

Lecture 9: Law of large numbers and estimation consistency

Lecturer: Yiqiao Zhong

Last update: 00:25:59, 2024-09-24

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

In this lecture, we will introduce basic asymptotic approach for statistical theory, and the underlying probability foundation behind it.

On notations. You will see different notations for the same thing in different literature. Expectation can be written as $\mathbb{E}, E, \mathbf{E}$, probability measure can be written as $\mathbb{P}, P, \mathbf{P}$, random variables are both uppercase/lowercase letters, real values can be either $\mathbb{R}, R$ or $\mathbf{R}$, indicator functions are written as $\mathbb{I}_{\{A\}}$ and $\mathbf{1}_A$, etc. For statistics papers, the standard notations are mostly consistent with Jun Shao’s textbook or an average paper in Annals of Statistics; for machine learning papers, a good reference is this page.

1 Exact inference vs. asymptotic approach

How do we measure the quality of an estimator? If the estimator has a simple closed form and data are drawn from a given parametric function, then we can calculate the bias, variance, and mean squared error explicitly. However, in general, data may come from complicated distributions, and estimators may not have a closed form.

1. Exact inference: often in simple problems. e.g. variance of sample mean, very complicated in complex problems.
For example, t -test is based on Gaussian assumption and exact calculation.
2. Asymptotic approach.
 - (1) Sample size is large, and dimension is fixed or moderately increasing. The theoretical foundations include Law of Large Numbers (LLN) and Central Limit Theorem (CLT)—topics that we will learn in the next 3 lectures.
 - (2) Sample size and dimension are both large. The theoretical foundations include random matrix theory and empirical processes. We will cover a few selected topics at the end of the course. *This is one of the active research areas in theoretical statistics.*
3. Other approaches. Below are some classical alternatives.
 - (1) Finite sample correction, edgeworth expansion. It is useful when the sample size is small and asymptotics are not accurate.
 - (2) Bootstrap. Useful when we have good computing power for subsampling data.
 - (3) Bayesian approach. Easy to incorporate into structured domain knowledge, widely used in the applied world, but can be computationally intensive.

2 Convergence in probability

Since estimators are functions of random variables (e.g., noisy measurements), we need the right language to talk about convergence when the sample size is large. Below, the first definition—almost sure (a.s.) convergence—is already covered in Lecture 4.

Definition 1. Let $X, X_1, X_2, \dots \in \mathbb{R}^k$ be random vectors defined on a probability space.

(i) We say that the sequence $\{X_n\}$ converges to X a.s. if and only if $\mathbb{P}(\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)) = 1$. We use the notations $X_n \xrightarrow{\text{a.s.}} X$ or $\lim_{n \rightarrow \infty} X_n = X$ a.s. to denote almost sure convergence.

(ii) We say that $\{X_n\}$ converges to X in probability if and only if, for every fixed $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\|X_n - X\|_2 > \epsilon) = 0.$$

We use the notation $X_n \xrightarrow{P} X$ to denote convergence in probability.

Almost sure convergence implies convergence in probability. If $\lim_{n \rightarrow \infty} X_n = X$ a.s., then the set $\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}$ has probability one. For ω in this set, in the ϵ - δ language, for any $\epsilon > 0$, there exists $N > 0$ such that $\|X_n(\omega) - X(\omega)\|_2 \leq \epsilon$ for all $n \geq N$. We convert this to set theory ($\exists \leftrightarrow \cup$, $\forall \leftrightarrow \cap$), $\lim_{n \rightarrow \infty} X_n = X$ a.s. holds iff

$$\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : \|X_n(\omega) - X(\omega)\|_2 \leq \epsilon\}$$

has probability one for any $\epsilon > 0$. Denote $A_{N,\epsilon} = \{\omega : \|X_n(\omega) - X(\omega)\|_2 \leq \epsilon\}$ and $B_{N,\epsilon} = \bigcap_{n=N}^{\infty} A_{n,\epsilon}$. Since $B_{1,\epsilon} \subseteq B_{2,\epsilon} \subseteq B_{3,\epsilon} \subseteq \dots$, we have (exercise: as a consequence of MCT)

$$P(\bigcup_{N=1}^{\infty} B_{N,\epsilon}) = 1 \iff \lim_{N \rightarrow \infty} P(B_{N,\epsilon}) = 1.$$

