Section: Module 5

MLE

Kentaro Nakamura

GOV 2003

February 27th, 2026

Today's Agenda

- Review of Tools in Asymptotic Statistics
- Review of Calculus
    - Taylor Expansion
    - Optimization Problem in Vector Space
    - Jensen's Inequality
- MLE
    - Basic Principle
    - Example (Normal Distribution)
- Asymptotic of MLE
    - Score Function
- Problem Set 2 is due next Friday before section

Tools in Asymptotic Statistics (AGAIN!)

- **Law of Large Numbers (LLN)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]$$

- **Central Limit Theorem (CLT)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[X_i])$$

- **Slutsky's Lemma**: If $X_n \xrightarrow{d} X$ for some random variable X and $Y_n \xrightarrow{p} c$ for some constant c,

$$X_n Y_n \xrightarrow{d} cX$$

Taylor Expansion

Theorem (Taylor's Theorem in $\mathbb{R}^d$)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be sufficiently smooth in a neighborhood of $\boldsymbol{a} \in \mathbb{R}^d$. Then for $\boldsymbol{x}$ near $\boldsymbol{a}$,

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^\top (\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^\top H_f(\boldsymbol{a})\,(\boldsymbol{x} - \boldsymbol{a}) + R_2(\boldsymbol{x}),$$

where $\nabla f(\boldsymbol{a}) \in \mathbb{R}^d$ is the ***gradient****, $H_f(\boldsymbol{a}) \in \mathbb{R}^{d \times d}$ is the* ***Hessian matrix****, and $R_2(\boldsymbol{x})$ is the remainder term.*

- Here $\nabla f(\boldsymbol{a}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\right)^\top$ and $\left[H_f(\boldsymbol{a})\right]_{ij} = \left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{a}}$

Optimization Problem in Vector Space

- In real line case, you checked both first-order condition (derivative is zero) and second-order condition (second derivative is negative)
  - This is generalized in the case of vector

- Suppose that first order condition is checked (i.e., $\nabla f(\boldsymbol{x}) = 0$). Then, Taylor expansion gives you

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{h}) &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top H_f(\boldsymbol{x})\boldsymbol{h} + R_2(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{h}^\top H_f(\boldsymbol{x})\boldsymbol{h} + R_2(\boldsymbol{x}) \end{aligned}$$

- If $f(\boldsymbol{x})$ attains the maximum, any small perturbation makes the function smaller, which means that $f(\boldsymbol{x}+\boldsymbol{h})$ should be smaller than $f(\boldsymbol{x})$
  - This is equivalent to the case when $\boldsymbol{h}^\top H_f(\boldsymbol{x})\boldsymbol{h}$ is negative for any $\boldsymbol{h}$
  - That is why we should check if hessian H_f is negative-definite or not

How to Check Positive/Negative Definiteness

- A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is positive definite ($A \succ 0$) iff any of the following equivalent conditions hold:
- **Sylvester's criterion**: All leading principal minors are positive:

$$a_{11} > 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} > 0, \quad \ldots, \quad \det(A) > 0$$

