ST5215 Advanced Statistical Theory, Lecture 21

HUANG Dongming

National University of Singapore

29 Oct 2020

Overview

Last time

- A Uniform Strong Law of Large Numbers
- Consistency of MLEs, M-estimators
- Kullback-Leibler information

Today

- Roots of the Likelihood Equation (RLE)
- Asymptotic Normality of RLEs and MLEs

Recap: Uniform Strong Law of Large Numbers

Theorem (USLLN, C)

Let $X_1, \ldots, X_n, \ldots,$ be i.i.d. sample from P and $U(x, \theta)$ be measurable on $\mathcal{X} \times \Theta$. Suppose

1. *$U(x, \theta)$ is a continuous in θ for any fixed x and for each θ, $\mu(\theta) = EU(X, \theta)$ is finite,*
2. *Θ is compact,*
3. *there exists a function $M(x)$ such that $EM(X) < \infty$ and $|U(x, \theta)| \leq M(x)$ for all x and θ.*

Then

$$P\left\{\lim_{n\to\infty} \sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{1}^{n} U\left(X_j, \theta\right) - \mu(\theta)\right| = 0\right\} = 1$$

Recap: KL Divergence

For two density functions f_0, f_1 defined on a common measure space, the *Kullback-Leibler information number* (or *KL divergence*) is defined as

$$K(f_0, f_1) = E_0 \log \frac{f_0(X)}{f_1(X)} = \int \log \frac{f_0(x)}{f_1(x)} f_0(x) d\nu(x).$$

Theorem

Shannon-Kolmogorov Information Inequality $K(f_0, f_1) \geq 0$ *with equality if and only if* $f_1(\omega) = f_0(\omega)$ ν*-a.e.*

Recap: Consistency of MLEs

Theorem (Continuous in θ)

Let $X_1, X_2, \ldots$ be i.i.d. from P, where P is in the family with density $f(x \mid \theta)$, $\theta \in \Theta$. Let θ_* denote the true value of θ.
Suppose

1. Θ *is compact,*
2. $f(x \mid \theta)$ *is continuous in* θ *for all* x*,*
3. *there exists a function* $M(x)$ *such that* $E_{\theta_*}|M(X)| < \infty$ *and*

$$\log f(x \mid \theta) - \log f(x \mid \theta_*) \leq M(x), \qquad \text{for all } x \text{ and } \theta$$

4. *(identifiability)* $f(x \mid \theta) = f(x \mid \theta_*)$ ν*-a.e.* $\Rightarrow \theta = \theta_*$.

Then, for any sequence of maximum-likelihood estimates $\hat{\theta}_n$ *of* θ*,*

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_*$$

Roots of the Likelihood Equation

- Assume the density $f(x \mid \theta)$ is differential in θ for any fixed x
- We often obtain the MLE by enumerating the roots of the likelihood equation (RLEs), i.e., θ that solves

$$\frac{\partial}{\partial \theta} \log L_n(\theta) = \mathbf{0} \tag{1}$$

- It turns out that there exists a sequence of RLEs that is strong consistent under some regularity condition.

A key observation:

- Let B be a small ball center around θ_* and ∂B is the boundary
  - Not to confuse the boundary with the partial differential operator
- If we can show that

$$\sup_{\theta \in \partial B} \log L_n(\theta) < \log L_n(\theta_*),$$

  then $\log L_n(\theta)$ must achieve a local maximum within the interior of B, which must be an RLE
- Since B is small, we ensure this RLE is close to θ_*

Basic Regularity conditions

Let $X_1, X_2, \ldots$ be i.i.d. sample from $P \in \mathcal{P}$ with p.d.f. $f(x \mid \theta)\}$ w.r.t. a σ-finite measure ν and $\theta \in \Theta$. Let θ_* denote the true value of θ.

