

Lecture 6: October 16, 2026

Lecturer: Wenlong Mou

1 M -estimation continued

For MLE, uniqueness of population minimizer is automatically satisfied when identifiable.

Non-example: Gaussian mixture model.

$$X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(-\mu, 1),$$

for some $\mu \in \mathbb{R}$. By flipping the sign of μ , we get the same distribution. So the model is not identifiable, and the population minimizer is not unique.

In such a case, we may consider estimation up to equivalence class, e.g., we can restrict the parameter space to $\Theta = [0, \infty)$, and then the population minimizer is unique.

On the other hand, if identifiability does hold, we have

Corollary 6.1. *If the parametric model is identifiable, and if Θ is compact, $p_\theta(x)$ is continuous in θ for every $x \in \mathcal{X}$, and $\mathbb{E}_{\theta^*}[\sup_{\theta \in \Theta} |\log p_\theta(X)|] < +\infty$, then the MLE is consistent:*

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow{p} \theta^* \quad \text{as } n \rightarrow \infty.$$

2 Z -estimation and asymptotic normality

On the one hand, we do not have a general consistency theory for Z -estimation. For M -estimation, at least we know the minimizer exists when Θ is compact and the loss is continuous. But for Z -estimation, the solution may not exist, and even if it exists, it may not be unique.

So for the Z -estimation theory, we start with the assumption that there exists a solution, and we assume it is already a consistent estimator. Locally, under certain regularity conditions, we can show asymptotic normality of the estimator. This is a local theory, and it does not guarantee existence or uniqueness of the solution.

Problem setup: We suppress standard moment and local regularity conditions and focus on the proof idea.

- Suppose there exists $\theta^* \in \Theta$ such that $H(\theta^*) = 0$.
- Suppose that for any $\varepsilon > 0$, we have

$$\mathbb{P}(\exists \theta \in \Theta \cap \mathbb{B}(\theta^*, \varepsilon), \text{ such that } H_n(\theta) = 0) \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

and when the solution exists, we call it $\hat{\theta}_n$.

Note: we additionally assume that the selected root $\hat{\theta}_n$ is consistent for θ^* .

- Assume $h(\theta; x)$ is continuously differentiable in θ (at least locally around θ^*) for every $x \in \mathcal{X}$.

- $\nabla_{\theta}H(\theta^*) \in \mathbb{R}^{d \times d}$ is non-singular (invertible).

Theorem 6.1. *Under above setup, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla_{\theta}H(\theta^*)^{-1} \Sigma^* (\nabla_{\theta}H(\theta^*))^{-\top}), \quad \text{as } n \rightarrow \infty,$$

where $\Sigma^* = \mathbb{E}[h(\theta^*, X)h(\theta^*, X)^{\top}]$.

Proof: by consistency of $\hat{\theta}_n$, we can apply Taylor expansion

$$0 = H_n(\hat{\theta}_n) = \underbrace{H_n(\theta^*)}_{\text{i.i.d. average}} + \underbrace{\nabla_{\theta}H_n(\theta^*)(\hat{\theta}_n - \theta^*)}_{\text{Involves quantity of interest}} + o_p(\|\hat{\theta}_n - \theta^*\|_2).$$

For the matrix pre-factor $\nabla H_n(\theta^*)$, we have

$$\nabla H_n(\theta^*) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}h(\theta^*, X_i) \xrightarrow{p} \mathbb{E}[\nabla_{\theta}h(\theta^*, X)] = \nabla_{\theta}H(\theta^*),$$

and for the first term $H_n(\theta^*)$, we have

$$\sqrt{n}H_n(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(\theta^*, X_i) \xrightarrow{d} \mathcal{N}(0, \Sigma^*) \quad \text{as } n \rightarrow \infty,$$

by multivariate central limit theorem. Putting them together, since $\nabla_{\theta}H_n(\theta^*) \rightarrow \nabla_{\theta}H(\theta^*)$ in probability and the limit is invertible, we have

$$(\hat{\theta}_n - \theta^*) = -\nabla_{\theta}H_n(\theta^*)^{-1} \cdot H_n(\theta^*) + o_p(\|\hat{\theta}_n - \theta^*\|_2).$$

