High Dimensional Probability Notes

Nong Minh Hieu[1]

[1] School of Physical and Mathematical Sciences, Nanyang Technological University (NTU - Singapore)

Contents

↺ *Return to Contents*

1 Random variables

1.1 Basic Inequalities

First, we revisit the definition of a random variable as well as some basic inequalities that we learned in introductory statistics.

Definition 1.1 (Random variable). —————————————————————————
Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. A random variable X is defined as a mapping from the sample space Ω to $\mathbb{R}$:

$$X : \Omega \to \mathbb{R} \tag{1}$$

Σ is the σ-algebra containing the possible events (collection of subsets of Ω) and $\mathbb{P}$ is a probability measure that assigns events with probabilities:

$$\mathbb{P} : \Sigma \to [0, 1] \tag{2}$$

For a given probability space $(\Omega, \Sigma, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, we will use the following basic notations throughout this note:

- $\|X\|_{L^p}$ - The pth root of the pth moment of the random variable X.

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}, \ p \in (0, \infty) \tag{3}$$
$$\|X\|_{L^\infty} = \operatorname{ess\,sup} |X| \tag{4}$$

- $L^p(\Omega, \Sigma, \mathbb{P})$ - The space of random variables X satisfying:

$$L^p(\Omega, \Sigma, \mathbb{P}) = \left\{ X : \Omega \to \mathbb{R} \middle| \|X\|_{L^p} < \infty \right\} \tag{5}$$

Some basic inequalities and identities:

- **1. Jensen's Inequality** - For a random variable X and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, we have:

$$\varphi(\mathbb{E}X) \leqslant \mathbb{E}\varphi(X) \tag{6}$$

- **2. Monotonicity of L^p norm** - For a random variable X:

$$\|X\|_{L^p} \leqslant \|X\|_{L^q}, \ 0 \leqslant p \leqslant q \leqslant \infty. \tag{7}$$

- **3. Minkowski's Inequality** - For $1 \leqslant p \leqslant \infty$ and two random variables X, Y in $L^p(\Omega, \Sigma, \mathbb{P})$ space:

$$\|X + Y\|_{L^p} \leqslant \|X\|_{L^p} + \|Y\|_{L^p}. \tag{8}$$

- **4. Holder's Inequality** - For $p, q \in [1, \infty]$ such that $1/p + 1/q = 1$. Then, for random variables $X \in L^p(\Omega, \Sigma, \mathbb{P})$ and $Y \in L^q(\Omega, \Sigma, \mathbb{P})$, we have:

$$|\mathbb{E}XY| \leqslant \|X\|_{L^p} \cdot \|Y\|_{L^q}. \tag{9}$$

- **5. Markov's Inequality** - For a non-negative random variable X and $t > 0$, we have:

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t}. \tag{10}$$

We can also generalize Markov's Inequality for p^{th} moment:

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}[|X|^p]}{t^p}, \forall t > 0, k \in [2, \infty). \tag{11}$$

- **6. Chebyshev's Inequality** - For a random variable X with mean μ and variance σ^2. Then, for any $t > 0$, we have:

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant \frac{\sigma^2}{t^2}. \tag{12}$$

- **7. Integral Identity** - Let X be a non-negative random variable, we have:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt. \tag{13}$$

Exercises

Exercise 1.1.1: Generalized Integral Identity

Let X be a random variable (not necessarily non-negative). Prove the following identity:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt - \int_{-\infty}^0 \mathbb{P}(X < t)dt. \tag{14}$$

Solution (Exercise 1.1.1). ───────────────────
For $x \in \mathbb{R}$, using the basic integral indentity, we have:

$$|x| = \int_0^\infty \mathbf{1}\{t < |x|\}dt$$

We consider the following cases:

- *When $x < 0 \implies x = -|x|$:*

$$x = -\int_0^\infty \mathbf{1}\{t < |x|\}dt = -\int_0^\infty \mathbf{1}\{t < -x\}dt = -\int_0^\infty \mathbf{1}\{-t > x\}dt = -\int_{-\infty}^0 \mathbf{1}\{t > x\}dt.$$

- *When $x \geqslant 0 \implies x = |x|$:*

$$x = \int_0^\infty \mathbf{1}\{t < |x|\}dt = \int_0^\infty \mathbf{1}\{t < x\}dt.$$

Therefore, for $x \in \mathbb{R}$, we can write:

$$x = \int_0^\infty \mathbf{1}\{t < x\}dt - \int_{-\infty}^0 \mathbf{1}\{t > x\}dt.$$

Therefore, for a random variable X not necessarily non-negative, we have:

$$\mathbb{E}X = \mathbb{E}\left[\int_0^\infty \mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbf{1}\{t > X\}dt\right]$$

$$= \mathbb{E}\int_0^\infty \mathbf{1}\{t < X\}dt - \mathbb{E}\int_{-\infty}^0 \mathbf{1}\{t > X\}dt$$

$$= \int_0^\infty \mathbb{E}\mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbb{E}\mathbf{1}\{t > X\}dt$$

$$= \int_0^\infty \mathbb{P}(t < X)dt - \int_{-\infty}^0 \mathbb{P}(t > X)dt.$$

$\square$.

Exercise 1.1.2: p^{th}-moments via tails

Let X be a random variable and $p \in (0, \infty)$. Show that:

$$\mathbb{E}|X|^p = \int_0^\infty pt^{p-1}\mathbb{P}(|X| > t)dt. \tag{15}$$

Solution (Exercise 1.1.2).
Let X be a random variable that is not necessarily non-negative. Using the integral identity, we have:

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(u < |X|^p)du.$$

Let $t^p = u \implies pt^{p-1}dt = du$. Since we integrate u from $0 \to \infty$, we also integrate t from $0 \to \infty$ when changing the variables. Hence, we have:

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(t^p < |X|^p)pt^{p-1}dt = \int_0^\infty \mathbb{P}(t < |X|)pt^{p-1}dt.$$

Hence, we obtained the desired identity. □.

1.2 Limit Theorems

1.2.1 Weak Law of Large Numbers

Theorem 1.1: Weak Law of Large Numbers (WLLN)

Let $X_1, \ldots, X_N$ be *i.i.d* random variables with mean μ. Consider the sum:

$$S_N = X_1 + \cdots + X_N$$

Then, the sample mean **converges to μ in probability** $(S_N/N \xrightarrow{p} \mu)$:

$$\lim_{N\to\infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = 0, \ \forall \epsilon > 0 \tag{16}$$

Proof (Weak Law of Large Numbers (**WLLN**)).
We split the proof into two sections corresponding to the assumptions of finite variance and non-finite variance.

1. **Finite variance case**: *Suppose that* $\mathrm{Var}X_i = \sigma^2 < \infty$ *for all* $1 \leqslant i \leqslant N$. *Let* $\bar{X} = S_N/N$. *Then, $\bar{X}$ is a random variable with the following mean and variance:*

$$\mathbb{E}\bar{X} = \mu \quad and \quad \mathrm{Var}\bar{X} = \frac{\sigma^2}{N}.$$

Hence, by the Chebyshev's inequality, we have:

$$\mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = \mathbb{P}\Big(|\bar{X} - \mu| > \epsilon\Big) \leqslant \frac{\sigma^2}{N\epsilon^2}.$$

Therefore, we have:

$$\lim_{N\to\infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) \leqslant \lim_{N\to\infty} \frac{\sigma^2}{N\epsilon^2} = 0.$$

Hence, we have $\lim_{N\to\infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = 0$ *and we obtained* (**WLLN**).

4

2. **Non-finite variance case**: *In this case, we rely on the Levy Continuity Theorem* (**LCT**), *which relies on the convergence of the characteristic function. For $n \geqslant 1$, define the sequence of random variable $Y_n = S_n/n$. Hence, we have:*

$$
\begin{aligned}
\varphi_{Y_n}(t) &= \varphi_{S_n/n}(t) \\
&= \varphi_{S_n}(t/n) \\
&= \prod_{i=1}^{n} \varphi_{X_i}(t/n) = \left[\varphi_X(t/n) \right]^n,
\end{aligned}
$$

Where $X = X_1 = \cdots = X_n$. By Taylor's expansion, we have:

$$
\varphi_X(t/n) = 1 + \frac{it\mathbb{E}[X]}{n} + \mathcal{O}(1/n^2) = 1 + \frac{it\mu}{n} + \mathcal{O}(1/n^2).
$$

Hence, we have:

$$
\lim_{n\to\infty} \varphi_{Y_n}(t) = \lim_{n\to\infty} \left(1 + \frac{it\mu}{n} + \mathcal{O}(1/n^2) \right)^n = e^{it\mu}.
$$

Therefore, by (**LCT**), *we have $Y_n \xrightarrow{p} \mu$.*

Remark 1.1 (Taylor expansion of Moment Generating and Characteristic Functions). ———
Given a random variable X. For reference, the following are the Taylor expansions of the Moment Generating Function $M_X(t)$ and the Characteristic Function $\varphi_X(t)$:

$$
\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] = 1 + \sum_{n=1}^{\infty} \frac{t^n}{n!}\mathbb{E}[X^n], \\
\varphi_X(t) &= \mathbb{E}[e^{itX}] = 1 + \sum_{n=1}^{\infty} \frac{(it)^n}{n!}\mathbb{E}[X^n].
\end{aligned}
\tag{17}
$$

For the sake of my laziness, here are the Taylor expansion for the first three terms of both the MGF and the CF:

$$
\begin{aligned}
M_X(t) &= 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \mathcal{O}(t^3), \\
\varphi_X(t) &= 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + \mathcal{O}(t^3).
\end{aligned}
\tag{18}
$$

▫.

Theorem 1.2: Levy Continuity Theorem (**LCT**)

Let $X_1, X_2, \ldots$ be *i.i.d* random variables. Then:

$$
\forall t \in \mathbb{R} : \lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t) \iff X_n \xrightarrow{d} X,
\tag{19}
$$

for some random variable X. In a special case where $X = c$ for some $c \in \mathbb{R}$, we have:

$$
\forall t \in \mathbb{R} : \lim_{n\to\infty} \varphi_{X_n}(t) = e^{itc} \iff X_n \xrightarrow{p} c.
\tag{20}
$$

Proof (Levy Continuity Theorem (**LCT**)). ——————————————————————
The proof for (**LCT**) *can be found in Gut 2004, Section 9.1, Theorem 9.1 and Collorary 9.1* ▫.

1.2.2 Strong Law of Large Numbers

> **Theorem 1.3: Strong Law of Large Numbers (SLLN)**
>
> Let $X_1, \ldots, X_N$ be $i.i.d$ random variables with mean $\mu < \infty$. Consider the sum:
>
> $$S_N = X_1 + \cdots + X_N$$
>
> Then, the sample mean **converges to μ almost surely** ($S_N/N \xrightarrow{a.s} \mu$):
>
> $$\mathbb{P}\left(\limsup_{N \to \infty} |S_N/N - \mu| > \epsilon \right) = 0, \ \forall \epsilon > 0 \tag{21}$$

Proof (Strong Law of Large Numbers (**SLLN**)). ───────────────

*For the sake of simplicity, we will present the proof for (**SLLN**) with an additional assumption that $\mathbb{E}[|X_n|^4] < \infty, \forall n \geqslant 1$. The proof for the general case of (**SLLN**) (also called the Kolmogorov Strong Law) can be found in Gut 2004, Section 6, Theorem 6.1. For convenience, we assume the following:*

1. *$\mathbb{E}[|X_n|^4] = K < \infty$.*

2. *$\mathbb{E}[X_n] = 0$. For non-zero mean case, we can set $Y_n = X_n - \mu$ and repeat the same arguments made below.*

We aim to prove that $\mathbb{P}\left(\limsup_{N \to \infty} |S_N/N| > \epsilon \right) = 0$ for any $\epsilon > 0$. Firstly, use the Multinomial formula to expand $\mathbb{E}[S_n]$. The expansion will contain the terms in the following forms:

$$X_i^2, X_i^3 X_j, X_i^2 X_j^2, X_i^2 X_j X_k, X_i X_j X_k X_\ell,$$

where i, j, k, ℓ are distinct indices. By independence, we have:

$$\mathbb{E}[X_i^3 X_j] = \mathbb{E}[X_i^2 X_j X_k] = \mathbb{E}[X_i X_l X_k X_\ell] = 0.$$

As a result, we have the following remaining terms by the Multinomial formula:

$$\mathbb{E}[S_n^4] = \sum_{i=1}^{n} \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1 \leqslant i < j \leqslant n} \mathbb{E}[X_i^2 X_j^2]$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i^4] + 6 \underbrace{\sum_{1 \leqslant i < j \leqslant n} \mathbb{E}[X_i^2 X_j^2]}_{n(n-1)/2 \ terms}$$

$$= nK + 3n(n-1)\mathbb{E}[X_i^2 X_j^2].$$

By independence, we have $\mathbb{E}[X_i^2 X_j^2] = \mathbb{E}[X_i^2]\mathbb{E}[X_j^2]$ and for any $1 \leqslant i \leqslant n$. Furthermore, we have $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2$. Therefore:

$$\mathbb{E}[S_n^4] = nK + 3n(n-1)(\sigma^2 + \mu^2) < nK + 3n^2(\sigma^2 + \mu^2).$$

Applying Markov's Inequality with the fourth moment, we have:

$$\mathbb{P}(|S_n/n| \geqslant \epsilon) = \mathbb{P}(|S_n| \geqslant n\epsilon)$$

$$\leqslant \frac{\mathbb{E}[S_n^4]}{n^4 \epsilon^4}$$

$$< \frac{K}{n^3 \epsilon^4} + \frac{3(\sigma^2 + \mu^2)}{n^2}.$$

6

Therefore, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}(|S_n/n| \geqslant \epsilon) < \frac{K}{\epsilon^4} \sum_{n=1}^{\infty} n^{-3} + 3(\sigma^2 + \mu^2) \sum_{n=1}^{\infty} n^{-2} < \infty \tag{22}$$

*Finally, by the Borel-Cantelli Lemma (**BCL**), we have:*

$$\mathbb{P}\left(\limsup_{n \to \infty} |S_n/n| \geqslant \epsilon \right) = 0, \quad \forall \epsilon > 0.$$

□.