Suppose

1. Θ is a open subset of $\mathcal{R}^k$,
2. $f(x \mid \theta)$ is twice continuously differentiable in θ for all x, and

$$\frac{\partial}{\partial \theta} \int f(x \mid \theta) \, \mathrm{d}\nu = \int \frac{\partial}{\partial \theta} f(x \mid \theta) \, \mathrm{d}\nu$$

$$\frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta^\top} f(x \mid \theta) \, \mathrm{d}\nu = \int \frac{\partial^2}{\partial \theta \partial \theta^\top} f(x \mid \theta) \, \mathrm{d}\nu$$

3. Let $\Psi(x, \theta) = \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(x \mid \theta)$. There exists a constant c and a nonnegative function H such that $EH(X) < \infty$ and

$$\sup_{\|\theta - \theta_*\| < c} \|\Psi(x, \theta)\| \leq H(x)$$

4. (Identifiability) $f(x \mid \theta) = f(x \mid \theta_*)$ ν-a.e. $\Rightarrow \theta = \theta_*$.

Theorem (Consistency of RLEs)

Under the basic regularity conditions, there exists a sequence of $\hat{\theta}_n$ such that $\frac{\partial}{\partial\theta}\log L_n(\hat{\theta}_n) = \mathbf{0}$ and $\hat{\theta}_n \xrightarrow{a.s.} \theta_*$

- Define $B(\rho) = \{\theta \in \mathcal{R}^k : \|\theta - \theta_*\| \leq \rho\}$. By condition (1), Θ is open, so for ρ small enough, $B(\rho) \subset \Theta$.
- Fixed a small ρ. Let $W(\rho) = \partial B(\rho) = \{\theta \in \mathcal{R}^k : \|\theta - \theta_*\| = \rho\}$.
- Based on the key observation, we only need to show that for $U(x,\theta) := \log f(x \mid \theta) - \log f(x \mid \theta_*)$,

$$P\left(\limsup_n \sup_{\theta\in W(\rho)} \frac{1}{n}\sum_{i=1}^n U(X_i,\theta) < 0\right) = 1 \qquad (2)$$

- It follows that

$$P\left(\forall k \in \mathcal{N},\ \ \exists N_k \in N,\ \ s.t.\ \ \sup_{n\geq N_k}\sup_{\theta\in W(1/k)} \frac{1}{n}\sum_{i=1}^n U(X_i,\theta) < 0\right) = 1,$$

which shows the existence of $\hat{\theta}_n$ that is strong consistent

Proof of Equation (2) by USLLN

- Note that $W(\rho)$ is a bounded close subset of $\mathcal{R}^k$ and thus is compact
- The log density ratio $U(x, \theta)$ is continuous. We are left to show $U(x, \theta)$ is dominated by an integrable function
- By Taylor expansion of $U(x, \gamma)$ around θ_*, for any $\gamma \in W(\rho)$

$$U(x,\gamma) = U(x,\theta_*) + \left[\frac{\partial}{\partial\theta}U(x,\theta_*)\right]^\top (\gamma-\theta_*)...$$
$$+\frac{1}{2}(\gamma-\theta_*)^\top \left[\frac{\partial^2}{\partial\theta\partial\theta^\top}\log f(x\mid\theta)\right]\Big|_{\theta=\tilde{\theta}}(\gamma-\theta_*),$$

  where $\tilde{\theta}$ is some convex combination of γ and θ_*
- The first term $= 0$, the second term is bounded by $\rho\left(\sum_{j\le k}|\partial/\partial\theta_i \log f(x\mid\theta)|\right)\Big|_{\theta=\theta_*}$, which is integrable by condition (2)
- By condition (3), if $\rho < c$, then the third term is bounded by $\rho^2 H(x)$, which is integrable

Proof (Cont.)