By Slutsky's theorem, we have

$$\sqrt{n}\nabla_{\theta}H_n(\theta^*)^{-1} \cdot H_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla_{\theta}H(\theta^*)^{-1} \Sigma^* (\nabla_{\theta}H(\theta^*))^{-\top}), \quad \text{as } n \rightarrow \infty,$$

This also implies that $\sqrt{n}(\hat{\theta}_n - \theta^*) = O_p(1)$, so the remainder term $o_p(\|\hat{\theta}_n - \theta^*\|_2) = o_p(1/\sqrt{n})$ is negligible. This completes the proof.

Statistical inference based on this result: Σ^* and $\nabla_{\theta}H(\theta^*)$ are unknown, but we can estimate them based on data.

$$\begin{aligned} \hat{\Sigma}_n &= \frac{1}{n} \sum_{i=1}^n h(\hat{\theta}_n, X_i)h(\hat{\theta}_n, X_i)^{\top}, \\ \hat{J}_n &= \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}h(\hat{\theta}_n, X_i). \end{aligned}$$

Assuming standard regularity conditions and $\Sigma^* \succ 0$, we have $\hat{\Sigma}_n \xrightarrow{p} \Sigma^*$ and $\hat{J}_n \xrightarrow{p} \nabla_{\theta}H(\theta^*)$. So we can use plug-in estimator to estimate the asymptotic covariance matrix, and we have

$$\sqrt{n} \left(\hat{J}_n^{-1} \hat{\Sigma}_n \hat{J}_n^{-\top} \right)^{-1/2} (\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, I_d), \quad \text{as } n \rightarrow \infty.$$

We can then build confidence sets and perform hypothesis testing based on this result.

Remark: this is just one method for inference. We can also use bootstrap to estimate the asymptotic covariance matrix, or we can use bootstrap to directly estimate the distribution of $\hat{\theta}_n$.

Implication to M -estimation: If the loss function is continuously differentiable, then an interior M -estimator is also a Z -estimator. So we can use this result to derive asymptotic normality of M -estimation.

Verifying the problem setup:

- Existence of θ^* : satisfied under differentiability when θ^* is in the interior of Θ .
- Local existence of $\hat{\theta}_n$: guaranteed by consistency of $\hat{\theta}_n$ (using the theory of M -estimation).
- Differentiability of $h(\theta, X) = \nabla_\theta f(\theta, X)$: satisfied when f is **twice** continuously differentiable in θ .
- Non-singularity of $\nabla_\theta H(\theta^*) = \nabla^2 F(\theta^*)$. Since θ^* is the minimizer of F , this condition is equivalent to the Hessian of F being positive definite at θ^* .

In such a case, we have

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 F(\theta^*)^{-1} \Sigma^* \nabla^2 F(\theta^*)^{-1}), \quad \text{as } n \rightarrow \infty,$$

where $\Sigma^* = \mathbb{E}[\nabla_\theta f(\theta^*, X) \nabla_\theta f(\theta^*, X)^\top]$.

Specialization to MLE: Under similar conditions as above, we can compute

$$\Sigma^* = \mathbb{E}_{\theta^*}[\nabla_\theta \log p_{\theta^*}(X) \nabla_\theta \log p_{\theta^*}(X)^\top], \quad \nabla^2 F(\theta^*) = -\mathbb{E}_{\theta^*}[\nabla_\theta^2 \log p_{\theta^*}(X)].$$

Theorem 6.2 (Fisher's identity). $\Sigma^* = \nabla^2 F(\theta^*)$, which is also called the Fisher information matrix, denoted as $I(\theta^*)$.