Theorem 1.4: Borel-Cantelli Lemma (BCL)

1. First Borel-Cantelli Lemma: Given a probability space $(X, \mathcal{S}, \mathbb{P})$ and a sequence $\{A_n\}_{n=1}^{\infty} \subset \mathcal{S}$. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, we have:

$$\mathbb{P}\left(\limsup_{n \to \infty} A_n \right) = 0. \tag{23}$$

2. Second Borel-Cantelli Lemma: On the other hand, if $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$, we have:

$$\mathbb{P}\left(\limsup_{n \to \infty} A_n \right) = 1. \tag{24}$$

Proof (Borel-Cantelli Lemma (**BCL**)). ────────────────────────────

We focus on proving the first Borel-Cantelli lemma. We define another sequence of $\mathcal{S}$-measurable sets $\{B_n\}_{n=1}^{\infty}$ such that:

$$B_n = \bigcup_{k=n}^{\infty} A_n.$$

Hence, we have $B_{\ell+1} \subset B_\ell$ for every $\ell \geqslant 1$. In other words, B_n is a decreasing sequence of $\mathcal{S}$-measurable sets. By continuity of measure, we have:

$$\mathbb{P}\left(\lim_{n \to \infty} B_n \right) = \lim_{n \to \infty} \mathbb{P}(B_n)$$

$$= \lim_{n \to \infty} \sum_{k=n}^{\infty} \mathbb{P}(A_n) \quad \textit{(By additivity)}$$

$$= \sum_{i=1}^{\infty} \mathbb{P}(A_i) - \lim_{n \to \infty} \sum_{k=1}^{n} \mathbb{P}(A_n)$$

$$= 0.$$

Furthermore, we have:

$$\mathbb{P}\left(\lim_{n \to \infty} B_n \right) = \mathbb{P}\left(\lim_{n \to \infty} \bigcup_{k=n}^{\infty} \right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_n \right) = \mathbb{P}\left(\limsup_{n \to \infty} A_n \right).$$

Hence proved the first Borel-Cantelli Lemma. To prove the second Borel-Cantelli Lemma, we prove the following:

$$1 - \mathbb{P}\left(\limsup_{n \to \infty} A_n \right) = \mathbb{P}\left(\left\{ \limsup_{n \to \infty} A_n \right\}^c \right)$$

$$= \mathbb{P}\left(\liminf_{n \to \infty} A_n^c \right) = 0.$$

7

1.2.3 Uniform Law of Large Numbers

The Uniform Law of Large Numbers (**ULLN**) provides a convergence result for collection of estimators where the convergence is uniform in the parameters space.

Theorem 1.5: Uniform Law of Large Numbers (ULLN)

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p_{\boldsymbol{\theta}_*}$ over $\mathcal{X}$ that depends on the true parameters $\boldsymbol{\theta}_*$. Let $\Theta \subset \mathbb{R}^m$ be the parameters space and $f_\theta : \mathcal{X} \to \mathbb{R}$ be a function indexed by $\theta \in \Theta$ that satisfies the following conditions:

1. $\boldsymbol{\theta}_* \in \Theta$ and Θ is compact.

2. $\mathbb{E}_{\boldsymbol{\theta}_*}[\sup_{\theta \in \Theta} |f_\theta(X)|] < \infty$[a].

3. $f_\theta(x)$ is continuous in θ for all $x \in \mathcal{X}$.

Then, we have:
$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(X_i) - \mathbb{E}_{\boldsymbol{\theta}_*}[f_\theta(X)] \right| \xrightarrow{p} 0. \tag{25}$$

---
[a]$\mathbb{E}_{\boldsymbol{\theta}_*}$ denotes the expectation taken over the distribution described by $p_{\boldsymbol{\theta}_*}$.

Proof (Theorem 1.5). _____

A nice proof is provided in Ferguson 1996, Theorem 16, Page 111. However, we will conduct our own version of the proof relying on results in learning theory. For $\theta \in \Theta$, define the following function:

$$\phi_\theta(\mathbf{X}) = \left| \frac{1}{n} \sum_{i=1}^n \left(f_\theta(X_i) - \mathbb{E}_{\boldsymbol{\theta}_*}[f_\theta(X)] \right) \right| \tag{26}$$

We have to prove that $\mathbb{P}(\sup_{\theta \in \Theta} |\phi_\theta(\mathbf{X})| \geq \epsilon) \to 0$ for all $\epsilon > 0$. For $\epsilon > 0$, we have:

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |\phi_\theta(\mathbf{X})| \geq \epsilon \right) = \mathbb{P}\left(\left\{ \sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \geq \epsilon \right\} \cup \left\{ \sup_{\theta \in \Theta}(-\phi_\theta(\mathbf{X})) \geq \epsilon \right\} \right)$$

$$\leq \mathbb{P}\left(\sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \geq \epsilon \right) + \mathbb{P}\left(\sup_{\theta \in \Theta}(-\phi_\theta(\mathbf{X})) \geq \epsilon \right)$$

$$\leq \frac{1}{\epsilon} \left(\mathbb{E}\left[\sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \right] + \mathbb{E}\left[\sup_{\theta \in \Theta}(-\phi_\theta(\mathbf{X})) \right] \right) \quad \textit{(Markov's Inequality)}$$

Now, we need to bound the expectations on the right-hand-side. To do so, we use the symmetrization trick. Specifically, given a sequence of i.i.d random variables $S = \{Z_1, \ldots, Z_n\} \sim \rho^n$ sampled from a distribution ρ and a class of functions $\mathcal{F}$. Let $S' = \{Z'_1, \ldots, Z'_n\} \sim \rho^n$ be a another sample

from the same distribution ρ (which we called the phantom sample), we have:

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(Z_i)-\mathbb{E}_S[f(Z_i)])\right]$$

$$=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(Z_i)-\mathbb{E}_{S'}[f(Z_i')])\right]$$

$$=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n f(Z_i)-\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{S'}[f(Z_i')]\right]$$

$$=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n(f(Z_i)-f(Z_i'))\right]\right]$$

$$\leqslant\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(Z_i)-f(Z_i'))\right]\quad(\textit{Jensen's Inequality})$$

$$=\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i(f(Z_i)-f(Z_i'))\right]\quad(\sigma\sim\mathrm{Rad}^n,\,f(Z_i)-f(Z_i')\textit{ is symmetric})$$

$$\leqslant\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(Z_i)\right]+\mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(-\sigma_i)f(Z_i')\right]$$

$$=\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(Z_i)\right]+\mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(Z_i')\right]\quad(\textit{Rademacher variables are symmetric})$$

$$=2\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(Z_i)\right]\quad(S\textit{ and }S'\textit{ are identically distributed})$$

$$=2\mathfrak{R}_n(\mathcal{F}). \tag{27}$$

Using the same argument, we can also have:

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(\mathbb{E}_S[f(Z_i)]-f(Z_i))\right]\leqslant 2\mathfrak{R}_n(\mathcal{F}). \tag{28}$$

Using equations 27 and 28 to bound $\mathbb{P}(\sup_{\theta\in\Theta}|\phi_\theta(S)|\geqslant\epsilon)$, we have:

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|\phi_\theta(S)|\geqslant\epsilon\right)\leqslant\frac{4\mathfrak{R}_n(\mathcal{F}_\Theta)}{\epsilon},\ \textit{where }\mathcal{F}_\Theta=\left\{f_\theta:\theta\in\Theta\right\}.$$

To complete the proof, we have to show that the Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_\Theta)\to 0$ as $n\to\infty$. By the definition of Rademacher Complexity, we have:

$$\mathfrak{R}_n(\mathcal{F}_\Theta)=\mathbb{E}_{\mathbf{X}\sim p_{\theta_*}^n}\mathbb{E}_\sigma\left[\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}_\Theta)\right],$$

It is sufficient to prove that the Empirical Rademacher Complexity $\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}_\Theta)\to 0$ as $n\to\infty$. Using Dudley's Entropy Integral Bartlett et al. 2017, Lemma A.5, we have:

$$\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}_\Theta)\leqslant\frac{12}{\sqrt{n}}\int_\alpha^M\sqrt{\log\mathcal{N}\left(\mathcal{F}_\Theta,\epsilon,L_2(\mathbf{X})\right)}d\epsilon,\quad\alpha>0, \tag{29}$$

where $M<\infty$ is a constant such that $|f_\theta(x)|\leqslant M$ p_{θ_*}-(almost everywhere) on $\mathcal{X}$ for all $\theta\in\Theta$[1]. Now, we need to construct a cover in L_2 norm for the class $\mathcal{F}_\Theta$. Since $f_\theta(x)$ is continuous in θ for all $x\in\mathcal{X}$, for all $\epsilon>0$, there exists $\delta_x>0$ such that $\|\theta-\bar{\theta}\|_2<\delta_x\implies|f_\theta(x)-f_{\bar\theta}(x)|<\epsilon$. Define $\delta>0$ as follows:

$$\delta=\min_{1\leqslant i\leqslant n}\delta_{X_i},\quad X_i\in\mathbf{X}. \tag{30}$$

---

[1]M exists because we have $\mathbb{E}_{\theta_*}[\sup_{\theta\in\Theta}|f_\theta(X)|]<\infty$.

Due to continuity, constructing an ϵ-cover in $\mathcal{F}_\Theta$ is equivalent to constructing a δ-cover for Θ with respect to the Euclidean norm. Since Θ is compact, there exists $N \in \mathbb{N}$ and a sequence of open balls $\{\mathcal{B}(y_j, r_j)\}_{j=1}^N$ where $y_j \in \Theta$ for all $1 \leqslant j \leqslant N$ such that:

$$\Theta \subseteq \bigcup_{j=1}^N \mathcal{B}(y_j, r_j). \tag{31}$$

By Long et al. 2020, Lemma A.8, we have:

$$\log \mathcal{N}\Big(\mathcal{B}(y_j, r_j), \delta, \|.\|_2\Big) \leqslant m \log\left(\frac{3r_j}{\delta}\right). \tag{32}$$

Therefore, we have:

$$\log \mathcal{N}\Big(\Theta, \delta, \|.\|_2\Big) \leqslant \sum_{j=1}^N \log \mathcal{N}\Big(\mathcal{B}(y_j, r_j), \delta, \|.\|_2\Big) \leqslant mN \log\left(\frac{\prod_{j=1}^N r_j}{\delta^N}\right). \tag{33}$$

From the above covering number bound, we can see that $\log \mathcal{N}\Big(\Theta, \delta, \|.\|_2\Big)$ does not grow with n and therefore, $\log \mathcal{N}\Big(\mathcal{F}_\Theta, \epsilon, L_2(\mathbf{X})\Big)$ does not grow with n. Therefore, we have $\hat{\mathfrak{R}}_\mathbf{X}(\mathcal{F}_\Theta) \in \mathcal{O}(1/\sqrt{n})$ and $\hat{\mathfrak{R}}_\mathbf{X}(\mathcal{F}_\Theta) \to 0$ as $n \to \infty$. □

1.2.4 Central Limit Theorem

> **Theorem 1.6: Central Limit Theorem (CLT)**
>
> Let $X_1, \dots X_n$ be a sequence of *i.i.d* random variables with expected value μ and finite variance σ^2. Then, we have:
>
> $$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \quad as \quad n \to \infty, \tag{34}$$
>
> where $\bar{X}_n = S_n/n$ and $\mathcal{N}(0, 1)$ is the standard normal distribution.