- We apply the USLLN to conclude

$$P\left(\limsup_n \sup_{\theta\in W(\rho)} \frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta) \le \sup_{\theta\in W(\rho)} \mu(\theta)\right) = 1,$$

  where $\mu(\theta)$ is $EU(X,\theta) = -K\left(f_{\theta_*}, f_\theta\right)$
- Note that $\mu(\theta)$ is continuous (as showed in the USLLN) and $\mu(\theta) < 0$ for all $\theta \in W(\rho)$
- Since $W(\rho)$ is compact, the continuous function $\mu(\theta)$ achieves its maximum on $W(\rho)$.
- Therefore $\sup_{\theta\in W(\rho)} EU(X,\theta) < 0$ and Equation (2) is proved

Remarks

- If for n large, the RLE is unique, then we can say that the RLE is consistent
- If the RLE is not unique, then since we don't know the true value θ_*, we may not be able to choose the RLE that lies in the ball around the true value

Compare with the consistency of MLE in the last lecture

Previously

- The log density is only assume to be continuous and dominated by an integrable function. But Θ need to be a compact set or can be compactified
- The result applies to any sequence of MLEs, or even a sequence that nearly maximizes $\log L_n(\theta)/n$

Here

- We allow Θ to be an open set but assume that the log density function has second order derivates that are dominated by an integrable function
- The result is about a sequence of RLEs. If the MLE is always the unique RLE, then it ensures the consistency of the MLE

Asymptotic Normality of RLEs

Theorem

Assume the basic regularity conditions and also the following

5. *The Fisher information*

$$I(\theta) = \int \frac{\partial}{\partial \theta} \log f(x \mid \theta) \left[\frac{\partial}{\partial \theta} \log f(x \mid \theta \right]^{\top} \, \mathrm{d}\nu(x)$$

   is positive definite at $\theta = \theta_*$.

Then for any consistent sequence $\{\tilde{\theta}_n\}$ of roots of the likelihood equation, it holds that

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_*\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, [I(\theta_*)]^{-1})$$

- The existence of strong consistent sequence of RLEs has been proved
- If the MLE is consistent, and if for any n large enough, the MLE is an RLE, then the result here shows the asymptotic normality of the MLE

Proof

- Note that $\frac{\partial}{\partial \theta} \log L_n(\tilde{\theta}_n) = \mathbf{0}$
- Using the mean-value theorem for vector-valued functions, we obtain that

$$\mathbf{0} - \frac{\partial}{\partial \theta} \log L_n(\theta_*) = \left[\underbrace{\int_0^1 \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta^\top} \log L_n \left(\theta_* + t(\tilde{\theta}_n - \theta_*) \right) dt}_{\text{def. } (-n) \times \tilde{J}_n} \right]^\top \left(\tilde{\theta}_n - \theta_* \right)$$

- By CLT,

$$n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \theta_*) \xrightarrow{\mathcal{D}} N\left[\mathbf{0}, I(\theta_*)\right]$$

because $\mathbf{0} = E\left(\frac{\partial}{\partial \theta} \log f(X_i \mid \theta_*)\right)$ by condition (2) and $I(\theta_*) = \mathrm{Cov}\left(\frac{\partial}{\partial \theta} \log f(X_i \mid \theta_*)\right)$ by condition (5)

We have

$$n^{-1/2} \frac{\partial}{\partial \theta} \log L_n(\theta_*) = \tilde{J}_n \cdot n^{1/2} \left(\tilde{\theta}_n - \theta_* \right) \tag{3}$$

and

$$n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \theta_*) \xrightarrow{\mathcal{D}} N\left[\mathbf{0}, I(\theta_*)\right] \tag{4}$$

Suppose we can show $\tilde{J}_n \xrightarrow{\mathcal{P}} I(\theta_*)$

- Since $I(\theta_*)$ is positive definite, any matrix in its neighbor is invertible
- By continuous mapping, we have $\tilde{J}_n^{-1} \xrightarrow{\mathcal{P}} [I(\theta_*)]^{-1}$
- The theorem then follows from Eq (3) and (4) and Slutsky's theorem