Proof: first, for the score function $\nabla \log p_\theta(X)$, we have

$$\mathbb{E}_\theta[\nabla_\theta \log p_\theta(X)] = \int p_\theta(x) \nabla_\theta \log p_\theta(x) dx = \int \nabla_\theta p_\theta(x) dx = \nabla_\theta \int p_\theta(x) dx = 0. \quad (6.1)$$

We can differentiate both sides of (6.1), for the i, j -th entry of the matrix, we have

$$\int \partial_{\theta_i} \partial_{\theta_j} \log p_\theta(x) \cdot p_\theta(x) dx + \int \partial_{\theta_j} \log p_\theta(x) \cdot \partial_{\theta_i} \log p_\theta(x) \cdot p_\theta(x) dx = 0.$$

By collecting these entries into a matrix, we have

$$\mathbb{E}_\theta[\nabla_\theta^2 \log p_\theta(X)] + \mathbb{E}_\theta[\nabla_\theta \log p_\theta(X) \nabla_\theta \log p_\theta(X)^\top] = 0.$$

Using the Fisher's identity in combination with the asymptotic normality result, we have

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta^*)^{-1}), \quad \text{as } n \rightarrow \infty,$$

for MLE, under the same regularity conditions as above.

Remarks:

- Intuitively, the Fisher information matrix $I(\theta^*)$ measures how much information the data contains about the unknown parameter θ^* . The larger the Fisher information, the smaller the asymptotic variance of the MLE, and the more accurate the estimation.
- There are cases where the Fisher information matrix is singular, and in such a case, the asymptotic normality may not hold. But MLE may still be consistent, and it may converge at a slower rate than $1/\sqrt{n}$.
- For inference, we just need to use plug-in method to estimate the Fisher information matrix.

2.1 Cramér–Rao lower bound

Fisher information is important because it captures the fundamental possibility of estimation. One example is the Cramér–Rao lower bound, which gives a lower bound on the variance of any unbiased estimator.

Idea: we want to find the best one among all possible estimators. This is generally not possible (decision theory), but we can restrict to unbiased estimators. Then we can show that the variance of any unbiased estimator is lower bounded by the inverse of Fisher information matrix.

Theorem 6.3 (Cramér–Rao lower bound). *Suppose that $\hat{\theta}_n$ is an unbiased estimator of θ , i.e., $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$. Then we have*

$$\text{Cov}_\theta(\hat{\theta}_n) \succeq \frac{1}{n} I(\theta)^{-1},$$

where $A \succeq B$ means that $A - B$ is positive semi-definite.

Proof (for simplicity, we consider the one-dimensional case): by unbiasedness, we have

$$\theta = \int \hat{\theta}_n(x_1, x_2, \dots, x_n) \prod_{i=1}^n p_\theta(x_i) dx_1 dx_2 \cdots dx_n, \quad \text{for all } \theta \in \Theta. \quad (6.2)$$

Differentiating both sides of (6.2) with respect to θ , we have

$$1 = \int \hat{\theta}_n(x_1, x_2, \dots, x_n) \cdot \sum_{i=1}^n \partial_\theta \log p_\theta(x_i) \cdot \prod_{i=1}^n p_\theta(x_i) dx_1 dx_2 \cdots dx_n \quad (6.3)$$

On the other hand, we have score identity (6.1):

$$0 = \int \sum_{i=1}^n \partial_\theta \log p_\theta(x_i) \cdot \prod_{i=1}^n p_\theta(x_i) dx_1 dx_2 \cdots dx_n. \quad (6.4)$$

(6.3) – $\theta \times$ (6.4), we have

$$\begin{aligned} 1 &= \int (\hat{\theta}_n - \theta) \cdot \sum_{i=1}^n \partial_\theta \log p_\theta(x_i) \cdot \prod_{i=1}^n p_\theta(x_i) dx_1 dx_2 \cdots dx_n = \mathbb{E}_\theta \left[\underbrace{(\hat{\theta}_n - \theta)}_{\text{estimation error}} \cdot \sum_{i=1}^n \partial_\theta \log p_\theta(X_i) \right] \\ &\leq \sqrt{\text{var}_\theta(\hat{\theta}_n - \theta)} \cdot \sqrt{\text{var}_\theta \left(\sum_{i=1}^n \partial_\theta \log p_\theta(X_i) \right)} = \sqrt{\text{var}_\theta(\hat{\theta}_n)} \cdot \sqrt{n I(\theta)}, \end{aligned}$$

which proves the Cramér–Rao lower bound.