Since $B_{N,\epsilon} \subseteq A_{n,\epsilon}$, then we have

$$\begin{aligned} \lim_{N \rightarrow \infty} P(B_{N,\epsilon}) = 1 &\implies \lim_{N \rightarrow \infty} P(A_{N,\epsilon}) = 1 \\ &\iff \lim_{n \rightarrow \infty} P(\|X_n(\omega) - X(\omega)\|_2 > \epsilon) = 0 \end{aligned} \tag{1}$$

where we take set complement and replace N by n in the last expression.

Remark 1. From this proof, we can see that almost sure convergence requires that all X_n to work together, which is stronger, yet convergence in probability is a property about each individual $X_n - X$, which is weaker.

When do we obtain almost sure convergence? If only we can inverse the direction in Eqn. 1! If a stronger property is satisfied, then it is true:

$$\lim_{N \rightarrow \infty} \sum_{n=N}^{\infty} P(A_{n,\epsilon}^c) = 0 \implies \lim_{N \rightarrow \infty} P(B_{N,\epsilon}^c) = 0 \iff \lim_{N \rightarrow \infty} P(B_{N,\epsilon}) = 1.$$

The above implication is a consequence of **Borel–Cantelli lemma**. You may also prove it using Fubini's theorem (exercise).

Introducing o_p and O_p notations. Recall the classical little- o and big- O notations: Let $\{a_n\}$ be a sequence of real numbers, and $\{b_n\}$ be a sequence of positive numbers, we denote

$$\begin{cases} a_n = O(b_n) & \text{if there exists } C > 0 \text{ such that } |a_n| \leq Cb_n & \forall n \\ a_n = o(b_n) & \text{if } a_n/b_n \rightarrow 0 & \text{as } n \rightarrow \infty \end{cases}$$

Definition 2. Let $X_1, X_2, \dots$ be random vectors and $Y_1, Y_2, \dots$ be random variables defined on a common probability space.

(i) $X_n = O_p(Y_n)$ if and only if, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that

$$\sup_n P(\|X_n\| \geq C_\epsilon |Y_n|) < \epsilon.$$

(ii) $X_n = o_p(Y_n)$ if and only if $\|X_n\|/Y_n \xrightarrow{P} 0$ as $n \rightarrow \infty$.

In particular, $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$.

Example 1. Suppose that $X_n \sim N(0, \sigma_n^2)$ for $n = 1, 2, 3, \dots$. If $\lim_{n \rightarrow \infty} \sigma_n^2 = \sigma^2 < \infty$, then $X_n = O_p(1)$; and if $\lim_{n \rightarrow \infty} \sigma_n^2 = \infty$, then $\{X_n\}$ is not bounded in probability (*exercise*).

The following result describes the relationship among $O_p(\cdot)$, $o_p(\cdot)$ and convergence modes.

Proposition 1. Let $X, X_1, X_2, \dots$ be random vectors and $Y_1, Y_2, \dots$ be random variables defined on a common probability space.

(i) If $X_n = o_p(Y_n)$, then $X_n = O_p(Y_n)$.

(ii) If $X_n = O_p(1)$ and $Y_n = O_p(1)$, then $X_n Y_n = o_p(1)$.

3 Estimation consistency

Suppose that we have some data $X^{(n)}$ for each n drawn from a distribution $P \in \mathcal{P}$, and an estimator $T_n(X^{(n)}) \in \mathbb{R}^p$ based on it. Usually we suppress the dependence of data on n , so we simply write $T_n(X)$.

Definition 3. (Consistency) Suppose that $T_n(X)$ is an estimator of unknown parameter $\nu \in \mathbb{R}^p$. Let $(a_n)_{n \geq 1}$ be a sequence of positive numbers with $a_n \rightarrow \infty$ as $n \rightarrow \infty$.