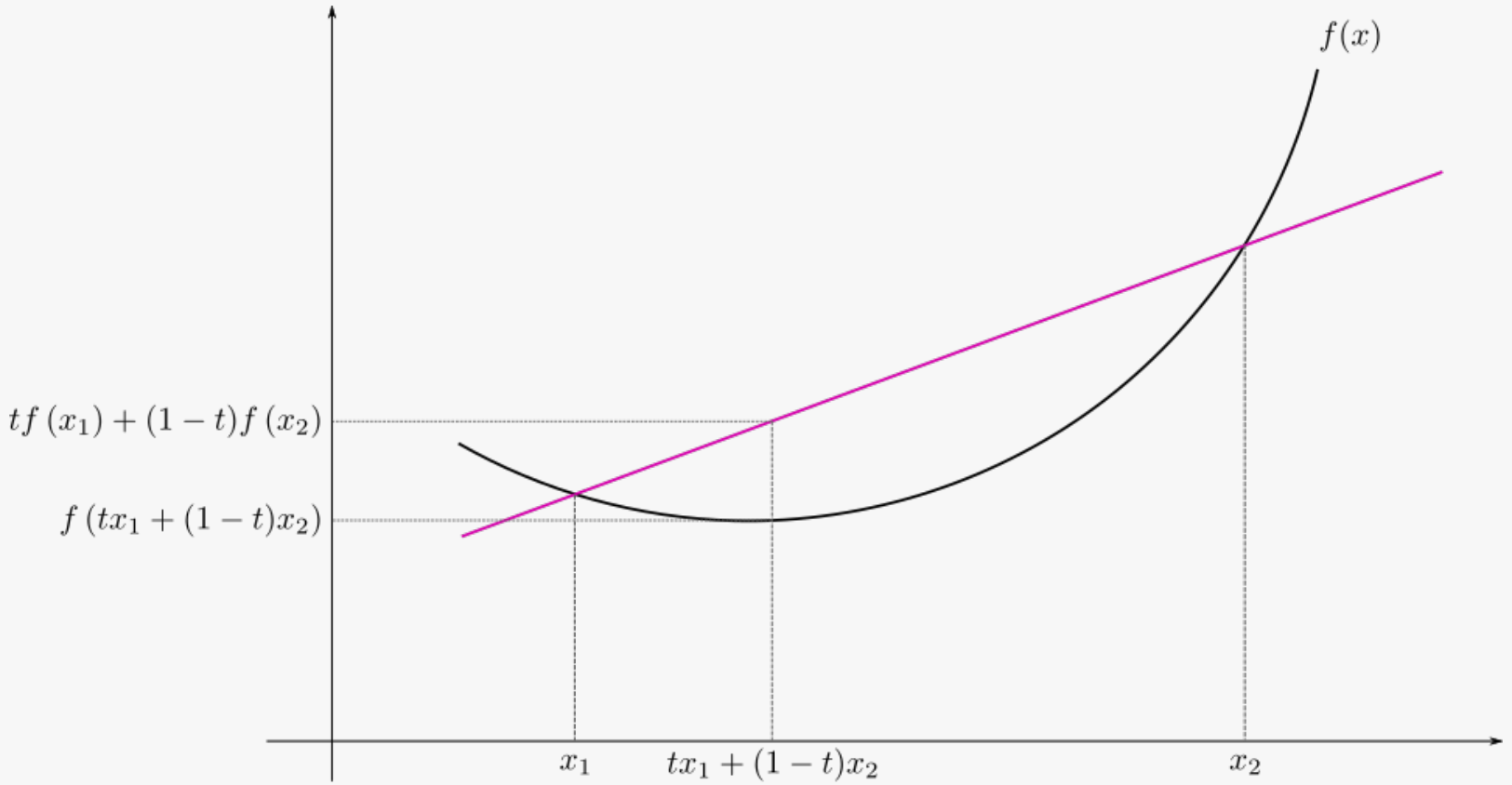
- For **negative definiteness** ($A \prec 0$): check leading principal minors alternate in sign
  - $a_{11} < 0$, $\det\left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right) > 0$, $\det(A_{3\times 3}) < 0, \ldots$
- We show this through example of MLE.

Jensen's Inequality

Lemma (Jensen's Inequality)

Let φ be a convex function on an interval I, and let X be a random variable taking values in I with $\mathbb{E}[|X|] < \infty$. Then

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$



MLE: Basic Principle

- **Setting**: Observe data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from a distribution with density (or pmf) $f_\theta(\boldsymbol{y})$, where $\theta \in \Theta$ is an unknown parameter.
- **Likelihood function**: Given observed data $\boldsymbol{y}$,

$$L(\theta) = L(\theta \mid \boldsymbol{y}) = f_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} f_\theta(y_i) \quad \text{(under i.i.d. assumption)}$$

- **Likelihood principle**: The likelihood function contains all the information about the unknown parameter of interest.
- **MLE**: The maximum likelihood estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta)$$

MLE: Example (Normal Distribution)

- **Problem (Example)**: Consider a sample drawn from $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ with unknown θ. How can we estimate θ?

Example

Let $f(X_i \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i-\theta)^2}{2}\right)$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right)$$

and the log-likelihood is

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2.$$

Setting $\ell'(\theta) = 0$ gives $\hat{\theta}_{\text{MLE}} = \bar{X}$.

