$I(\theta) := \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(\boldsymbol{Y})\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(\boldsymbol{Y})\right]$$

- Note: the second equality holds under the assumption that derivative and integral are exchangeable.
  - Proof of this is in next page.
  - This is called **information equality**

Proof of Information Equality

- By taking derivative, we get

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f_\theta(\boldsymbol{Y}) &= \frac{\partial}{\partial\theta}\frac{f'_\theta(\boldsymbol{Y})}{f_\theta(\boldsymbol{Y})} \\ &= \frac{f''_\theta(\boldsymbol{Y})\, f_\theta(\boldsymbol{Y}) - \left(f'_\theta(\boldsymbol{Y})\right)^2}{f_\theta(\boldsymbol{Y})^2} \\ &= \frac{f''_\theta(\boldsymbol{Y})}{f_\theta(\boldsymbol{Y})} - \left(\frac{f'_\theta(\boldsymbol{Y})}{f_\theta(\boldsymbol{Y})}\right)^2\end{aligned}$$

- Taking expectation yields

$$\mathbb{E}\left[\frac{f''_\theta(\boldsymbol{Y})}{f_\theta(\boldsymbol{Y})}\right] = \int f''_\theta(\boldsymbol{y})\, d\boldsymbol{y} = \frac{\partial^2}{\partial\theta^2}\int f_\theta(\boldsymbol{y})\, d\boldsymbol{y} = 0$$

- Hence

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(\boldsymbol{Y})\right] = \mathbb{E}\left[\left(\frac{f'_\theta(\boldsymbol{Y})}{f_\theta(\boldsymbol{Y})}\right)^2\right] = I(\theta)$$

Asymptotic Normality of MLE

Theorem (Fisher's Theorem / MLE Asymptotic Normality)

Let $Y_1, \cdots, Y_n \sim f_{\theta_0}$. Under the regularity conditions, the MLE estimator $\hat{\theta}_n$ achieves the asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

where $I(\theta_0)$ is the fisher information.

- This is equivalent to

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta_0, \frac{1}{n} I(\theta)^{-1}\right)$$

  - We can use this to statistical inference
  - **Note**: This assumes that the model is correctly specified.

- We need some mathematical conditions (called regularity conditions) to prove this rigorously
  - This is the contents of STAT211 (beyond the scope of this course)

Proof Strategy: Taylor Expansion (1)

- By Taylor expansion of score function, we get

$$S_n(\hat{\theta}_n) = S_n(\theta_0) + S_n'(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0)^2$$

  where θ_n^* is between $\hat{\theta}_n$ and θ_0.

- Now, notice that $S_n(\hat{\theta}_n) = 0$ by the definition of MLE

- As a result, you can show that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}S_n(\theta_0)}{-S_n'(\theta_0) - \frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0)}$$

Proof Strategy: Taylor Expansion (2)

- Under i.i.d. assumption, by CLT,

$$\sqrt{n}S_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f_\theta(Y_i)}{\partial\theta} \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$$

- Under i.i.d., by LLN,

$$-S_n'(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_\theta(Y_i)}{\partial\theta^2} \xrightarrow{p} \mathbb{E}\left[-\frac{\partial^2 \log f_\theta(Y_i)}{\partial\theta^2}\right] = I(\theta_0)$$

- As $\hat{\theta}_n - \theta_0 \xrightarrow{p} 0$, by Slutsky's theorem,[2]

$$-\frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0) \xrightarrow{p} 0$$

[2]You need to show $S_n''(\theta_n^*)$ is bounded in probability, strictly speaking. You can show it as θ_n^* is bounded between $\hat{\theta}_n$ and θ_0 and $\hat{\theta}_n \xrightarrow{p} \theta_0$ but this is beyond the scope of this course.