

STA355F: Theory of Statistical Practice

Fall 2026

Lecture 5: October 9, 2026

Lecturer: Wenlong Mou

1 Bootstrap continued

The bootstrap discussed last time draws samples $X_1^*, X_2^*, \dots, X_n^*$ from the empirical distribution $\hat{F}_n$ with replacement. This is called the nonparametric bootstrap. There are also other variants of bootstrap, including parametric bootstrap.

Suppose that we have a parametric model $\{F_\theta : \theta \in \Theta\}$, and we assume that the true distribution F is in this parametric family, i.e., $F = F_{\theta_0}$ for some unknown $\theta_0 \in \Theta$. In that case, instead of drawing from the empirical distribution $\hat{F}_n$, we can draw samples from the estimated parametric distribution $F_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is an estimator of θ_0 based on the original samples $X_1, X_2, \dots, X_n$.

In other words, in the parametric bootstrap world, we have

$$X_i^* \sim F_{\hat{\theta}_n}, \quad i = 1, 2, \dots, n,$$

conditionally on the original samples $X_1, X_2, \dots, X_n$. Then we can compute the bootstrap estimator $\hat{T}_n^* = g(X_1^*, X_2^*, \dots, X_n^*)$ and proceed with the bootstrap procedure as before. We can still perform Monte Carlo simulation to approximate the conditional distribution of $\hat{T}_n^*$. Just as before, the hope is that the conditional distribution of $\hat{T}_n^*$ given the original samples $X_1, X_2, \dots, X_n$ is a good approximation of the marginal distribution of T_n . We can also construct confidence intervals using normal, pivotal, or percentile methods as before.

Remark:

- If the parametric model is well-specified (the family $(F_\theta : \theta \in \Theta)$ contains the true distribution), parametric bootstrap can be more accurate.
- It relies on parametric model being correct. There are important cases where we care about inference about a functional $T(F)$, but we do not want to assume a parametric model for F (e.g. average treatment effect in causal inference; or prediction model in statistical learning). In that case, nonparametric bootstrap is more appropriate.

2 Z-estimation and M-estimation

Roughly speaking, Z-estimation is solving an estimation equation, while M-estimation is solving an optimization problem.

Z-estimators: Suppose that $\Theta \in \mathbb{R}^d$, and you have $h : \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}^d$. An Z-estimator $\hat{\theta}_n$ solves the following estimating equation:

$$\frac{1}{n} \sum_{i=1}^n h(\theta, X_i) = 0,$$

where $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ for some unknown distribution $\mathbb{P}$.

Note:

- Solution may not always exist, and if it exists, it may not be unique. We usually need to impose some conditions to show existence of the solution (or just assume existence). And in the non-unique case, we can choose one of the solutions as the estimator.
- Example: method of moments. Given a parametric model $\{F_\theta : \theta \in \Theta\}$, and data

$$X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F_{\theta_0}$$

for some unknown $\theta_0 \in \Theta$. When $X_i \in \mathbb{R}$, one estimating strategy is by moment matching. For $j \in \mathbb{N}_+$ and $\theta \in \Theta$, we define

$$\alpha_j(\theta) = \mathbb{E}_{F_\theta}[X^j].$$

Consider the first d moments, we can define the estimating function $h : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$ as

$$h(\theta, x) = \begin{pmatrix} x - \alpha_1(\theta) \\ x^2 - \alpha_2(\theta) \\ \vdots \\ x^d - \alpha_d(\theta) \end{pmatrix}.$$

The estimating equation is then

$$\frac{1}{n} \sum_{i=1}^n X_i^j = \alpha_j(\hat{\theta}_n), \quad j = 1, 2, \dots, d.$$

A canonical example is Gaussian distribution, where we match the mean and variance to estimate the parameters.

In general, Z -estimator does not impose any parametric assumption on the data distribution. So where does $\hat{\theta}_n$ converge to?

By law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^n h(\theta, X_i) \xrightarrow{p} \mathbb{E}_{\mathbb{P}}[h(\theta, X)] := H(\theta) \quad \text{as } n \rightarrow \infty,$$

for each fixed $\theta \in \Theta$. By Z -estimation, we hope that $\hat{\theta}_n$ converges to the solution of the population estimating equation

$$H(\theta) = \mathbb{E}_{\mathbb{P}}[h(\theta, X)] = 0,$$

the solution of which is denoted as θ^* (assuming that the solution exists and is unique).

