High Dimensional Probability Notes

Nong Minh Hieu[1]

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

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 p^{th} root of the p^{th} 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:

$$\begin{aligned}
\mathbb{E}X &= \mathbb{E}\left[\int_0^\infty \mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbf{1}\{t > X\}dt\right] \\
&= \mathbb{E}\int_0^\infty \mathbf{1}\{t < X\}dt - \mathbb{E}\int_{-\infty}^0 \mathbf{1}\{t > X\}dt \\
&= \int_0^\infty \mathbb{E}\mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbb{E}\mathbf{1}\{t > X\}dt \\
&= \int_0^\infty \mathbb{P}(t < X)dt - \int_{-\infty}^0 \mathbb{P}(t > X)dt.
\end{aligned}$$

□.

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:*

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

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**).

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* **book:allen2004** □.

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* **book:allen2004***. 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 $\mathbb{P}\Big(\limsup_{N\to\infty} |S_N/N| > \epsilon\Big) = 0$ *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:

$$\begin{aligned}
\mathbb{E}[S_n^4] &= \sum_{i=1}^n \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1\leqslant i<j\leqslant n} \mathbb{E}[X_i^2 X_j^2] \\
&= \sum_{i=1}^n \mathbb{E}[X_i^4] + 6 \underbrace{\sum_{1\leqslant i<j\leqslant n} \mathbb{E}[X_i^2 X_j^2]}_{n(n-1)/2 \text{ terms}} \\
&= nK + 3n(n-1)\mathbb{E}[X_i^2 X_j^2].
\end{aligned}$$

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:

$$\begin{aligned}
\mathbb{P}(|S_n/n| \geqslant \epsilon) &= \mathbb{P}(|S_n| \geqslant n\epsilon) \\
&\leqslant \frac{\mathbb{E}[S_n^4]}{n^4\epsilon^4} \\
&< \frac{K}{n^3\epsilon^4} + \frac{3(\sigma^2+\mu^2)}{n^2}.
\end{aligned}$$

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:

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} B_n \right) &= \lim_{n\to\infty} \mathbb{P}(B_n) \\
&= \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_n) \quad \textit{(By additivity)} \\
&= \sum_{i=1}^{\infty} \mathbb{P}(A_i) - \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(A_n) \\
&= 0.
\end{aligned}$$

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:

$$\begin{aligned}
1 - \mathbb{P}\left(\limsup_{n\to\infty} A_n \right) &= \mathbb{P}\left(\left\{ \limsup_{n\to\infty} A_n \right\}^c \right) \\
&= \mathbb{P}\left(\liminf_{n\to\infty} A_n^c \right) = 0.
\end{aligned}$$

□.

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 $\mathbf{X} = (X_1, \dots, X_n)$ 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 **book:ferguson1996**. *However, we will conduct our own version of the proof relying on results in learning theory. For $\theta \in \Theta$, define the following function:*

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

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

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\theta \in \Theta} |\phi_\theta(\mathbf{X})| \geqslant \epsilon \Big) &= \mathbb{P}\Big(\Big\{ \sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \geqslant \epsilon \Big\} \cup \Big\{ \sup_{\theta \in \Theta} (-\phi_\theta(\mathbf{X})) \geqslant \epsilon \Big\} \Big) \\
&\leqslant \mathbb{P}\Big(\sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \geqslant \epsilon \Big) + \mathbb{P}\Big(\sup_{\theta \in \Theta} (-\phi_\theta(\mathbf{X})) \geqslant \epsilon \Big) \\
&\leqslant \frac{1}{\epsilon} \Big(\mathbb{E}\Big[\sup_{\theta \in \Theta} \phi_\theta(\mathbf{X}) \Big] + \mathbb{E}\Big[\sup_{\theta \in \Theta} (-\phi_\theta(\mathbf{X})) \Big] \Big) \quad \textit{(Markov's Inequality)}
\end{aligned}$$