Show $\tilde{J}_n \xrightarrow{\mathcal{P}} I(\theta_*)$

- Recall that $\Psi(x, \theta) = \frac{\partial^2}{\partial\theta\partial\theta^\top} \log f(x \mid \theta)$. Let $J_n(\theta) = \frac{1}{n}\sum_{j=1}^n -\Psi(X_j, \theta)$
- Then $\tilde{J}_n = \int_0^1 J_n\left(\theta_* + t(\tilde{\theta}_n - \theta_*)\right) dt$
- Since $\Psi(x, \theta)$ is continuous in θ (condition (2)) and each element of $\Psi(x, \theta)$ is dominated by an integrable function (condition (3) and $\rho < c$), we have that $J(\theta) := E\left[-\Psi(X, \theta)\right]$ is continuous in θ
- Note that $J(\theta_*) = I(\theta_*)$ because θ_* is the true value
- For any $\epsilon > 0$, $\exists \rho \in (0, c)$, s.t. for any $\theta \in B(\rho) := \{\theta : \|\theta - \theta_*\| \leq \rho\}$, we have $\|J(\theta) - I(\theta_*)\| < \epsilon$
- Since $B(\rho)$ is compact, we can apply USLLN to each elements of $\Psi(x, \theta)$ and conclude that

$$\lim_n \sup_{\theta \in B(\rho)} \|J_n(\theta) - J(\theta)\| = 0 \text{ a.s.}$$

- When $\tilde{\theta}_n \in B(\rho)$,

$$\tilde{J}_n - I(\theta_*) = \int_0^1 J_n\left(\theta_* + t(\tilde{\theta}_n - \theta_*)\right) - J\left(\theta_* + t(\tilde{\theta}_n - \theta_*)\right) dt + \int_0^1 J\left(\theta_* + t(\tilde{\theta}_n - \theta_*)\right) - I(\theta_*) dt,$$

  which follows that $\|\tilde{J}_n - I(\theta_*)\| \leq \sup_{\theta \in B(\rho)} \|J_n(\theta) - J(\theta)\| + \epsilon$

- Hence

$$\begin{aligned} & P(\|\tilde{J}_n - I(\theta_*)\| > 2\epsilon) \\ \leq & P\left(\tilde{\theta}_n \notin B(\rho)\right) + P\left(\sup_{\theta \in B(\rho)} \|J_n(\theta) - J(\theta)\| > \epsilon\right) \to 0 \end{aligned}$$

Example: Natural Exponential Families

Suppose that X_i's are i.i.d. sample from a population in a natural exponential family, i.e., the p.d.f. of X_i is

$$f_\eta(x_i) = \exp\{\eta^\tau T(x_i) - \zeta(\eta)\} h(x_i)$$

- Conditions (1,2,4) hold because of the properties of exponential families (Proposition 3.2)
- Since $\partial^2 \log f_\eta(x)/\partial\eta\partial\eta^\tau = -\partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau$ does not depend on x, Condition (3) holds
- Define the *score function* to be the gradient of the log likelihood

$$s_n(\eta) := \frac{\partial}{\partial\eta} \log L_n(\eta) = \sum_{i=1}^n \left[T(X_i) - \frac{\partial\zeta(\eta)}{\partial\eta} \right]$$

- Let $g(\eta) = \frac{\partial\zeta(\eta)}{\partial\eta}$ and $\hat{\mu}_n = n^{-1}\sum_{i=1}^n T(X_i)$
- The theorem says that when n is large, there exists $\hat{\eta}_n$ such that $g(\hat{\eta}_n) = \hat{\mu}_n$ and $\hat{\eta}_n \xrightarrow{a.s.} \eta$, $\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow{\mathcal{D}} N(0, [\frac{\partial^2}{\partial\eta\partial\eta^\top}\zeta(\eta)]^{-1})$

Tutorial

1. Exercise 3.6.34 in JS
2. Exercise 3.6.37 in JS
3. Prove the second Borel-Cantelli lemma: For a sequence of pairwisely independent events $\{A_n\}_{n=1}^{\infty}$, if $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \quad i.o.) = 1$.
4. Suppose X and $\{X_n\}$ are r.v.s. defined on a common probability space. Prove that if for any sub-sequence $\{X_{n_k}\}_{k=1}^{\infty}$ there exists a sub-sub-sequence $\{X_{n_{k_i}}\}_{i=1}^{\infty}$ such that $X_{n_{k_i}} \overset{a.s.}{\to} X$, then $X_n \overset{\mathcal{P}}{\to} X$.