(i) $T_n(X)$ is consistent for ν if and only if $T_n(X) \xrightarrow{P} \nu$ holds for any $P \in \mathcal{P}$.

(ii) $T_n(X)$ is a_n -consistent for ν if and only if $a_n[T_n(X) - \nu] = O_p(1)$ holds for any $P \in \mathcal{P}$.

(iii) $T_n(X)$ is strongly consistent if and only if $T_n(X) \xrightarrow{a.s.} \nu$ holds for any $P \in \mathcal{P}$.

Note that in (non-Bayesian) statistical theory, $P \in \mathcal{P}$ is unknown. Consistency requires that convergence holds for every $P \in \mathcal{P}$.

Remark 2. Consistency may not be an obvious notation when we have a dataset of a fixed sample size n . When people talk about consistency, usually we imagine there is a sequence of similar problems with increasing sample sizes n , and the same method or algorithm yields an estimator $T_n(X)$ for each n .

Notably, the **mindset is different** in computer science, where it is uncommon to adopt this asymptotic view. Instead, people want to know the guarantee for fixed n using nonasymptotic statements, e.g., the error being small than ϵ holds with probability at least $1 - 2/n$.

In classical statistics, consistency is usually a basic expectation for an estimation procedure to be useful. In later lectures, we will learn different but important philosophies such as in shrinkage estimation and high-dimensions statistics.

4 Law of large numbers

First we state a version of LLN that is simple to prove. It requires a stronger moment assumption (finite second moment) compared with the standard LLN, but no independence is required.

Consider a sequence of random variables $X_1, X_2, X_3 \dots \in \mathbb{R}$ defined on a probability space.

Theorem 1 (LLN: simple version). *Suppose that $X_1, X_2, X_3 \dots$ all have finite second moment with the same mean $\mu = \mathbb{E} X_i$ and bounded variance $\sigma_i^2 = \text{Var}(X_i) \leq C$ for all i . Further assume that any pair of two random variables are uncorrelated. Then*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)^2 \right] = 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu.$$

Proof. We use linearity of the expectation to expand the sum.

$$\text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] + \frac{1}{n^2} \sum_{i \neq j} \mathbb{E}[(X_i - \mu)(X_j - \mu)] = \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \leq \frac{C}{n} \quad (2)$$

where we use the assumption $\mathbb{E}[(X_i - \mu)(X_j - \mu)] = 0$ above. This proves the first result. Now denote $Y = n^{-1} \sum_{i=1}^n X_i - \mu$. The second result follows from the first result and Chebyshev's inequality introduced below. $\square$

Theorem 2 (Chebyshev's inequality). *If Y is a random variable satisfying $\mathbb{E}(Y^2) \leq \infty$ and $a > 0$, then*

$$\mathbb{P}(|Y - \mathbb{E} Y| \geq a) \leq \frac{\text{Var}(Y)}{a^2}.$$

Proof of Theorem 2.

$$\mathbb{P}(|Y - \mathbb{E} Y| \geq a) = \mathbb{P}(|Y - \mathbb{E} Y|^2 \geq a^2) \leq \frac{\mathbb{E}(|Y - \mathbb{E} Y|)^2}{a^2}$$

where the inequality is due to Markov's inequality. $\square$

Remark 3. (i) Recall that we implicitly used a fact stated earlier: the moment of a random variable is finite if it has a higher finite moment. Namely,

$$\left(\mathbb{E} \|X\|_p^p \right)^{1/p} \leq \left(\mathbb{E} \|X\|_q^q \right)^{1/q}, \quad \text{for all } 1 \leq p \leq q \leq \infty.$$

(ii) Under the functional analysis perspective, each random variable X with a finite second moment is treated as an element in the square integrable space, which is denoted by $X \in L^2(\Omega)$. This space is equipped with the L_2 norm $\|X\|_{L^2(\Omega)} = (\mathbb{E}[|X|^2])^{1/2}$, or simply denoted by $\|X\|_2$. Then Theorem 1 says $n^{-1} \sum_{i=1}^n X_i \xrightarrow{L^2} \mu$.