Note: the target θ^* here does not have to be the data-generating mechanism. We may work without parametric model, and still have a well-defined target θ^* as the solution of the population estimating equation. In that case, we can think of θ^* as a population-level parameter that we want to estimate.

M -estimators: Suppose that $\Theta \in \mathbb{R}^d$, and you have loss function $f : \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}$, an M -estimator $\hat{\theta}_n$ solves the following optimization problem:

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n f(\theta, X_i).$$

Remarks:

- Existence hold true under mild assumptions, e.g., if Θ is compact and f is continuous. Uniqueness may not hold, but we can choose one of the minimizers as the estimator.
- The target of M -estimation is the minimizer of the population risk:

$$\theta^* = \arg \min_{\theta \in \Theta} \underbrace{\mathbb{E}_{\mathbb{P}}[f(\theta, X)]}_{=: F(\theta)}.$$

- When $f(\cdot, x)$ is differentiable for every $x \in \mathcal{X}$, and when the minimizers are in the interior of Θ (e.g. when $\Theta = \mathbb{R}^d$), then the first-order optimality condition gives

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} f(\hat{\theta}_n, X_i) = 0, \quad \mathbb{E}_{\mathbb{P}}[\nabla_{\theta} f(\theta^*, X)] = 0.$$

In such a case, by taking $h(\theta, x) = \nabla_{\theta} f(\theta, x)$, we can see that the M -estimator is also a Z -estimator. But in general (with non-differentiable loss function or constrained optimization), M estimation may not be formulated as Z -estimation. Z and M estimation are not comparable in general.

An important sub-class of M -estimation in parametric models is the maximum likelihood estimation (MLE). Suppose that we have a parametric model $\{F_{\theta} : \theta \in \Theta\}$, and we get i.i.d. samples $X_1, X_2, \dots, X_n \sim F_{\theta^*}$ for some unknown $\theta^* \in \Theta$. We further assume that F_{θ} has a density function p_{θ} (in continuous or discrete sense). In such a case, the maximal likelihood estimator is defined as

$$\hat{\theta}_n^{\text{MLE}} = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i)$$

Note that here the joint likelihood of data is

$$p_{\theta}(X_1, X_2, \dots, X_n) = \prod_{i=1}^n p_{\theta}(X_i),$$

so MLE is maximizing the joint likelihood of the data. MLE is a general idea – Θ does not need to be finite-dimensional, and p_{θ} does not need to be a parametric model. Here, we will mostly focus on the parametric case.

MLE is an M -estimator with loss function $f(\theta, x) = -\log p_{\theta}(x)$.

Examples:

- $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$. Then the MLE is

$$\hat{\theta}_n^{\text{MLE}} = \arg \max_{\theta \in [0, 1]} \frac{1}{n} \sum_{i=1}^n (X_i \log \theta + (1 - X_i) \log(1 - \theta)) = \frac{1}{n} \sum_{i=1}^n X_i.$$

Actually, this applies to any categorical distribution, where the MLE is just the empirical frequency of each category.

- $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with the unknown parameter $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Then the MLE is

$$(\hat{\mu}_n^{\text{MLE}}, \hat{\sigma}_n^{2, \text{MLE}}) = \arg \max_{(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)} \frac{1}{n} \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

By first-order optimality condition, we can compute the MLE as

$$\hat{\mu}_n^{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma}_n^{2,\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_n^{\text{MLE}})^2.$$

- $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([0, \theta])$, for some unknown $\theta > 0$. Then the MLE is

$$\hat{\theta}_n^{\text{MLE}} = \arg \max_{\theta > 0} \frac{1}{n} \sum_{i=1}^n \log \left(\frac{1}{\theta} \mathbf{1}_{X_i \in [0, \theta]} \right) = \max_{i=1,2,\dots,n} X_i.$$

For this estimator, we have

$$\mathbb{P}(\hat{\theta}_n^{\text{MLE}} \leq t) = \mathbb{P}(X_1 \leq t, X_2 \leq t, \dots, X_n \leq t) = (t/\theta)^n, \quad 0 < t < \theta.$$

Integrating the tail probability, we know that

$$\mathbb{E}[|\hat{\theta}_n^{\text{MLE}} - \theta|^2] = O(1/n^2).$$

So this estimator is not asymptotically normal, but it converges to the true parameter θ at a faster rate than $1/\sqrt{n}$.