Proof (Central Limit Theorem (**CLT**)).
We prove this via the Characteristic Function. Let $\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$, notice that:

$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma},$$

Let $Z_i = X_i - \mu$ for $1 \leqslant i \leqslant n$ and suppose $Z = Z_1 = \cdots = Z_n$, we have:

$$\begin{aligned}
\varphi_{\bar{Z}_n}(t) &= \varphi_{\sum_{i=1}^n Z_i}\left(\frac{t}{\sqrt{n}}\right) = \left[\varphi_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{it\mathbb{E}[Z]}{\sqrt{n}} - \frac{t^2}{2n}\mathbb{E}[Z^2] + \mathcal{O}(1/n)\right]^n \quad \textit{(Taylor's Expansion)} \\
&= \left[1 - \frac{t^2}{2n} + \mathcal{O}(1/n)\right]^n.
\end{aligned}$$

The final equality comes from the fact that $\mathbb{E}[Z] = 0$ and $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \mathrm{Var}(Z) = 1$. Finally, we have:

$$\lim_{n \to \infty} \varphi_{\bar{Z}_n}(t) = \lim_{n \to \infty} \left[1 - \frac{t^2}{2n} + \mathcal{O}(1/n)\right]^n = e^{-t^2/2}.$$

Since $e^{-t^2/2}$ is the Characteristic Function of the standard normal distribution, by (**LCT**), we have $\bar{Z}_n \xrightarrow{d} \mathcal{N}(0,1)$. □.

1.3 Convergence of Random Variables

1.3.1 Convergence in Distribution

Definition 1.2 (Convergence in Distribution). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯
Given a sequence of real-valued random variables $X_1, X_2, \ldots$ with CDFs $F_1, F_2, \ldots$. We say that the sequence converges in distribution to a random variable X with CDF F, denoted $X_n \xrightarrow{d} X$ if:

$$\lim_{n \to \infty} F_n(x) = F(x), \tag{35}$$

for all $x \in \mathbb{R}$ at which F is continuous. Convergence in distribution can also be referred to as <u>weak convergence</u> *in measure theory.*

Theorem 1.7: Slutsky's Theorem (SLUTSKY)

Let X_n and Y_n be two sequences of random variables such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c^a$ where $c < \infty$ is a constant. Then, we have:

1. $X_n + Y_n \xrightarrow{d} X + c$.

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

3. $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$.

---
^{a}In the next section, we will see that $Y_n \xrightarrow{d} c$ implies $Y_n \xrightarrow{p} c$ for a constant c.

Proof (Theorem 1.7). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

1. $X_n + Y_n \xrightarrow{d} X + c$. Let $\epsilon > 0$ be any positive constant, we have:

$$\begin{aligned}
F_{X_n + Y_n}(t) &= \mathbb{P}(X_n + Y_n \leqslant t) \\
&= \mathbb{P}(X_n + Y_n \leqslant t, |Y_n - c| \leqslant \epsilon) + \mathbb{P}(X_n + Y_n \leqslant t, |Y_n - c| > \epsilon) \\
&\leqslant \mathbb{P}(X_n + Y_n \leqslant t, |Y_n - c| \leqslant \epsilon) + \underbrace{\mathbb{P}(|Y_n - c| > \epsilon)}_{approaches\ 0\ as\ n \to \infty}.
\end{aligned}$$

In the event that $|Y_n - c| \leqslant \epsilon$, we have $Y_n \geqslant c - |Y_n - c| \geqslant c - \epsilon$. Therefore, we have:

$$F_{X_n + Y_n}(t) \leqslant \mathbb{P}(X_n \leqslant t - c + \epsilon) + \mathbb{P}(|Y_n - c| > \epsilon).$$

Similarly, we have:

$$F_{X_n + Y_n}(t) \geqslant \mathbb{P}(X_n \leqslant t - c - \epsilon) - \mathbb{P}(|Y_n - c| > \epsilon).$$

Taking limits of both inequalities, we have:

$$F_X(t - c - \epsilon) \leqslant \lim_{n \to \infty} F_{X_n + Y_n}(t) \leqslant F_X(t - c + \epsilon).$$

Let $\epsilon \to 0$, we have $\lim_{n \to \infty} F_{X_n + Y_n}(t) \to F_X(t - c) = F_{X+c}(t)$ as $n \to \infty$.

2. $X_n Y_n \xrightarrow{d} cX$. (Apply the same proof method as (1)).

3. $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$. (Apply the same proof method as (1)).

□.

Connection to Weak Convergence of Measures

Definition 1.3 (Continuity Sets). ─────────────────────────────
Let X be a metric space and $\mathcal{B}$ be the Borel-σ-algebra generated from open subsets of X. Let $A \in \mathcal{B}$ be a Borel set. Then, we say that A is a continuity set with respect to a measure μ, or A is a μ-continuity set, if:

$$\mu(\partial A) = 0, \ or \ \mu(\mathrm{cl}_X(A)) = \mu(A), \tag{36}$$

where ∂A is the topological boundary of A, defined as:

$$\partial A = \mathrm{cl}_X(A) \backslash \mathrm{in}_X(A), \tag{37}$$

where $\mathrm{cl}_X(A)$ is the closure, which contains all limit points of A and $\mathrm{in}_X(A)$ is the set of all interior points of A. Intuitively, a set is continuous with respect to a measure if its closure does not change its measure.

Remark 1.2 (Non-examples of Continuity Sets). ─────────────────
Some examples of the non-continuity sets are as follows:

1. *The set of rational numbers $\mathbb{Q}$ is not a continuity set with respect to the Lebesgue measure λ because its closure is the set of real numbers. Hence, we have:*

$$\lambda(\partial \mathbb{Q}) = \lambda(\mathbb{R} \backslash \mathbb{Q}) = \infty.$$

2. *Given an experiment of rolling an unfair 6-sided die with probabilities $p_1, \ldots, p_6$. Let the sample space $\Omega = [1, 6]$ and $\mathcal{B}(\Omega)$ be the Borel-σ-algebra generated from open subsets of the sample space. Then, we have the corresponding probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where μ is defined as:*

$$\mu((a, b)) = \sum_{i = \lceil a \rceil}^{\lfloor b \rfloor} p_i, \qquad \forall (a, b) \in \mathcal{B}(\Omega),$$

which denotes the probability that the rolled die lands on a face in the range between a and b. Then, any set (a, b) in $\mathcal{B}(\Omega)$ where a or b is integer is not μ-continuous. For example:

$$\mu((2, 4.5)) = p_3 + p_4,$$
$$\mu(\mathrm{cl}((2, 4.5))) = \mu([2, 4.5]) = p_2 + p_3 + p_4 > \mu((2, 4.5)).$$

Definition 1.4 (Weak Convergence of Measures). ─────────────────
Let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of measures and μ be a measure on a measurable space $(X, \mathcal{F})$. We say that μ_n converges weakly to a measure μ if:

$$\lim_{n \to \infty} \mu_n(A) = \mu(A), \tag{38}$$

for all μ-continuity sets $A \in \mathcal{F}$.

> **Theorem 1.8: Portmanteau's Theorem** (PORTMANTEAU)
>
> Let $\{\mathbb{P}_n\}_{n=1}^{\infty}, \mathbb{P}$ be probability measures on a measurable space $(X, \mathcal{F})$. Then, the following statements are equivalent:
>
> 1. $\mathbb{P}_n$ converges weakly to $\mathbb{P}$.
>
> 2. $\int f d\mathbb{P}_n \to \int f d\mathbb{P}$ for all $f \in C_b(X)^a$.
>
> 3. $\limsup_{n\to\infty} \mathbb{P}_n(F) \leqslant \mathbb{P}(F)$ for all closed $F \in \mathcal{F}$.
>
> 4. $\liminf_{n\to\infty} \mathbb{P}_n(G) \geqslant \mathbb{P}(G)$ for all open $G \in \mathcal{F}$.
>
> ———————————————————————
> $^a C_b(X)$ denotes the space of bounded continuous $f : X \to \mathbb{C}$ functions.

Proof (Theorem 1.8). ——————————————————————————————————

Assume that for all Borel sets $A \in \mathcal{F}$ that is $\mathbb{P}$-continuity, we have $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A)$. We prove the following:

1. *$\int f d\mathbb{P}_n \to \int f d\mathbb{P}$ for all $f \in C_b(X)$: Let $b > 0$, without loss of generality, suppose that $0 \leqslant f \leqslant b$ for all $f \in C_b(X)$. By the integral identity, we have:*

$$f(x) = \int_0^b \mathbf{1}\{t < f(x)\}dt, \quad \forall x \in X.$$

   Hence, for $n \geqslant 1$, we have:

$$\int f d\mathbb{P}_n = \int \int_0^b \mathbf{1}\{t < f\}dt d\mathbb{P}_n$$
$$= \int_0^b \int \mathbf{1}\{t < f\}d\mathbb{P}_n dt \quad \textit{(Fubini's Theorem)}$$
$$= \int_0^b \mathbb{P}_n(f^{-1}(t, b))dt.$$

   We have $\partial f^{-1}(t, b) \subseteq f^{-1}(\{t, b\})$. Hence, we have $\mathbb{P}(\partial f^{-1}(t, b)) \leqslant \mathbb{P}(f^{-1}(\{t, b\})) = 0$, which means that $f^{-1}(t, b)$ is a $\mathbb{P}$-continuity set. Therefore, by the initial assumption, we have:

$$\lim_{n\to\infty} \mathbb{P}_n(f^{-1}(t, b)) = \mathbb{P}(f^{-1}(t, b)).$$

   By Dominated Convergence Theorem, we have $\int_0^b \mathbb{P}_n(f^{-1}(t, b))dt \to \int_0^b \mathbb{P}(f^{-1}(t, b))dt$. Hence, we obtain the desired result.

2. *$\liminf_{n\to\infty} \mathbb{P}_n(G) \geqslant \mathbb{P}(G)$ for all open $G \in \mathcal{F}$: Let $G \in \mathcal{F}$ be an open set and let $U = G^c$. Consider the following sequence of continuous, bounded functions:*

$$f_m(s) = \min(1, m \cdot d_X(s, U)),$$

   where we denote:

$$d_X(s, U) = \inf_{x \in U} d_X(s, x),$$

   which is the projection of a point s onto the set U. We can re-write f_m as follows:

$$f_m(s) = \begin{cases} 1, & \text{if } d(s, U) \geqslant \frac{1}{m} \\ m \cdot d(s, U), & \text{if } d(s, U) < \frac{1}{m} \end{cases}.$$

In other words, $f_m(s) = 1$ when the point s is sufficiently far away from the complement U and as m increases, the distance threshold required for $f_m(s) = 1$ becomes smaller, making f_m approaches the indicator function $\mathbf{1}_G$. Hence, we have:

$$\mathbb{P}_n(G) = \int \mathbf{1}_G d\mathbb{P}_n \geqslant \int f_m d\mathbb{P}_n \to \int f_m d\mathbb{P}, \qquad (By \ (1)).$$

Hence, we have $\liminf_{n\to\infty} \mathbb{P}_n(G) \geqslant \int f_m d\mathbb{P}$[2]. Taking $m \to \infty$ and using Monotone Convergence Theorem, we have:

$$\liminf_{n\to\infty} \mathbb{P}_n(G) \geqslant \lim_{m\to\infty} \int f_m d\mathbb{P} = \int \lim_{m\to\infty} f_m d\mathbb{P} = \int \mathbf{1}_G d\mathbb{P} = \mathbb{P}(G).$$

3. $\limsup_{n\to\infty} \mathbb{P}_n(F) \leqslant \mathbb{P}(F)$ for all closed $F \in \mathcal{F}$: Taking the complement of F and apply (2), we obtain the desired result.

$\square$.

1.3.2   Convergence in Probability

Definition 1.5 (Convergence in Probability). ────────────────────────────

Given a sequence of real-valued random variables $X_1, X_2, \ldots$. We say that the sequence converges in probability to a random variable X, denoted $X_n \overset{p}{\to} X$ if:

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| \geqslant \epsilon\Big) = 0, \quad \forall \epsilon > 0. \tag{39}$$

We also refer to convergence in probability as convergence in measure *in measure theory.*

Proposition 1.1: $X_n \overset{p}{\to} X \implies X_n \overset{d}{\to} X$

Let X and the sequence $X_1, X_2, \ldots$ be real-valued random variables. If $X_n \overset{p}{\to} X$, then $X_n \overset{d}{\to} X$.