MLE as Optimization Problem

- For MLE, we need to maximize the log-likelihood (objective function)

- To check whether the solution attains maximum, we need to check the first-order condition and the second-order condition
  - First order condition: first derivative is zero

$$\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

  - Second order condition: second derivative is negative (or negative definite for multivariate θ)

$$\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$$

MLE Example: Normal with Unknown μ and σ^2

- **Problem**: Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)$ both unknown.
- The log-likelihood is:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

- **First-order conditions** (setting partial derivatives to zero):

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \quad \Longrightarrow \quad \hat{\mu} = \bar{X}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

MLE Example: Checking Second-Order Condition

- **Second-order condition**: We need the Hessian to be negative definite at $(\hat{\mu}, \hat{\sigma}^2)$.

- The Hessian matrix of $\ell(\mu, \sigma^2)$ is:

$$H = \begin{pmatrix} \dfrac{\partial^2 \ell}{\partial \mu^2} & \dfrac{\partial^2 \ell}{\partial \mu \, \partial \sigma^2} \\ \dfrac{\partial^2 \ell}{\partial \sigma^2 \, \partial \mu} & \dfrac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\dfrac{n}{\sigma^2} & 0 \\ 0 & -\dfrac{n}{2\sigma^4} \end{pmatrix} \Bigg|_{(\hat{\mu}, \hat{\sigma}^2)}$$

- Both diagonal entries are negative, and the off-diagonal is zero, so H is negative definite.
- Therefore $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \ \frac{1}{n}\sum (X_i - \bar{X})^2)$ is indeed the MLE.

Convergence in Probability

Definition (Convergence in Probability)

A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges in probability to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

We write $X_n \xrightarrow{P} X$.

- **Key tool 1**: Law of Large Numbers (LLN)
  - If $X_1, \cdots, X_n$ are i.i.d.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]$$

- **Key tool 2**: Chebyshev's inequality (proof in appendix)
  - For any random variable Y with finite variance and any $\varepsilon > 0$,

$$\mathbb{P}\Big(\,|Y - \mathbb{E}[Y]| > \varepsilon\,\Big) \le \frac{\mathrm{Var}(Y)}{\varepsilon^2}$$

Score Function

Definition (Score Function)

The score function $S_n(\boldsymbol{Y}, \theta)$ (or simply $S_n(\theta)$) for the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is defined as

$$S_n(\boldsymbol{Y}, \theta) = \frac{\partial \log f_\theta(\boldsymbol{Y})}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta}$$

where $\ell(\theta)$ is the log-likelihood function.

- Note: Even though it is often called score statistic, this is not a statistic (a quantity calculable from data alone).
  - Indeed, the score function depends on both data $\boldsymbol{Y}$ and unknown parameter θ.

- **Property**: The expected score is zero. That is, $\mathbb{E}_\theta\big[S_n(\boldsymbol{Y}, \theta)\big] = 0$
  - $\mathbb{E}_\theta[S_n(\boldsymbol{Y}, \theta)] = \int \frac{f'_\theta(\boldsymbol{y})}{f_\theta(\boldsymbol{y})} f_\theta(\boldsymbol{y})\, d\boldsymbol{y} = \frac{\partial}{\partial \theta} \int f_\theta(\boldsymbol{y})\, d\boldsymbol{y} = \frac{\partial}{\partial \theta} 1 = 0$

Appendix: Proof of Chebyshev's inequality

- We first prove **Markov inequality**: Let X be a nonnegative random variable. For any $a > 0$,

$$\Pr(X \geq 0) \leq \frac{\mathbb{E}[X]}{a}$$

- **Proof**: Now, notice that $X \geq a \cdot \mathbf{1}\{X \geq a\}$. By taking expectation,

$$\mathbb{E}[X] \geq a \cdot \mathbb{E}[\mathbf{1}\{X \geq a\}] = a \cdot \Pr(X \geq a).$$

- Therefore,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

- By applying this Markov inequality to $X = |Y - \mathbb{E}[Y]|^2$, we get

$$\Pr(|Y - \mathbb{E}[Y]|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[|Y - \mathbb{E}[Y]]}{\epsilon^2}$$

$$\Leftrightarrow \Pr(|Y - \mathbb{E}[Y]| \geq \epsilon) \leq \frac{\mathbb{E}[|Y - \mathbb{E}[Y]]}{\epsilon^2}$$