Note: in the first two examples, MLE is also a Z -estimator, but in the last example, MLE is not a Z -estimator. In general, we have different theories for Z -estimation and M -estimation.

- M estimation requires less assumptions, and we can easily get consistency results under very weak assumptions.
- Z estimation requires more assumptions, but we can get asymptotic normality. We can also use this to derive asymptotic normality of M -estimation, by using the first-order optimality condition.

3 Theoretical properties of M -estimation and Z -estimation

3.1 Consistency theory of M -estimators

Recall that

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} F_n(\theta) = \frac{1}{n} \sum_{i=1}^n f(\theta, X_i), \quad \theta^* = \arg \min_{\theta \in \Theta} F(\theta) = \mathbb{E}_{\mathbb{P}}[f(\theta, X)].$$

If F_n is close to F , and if F has a unique minimizer, with some regular geometric conditions around its minimizer, then we can expect that $\hat{\theta}_n$ is close to θ^* . This consists of two steps:

- Showing that $F(\hat{\theta}_n)$ is close to $F(\theta^*)$ (i.e., the function value is close to the optimal value).
- Use this together with the geometric conditions to show that $\hat{\theta}_n$ is close to θ^* . If F is flat around θ^* , then it could be possible that $\hat{\theta}_n$ is far away from θ^* , but still has a function value close to the optimal value. So we need conditions to avoid this.

Theorem 5.1. *Assume that:*

1. *Uniform law of large numbers:*

$$\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty.$$

2. *For any $\varepsilon > 0$, we have*

$$\inf_{\theta \in \Theta: \|\theta - \theta^*\|_2 \geq \varepsilon} F(\theta) > F(\theta^*).$$

Then we have $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \rightarrow \infty$.

Remark:

- The first condition requires uniform convergence of the loss functions (similar to Glivenko–Cantelli theorem), and the second condition requires that the population risk F is not flat around its minimizer θ^* .
- Verifying these conditions can be challenging, and we will discuss simple sufficient conditions.

Proof of the theorem:

Convergence of function value: We note that

$$0 \leq F(\hat{\theta}_n) - F(\theta^*) = \underbrace{F(\hat{\theta}_n) - F_n(\hat{\theta}_n)}_{\text{That is new here}} + \underbrace{F_n(\hat{\theta}_n) - F_n(\theta^*)}_{\leq 0} + \underbrace{F_n(\theta^*) - F(\theta^*)}_{\xrightarrow{p} 0, \text{ by LLN}}.$$

Note: we **cannot** directly use (the non-uniform version of) law of large numbers to show that $F_n(\hat{\theta}_n) - F(\hat{\theta}_n) \xrightarrow{p} 0$, because $\hat{\theta}_n$ is random and depends on the data. Evaluating a random function at a random point (with correlated randomness) is outside the scope of standard law of large numbers. But we can use the uniform law of large numbers to show that

$$|F(\hat{\theta}_n) - F_n(\hat{\theta}_n)| \leq \sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{p} 0.$$

So we have $F(\hat{\theta}_n) - F(\theta^*) \xrightarrow{p} 0$ as $n \rightarrow \infty$.

Convergence of the minimizer: For fixed $\varepsilon > 0$, we let

$$\delta := \inf_{\theta \in \Theta: \|\theta - \theta^*\|_2 \geq \varepsilon} F(\theta) - F(\theta^*) > 0.$$

By convergence in probability of the function value, we have

$$\mathbb{P}(F(\hat{\theta}_n) - F(\theta^*) \geq \delta) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

By assumption, if $\|\hat{\theta}_n - \theta^*\|_2 \geq \varepsilon$, then $F(\hat{\theta}_n) - F(\theta^*) \geq \delta$. So we have

$$\mathbb{P}(\|\hat{\theta}_n - \theta^*\|_2 \geq \varepsilon) \leq \mathbb{P}(F(\hat{\theta}_n) - F(\theta^*) \geq \delta) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

This shows that $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \rightarrow \infty$.

Additional remark: the gap $F(\hat{\theta}_n) - F_n(\hat{\theta}_n)$ is the generalization gap in statistical machine learning. $F(\hat{\theta}_n)$ is the population (test) loss, and $F_n(\hat{\theta}_n)$ is the training loss on the training set. In machine learning, it happens quite often that your model could perfectly fit training data, but it does not generalize well to unseen test data. This is the overfitting phenomenon in machine learning. The uniform law of large number is basically saying that: “if you use larger and larger sample size, with a fixed model class, the amount of overfitting will vanish”. Of course, this is an assumption, which may or may not hold depending on the model class.