Proof (Proposition 1.1). ────────────────────────────

We first prove the following claim: Let X, Y be random variables, $a \in \mathbb{R}$ and $\epsilon > 0$, the inequality $\mathbb{P}(Y \leqslant a) \leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(|Y - X| \geqslant \epsilon)$ holds. We have:

$$\begin{aligned}
\mathbb{P}(Y \leqslant a) &= \mathbb{P}(Y \leqslant a, X \leqslant a + \epsilon) + \mathbb{P}(Y \leqslant a, X \geqslant a + \epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant a - X, a - X \leqslant -\epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant -\epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant -\epsilon) + \mathbb{P}(Y - X \geqslant \epsilon) \\
&= \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(|Y - X| \geqslant \epsilon).
\end{aligned}$$

Using the above inequality, we have:

$$\mathbb{P}(X \leqslant a - \epsilon) - \mathbb{P}(|X_n - X| \geqslant \epsilon) \leqslant \mathbb{P}(X_n \leqslant a) \leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(|X_n - X| \geqslant \epsilon).$$

Taking limits as $n \to \infty$ from both sides, we have:

$$F_X(a - \epsilon) \leqslant \lim_{n\to\infty} F_{X_n}(a) \leqslant F_X(a + \epsilon).$$

Taking $\epsilon \to 0^+$, we have $\lim_{n\to\infty} F_{X_n}(a) = F_X(a)$.

$\square$.

---

[2]From $\mathbb{P}_n(G) \geqslant \int f_m d\mathbb{P}_n$, we have $\liminf_{n\to\infty} \mathbb{P}_n(G) \geqslant \liminf_{n\to\infty} \int f_m d\mathbb{P}_n = \lim_{n\to\infty} \int f_m d\mathbb{P}_n = \int f_m d\mathbb{P}$.

Proposition 1.2: $X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$

Let $c \in \mathbb{R}$ be a constant and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Then, $X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$.

Proof (Proposition 1.2, Pishro-Nik 2014).

Since $X_n \xrightarrow{d} c$, we immediately have the following:

$$\lim_{n \to \infty} F_{X_n}(c - \epsilon) = 0,$$
$$\lim_{n \to \infty} F_{X_n}(c + \epsilon/2) = 1.$$

Then, for any $\epsilon > 0$, we have:

$$\begin{aligned}
\lim_{n \to \infty} (|X_n - c| \geq \epsilon) &= \lim_{n \to \infty} \mathbb{P}\Big[\mathbb{P}(X_n \leq c - \epsilon) + \mathbb{P}(X_n \geq c + \epsilon)\Big] \\
&= \underbrace{\lim_{n \to \infty} F_{X_n}(c - \epsilon)}_{=0} + \lim_{n \to \infty} \mathbb{P}(X_n \geq c + \epsilon) \\
&\leq \lim_{n \to \infty} \mathbb{P}(X_n \geq c + \epsilon/2) \\
&= 1 - \underbrace{\lim_{n \to \infty} F_{X_n}(c + \epsilon/2)}_{=1} \\
&= 0.
\end{aligned}$$

From the above, we have $\lim_{n \to \infty} \mathbb{P}(|X_n - c| \geq \epsilon) = 0$ and $X_n \xrightarrow{p} c$. □.

1.3.3 Convergence in L^p norm

Definition 1.6 (Convergence in L^p norm).
Given a sequence of random variables $X_1, X_2, \ldots$ and a real number $p \in [1, \infty)$. We say that the sequence converges in L^p norm to a random variable X, denoted as $X_n \xrightarrow{L^p} X$ if:

$$\lim_{n \to \infty} \mathbb{E}|X_n - X|^p = 0. \tag{40}$$

Proposition 1.3: $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$

Let $p \geq 1$ and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Let X be a random variable, then, $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$.

Proof (Proposition 1.3).
Let $\epsilon > 0$, we have:

$$\begin{aligned}
\mathbb{P}(|X_n - X| \geq \epsilon) &= \mathbb{P}(|X_n - X|^p \geq e^p) \quad (p \geq 1) \\
&\leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p}. \quad \text{(Markov's Inequality)}
\end{aligned}$$

Taking the limits from both sides, we have $\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$ and $X_n \xrightarrow{p} X$. □.

15

1.3.4 Almost-sure Convergence

Definition 1.7 (Convergence almost-surely). ────────────────
Let $X_1, X_2, \ldots$ be a sequence of real-valued random variables that map from a sample space Ω. Let X also be a real-valued random variable. We say that X_n converges almost surely to X, denoted as $X_n \xrightarrow{a.s} X$, if $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for almost all $\omega \in \Omega$:

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right) = 1. \tag{41}$$

In other words, we can write:

$$\mathbb{P}\left(\limsup_{n \to \infty} E_n\right) = 0 \quad where \quad E_n = \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\right\}. \tag{42}$$

Remark 1.3 (Consequence of (**BCL**)). ────────────────
Let $X, X_1, X_2, \ldots$ be random variables defined over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For $\epsilon > 0$ be chosen arbitrarily, let $E_n = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\}$. Then, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \implies X_n \xrightarrow{a.s} X. \tag{43}$$

In other words, if the sequence $\{\mathbb{P}(E_n)\}_{n=1}^{\infty}$ converges, then X_n converges almost surely to X.

Proposition 1.4: $X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{p} X$

Let $X_1, X_2, \ldots$ be a sequence of real-valued random variables and also let X be a real valued random variables. If $X_n \xrightarrow{a.s} X$ then $X_n \xrightarrow{p} X$.

Proof (Proposition 1.4). ────────────────
Let $f_n : \Omega \to \mathbb{R}_+$ be a sequence of nonnegative Borel-measurable functions such that $f_n(\omega) = |X_n(\omega) - X(\omega)|$. By Fatou's Lemma (reverse), we have:

$$\underbrace{\mathbb{P}\left(\limsup_{n \to \infty}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\}\right)}_{=0} = \int f_n d\mathbb{P}$$

$$\geqslant \limsup_{n \to \infty} \mathbb{P}(|X_n - X| \geqslant \epsilon)$$

$$\geqslant \lim_{n \to \infty} \mathbb{P}(|X_n - X| \geqslant \epsilon).$$

Hence, we have $\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geqslant \epsilon) = 0$ and $X_n \xrightarrow{p} X$. □.

Theorem 1.9: Continuous Mapping Theorem (CMT)

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous almost everywhere function and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Then, the following statements hold:

1. $X_n \xrightarrow{d} X \implies f(X_n) \xrightarrow{d} f(X)$.

2. $X_n \xrightarrow{p} X \implies f(X_n) \xrightarrow{p} f(X)$.

3. $X_n \xrightarrow{a.s} X \implies f(X_n) \xrightarrow{a.s} f(X)$.

16

Proof (Continuous Mapping Theorem (**CMT**)). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯
We handle each statement separately:

1. *Let $\mathbb{P}_n, \mathbb{P}$ be the probability measures induced by X_n and X, respectively. By (**PORTMANTEAU**), $\mathbb{P}_n$ converges weakly to $\mathbb{P}$ is equivalent to $\int g d\mathbb{P}_n \to \int g d\mathbb{P}$ for all continuous, bounded $g : \mathbb{R} \to \mathbb{R}$. For all g continuous and bounded, $g \circ f$ is also continuous and bounded. Hence:*

$$\int g \circ f d\mathbb{P}_n \to \int g \circ f d\mathbb{P} \text{ as } n \to \infty.$$

   Hence, let $\mathbb{P}_n f$ and $\mathbb{P}f$ be the probability measures induced by $f(X_n)$ and $f(X)$, respectively. We have $\mathbb{P}_n f$ converges $\mathbb{P}f$ weakly. Therefore, $f(X_n) \xrightarrow{d} f(X)$.

2. *Let C_f denotes the set of points on $\mathbb{R}$ where f is continuous. By assumption, we have $\mathbb{P}(x)$. Let $\epsilon > 0$ be arbitrary. For all $\delta > 0$, we denote the following set:*

$$B_\delta = \left\{ x \in C_f : \exists y \in \mathbb{R} \ s.t \ |x - y| < \delta \ but \ |f(x) - f(y)| \geqslant \epsilon \right\}$$

   In other words, B_δ denotes the set of all conitnuous points x of f such that we can find a point close to x but its output is not close to $f(x)$. We have:

$$\mathbb{P}\Big(|f(X_n) - f(X)| \geqslant \epsilon\Big) = \mathbb{P}\Big(\big\{X \notin C_f\big\} \cup \big\{|X_n - X| \geqslant \delta\big\} \cup \big\{X \in B_\delta\big\}\Big)$$

$$\leqslant \underbrace{\mathbb{P}(X \notin C_f)}_{=0} + \mathbb{P}(|X_n - X| \geqslant \delta) + \mathbb{P}(X \in B_\delta)$$

$$= \mathbb{P}(|X_n - X| \geqslant \delta) + \mathbb{P}(X \in B_\delta).$$

   Since $\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geqslant \delta) = 0$ for all $\delta > 0$ by assumption and $\mathbb{P}(X \in B_\delta) = 0$ when $\delta \to 0$, we have:

$$\lim_{n \to \infty} \mathbb{P}\Big(|f(X_n) - f(X)| \geqslant \epsilon\Big) = 0 \ or \ f(X_n) \xrightarrow{p} f(X).$$

3. *Since f is continuous, for any $\omega \in \Omega$ such that $X_n(\omega) \to X(\omega)$, we have $f(X_n(\omega)) \to f(X(\omega))$. Therefore, we have:*

$$\Big\{\omega \in \Omega : X_n(\omega) \to X(\omega)\Big\} \subseteq \Big\{\omega \in \Omega : f(X_n(\omega)) \to f(X(\omega))\Big\}.$$

   Therefore, for all $\epsilon > 0$, we have:

$$\mathbb{P}\Big(\big\{\omega \in \Omega : \lim_{n \to \infty} f(X_n(\omega)) = f(X(\omega))\big\}\Big)$$

$$\geqslant \mathbb{P}\Big(\big\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\big\}\Big) = 1,$$

   Hence, we have $f(X_n) \xrightarrow{a.s} f(X)$.

$\square$.

2 Statistical Inference

2.1 Sufficiency & Likelihood Principles

2.1.1 Sufficiency

Definition 2.1 (Sufficient Statistics).

Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p_{\boldsymbol{\theta}_*}$ be a random sample drawn i.i.d from a distribution with parameters $\boldsymbol{\theta}_*$. Let $\mathbf{U} = T(\mathbf{X})$ be a statistic, then it is called a <u>sufficient statistic</u> if the conditional distribution $p_{\mathbf{X}|\mathbf{U}}$ does not depend on $\boldsymbol{\theta}_*$.

Example 2.1 (Bernoulli random variables).

Let $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{Bernoulli}(\theta)$ be a random sample from the Bernoulli distribution. Let $\mathbf{U} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\mathbf{U}$ is a sufficient statistic of θ. To illustrate this, suppose that $\mathbf{x} = (x_1, \ldots, x_n)$ is an observation of the random sample $\mathbf{X}$ and $\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We have:

$$
\begin{aligned}
\mathbb{P}(\mathbf{X} = \mathbf{x}|\mathbf{U} = \mathbf{u}) &= \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u})}{\mathbb{P}(\mathbf{U} = \mathbf{u})} \\
&= \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i)}{\mathbb{P}(\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i)} \\
&= \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}(\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i)} \\
&= \frac{\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}}{\mathbb{P}(\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i)}.
\end{aligned}
$$

Now, setting $k = \sum_{i=1}^{n} x_i$, The denominator is basically the probability that the Bernoulli variables sums up to k. Hence, we can calculate the denominator as follows:

$$
\mathbb{P}\left(\sum_{i=1}^{n} X_i = k\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.
$$

Therefore, we have:

$$
\mathbb{P}(\mathbf{X} = \mathbf{x}|\mathbf{U} = \mathbf{u}) = \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \frac{1}{\binom{n}{k}}.
$$

Therefore, the conditional distribution does not depend on θ and $\mathbf{U}$ is a sufficient statistic.

Definition 2.2 (Sufficiency Principle).