We assumed finite second moments in Theorem 1, which imply the existence of first moments, but the converse is not true (exercise). It turns out that we can establish the same result without assuming second moments.

Theorem 3. *Let $X_1, X_2, \dots$ be i.i.d. random variables, and $a_n = \mathbb{E}[X_1 \mathbb{I}_{\{|X_1| \leq n\}}]$. Suppose $n P(|X_n| > n) \rightarrow 0$ as $n \rightarrow \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^n X_i - a_n \xrightarrow{p} 0. \quad (3)$$

As a consequence, we obtain LLN: if $\mathbb{E}(|X_1|) < \infty$, then $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}(X_1)$.

Proof. The proof of the ‘as a consequence’ part is easy. In fact, we can use DCT to prove $\lim_n a_n = \mathbb{E}(|X_1|)$ and $\lim_n nP(|X_n| > n) = 0$; **exercise**: try proving this by using the observations $|X_1|\mathbb{I}_{\{|X_1| \leq n\}} \leq |X_1|$ and $n\mathbb{I}_{\{|X_n| > n\}} \leq |X_n|$.

For the major part, we use the **truncation argument** commonly used in probability theory. Consider a sequence of random variables obtained by truncating X_j ’s at n : $Y_{nj} = X_j\mathbb{I}_{\{|X_j| \leq n\}}$. Then

$$P\left(\underbrace{\frac{1}{n} \sum_{j=1}^n X_j}_{S_n} \neq \underbrace{\frac{1}{n} \sum_{j=1}^n Y_{nj}}_{T_n}\right) \stackrel{(i)}{\leq} \sum_{j=1}^n P(X_j \neq Y_{nj}) \stackrel{(ii)}{=} nP(|X_j| \geq n) \rightarrow 0. \quad \text{as } n \rightarrow \infty. \quad (4)$$

where (i) holds because of **union bound**: for events $A_1, \dots, A_n$, we have $P(\cup_{j=1}^n A_j) \leq \sum_{j=1}^n P(A_j)$ and (ii) holds because of the i.i.d. assumption.

Thus, we only need to work on the truncated variables, namely showing $\lim_n P(|T_n - \mathbb{E} T_n| \geq \epsilon) = 0$, where we denote $T_n = n^{-1} \sum_{j \leq n} Y_{nj}$. For any $\epsilon > 0$,

$$\begin{aligned} P(|T_n - \mathbb{E} T_n| \geq \epsilon) &\leq \frac{\text{Var}(Y_{nj})}{n\epsilon^2} \leq \frac{\mathbb{E}(Y_{nj}^2)}{n\epsilon^2} \\ &\leq \frac{1}{n\epsilon^2} \mathbb{E}[\min\{|X_j|, n\}^2] = \frac{1}{n\epsilon^2} \mathbb{E} \left[\int_0^n 2t \mathbb{I}_{\{|X_j| \geq t\}} dt \right] \\ &= \frac{2}{n\epsilon^2} \int_0^n t P(|X_j| \geq t) dt \rightarrow 0. \quad (\text{by Fubini's theorem and the assumption}) \end{aligned}$$

(**exercise**: Please check each step and show why it is true.) This proves that $T_n - \mathbb{E}(T_n) \xrightarrow{P} 0$, which together with (1), (2) and the fact that $\mathbb{E}(T_n) = \mathbb{E}(Y_{1j}) = \mathbb{E}(|X_1|\mathbb{I}_{\{|X_1| \leq n\}})$ imply the result. $\square$

Remark 4 (Why the proof works?). Let’s decipher the proof strategy of LLN. Our main plan stays the same as before, which is calculating the variance of an averaged random variable so that the cross terms cancel out due to independence. To solve the issue of lack of finite second moments, we used several observations.