Verifying the two assumptions: For the first assumption, we use

Theorem 5.2 (Wald's theorem). *Suppose that Θ is compact (in finite-dimensional case, compactness means bounded and closed), and $f(\theta, x)$ is continuous in θ for every $x \in \mathcal{X}$. Furthermore, assume that*

$$\mathbb{E} \left[\sup_{\theta \in \Theta} |f(\theta, X)| \right] < +\infty. \quad (5.1)$$

Then we have

$$\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty.$$

Remarks:

- Actually, most of the works are done in this theorem, and the consistency theorem is just a direct corollary. Indeed, these two combined together are sometimes called Wald's consistency theorem.
- The condition (5.1) is coming from dominance convergence theorem. The class of functions $\{f(\theta, \cdot) : \theta \in \Theta\}$ needs to be dominated by an integrable function. In practice, you can manually verify this by taking upper bound of $|f(\theta, x)|$ over $\theta \in \Theta$.
- In finite-dimensional Euclidean spaces, compactness is equivalent to boundedness and closedness. In many practical problems, we minimize the loss function over unbounded domain, such as $\Theta = \mathbb{R}^d$. In such a case, uniform convergence may not hold. Usually, we first show convergence of a constrained estimator $\hat{\theta}_n^{(R)} = \arg \min_{\theta \in \Theta: \|\theta\|_2 \leq R} F_n(\theta)$ for some very large R , and then use the fact that the loss grows to infinity as $\|\theta\|_2 \rightarrow \infty$ to show $\mathbb{P}(\hat{\theta}_n^{(R)} = \hat{\theta}_n) \rightarrow 1$ as $n \rightarrow \infty$. In that case, we can show that $\hat{\theta}_n$ is consistent.
- If Θ is infinite-dimensional, compactness roughly requires that Θ can be approximated by finite-dimensional structures. The theorem still holds true in such a case.

Proof idea: starting with a simple case – if $\Theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ is finite. Then we have

$$\mathbb{P} \left(\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| > \varepsilon \right) \leq \sum_{j=1}^M \mathbb{P} \left(|F_n(\theta_j) - F(\theta_j)| > \varepsilon \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

For general compact Θ , we can use discrete approximation of Θ to reduce it to the finite case, e.g., by taking ε -covering of Θ :

$$\Theta \subseteq \bigcup_{j=1}^M B(\theta_j, \varepsilon) \quad \text{for some } \theta_1, \theta_2, \dots, \theta_M \in \Theta.$$

Compactness ensures the existence of such a finite covering. Then we can use continuity of f and the condition (5.1) to show that the discretization error vanishes as $\varepsilon \rightarrow 0$.

Verifying the second assumption:

Proposition 5.1. *Suppose that F is continuous and Θ is compact. If θ^* is the unique minimizer of F , then for any $\varepsilon > 0$, we have*

$$\inf_{\theta \in \Theta: \|\theta - \theta^*\|_2 \geq \varepsilon} F(\theta) > F(\theta^*).$$

Proof: F is a continuous function on the domain

$$\Theta_\varepsilon := \{\theta \in \Theta : \|\theta - \theta^*\|_2 \geq \varepsilon\},$$

which is also compact. So F achieves its minimum on Θ_ε , and the minimum is strictly larger than $F(\theta^*)$ by uniqueness of the minimizer.

Specializing to MLE:

$$\begin{aligned} F(\theta) &= -\mathbb{E}_{\theta^*}[\log p_\theta(X)] \\ &= \underbrace{\mathbb{E}_{\theta^*}[\log p_{\theta^*}(X)]}_{\text{constant, independent of } \theta} + \underbrace{\mathbb{E}_{\theta^*}\left[\log \frac{p_{\theta^*}(X)}{p_\theta(X)}\right]}_{D_{\text{KL}}(\mathbb{P}_{\theta^*} \parallel \mathbb{P}_\theta)}. \end{aligned}$$

Kullback-Leibler divergence is always non-negative, and it is zero if and only if $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$. So if the parametric model is identifiable, i.e., $\theta_1 \neq \theta_2$ implies $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$, then this assumption is satisfied.