If $\mathbf{U} = T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}_*$, then any inference about $\boldsymbol{\theta}_*$ should only depend on the sample $\mathbf{X}$ through $\mathbf{U}$. *In other words, if we estimate $\boldsymbol{\theta}_*$ using an estimator $\hat{\boldsymbol{\theta}}_*$, only $\mathbf{U}$ shows up in the formula of $\hat{\boldsymbol{\theta}}_*$, not the sample $\mathbf{X}$ itself. We will see why this is the case in the Factorisation Theorem (**FacT**), which states that we can factorise the density function into a function of $\mathbf{U}, \boldsymbol{\theta}_*$ and a function of the observations $\mathbf{x}$ and thus, the inference about $\boldsymbol{\theta}_*$ is independent of the observations $\mathbf{x}$.*

> **Theorem 2.1: Factorisation Theorem (FacT)**
>
> Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with joint density function $p_{\boldsymbol{\theta}_*}$ over $\mathcal{X}^n$. The statistic $\mathbf{U} = T(\mathbf{X})$ is sufficient for the parameters $\boldsymbol{\theta}_*$ if and only if we can find functions h, g such that:
>
> $$p_{\boldsymbol{\theta}_*}(\mathbf{x}) = g(T(\mathbf{x}), \boldsymbol{\theta}_*)h(\mathbf{x}),$$
>
> for all $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\theta}_* \in \Theta$.

Proof (Factorisation Theorem (**FacT**)). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯
We have to conduct the proof in both directions.

- $T(\mathbf{X})$ *is sufficient* $\implies$ *Factorisation exists: Let* $\mathbf{U} = T(\mathbf{X})$ *be a sufficient statistics and* $\mathbf{u} = T(\mathbf{x})$ *be the statistics evaluated on the observations* $\mathbf{x}$*. Then, we have:*

$$
\begin{aligned}
p_{\boldsymbol{\theta}_*}(\mathbf{x}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_*) \\
&= \mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*).
\end{aligned}
$$

  Since $\mathbf{U} = T(\mathbf{X})$ *is a sufficient statistics,* $\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)$ *does not depend on* $\boldsymbol{\theta}_*$*. Hence, we denote* $h(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)$*. Furthermore,* $\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)$ *is a function of* $\mathbf{u}$ *and* $\boldsymbol{\theta}_*$*. We denote this function as* $g(\mathbf{u}, \boldsymbol{\theta}_*)$ *and conclude that the factorisation* $p_{\boldsymbol{\theta}_*}(\mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x}), \boldsymbol{\theta}_*)$ *indeed exists.*

- *Factorisation exists* $\implies$ $T(\mathbf{X})$ *is sufficient: Suppose that there exists* g, h *such that we have the factorisation* $p_{\boldsymbol{\theta}_*}(\mathbf{x}) = g(T(\mathbf{x}), \boldsymbol{\theta}_*)h(\mathbf{x})$*. We then have:*

$$
\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*) = \frac{p_{\boldsymbol{\theta}_*}(\mathbf{x})}{\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)} = \frac{g(\mathbf{u}, \boldsymbol{\theta}_*)h(\mathbf{x})}{\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*)}.
$$

  We denote $A_{\mathbf{u}} = \left\{ \tilde{\mathbf{x}} \in \mathcal{X}^n : T(\tilde{\mathbf{x}}) = \mathbf{u} \right\}$*. We have:*

$$
\begin{aligned}
\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*) &= \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} \mathbb{P}(\mathbf{X} = \tilde{\mathbf{x}}) \\
&= \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} p(\tilde{\mathbf{x}}; \boldsymbol{\theta}_*) = \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} g(T(\tilde{\mathbf{x}}), \boldsymbol{\theta}_*)h(\tilde{\mathbf{x}}) \\
&= g(\mathbf{u}, \boldsymbol{\theta}_*) \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} h(\tilde{\mathbf{x}}).
\end{aligned}
$$

  From the above, we have:

$$
\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}_*) = \frac{h(\mathbf{x})}{\sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} h(\tilde{\mathbf{x}})},
$$

  and the above expression does not depend on $\boldsymbol{\theta}_*$*. Hence,* $T(\mathbf{X})$ *is a sufficient statistics.*

$\square$.

2.1.2 Likelihood

Definition 2.3 (Likelihood Function). _____

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample whose distribution belongs to a family of distributions $\mathcal{P} = \left\{ p_\theta : \theta \in \Theta \right\}$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an observation of the random sample $\mathbf{X}$. Then, the likelihood function $L(\theta; \mathbf{x})$ is defined as follows:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i), \quad \theta \in \Theta. \tag{44}$$

In some cases, we also use the log-likelihood function:

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log p_\theta(x_i), \quad \theta \in \Theta. \tag{45}$$

Essentially, $L(\theta; \mathbf{x})$ quantifies the likelihood that θ generates the observations $\mathbf{x}$. In a way, it is the inverse of probability density (mass) functions, we can see the contrast as follows:

- ***Probability Density Function**: The parameters are fixed but the observations are random.*

- ***Likelihood Function**: The observations are fixed but the parameters are variable.*

Definition 2.4 (Maximum Likelihood Estimator). _____

Given $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample whose distribution belongs to a family of distributions $\mathcal{P} = \left\{ p_\theta : \theta \in \Theta \right\}$ and let $\mathbf{x} = (x_1, \ldots, x_n)$ be an observation of the random sample $\mathbf{X}$. The <u>Maximum Likelihood Estimator</u> $\theta_{MLE} \in \Theta$ is the parameter that maximizes the likelihood function:

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x}). \tag{46}$$

In the subsequent propositions, we will discuss some of the key properties of MLE.

Proposition 2.1: Consistency of MLE

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p_{\boldsymbol{\theta}_*}$ over $\mathcal{X}$ dependent on a true set of parameters $\boldsymbol{\theta}_*$. Let Θ be the parameters space. Then, the Maximum Likelihood Estimator $\Theta_{MLE} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{X})$, which is a random variable, is consistent, meaning $\Theta_{MLE} \xrightarrow{p} \boldsymbol{\theta}_*$, provided that the following conditions are met:

1. $\boldsymbol{\theta}_* \in \Theta$ and Θ is a compact space.

2. $\log p_\theta(x)$ is continuous in θ for almost all $x \in \mathcal{X}$.

3. $\mathbb{E}_{\boldsymbol{\theta}_*}[\sup_{\theta \in \Theta} |\log p_\theta(X)|] < \infty$.

4. The mapping $\xi \mapsto p_\xi, \quad \xi \in \Theta$ is one-to-one (Identifiability).[a]

Furthermore, we can also show that Θ_{MLE} is asymptotically unbiased. In other words, $\lim_{n \to \infty} \mathbb{E}[\Theta_{MLE}] = \boldsymbol{\theta}_*$.

---
[a]In general, it is required that the model is strongly identifiable. However, since the parameters space Θ is compact, this requirement is satisfied.

Proof (Proposition 2.1). _____

A proof for consistency of MLE can be found in "Chapter 36 Large sample estimation and hypoth-

esis testing" 1994, Theorem 2.5 but we attempt our own proof anyway. The general proof strategy is listed below:

1. *First, prove that $\boldsymbol{\theta}_* = \arg\max_{\xi \in \Theta} \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\xi(X)]$.*

2. *Then, by (**ULLN**): $\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \overset{p}{\to} \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\theta(X)], \quad \forall \theta \in \Theta$.*

3. *Prove that if a stochastic process converges in probability to a deterministic process, then the maximizers of the stochastic process converges in probability to the maximizer of the deterministic process.*

To complete the proof, it is sufficient to prove the first point. For any $\theta \in \Theta$, we have:

$$\mathbb{E}_{\boldsymbol{\theta}_*}\left[\log \frac{p_\theta(X)}{p_{\boldsymbol{\theta}_*}(X)}\right] \leqslant \log \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{p_\theta(X)}{p_{\boldsymbol{\theta}_*}(X)}\right] = \log \int_{\mathcal{X}} \frac{p_\theta(x)}{p_{\boldsymbol{\theta}_*}(x)} p_{\boldsymbol{\theta}_*}(x)dx = \log 1 = 0.$$

Therefore, for all $\theta \in \Theta : \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\theta(X)] \leqslant \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_{\boldsymbol{\theta}_*}(X)]$. Hence, $\boldsymbol{\theta}_* = \arg\max_{\xi \in \Theta} \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\xi(X)]$ as desired. Define the following continuous mappings:

$$M_n = \xi \mapsto \frac{1}{n}\sum_{i=1}^n \log p_\xi(X_i),$$
$$\text{and } M = \xi \mapsto \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\xi(X)].$$

*Then, by (**ULLN**), we have $\|M_n - M\|_\infty \overset{p}{\to} 0$. This means that for any fixed $\epsilon > 0$ and $\delta \in (0,1)$, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, with probability of at least $1 - \delta$, we have:*

$$|M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| < \frac{\epsilon}{2}, \quad \text{and} \quad |M_n(\boldsymbol{\theta}_*) - M(\boldsymbol{\theta}_*)| < \frac{\epsilon}{2},$$

simultaneously, where $\hat{\theta}_n = \arg\max_{\theta \in \Theta} M_n(\theta)$. From the above, with probability of at least $1 - \delta$, we have:

$$M(\hat{\theta}_n) \geqslant M_n(\hat{\theta}_n) - \frac{\epsilon}{2} \geqslant M_n(\boldsymbol{\theta}_*) - \frac{\epsilon}{2} \geqslant M(\boldsymbol{\theta}_*) - \epsilon.$$

Hence, we have $M(\boldsymbol{\theta}_*) - M(\hat{\theta}_n) < \epsilon$ with high probability. Since M is continuous on a compact space, for all $\epsilon > 0$, there exists a constant $\xi > 0$ such that $|\theta - \boldsymbol{\theta}_*| < \xi \implies |M(\boldsymbol{\theta}_*) - M(\theta)| < \epsilon$. Hence, we have:

$$\mathbb{P}(|\boldsymbol{\theta}_* - \hat{\theta}_n| < \xi) = \mathbb{P}(|M(\boldsymbol{\theta}_*) - M(\hat{\theta}_n)| < \epsilon) \geqslant 1 - \delta.$$

Hence, we have $|\boldsymbol{\theta}_* - \hat{\theta}_n| < \xi$ with high probability. Since δ is chosen arbitrarily, we can take $\delta \to 0$ and for all $\xi > 0$, we have:

$$\lim_{n \to \infty} \mathbb{P}(|\boldsymbol{\theta}_* - \hat{\theta}_n| < \xi) = 1 \implies \hat{\theta}_n \overset{p}{\to} \boldsymbol{\theta}_*.$$

$\square$.

Proposition 2.2: Asymptotic Normality of MLE

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p_{\boldsymbol{\theta}_*}$ dependent on a true set of parameters $\boldsymbol{\theta}_*$. Suppose that the MLE is <u>consistent</u>. Then, the Maximum Likelihood Estimator $\Theta_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{X})$ is asymptotically normal:

$$\sqrt{n}(\Theta_{\mathrm{MLE}} - \boldsymbol{\theta}_*) \overset{d}{\to} \mathcal{N}(0, \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)^{-1}). \tag{47}$$

Proof (Proposition 2.2).