Exercise 3.6.34 in JS

(Example 3.9 in JS) Let $X_1, \ldots, X_n$ be i.i.d. with the Lebesgue p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $f(x) > 0$ and $f'(x)$ exists for all $x \in \mathcal{R}$, $\mu \in \mathcal{R}$, and $\sigma > 0$ location-scale family). Let $\theta = (\mu, \sigma)$. Then, the Fisher information about θ contained in $X_1, \ldots, X_n$ is

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \int \frac{\left[f'(x)\right]^2}{f(x)} dx & \int \frac{f'(x)\left[xf'(x)+f(x)\right]}{f(x)} dx \\ \int \frac{f'(x)\left[xf'(x)+f(x)\right]}{f(x)} dx & \int \frac{\left[xf'(x)+f(x)\right]^2}{f(x)} dx \end{pmatrix}$$

Remark. We need to assume the integrals in the expression are finite and $\int |f'| \, \mathrm{d}x < \infty$, then we can check that the regularity conditions hold.

Proof:

- Since the data are i.i.d. and the , we only need to compute the Cov matrix of the gradient of the log likelihood of one sample
- Let $g(\mu, \sigma, x) = \log \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$. Then

$$\frac{\partial}{\partial \mu} g(\mu, \sigma, x) = -\frac{f'\left(\frac{x-\mu}{\sigma}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)}$$

and

$$\frac{\partial}{\partial \sigma} g(\mu, \sigma, x) = -\frac{(x-\mu) f'\left(\frac{x-\mu}{\sigma}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)} - \frac{1}{\sigma}$$

Exercise 1 (Cont.)

$$E\left[\frac{\partial}{\partial\mu}g\left(\mu,\sigma,X_1\right)\right]^2=\frac{1}{\sigma^2}\int\left[\frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)}\right]^2\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)dx$$

$$=\frac{1}{\sigma^2}\int\frac{\left[f'\left(\frac{x-\mu}{\sigma}\right)\right]^2}{f\left(\frac{x-\mu}{\sigma}\right)}d\left(\frac{x}{\sigma}\right)$$

$$=\frac{1}{\sigma^2}\int\frac{\left[f'(x)\right]^2}{f(x)}dx$$

$$E\left[\frac{\partial}{\partial\sigma}g\left(\mu,\sigma,X_1\right)\right]^2=\frac{1}{\sigma^2}\int\left[\frac{x-\mu f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)}+1\right]^2\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)dx$$

$$=\frac{1}{\sigma^2}\int\left[x\frac{f'(x)}{f(x)}+1\right]^2f(x)dx$$

$$=\frac{1}{\sigma^2}\int\frac{\left[xf'(x)+f(x)\right]^2}{f(x)}dx$$

Exercise 1 (Cont.)

$$
\begin{aligned}
& E\left[\frac{\partial}{\partial \mu} g\left(\mu, \sigma, X_1\right) \frac{\partial}{\partial \sigma} g\left(\mu, \sigma, X_1\right)\right] \\
=& \frac{1}{\sigma^2} \int \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)}\left[\frac{x-\mu}{\sigma} \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)}+1\right] \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \\
=& \int \frac{f'(x)\left[x f'(x)+f(x)\right]}{f(x)} dx
\end{aligned}
$$

Exercise 2

Let $X_1, \ldots, X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$ with $\theta > 0$
(a) Show that condition (3.3) does not hold for $h(X) = X_{(n)}$.
(b) Show that the inequality in (3.6) does not hold for the UMVUE of θ.