- The existence of the first moment implies decay of tail probability $P(|X_n| > n) \ll n^{-1}$.
- Truncation does not alter the probability too much as shown in Eqn. 4.
- Truncation results in finite second moment $\mathbb{E}(Y_{nj}^2)$ but the variance is increasing in n . We need calculation to show that $\mathbb{E}(Y_{nj}^2) \ll n$.

Remark 5. LLN can be easily extended from random variables to random vectors, under the assumption $\mathbb{E}(\|X_1\|_2) < \infty$.

5 Strong LLN

A question: Can we strengthen *convergence in probability* to *a.s. convergence*?

A motivating phenomenon: Kolmogorov’s 0-1 law*. First let us convince ourselves that a.s. convergence is plausible. Let $X_1, X_2, \dots$ be random variables, define $\mathcal{F}'_n = \sigma(X_n, X_{n+1}, \dots)$ ¹. $\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{F}'_n$ is called tail σ -algebra. Intuitively, an event is in the tail σ -algebra if it does not depend on any finite number of variables. A limiting variable is often in this σ -algebra. We give an example to help explain the definition.

¹Recall $\sigma(X_n) = \{X_n^{-1}(B) : B \in \mathcal{B}\}$. We may put an infinite number of random variables to define a sigma-algebra generated by individual sigma-algebras.

Example 2. Let $X_1, X_2, \dots$ be independent random variables, then

$$\left\{ \limsup_n \frac{X_1 + X_2 + \dots + X_n}{n} \geq t \right\} \in \mathcal{T}.$$

A famous result by Kolmogorov reveals surprising properties about tail σ -algebra.

Theorem 4 (Kolmogorov's 0-1 Law). If $X_1, X_2, \dots$ are independent random variables, and $A \in \mathcal{T}$, then $\mathbb{P}(A) = 0$ or 1.

*Proof idea.*² We can show that A is independent of itself (what?!), so $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A^2)$, then $\mathbb{P}(A) = 0$ or 1. $\square$

Kolmogorov's 0-1 Law provides good evidence that convergence in probability can be improved to a.s. convergence, because events

$$\left\{ \limsup_n \frac{X_1 + X_2 + \dots + X_n}{n} \geq \mathbb{E}[X_1] + \varepsilon \right\}, \quad \left\{ \liminf_n \frac{X_1 + X_2 + \dots + X_n}{n} \leq \mathbb{E}[X_1] - \varepsilon \right\}$$

should have probabilities 0 or 1 for every $\varepsilon > 0$. We can reasonably expect the probability is 1 instead of 0, and if so, choosing a sequence (ε_M) with $\lim_{M \rightarrow \infty} \varepsilon_M = 0$ and setting $M \rightarrow \infty$ yields the a.s. convergence ([exercise](#)).

Strong LLN and proof idea.

Theorem 5 (Strong LLN). Let $X_1, X_2, \dots$ be i.i.d. random variables, Suppose $\mathbb{E}(|X_1|) < \infty$ and $c_i \geq 1$ is bounded for i , i.e. $\sup_i |c_i| < \infty$. Then

$$\frac{1}{n} \sum_{i=1}^n c_i (X_i - \mathbb{E}(X_i)) \xrightarrow{a.s.} 0.$$

Remark 6. $\mathbb{E}(|X_1|) < \infty$ can not be removed.³

Proof sketch^{*}. Use truncation + subsequencing + BC Lemma. Assume $c_i \equiv 1$ for simplicity.

For any $\alpha > 1$, let $k_n = \lfloor \alpha^n \rfloor$, $Y_n = X_n \mathbb{I}_{\{|X_n| \leq n\}}$, $T_n = \frac{1}{n} \sum_{i=1}^n Y_n$, then

$$\sum_{n=1}^{\infty} \mathbb{P}(|Y_{k_n} - \mathbb{E}(Y_{k_n})| > \epsilon k_n) \leq \sum_{n=1}^{\infty} \frac{\text{Var}(Y_{k_n})}{\epsilon^2 k_n^2} \quad (\text{by Chebeshev's inequality}). \quad (5)$$