By the Mean Value Theorem, for a continuous function $f : [a, b] \to \mathbb{R}$, we have:

$$\frac{f(b) - f(a)}{b - a} = f'(c), \quad \text{for some } c \in [a, b].$$

Apply MVT for the log-likelihood function, for some φ_n within the interval formed by $\hat{\Theta}_{\mathrm{MLE}}$ and $\boldsymbol{\theta}_*$ (hence, φ_n is a random variable since $\hat{\Theta}_{\mathrm{MLE}}$ is random), we have:

$$\ell''(\varphi_n; \mathbf{X}) = \frac{\ell'(\hat{\Theta}_{\mathrm{MLE}}; \mathbf{X}) - \ell'(\boldsymbol{\theta}_*; \mathbf{X})}{\hat{\Theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_*} = -\frac{\ell'(\boldsymbol{\theta}_*; \mathbf{X})}{\hat{\Theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_*} \qquad \left(\ell'(\hat{\Theta}_{\mathrm{MLE}}; \mathbf{X}) = 0 \right).$$

Therefore, we have:

$$\hat{\Theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_* = -\frac{\ell'(\boldsymbol{\theta}_*; \mathbf{X})}{\ell''(\varphi_n; \mathbf{X})} \implies \sqrt{n}(\hat{\Theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_*) = -\frac{\frac{1}{\sqrt{n}}\ell'(\boldsymbol{\theta}_*; \mathbf{X})}{\frac{1}{n}\ell''(\varphi_n; \mathbf{X})}. \tag{A}$$

We have:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\ell'(\boldsymbol{\theta}_*; \mathbf{X}) &= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\xi}\ln p_\xi(X_i)\Big|_{\xi=\boldsymbol{\theta}_*} \right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{\theta}_*; X_i) \right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{\theta}_*; X_i) - \underbrace{\mathbb{E}_{\boldsymbol{\theta}_*}\left[s(\boldsymbol{\theta}_*; X) \right]}_{0} \right] \\
&\xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_{\boldsymbol{\theta}_*}[s(\boldsymbol{\theta}_*; X)] \right) \qquad \textbf{(CLT)} \\
&= \mathcal{N}(0, \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)).
\end{aligned} \tag{B}$$

We also have:

$$\frac{1}{n}\ell''(\varphi_n; \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\xi^2}\ln p_\xi(X_i)\Big|_{\xi=\varphi_n} \xrightarrow{p} \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{d^2}{d\xi^2}\ln p_\xi(X)\Big|_{\xi=\varphi_n} \right]. \qquad \textbf{(WLLN)}$$

By the initial assumption, $\hat{\Theta}_{\mathrm{MLE}} \xrightarrow{p} \boldsymbol{\theta}_*$ and φ_n lies between $\hat{\Theta}_{\mathrm{MLE}}$ and $\boldsymbol{\theta}_*$. Therefore, $\varphi_n \xrightarrow{p} \boldsymbol{\theta}_*$. As a result, by $(\textbf{CMT})$[3], we have $\mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{d^2}{d\xi^2}\ln p_\xi(X)\Big|_{\xi=\varphi_n} \right] \xrightarrow{p} \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{d^2}{d\xi^2}\ln p_\xi(X)\Big|_{\xi=\boldsymbol{\theta}_*} \right] = -\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)$. Therefore, we have:

$$\frac{1}{n}\ell''(\varphi_n; \mathbf{X}) \xrightarrow{p} \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{d^2}{d\xi^2}\ln p_\xi(X)\Big|_{\xi=\varphi_n} \right] \xrightarrow{p} -\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*). \tag{C}$$

From Eqns. A, B, C and $(\textbf{SLUTSKY})$ lemma, we have $\sqrt{n}(\hat{\Theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)^{-1})$, as desired. □.

2.2 Point Estimation

2.2.1 Bias, Variance, Consistency and MSE

2.2.2 Sufficient Statistics & Rao-Blackwell Theorem

> **Theorem 2.2: Rao-Blackwell Theorem (RB)**

---

[3]The mapping $p \mapsto \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{d^2}{d\xi^2}p_\xi(X)\Big|_{\xi=p} \right]$ is a continuous mapping by the initial assumption that p_θ is continuous in θ for $\theta \in \Theta$.

2.2.3 Estimator Variance & Cramer-Rao Lower Bound

Definition 2.5 (Fisher Information).
Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p_{\boldsymbol{\theta}_*}$ be a random sample from a distribution $p_{\boldsymbol{\theta}_*}$ with the true unknown parameter $\boldsymbol{\theta}_*$. The Fisher Information about $\boldsymbol{\theta}_*$ in the random sample $\mathbf{X}$ is defined as follows:

$$\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) = \mathbb{E}_{\boldsymbol{\theta}_*}\left[\left(\frac{\partial}{\partial \xi} \log L(\xi; \mathbf{X})\bigg|_{\xi = \boldsymbol{\theta}_*}\right)^2\right], \tag{48}$$

where $\mathbb{E}_{\boldsymbol{\theta}_*}$ denotes the expectation conditioned on $\boldsymbol{\theta}_*$. The Fisher Information is the total information about $\boldsymbol{\theta}_*$ contained in the sample $\mathbf{X}$.

Remark 2.1.
Some Notes on Fisher Information:

1. *The expectation $\mathbb{E}_\theta$ conditioned on θ means that the expectation is taken over the sample $\mathbf{X}$ under the hypothesis that θ is the true parameter of all the X_i's.*

2. *Inside the expectation is the score function evaluated at $\boldsymbol{\theta}_*$.*

3. *From two points above, the Fisher Information is essentially calculated as **the variance conditioned on $\boldsymbol{\theta}_*$ of the score evaluated at $\boldsymbol{\theta}_*$.***

Proposition 2.3: Alternative Representation of Fisher Information

Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p_{\boldsymbol{\theta}_*}$ be a random sample from a distribution parameterized by a true unknown parameter $\boldsymbol{\theta}_*$. Let $\ell(\theta; \mathbf{X}) = \log L(\theta; \mathbf{X})$ be the log-likelihood function. Then, we have:

$$\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) = \mathrm{Var}(s(\boldsymbol{\theta}_*; \mathbf{X})), \tag{49}$$

where $s(\boldsymbol{\theta}_*; \mathbf{X}) = \frac{\partial \ell(\boldsymbol{\theta}_*; \mathbf{X})}{\partial \boldsymbol{\theta}_*}$ is the score evaluated at the true parameter $\boldsymbol{\theta}_*$. Furthermore, we have:

$$\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) = -\mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{\partial^2}{\partial \xi^2} \ell(\xi; \mathbf{X})\bigg|_{\xi = \boldsymbol{\theta}_*}\right]. \tag{50}$$

Proof (Proposition 2.3).
Denote $s(\boldsymbol{\theta}_*; \mathbf{X})$ as the score evaluated at the true parameter $\boldsymbol{\theta}_*$, we have:

$$\begin{aligned}
\mathrm{Var}(s(\boldsymbol{\theta}_*; \mathbf{X})) &= \mathbb{E}_{\boldsymbol{\theta}_*}\left[\left(\frac{\partial}{\partial \xi} \ell(\xi; \mathbf{X})\bigg|_{\xi = \boldsymbol{\theta}_*}\right)^2\right] - \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{\partial}{\partial \xi} \ell(\xi; \mathbf{X})\bigg|_{\xi = \boldsymbol{\theta}_*}\right]^2 \\
&= \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) - \mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \log p_\xi(X_i)}{\partial \xi}\bigg|_{\xi = \boldsymbol{\theta}_*}\right] \\
&= \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) - \mathbb{E}_{\boldsymbol{\theta}_*}[s(\boldsymbol{\theta}_*; X)]^2,
\end{aligned}$$

where X is identically distributed as $X_1, \ldots, X_n$. We claim that $\mathbb{E}_{\boldsymbol{\theta}_*}[s(\boldsymbol{\theta}_*; X)] = 0$. Let χ be the

sample space, we have:

$$\mathbb{E}_{\boldsymbol{\theta}_*}[s(\boldsymbol{\theta}_*; X)] = \int_\chi p_{\boldsymbol{\theta}_*}(x) \frac{\partial \log p_\xi(x)}{\partial \xi}\bigg|_{\xi=\boldsymbol{\theta}_*} dx$$

$$= \int_\chi p_{\boldsymbol{\theta}_*}(x) \cdot \frac{1}{p_{\boldsymbol{\theta}_*}(x)} \frac{\partial p_{\boldsymbol{\theta}_*}(x)}{\partial \boldsymbol{\theta}_*} dx$$

$$= \int_\chi \frac{\partial p_{\boldsymbol{\theta}_*}(x)}{\partial \boldsymbol{\theta}_*} dx$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}_*} \int_\chi p_{\boldsymbol{\theta}_*}(x) dx = \frac{\partial}{\partial \boldsymbol{\theta}_*} 1 = 0. \qquad \textit{(Leibniz's Differentiation Rule)}$$

Therefore, we have $\boxed{\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) = \mathrm{Var}(s(\boldsymbol{\theta}_*; \mathbf{X}))}$. *Now, we prove the second point:* $\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*) = -\mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{\partial^2}{\partial \xi^2} \ell(\xi; \mathbf{X})\big|_{\xi=\boldsymbol{\theta}_*}\right]$.
Expanding $\frac{\partial^2}{\partial \xi^2} \ell(\xi; \mathbf{X})\big|_{\xi=\boldsymbol{\theta}_*}$, *we have:*

$$\frac{\partial^2}{\partial \xi^2} \ell(\xi; \mathbf{X})\bigg|_{\xi=\boldsymbol{\theta}_*} = \frac{\partial}{\partial \boldsymbol{\theta}_*}\left[\frac{\partial}{\partial \boldsymbol{\theta}_*} \ell(\boldsymbol{\theta}_*; \mathbf{X})\right]$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}_*}\left[\frac{1}{L(\boldsymbol{\theta}_*; \mathbf{X})} \frac{\partial}{\partial \boldsymbol{\theta}_*} L(\boldsymbol{\theta}_*; \mathbf{X})\right]$$

$$= -\frac{1}{L(\boldsymbol{\theta}_*; \mathbf{X})^2}\left[\frac{\partial}{\partial \boldsymbol{\theta}_*} L(\boldsymbol{\theta}_*; \mathbf{X})\right]^2 + \frac{1}{L(\boldsymbol{\theta}_*; \mathbf{X})} \frac{\partial^2 L(\boldsymbol{\theta}_*; \mathbf{X})}{\partial \boldsymbol{\theta}_*{}^2}$$

$$= -\left[\frac{\partial}{\partial \boldsymbol{\theta}_*} \ell(\boldsymbol{\theta}_*; \mathbf{X})\right]^2 + \frac{1}{L(\boldsymbol{\theta}_*; \mathbf{X})} \frac{\partial^2 L(\boldsymbol{\theta}_*; \mathbf{X})}{\partial \boldsymbol{\theta}_*{}^2}.$$

Now, let $P_{\boldsymbol{\theta}_*}$ *be the probability density function for the random sample* $\mathbf{X}$. *Then, essentially,*
$P_{\boldsymbol{\theta}_*}(\mathbf{X}) = L(\boldsymbol{\theta}_*; \mathbf{X})$. *Therefore, we have:*

$$\mathbb{E}_{\mathbf{X} \sim P_{\boldsymbol{\theta}_*}}\left[\frac{1}{L(\boldsymbol{\theta}_*; \mathbf{X})} \frac{\partial^2 L(\boldsymbol{\theta}_*; \mathbf{X})}{\partial \boldsymbol{\theta}_*{}^2}\right] = \int_{\chi^n} \frac{1}{L(\boldsymbol{\theta}_*; \mathbf{x})} \frac{\partial^2 L(\boldsymbol{\theta}_*; \mathbf{x})}{\partial \boldsymbol{\theta}_*{}^2} P_{\boldsymbol{\theta}_*}(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\chi^n} \frac{\partial^2 L(\boldsymbol{\theta}_*; \mathbf{x})}{\partial \boldsymbol{\theta}_*{}^2} d\mathbf{x}$$

$$= \frac{\partial^2}{\partial \boldsymbol{\theta}_*{}^2} \int_{\chi^n} L(\boldsymbol{\theta}_*; \mathbf{x}) d\mathbf{x} \qquad \textit{(Leibniz's Differentiation Rule)}$$

$$= \frac{\partial^2}{\partial \boldsymbol{\theta}_*{}^2} 1 = 0.$$

Therefore, we have:

$$-\mathbb{E}_{\boldsymbol{\theta}_*}\left[\frac{\partial^2}{\partial \xi} \ell(\xi; \mathbf{X})\bigg|_{\xi=\boldsymbol{\theta}_*}\right] = \mathbb{E}_{\boldsymbol{\theta}_*}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}_*} \ell(\boldsymbol{\theta}_*; \mathbf{X})\right)^2\right] = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*),$$

as desired. □.

Remark 2.2 (Exchanging Derivative and Integral). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯
*Note that in the proof of proposition 2.3, we exchanged the integration and the partial derivative
to prove that* $\mathbb{E}_{\boldsymbol{\theta}_*}[s(\boldsymbol{\theta}_*; X)] = 0$. *This exchange of differentiation and integral can be done under
certain conditions, which are listed in the appendix, section I.1.*

Theorem 2.3: Cramer-Rao Lower Bound (CRLB)

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a random sample where $X_i \sim p_{\boldsymbol{\theta}_*}$. If $\hat{\Theta}(\mathbf{X})$ is an estimator (biased or unbiased) for $\boldsymbol{\theta}_*$ based on the random sample $\mathbf{X}$, then, we have:

$$\operatorname{Var}(\hat{\Theta}(\mathbf{X})) \geqslant \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)^{-1} \left[\frac{d}{d\boldsymbol{\theta}_*} \mathbb{E}[\hat{\Theta}(\mathbf{X})] \right]^2. \tag{51}$$

When $\hat{\Theta}$ is unbiased, we have $\mathbb{E}[\hat{\Theta}(\mathbf{X})] = \boldsymbol{\theta}_*$. Hence, we have $\boxed{\operatorname{Var}(\hat{\Theta}(\mathbf{X})) \geqslant \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)^{-1}}$.