Proof: Part (a)

- Denote $f_\theta(x) = \theta^{-n} \prod_{i \le n} I_{x_i \le \theta}$ for $x = (x_1, \ldots, x_n)$ and ν be the Lebesgue measure on $\mathcal{R}^k$
- Using $P(X_{(n)} \le t) = \prod_i P(X_i \le t) = (t/\theta)^n$, we know that the p.d.f. of $X_{(n)}$ is $g_\theta(t) = n\theta^{-n} t^{n-1} I_{(0,\theta)}(t)$
- Then $E_\theta[h(X)] = \frac{n\theta}{n+1}$, and

$$\frac{d}{d\theta} \int h(x) f_\theta(x) \, \mathrm{d}\nu = \frac{n}{n+1}$$

- Note that $\frac{d}{d\theta} f_\theta(x) = \frac{-n}{\theta^{n+1}} \prod_{i \le n} I_{x_i \le \theta} = \frac{-n}{\theta} f_\theta(x)$, ν-a.e.
- So

$$\int h(x) \frac{d}{d\theta} f_\theta(x) \, \mathrm{d}\nu = -\frac{n}{\theta} E_\theta h(X) = -\frac{n^2}{n+1}$$

Part (b)

- The UMVUE of θ is $(n+1)X_{(n)}/n$ (proved in Lecture)
- It is straightfoward to compute $E_\theta X_{(n)}^2 = \frac{n\theta^2}{n+2}$, so the variance of the UMVUE is

$$\theta^2/[n(n+2)].$$

- On the other hand, $\frac{d}{d\theta}\log f_\theta(x) = \frac{-n}{\theta}$ ν-a.e. and thus the Fisher information is $I(\theta) = n^2/\theta^2$
- $[I(\theta)]^{-1} = \theta^2/n^2$ is larger than the variance of the UMVUE

Remark. Here the Fisher information is computed using the definition. Since the regularity condition does not hold, common properties of the Fisher information cannot be used.

Q3

Prove the second Borel-Cantelli lemma: For a sequence of pairwisely independent events $\{A_n\}_{n=1}^{\infty}$, if $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \quad i.o.) = 1$.

The key is use Chebyshev's inequality.

- Denote by $S_n = \sum_{i=1}^{n} I_{A_i}$ the number of events happens among the first n
- Let $p_i = P(A_i)$
- Let $\mu_n = ES_n = \sum_{i \le n} p_i$. We have $\mu_n \to \infty$
- By the pairwise independence, $\mathrm{Var}(S_n) = \sum_{i \le n} p_i(1 - p_i) \le \sum_{i \le n} p_i = \mu_n$
- For any $x > 0$, when n is sufficient large, $x < \mu_n$. By Chebyshev's inequality

$$P(S_n \le x) = P(\mu_n - S_n \ge \mu_n - x) \le \frac{\mathrm{Var}(S_n)}{(\mu_n - x)^2} \le \frac{\mu_n}{(\mu_n - x)^2},$$

  which goes to 0
- Note that $S_n \uparrow$. So we have $P(\lim_n S_n \le x) = 0$ for all $x > 0$, and thus $P(\lim_n S_n < \infty) = 0$
- That is $1 = P(\sum_{n=1}^{\infty} I_{A_n} = \infty) = P(A_n \quad i.o.)$

Q4

Suppose X and $\{X_n\}$ are r.v.s. defined on a common probability space. Prove that if for any sub-sequence $\{X_{n_k}\}_{k=1}^{\infty}$ there exists a sub-sub-sequence $\{X_{n_{k_i}}\}_{i=1}^{\infty}$ such that $X_{n_{k_i}} \overset{a.s.}{\to} X$, then $X_n \overset{\mathcal{P}}{\to} X$.

- Proof by contradiction
- If X_n does not converge to X in probability, there exists a subsequence X_{n_k} such that for some $\epsilon > 0$, $P(|X_{n_k} - X| > \epsilon) > \epsilon$ for all k
- But then there exists a sub-sub-sequence $\{X_{n_{k_i}}\}_{i=1}^{\infty}$ such that $X_{n_{k_i}} \overset{a.s.}{\to} X$. This implies that for i large enough, $P(|X_{n_{k_i}} - X| > \epsilon) < \epsilon$, which contradicts with the last bullet point

Remark. Now we can say that: $X_n \overset{\mathcal{P}}{\to} X$ if and only if for any sub-sequence, there exists a sub-sub-sequence that converges to X almost surely