By direction calculation and analysis (which we skip), we can show the RHS of (5) is finite.⁴ Then by BC lemma, we have:

$$\frac{T_{k_n} - \mathbb{E}(T_{k_n})}{k_n} \xrightarrow{a.s.} 0.$$

This implies

$$\frac{1}{\alpha} \mathbb{E}(X_1) \leq \liminf_n \frac{T_n}{n} \leq \limsup_n \frac{T_n}{n} \leq \alpha \mathbb{E}(X_1) \quad \text{a.s.}$$

The proof is complete by letting $\alpha \rightarrow 1$. $\square$

Remark 7. This proof contains many standard techniques in probability theory: truncation is often used to weaken moments assumptions, BC lemma is usually used to prove a.s. convergence, etc.

²Note: See Durrett[1] Thm 2.5.3. for more details.

³Note: See Durrett Thm 2.3.7. for more details.

⁴Note: See Durrett Lemma 2.4.2-2.4.4 for more details.

6 Uniform LLN*

First, we show a classic result which states that convergence in LLN is uniform in terms of (empirical) distribution functions.

Theorem 6 (Glivenko-Cantelli Theorem). *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with cdf $F(x)$, the empirical distribution function of $X_1, X_2, \dots, X_n$ is defined as $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}$ (which is random), then*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty.$$

Proof for continuous $F(x)$ case. The main idea is to apply SLLN on a discretized set over $\mathbb{R}$. Fix any $\epsilon > 0$. Pick $x_1 < x_2 < \dots < x_k \in \mathbb{R}$ such that $F(x_1) < \epsilon$, $1 - F(x_k) < \epsilon$, and $F(x_j) - F(x_{j-1}) < \epsilon$ for $j = 2, 3, \dots, k$. For any x_j , apply SLLN:

$$\begin{aligned} F_n(x_j) &= \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x_j\}} \xrightarrow{a.s.} \mathbb{E}(\mathbb{I}_{\{X_1 \leq x_j\}}) = F(x_j) \\ \implies \sup_{x \in \{x_1, x_2, \dots, x_k\}} |F_n(x) - F(x)| &\xrightarrow{a.s.} 0. \end{aligned}$$

If $x \in (x_j, x_{j+1})$, then

$$F_n(x) - F(x) \leq F_n(x_{j+1}) - F(x_{j+1}) + \sup_{x \in (x_j, x_{j+1})} |F(x) - F(x_{j+1})| \leq 2\epsilon.$$

Similarly, we can obtain $F(x) - F_n(x) \leq 2\epsilon$. If $x \in (-\infty, x_1)$ or $x \in (x_k, \infty)$, similarly we have

$$|F_n(x) - F(x)| \leq \sup_{x \in \{x_1, x_2, \dots, x_k\}} |F_n(x) - F(x)| + \epsilon.$$

Thus, we have

$$\limsup_x |F_n(x) - F(x)| \leq 2\epsilon \quad \text{a.s.}$$

The proof is complete since ϵ can be arbitrary. $\square$

Remark 8. *The above theorem provides a rigorous statement for the intuition that “empirical histogram converges to a limit”.*

Theorem 7 (Uniform LLN). *Let $g : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}$ be a function such that: $g(x, \theta)$ is continuous of θ for all x 's, and $g(x, \theta)$ is a measurable function on x for all θ 's, where $\Theta \subset \mathbb{R}$ is compact. Let $X_1, X_2, \dots \stackrel{\text{i.i.d.}}{\sim} X$. If there exists a function h such that: $|g(x, \theta)| \leq h(x)$ and $\mathbb{E}(|h(X)|) < \infty$, then*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n g(X_i, \theta) - \mathbb{E}[g(X, \theta)] \right| \xrightarrow{a.s.} 0. \quad (6)$$