Proof (Cramer-Rao Lower Bound (**CRLB**)). ─────────────────────

We make use of the following covariance inequality: For two random variables X, Y, the following inequality holds

$$\operatorname{Cov}(X, Y)^2 \leqslant \operatorname{Var}(X) \cdot \operatorname{Var}(Y). \tag{52}$$

We denote X, Y as follows:

$$X = \hat{\Theta}(\mathbf{X}); \qquad Y = s(\boldsymbol{\theta}_*; \mathbf{X}).$$

Applying the above inequality, we have:

$$\operatorname{Var}(\hat{\Theta}(\mathbf{X})) \geqslant \frac{\operatorname{Cov}(\hat{\Theta}(\mathbf{X}), s(\boldsymbol{\theta}_*; \mathbf{X}))}{\operatorname{Var}(s(\boldsymbol{\theta}_*; \mathbf{X}))} = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}_*)^{-1} \operatorname{Cov}(\hat{\Theta}(\mathbf{X}), s(\boldsymbol{\theta}_*; \mathbf{X})).$$

Now, we derive the lower bound for $\operatorname{Cov}(\hat{\Theta}(\mathbf{X}), s(\boldsymbol{\theta}_*; \mathbf{X}))$. Suppose that the random variables X_i maps from a sample space to χ and $P_{\boldsymbol{\theta}_*} : \chi^n \to [0, 1]$ is the probability density for the random sample $\mathbf{X}$, we have:

$$\operatorname{Cov}(\hat{\Theta}(\mathbf{X}), s(\boldsymbol{\theta}_*; \mathbf{X})) = \mathbb{E}[\hat{\Theta}(\mathbf{X}) s(\boldsymbol{\theta}_*; \mathbf{X})] - \underbrace{\mathbb{E}[\hat{\Theta}(\mathbf{X})] \mathbb{E}[s(\boldsymbol{\theta}_*; \mathbf{X})]}_{Equals\ 0}$$

$$= \int_{\chi^n} \hat{\Theta}(\mathbf{x}) s(\boldsymbol{\theta}_*; \mathbf{x}) P_{\boldsymbol{\theta}_*}(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\chi^n} \hat{\Theta}(\mathbf{x}) \frac{d}{d\boldsymbol{\theta}_*} \ln L(\boldsymbol{\theta}_*; \mathbf{x}) P_{\boldsymbol{\theta}_*}(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\chi^n} \hat{\Theta}(\mathbf{x}) \frac{\frac{d}{d\boldsymbol{\theta}_*} L(\boldsymbol{\theta}_*; \mathbf{x})}{L(\boldsymbol{\theta}_*; \mathbf{x})} P_{\boldsymbol{\theta}_*}(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\chi^n} \hat{\Theta}(\mathbf{x}) \frac{d}{d\boldsymbol{\theta}_*} L(\boldsymbol{\theta}_*; \mathbf{x}) d\mathbf{x} \qquad (L(\boldsymbol{\theta}_*; \mathbf{x}) = P_{\boldsymbol{\theta}_*}(\mathbf{x}))$$

$$= \frac{d}{d\boldsymbol{\theta}_*} \int_{\chi^n} \hat{\Theta}(\mathbf{x}) L(\boldsymbol{\theta}_*; \mathbf{x}) d\mathbf{x} \qquad (Leibniz\ Differentiation\ Rule)$$

$$= \frac{d}{d\boldsymbol{\theta}_*} \mathbb{E}[\hat{\Theta}(\mathbf{X})].$$

Hence, we obtain the desired lower bound. □.

3 Concentration Inequalities

3.1 Sub-Gaussian Distributions

A random variable X is called "sub-Gaussian" if its distribution has a strong tail decay similar to that of a Gaussian distribution. More precisely, the tail of X's probability density/mass function is dominated by the tail of a Gaussian distribution, hence the name "sub-Gaussian".

Sub-Gaussian is a particularly interesting family of random variables because by modelling random events using sub-Gaussian distributions, we can bound the probability of very rare events happening at the tail. Furthermore, sub-Gaussianity implies the following properties:

  (i) Bounded **tail probability**.

 (ii) Bounded **moment**.

(iii) Bounded **moment generating function** (for X and X^2).

 (iv) Bounded **sub-Gaussian norm***.

Definition 3.1 (Sub-Gaussian Random Variable). ─────────────────────
A random variable X is called sub-Gaussian if there exists a constant $C > 0$ such that:

$$\mathbb{P}(|X| \geqslant t) \leqslant 2\exp(-t^2/C^2). \tag{53}$$

Definition 3.2 (Sub-Gaussian Norm). ─────────────────────
Let X be a random variable. Then, we denote the sub-Gaussian norm of X as:

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(X^2/t^2\right)\right] \leqslant 2\right\}. \tag{54}$$

*From the above definition and the definition of sub-Gaussian variables, we can deduce that X is sub-Gaussian if and only if $\|X\|_{\psi_2} < \infty$. Furthermore, the intuition behind $\|.\|_{\psi_2}$ norm is that it measures the light-tailed-ness of a random variable. In other words, **the smaller the sub-Gaussian norm, the faster the tail decays**.*

Definition 3.3 (Sub-Gaussian Variance Proxy). ─────────────────────
Let X be a sub-Gaussian random variable. Then, we say that X is sub-Gaussian with variance proxy σ^2 if:

$$M_X(\lambda) \leqslant \exp(\lambda^2\sigma^2/2), \quad \forall \lambda \in \mathbb{R}. \tag{55}$$

We denote that $X \in \mathcal{G}(\sigma^2)$.

> **Proposition 3.1: Properties of Sub-Gaussian Variables**
>
> Let X be a random variable with mean μ. Then, the following properties are equivalent:
>
> (i) **Sub-Gaussianity (bounded tail probability)**: There exists $K_1 > 0$ such that
>
> $$\mathbb{P}(|X| \geqslant t) \leqslant 2 \exp\left(-\frac{t^2}{K_1^2}\right). \tag{56}$$
>
> (ii) **Bounded p-moment**: There exists $K_2 > 0$ such that
>
> $$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leqslant K_2 \sqrt{p}, \quad \forall p \geqslant 1. \tag{57}$$
>
> (iii) **Bounded MGF (of X^2)**: There exists $K_3 > 0$ such that
>
> $$M_{X^2}(\lambda^2) \leqslant \exp(K_3^2 \lambda^2), \quad \forall \lambda \in \mathbb{R} \text{ and } |\lambda| < \frac{1}{K_3}. \tag{58}$$
>
> (iv) **Bounded MGF (of X)**: There exists $K_4 > 0$ such that
>
> $$M_{X-\mu}(\lambda) \leqslant \exp(K_4^2 \lambda^2), \quad \forall \lambda \in \mathbb{R}. \tag{59}$$
>
> In other words, $X - \mu$ has a variance proxy of $\sigma^2 \leqslant 2K_4^2$.
>
> (v) **Bounded sub-Gaussian norm**: There exists $K_5 > 0$ such that
>
> $$\mathbb{E}\left[\exp\left(X^2/K_5^2\right)\right] \leqslant 2. \tag{60}$$
>
> In other words, $\|X\|_{\psi_2} \leqslant K_5$.
>
> The parameters $K_1, \ldots, K_5$ differ from each other by at most an absolute constant. Meaning, there exists a constant C independent of $K_1, \ldots, K_5$ such that $K_i \leqslant C K_j$ for any two $i, j \in \{1, \ldots, 5\}$.

Proof (Proposition 3.1). ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

We prove that **(i)** $\implies$ **(ii)** $\implies$ **(iii)** $\implies$ **(v)** *then prove that* **(v)** $\implies$ **(i)**. *This means that the statements* **(i)**, **(ii)**, **(iii)**, **(v)** *are equivalent. Then, we prove that* **(ii)** $\iff$ **(iv)**. *Without loss of generality, assume that $K_1 = 1$ because we can rescale X to X/K_1.*

1. **(i)** $\implies$ **(ii)**: *By the integral identity for p^{th} moments, we have*

$$\mathbb{E}|X|^p = \int_0^\infty p t^{p-1} \mathbb{P}(|X| > t) dt$$

$$\leqslant 2p \int_0^\infty t^{p-1} e^{-t^2} dt.$$

Recall that $\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$. By the change of variable, let $u = t^2$ (then, we have $du = 2t dt$). We have:

$$\mathbb{E}|X|^p \leqslant p \int_0^\infty (\sqrt{u})^{p-2} e^{-u} du$$

$$= p \int_0^\infty u^{p/2-1} e^{-u} du = p\Gamma(p/2).$$

Since $\Gamma(z) < 3z^z$ for all $z \geqslant 1/2$, we have $\mathbb{E}|X|^p \leqslant 3p(p/2)^{p/2}$. Taking p^{th} root from both sides, we can have $K_2 = 3$ since $[3p(p/2)^{p/2}]^{1/p} \leqslant 3\sqrt{p}$.

2. **(ii)** $\implies$ **(iii)**: *Using the identity $e^x = 1 + \sum_{k=1}^{\infty} \frac{x^k}{k!}$, we have:*

$$M_{X^2}(\lambda^2) = \mathbb{E} \exp(X^2 \lambda^2)$$
$$= \mathbb{E}\left[1 + \sum_{k=1}^{\infty} \frac{(\lambda^2 X^2)^k}{k!} \right]$$
$$= 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k} \mathbb{E}[X^{2k}]}{k!}.$$

By property **(ii)***, we have $\mathbb{E}[X^{2k}] \leqslant K_2^{2k}(2k)^k$. By Stirling's approximation, we have $k! \geqslant (k/e)^k$. Combining the bounds, we have:*

$$M_{X^2}(\lambda^2) \leqslant 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k} K_2^{2k} (2k)^k}{(k/e)^k} = 1 + \sum_{k=1}^{\infty} \frac{(2k\lambda^2 K_2^2)^k}{(k/e)^k} = 1 + \sum_{k=1}^{\infty} (2e\lambda^2 K_2^2)^k.$$

When $2e\lambda^2 K_2^2 < 1$, we have the sum of geometric series that converges to $\frac{1}{1-2e\lambda^2 K_2^2}$. Hence, we need $|\lambda| < \frac{1}{K_2\sqrt{2e}}$. Therefore, property **(iii)** *holds with $K_3 = K_2\sqrt{2e}$.*

3. **(iii)** $\implies$ **(v)**: *Let $K_5 = \frac{K_3}{\sqrt{\ln 2}}$ and set $\lambda = \frac{1}{K_5}$. Since $K_5^{-1} < \frac{1}{K_3}$, statement* **(iii)** *holds with $\lambda = K_5^{-1}$. Therefore:*

$$\mathbb{E}\exp\left[X^2/K_5^2 \right] \leqslant \exp(K_3^2/K_5^2) = \exp(\ln 2) = 2.$$

4. **(v)** $\implies$ **(i)**: *Let $K_5 > 0$ be the constant that makes* **(v)** *hold. We have*

$$\mathbb{P}(|X| \geqslant t) = \mathbb{P}(X^2 \geqslant t^2)$$
$$= \mathbb{P}\left(e^{X^2/K_5^2} \geqslant e^{t^2/K_5^2} \right)$$
$$\leqslant \exp\left(-\frac{t^2}{K_5^2} \right) \mathbb{E}\left[e^{X^2/K_5^2} \right] \quad \text{(Markov's Inequality)}$$
$$\leqslant 2\exp\left(-\frac{t^2}{K_5^2} \right).$$

Hence, **(i)** *holds with $K_1 \geqslant K_5$. From here on, we have proved that the statements* **(i)**, **(ii)**, **(iii)** *and* **(v)** *are equivalent.*

5. **(i)** $\iff$ **(iv)**: *This relationship is proven below in theorem 3.1.*

$\square$.

Theorem 3.1: Bounded Tail Probability $\iff$ $\mathcal{SG}$ (BTP – SG)

Let X be a random variable with mean $\mathbb{E}[X] = \mu$ and Z be the centered random variable $Z = X - \mu$. Then, for all $t \in \mathbb{R}$,

(i) $Z \in \mathcal{SG}(\sigma^2) \implies \mathbb{P}(|X - \mu| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2} \right)$ for some $\sigma > 0$.

(ii) $\mathbb{P}(|X - \mu| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2\xi^2} \right) \implies Z \in \mathcal{SG}(16\xi^2)$ for some $\xi > 0$.

Proof (Theorem (**BTP − SG**)). ─────────────────────────────────
We prove each bullet point one by one.