Remarks:

- Unbiasedness is very restrictive, and MLE itself is not unbiased in general. So this lower bound itself does not prove optimality of MLE.
- However, Fisher information is useful beyond the context of unbiasedness. In particular, people have discovered that $\mathcal{N}(0, I(\theta^*)^{-1})$ is the optimal asymptotic distribution of any regular estimator. This asymptotic optimality result does not require unbiasedness, and it is a much stronger result than Cramér–Rao lower bound. So we actually know that MLE is asymptotically optimal among all regular estimators, and it achieves the best possible asymptotic variance. This theory is known as Hájek–Le Cam local asymptotic minimax theorem, which is beyond the scope of this course.

2.2 Additional topics

Computational algorithms for M and Z estimation:

- In some cases, we can solve them with very high precision.
- In some cases, we want to build statistical theory for approximate solutions that are computed from algorithms.

Here I present two classical examples.

Newton's method: Suppose we want to solve $H_n(\theta) = 0$. We can use Newton's method to iteratively update the solution:

$$\theta^{(t+1)} = \theta^{(t)} - \nabla_{\theta} H_n(\theta^{(t)})^{-1} H_n(\theta^{(t)}).$$

(under certain conditions, this converges at super-linear rate locally around the true solution).

On the other hand, for statistical analysis, sometimes just one-step Newton's method is enough.

Theorem 6.4. *Starting from estimator $\tilde{\theta}_n$, and suppose that $\|\tilde{\theta}_n - \theta^*\|_2 = O_p(1/\sqrt{n})$, then we can define one-step Newton's method estimator as*

$$\hat{\theta}_n = \tilde{\theta}_n - \nabla_{\theta} H_n(\tilde{\theta}_n)^{-1} H_n(\tilde{\theta}_n).$$

Then we have

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla_{\theta} H(\theta^*)^{-1} \Sigma^* (\nabla_{\theta} H(\theta^*))^{-\top}), \quad \text{as } n \rightarrow \infty.$$

Operationally, this allows us to find a computable estimator that is asymptotically equivalent to the Z -estimator.

Proof idea:

$$\begin{aligned} & \tilde{\theta}_n - \nabla_{\theta} H_n(\tilde{\theta}_n)^{-1} H_n(\tilde{\theta}_n) \\ &= \tilde{\theta}_n - \nabla_{\theta} H_n(\tilde{\theta}_n)^{-1} \left\{ H_n(\theta^*) + \nabla H_n(\theta^*)(\tilde{\theta}_n - \theta^*) + o_p(\|\tilde{\theta}_n - \theta^*\|_2) \right\} \end{aligned}$$

The terms in blue asymptotically cancels, and the residual is negligible under our assumption. This gives

$$\hat{\theta}_n = \theta^* - \nabla H_n(\theta^*)^{-1} H_n(\theta^*) + o_p(1/\sqrt{n}),$$

which proves the asymptotic normality of $\hat{\theta}_n$.

Stochastic gradient descent:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} f(\theta_t, X_{t+1}),$$

for $t = 0, 1, \dots, n-1$.

Polyak–Juditsky–Ruppert averaging: $\hat{\theta}_n = \frac{1}{n} \sum_{t=1}^n \theta_t$.

Under certain condition, we can show that $\hat{\theta}_n$ is asymptotically normal and achieves the same asymptotic covariance as the M -estimator.