Proof idea. It is clear that (4) holds for any $\Theta_0 \subset \Theta$ with finite cardinality. To establish a continuity result, let $B_r(\theta_0)$ be the ball of radius r and center θ_0 , and $\mathcal{N}_r(\theta_0) = B_r(\theta_0) \cap \Theta$. Let $\bar{\theta}_r(x)$ be a maximizer of $g(x, \theta)$ over $\mathcal{N}_r(\theta_0)$ for fixed x . Apply SLLN, we have

$$\frac{1}{n} \sum_{i=1}^n g(X_i, \bar{\theta}_r(X_i)) \xrightarrow{a.s.} \mathbb{E}[g(X, \bar{\theta}_r(X))]. \quad (7)$$

By DCT, the RHS of (5) converges to $\mathbb{E}[g(x, \theta_0)]$ as $r \rightarrow 0$.

Use this continuity idea and compactness of Θ , we can obtain ULLN. $\square$

7 Consistency of MoM and MLE

A simple consequence of LLN and SLLN is the following example.

Consistency of MoM.

Example 3 (Example 2.33 of [3] continued). Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P \in \mathcal{P}$. Assume that the first moment of P exists. Let μ be the mean of P namely $\mu = \mathbb{E}_{X \sim P}(X)$.

1. By the LLN, $\bar{X}$ is consistent for μ .
2. By the SLLN, $\bar{X}$ is strongly consistent for μ .
3. Further assume that the second moment of P exists. Denote the variance $\sigma^2 = \text{Var}_{X \sim P}(X)$. Then, both the sample variance estimator and empirical variance estimator

$$S_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \frac{n-1}{n} S_n$$

are (strongly) consistent for σ^2 (*exercise*).

This example is a special case of consistency of method-of-moments estimators. Let $\mu \in \mathbb{R}^k$ be the first k -th moments of P_θ (assuming existence) and $h^{-1} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ be a continuous bijection mapping moments to parameters: $\theta = h^{-1}(\mu)$. Denote by $\hat{\mu}_n$ the empirical moments based on n i.i.d. random variables drawn from P_θ . Then under the assumption on finite moments,

$$\hat{\mu}_n \xrightarrow{a.s.} \mu \quad \implies \quad h^{-1}(\hat{\mu}_n) \xrightarrow{a.s.} h^{-1}(\mu) = \theta \quad \text{as } n \rightarrow \infty$$

Consistency of MLE. In general, MLE is consistent under some conditions, which will be a topic in STAT 710. Here we give a high-level argument. For i.i.d. data, the negative log-likelihood function (with rescaling) is the sum of n independent terms for every $\theta \in \Theta$,

$$\ell_n(\theta; y) = \frac{1}{n} \sum_{i=1}^n g(X_i, \theta).$$

For example, for linear regression with independent Gaussian observations $X_i \sim N(\theta^\top Z_i, \sigma^2)$ where $z_i \in \mathbb{R}^p$ is fixed and deterministic covariates, we have $g(X_i, \theta) = \|X_i - \theta^\top Z_i\|_2^2$ and $\Theta = \mathbb{R}^p$. For general MLE, under some conditions,

$$\hat{\theta}_n(X) = \arg \min_{\theta \in \Theta} \ell_n(\theta; y) \xrightarrow{P} \arg \min_{\theta \in \Theta} \mathbb{E} [\ell_n(\theta; y)] \quad \text{as } n \rightarrow \infty$$

as a consequence of Uniform LLN and (locally) strict convex properties of the expected function.

Under some regularity conditions, MLE is $\sqrt{n}$ -consistent. However, the following shows that faster consistency is possible (in fact, lower rate is also possible).

Example 4 (MLE: Example 2.34 of [3]). Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([0, \theta])$. We can show that $X_{(n)} = \max_{1 \leq i \leq n} X_i$ is the MLE for estimating θ , and $X_{(n)} - \theta = O_P(\frac{1}{n})$ (*exercise*). So, X_n is n -consistent.

8 Further reading

LLN is a classic topic in probability theory. A reference is [1][Chap. 2.2—2.4]. Kolmogorov's 0-1 law can be found in [1][Chap. 2.5]. The statement of ULLN can be found in [2].

References

- [1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [2] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.
- [3] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.