(i) *From the definition, for all $\lambda \in \mathbb{R}$, we have $M_Z(\lambda) = \mathbb{E}\exp[\lambda Z] \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$. For $\lambda > 0$ and for all $t > 0$, we have:*

$$\mathbb{P}(|X - \mu| \geqslant t) = \mathbb{P}\left(\left\{X - \mu \geqslant t\right\} \vee \left\{X - \mu \leqslant -t\right\}\right)$$
$$\leqslant \mathbb{P}(Z \geqslant t) + \mathbb{P}(Z \leqslant -t)$$
$$= \mathbb{P}(e^{\lambda Z} \geqslant e^{\lambda t}) + \mathbb{P}(Z \leqslant -t)$$

Then, we have the following inequalities:

$$\begin{cases} \mathbb{P}(e^{\lambda Z} \geqslant e^{\lambda t}) & \leqslant e^{-\lambda t}M_Z(\lambda) & \text{(Markov's Inequality)} \\ \mathbb{P}(Z \leqslant -t) & \leqslant e^{-\lambda t}M_Z(-\lambda) & \text{(Chernoff's Bound)} \end{cases},$$

where the second inequality comes from the left-tail Chernoff's Bound[4]. As a result, we have:

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant e^{-\lambda t}\left[M_Z(\lambda) + M_Z(-\lambda)\right]$$
$$\leqslant 2e^{-\lambda t}\exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$
$$= 2\exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right).$$

Using Lagrange multiplier to solve for λ, we have $\lambda = \frac{t}{\sigma^2}$. Hence, we have:

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

(ii) *Firstly, we prove that bounded tail probability implies that $\mathbb{E}|Z|^{2q} \leqslant q!(4\xi^2)^q$ for all integers $q \geqslant 1$. Using the identity $\mathbb{E}|Z|^q = \int_0^\infty qt^{q-1}\mathbb{P}(|Z| \geqslant t)dt$, we have:*

$$\mathbb{E}|Z|^{2q} = 2q\int_0^\infty t^{2q-1}\mathbb{P}(|Z| \geqslant t)dt$$
$$\leqslant 4q\int_0^\infty t^{2q-1}\exp\left(-\frac{t^2}{2\xi^2}\right)dt.$$

Letting $u = \frac{t^2}{2\xi^2}$, hence $t^2 = 2u\xi^2$ and $dt = \frac{\xi^2 du}{t}$, the above integral becomes:

$$\mathbb{E}|Z|^{2q} \leqslant 4q\xi^2\int_0^\infty t^{2q-2}e^{-u}du$$
$$= 4q\xi^2\int_0^\infty (2u\xi^2)^{q-1}e^{-u}du$$
$$= 2q \cdot (2\xi^2)^q\underbrace{\int_0^\infty u^{q-1}e^{-u}du}_{\Gamma(q)}$$
$$= 2q!(2\xi^2)^q \leqslant q!(4\xi^2)^q.$$

───
[4] $\mathbb{P}(X \leqslant a) \leqslant e^{-sa}M_X(s)$ for all $s < 0$. Hence, $\mathbb{P}(Z \leqslant -t) \leqslant e^{st}M_Z(s)$ for all $s < 0$. Setting $s = -\lambda$, we have $\mathbb{P}(Z \leqslant -t) \leqslant e^{-\lambda t}M_Z(-\lambda)$.

Let $\tilde{Z}$ be the i.i.d copy of Z. Hence, $Z - \tilde{Z}$ is symmetric about 0, which means that $\mathbb{E}[(Z - \tilde{Z})^p] = 0$ for odd-order p-moments. Therefore, For all $\lambda > 0$, we have:

$$M_Z(\lambda)M_{-\tilde{Z}}(\lambda) = M_{Z-\tilde{Z}}(\lambda) \quad (\text{Due to independence})$$
$$= \mathbb{E}\exp\left(\lambda(Z - \tilde{Z})\right)$$
$$= 1 + \sum_{q=1}^{\infty} \frac{\lambda^{2q}\mathbb{E}\left[(Z - \tilde{Z})^{2q}\right]}{(2q)!}.$$

By the convexity of $f(z) = z^{2q}$, for all $t \in (0, 1)$, we have:

$$\left[tZ + (1-t)(-\tilde{Z})\right]^{2q} \leqslant tZ^{2q} + (1-t)\tilde{Z}^{2q}.$$

Setting $t = \frac{1}{2}$, we have:

$$\left[\frac{Z - \tilde{Z}}{2}\right]^{2q} \leqslant \frac{Z^{2q} + \tilde{Z}^{2q}}{2} \implies (Z - \tilde{Z})^{2q} \leqslant 2^{2q-1}(Z^{2q} + \tilde{Z}^{2q}).$$

As a result, we have $\mathbb{E}[(Z - \tilde{Z})^{2q}] \leqslant 2^{2q-1}(\mathbb{E}[Z^{2q}] + \mathbb{E}[\tilde{Z}^{2q}]) = 2^{2q}\mathbb{E}[Z^{2q}]$. Plugging this back to the formula of $M_{Z-\tilde{Z}}(\lambda)$, we have:

$$\mathbb{E}[e^{\lambda Z}]\mathbb{E}[e^{-\lambda\tilde{Z}}] \leqslant 1 + \sum_{q=1}^{\infty} \frac{\lambda^{2q}2^{2q}\mathbb{E}[Z^{2q}]}{(2q)!}$$
$$\leqslant 1 + \sum_{q=1}^{\infty} \frac{\lambda^{2q}2^{2q}(4\xi^2)^q q!}{(2q)!}.$$

Since $\mathbb{E}[e^{-\lambda\tilde{Z}}] \geqslant 1$ for all $\lambda > 0$ and

$$\frac{(2q)!}{q!} = \prod_{j=1}^{q}(q + j) \geqslant \prod_{j=1}^{q}(2j) = 2^q q!,$$

We have:

$$\mathbb{E}[e^{\lambda Z}] \leqslant 1 + \sum_{q=1}^{\infty} \frac{(2\lambda^2 \cdot 4\xi^2)^q}{q!} = e^{8\lambda^2\xi^2}.$$

Hence, $Z \in \mathcal{SG}(16\xi^2)$.

$\square$.

I . Appendix

I.1   Leibniz Differentiation Rule

Leibniz Differentiation/Integral Rule refers to the conditions by which differentiation and integration can be exchanged. Essentially, let Ω be a measure space and $A \subset \mathbb{R}^n$ be an open subset. Then, if a function $f : A \times \Omega \to \mathbb{R}$ satisfies:

1. For all $x \in A$, $\omega \mapsto f(x, \omega)$ is a measurable function in ω (f is a <u>Caratheodory Function</u>).

2. For almost all $\omega \in \Omega$, $\frac{\partial f(x,\omega)}{\partial x}$ exists for all $x \in A$.

3. For almost all $\omega \in \Omega$, there is an integrable function $h : \Omega \to \mathbb{R}$ such that $|f(x, \omega)| \leqslant h(\omega)$ for all $x \in A$.

Then, we have:

1. $g(x) = \int_\Omega f(x, \omega) d\mu(\omega)$ is continuous.

2. $D_i g(x) = \int_\Omega D_i f(x, \omega) d\mu(\omega)$.

Definition I.1 (Caratheodory Function). ─────────────────────────────
Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $A \subset \mathbb{R}^n$. We say $f : A \times \Omega \to \mathbb{R}$ is a Caratheodory Function if it satisfies:

1. *For all $x \in A$: $\omega \mapsto f(x, \omega)$ is $\mathcal{F}$-measurable.*

2. *For all $\omega \in \Omega$: $x \mapsto f(x, \omega)$ is continuous.*

We also say that f is a Caratheodory function if it is measurable in ω and continuous in x.

> **Proposition I.1: Leibniz Differentiation Rule**
>
> Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $A \subset \mathbb{R}^n$ be open. Let the function $f : A \times \Omega \to \mathbb{R}$ satisfy the following:
>
> 1. f is a Caratheodory function.
>
> 2. There exists $h : \Omega \to \mathbb{R}$ such that $|f(x, \omega)| \leqslant h(\omega)$ for almost all $\omega \in \Omega$ for all $x \in A$.
>
> Then, if we define $g(x) = \int_\Omega f(x, \omega) d\mu(\omega)$, we have:
>
> (i) g is continuous.
>
> (ii) $D_i g(x) = \int_\Omega D_i f(x, \omega) d\mu(\omega)$.

Proof (Proposition I.1). ─────────────────────────────
We prove the above points one by one:

(i) *Suppose that $\{x_n\}_{n=1}^\infty \subset A$ is a sequence such that $x_n \to x \in A$. We have to prove that $g(x_n) \to g(x)$. Since f is a Caratheodory function, we have $x \mapsto f(x, \omega)$ is continuous for almost all $\omega \in \Omega$. Therefore, $f(x_n, \omega) \to f(x, \omega)$ for almost all $\omega \in \Omega$. Then, by the Dominated Convergence Theorem[5]:*

$$\int_\Omega f(x_n, \omega) d\mu(\omega) \to \int_\Omega f(x, \omega) d\mu(\omega).$$

Hence, we have $g(x_n) \to g(x)$ as desired.

---
[5]DCT is applied for the sequence of functions $\{k_n\}_{n=1}^\infty$ where $k_n : \Omega \to \mathbb{R}$ and $k_n(\omega) = f(x_n, \omega)$.

(ii) *Let e_i denotes the ith standard basis in $\mathbb{R}^n$. We have:*

$$D_i f(x, \omega) = \lim_{\xi \to 0} \frac{f(x + \xi e_i, \omega) - f(x, \omega)}{\xi}.$$

Furthermore, we have:

$$\begin{aligned}
D_i g(x) &= \lim_{\xi \to 0} \frac{g(x + \xi e_i) - g(x)}{\xi} \\
&= \lim_{\xi \to 0} \xi^{-1} \left[\int_\Omega f(x + \xi e_i, \omega) d\mu(\omega) - \int_\Omega f(x, \omega) d\mu(\omega) \right] \\
&= \lim_{\xi \to 0} \int_\Omega \frac{f(x + \xi e_i, \omega) - f(x, \omega)}{\xi} d\mu(\omega)
\end{aligned}$$

By the Mean Value Theorem, there exists a point $y(\xi, \omega)$ between x and $x + \xi e_i$ such that:

$$f(x + \xi e_i, \omega) - f(x, \omega) = \xi D_i f(y(\xi, \omega), \omega).$$

Therefore, we have:

$$\begin{aligned}
D_i g(x) &= \lim_{\xi \to 0} \int_\Omega D_i f(y(\xi, \omega), \omega) d\mu(\omega) \\
&= \lim_{n \to \infty} \int_\Omega D_i f(y(n^{-1}, \omega)) d\mu(\omega) \\
&= \int_\Omega \lim_{n \to \infty} D_i f(y(n^{-1}, \omega)) d\mu(\omega) \qquad \text{(Dominated Convergence Theorem)}.
\end{aligned}$$

As $n \to \infty$, we have $y(n^{-1}, \omega) \to x$ since it lies between x and $x + \frac{e_i}{n}$. Hence, we have $D_i f(y(n^{-1}, \omega), \omega) \to D_i f(x, \omega)$. Therefore, $D_i g(x) = \int_\Omega D_i f(x, \omega) d\mu(\omega)$ as desired.

$\square$.

II . List of Definitions

III   . Important Theorems

IV   . Important Propositions

↩ ***Return to Contents***

V . To-do List

1. State and prove the Leibniz Differentiation Rule.

2. State and prove the Cramer-Rao Lower Bound (CRLB).

3. State and prove the Rao-Blackwell Theorem + Work on sufficiency principle.

4. Prove the Asymptotic Normality of the MLE.

5. Start the chapter on Sub-Gaussian variables.

VI . References

References

Bartlett, Peter L., Dylan J. Foster, and Matus Telgarsky (2017). "Spectrally-normalized margin bounds for neural networks". In: *Conference on Neural Information Processing Systems*.

"Chapter 36 Large sample estimation and hypothesis testing" (1994). In: vol. 4. Handbook of Econometrics. Elsevier, pp. 2111–2245. DOI: https://doi.org/10.1016/S1573-4412(05)80005-4. URL: https://www.sciencedirect.com/science/article/pii/S1573441205800054.

Durrett, Rick (2010). *Probability: Theory and Examples*. 4th. USA: Cambridge University Press. ISBN: 0521765390.

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.

Gut, Allan (2004). *A Graduate Course in Probability*. Graduate Text in Mathematics.

Long, Philip M. and Hanie Sedghi (2020). "Generalization Bounds for Deep Convolutional Neural Networks". In: *International Conference on Learning Representation*.

Pishro-Nik, Hossein (2014). *Introduction to Probability, Statistics and Random Processes*. Kappa Research, LLC.

undefinedinlar, Erhan (2011). *Probability and Stochastics*. Springer New York. ISBN: 9780387878591. URL: http://dx.doi.org/10.1007/978-0-387-87859-1.

Wikipedia (2023). *Vitali set — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Vitali%20set&oldid=1187241923. [Online; accessed 24-December-2023].