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 $S = \{Z_1, \dots, Z_n\} \sim \rho^n$ sampled from a distribution ρ and a class of functions $\mathcal{F}$. Let $S' = \{Z_1', \dots, Z_n'\} \sim \rho^n$ be a another sample *from the same distribution ρ (which we called the phantom sample), we have:*

$$\begin{aligned}
&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\mathbb{E}_S[f(Z_i)])\right]\\
&=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\mathbb{E}_{S'}[f(Z_i')])\right]\\
&=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S'}[f(Z_i')]\right]\\
&=\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-f(Z_i'))\right]\right]\\
&\leqslant\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-f(Z_i'))\right]\quad\textit{(Jensen's Inequality)}\\
&=\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(f(Z_i)-f(Z_i'))\right]\quad(\sigma\sim\mathrm{Rad}^n, f(Z_i)-f(Z_i')\ \textit{is symmetric})\\
&\leqslant\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right]+\mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)f(Z_i')\right]\\
&=\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right]+\mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i')\right]\quad\textit{(Rademacher variables are symmetric)}\\
&=2\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right]\quad(S\ \textit{and}\ S'\ \textit{are identically distributed})\\
&=2\mathfrak{R}_n(\mathcal{F}).
\end{aligned}\tag{27}$$

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:

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

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:

$$\mathfrak{R}_n(\mathcal{F}_\Theta)=\mathbb{E}_{\mathbf{X}\sim p^n_{\boldsymbol{\theta}_*}}\mathbb{E}_\sigma\Big[\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}_\Theta)\Big],$$

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 **article:bartlett2017**, *we have:*

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

where $M<\infty$ is a constant such that $|f_\theta(x)|\leqslant M$ $p_{\boldsymbol{\theta}_*}$-(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 $\mathbb{E}_{\boldsymbol{\theta}_*}[\sup_{\theta\in\Theta}|f_\theta(X)|]<\infty$.

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 **article:long2020***, 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 $X_1, \ldots X_n$ 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:*

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

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 $\mathcal{B}$ be the Borel-σ-algebra and $A \in \mathcal{B}$ 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:

$$\mu(\partial A) = 0, \tag{36}$$

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.

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

$$\int f d\mu_n \to \int f d\mu, \tag{38}$$

for all f continuous and bounded.

Theorem 1.8: Portmanteau's Theorem (PORTMANTEAU)

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 $X_n \xrightarrow{p} X$ 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: $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$

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

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 $\lim_{n\to\infty} F_{X_n}(a) = F_X(a)$. □.

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, **book:hossien2014**).

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

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

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

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

From the above, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - c| \geqslant \epsilon) = 0$ 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^pnorm 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 $p \geqslant 1$ 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:

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

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

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 $\omega \in \Omega$:

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

In other words, we can write:

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

Remark 1.2 (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:

$$\begin{aligned}\underbrace{\mathbb{P}\Big(\limsup_{n\to\infty}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\}\Big)}_{=0} &= \int f_n d\mathbb{P} \\ &\geqslant \limsup_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon) \\ &\geqslant \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon).\end{aligned}$$

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

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

Proof (Continuous Mapping Theorem (**CMT**)).

We handle each statement separately:

1. *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:*

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

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:

$$\begin{aligned}\mathbb{P}\Big(|f(X_n) - f(X)| \geqslant \epsilon\Big) &= \mathbb{P}\Big(\Big\{X \notin C_f\Big\} \cup \Big\{|X_n - X| \geqslant \delta\Big\} \cup \Big\{X \in B_\delta\Big\}\Big)\\ &\leqslant \underbrace{\mathbb{P}(X \notin C_f)}_{=0} + \mathbb{P}(|X_n - X| \geqslant \delta) + \mathbb{P}(X \in B_\delta)\\ &= \mathbb{P}(|X_n - X| \geqslant \delta) + \mathbb{P}(X \in B_\delta).\end{aligned}$$

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:

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

2. *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:

$$\begin{aligned}&\mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} f(X_n(\omega)) = f(X(\omega))\Big\}\Big)\\ &\geqslant \mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1,\end{aligned}$$

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 $\mathbf{X} = (X_1, \dots, X_n) \sim p_{\boldsymbol{\theta}_*}$ 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 sufficient statistic if the conditional distribution $p_{\mathbf{X}|\mathbf{U}}$ does not depend on $\boldsymbol{\theta}_*$.

Example 2.1 (Bernoulli random variables).
Let $\mathbf{X} = (X_1, \dots, X_n) \sim \text{Bernoulli}(\theta)$ 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 $\mathbf{x} = (x_1, \dots, x_n)$ is an observation of the random sample $\mathbf{X}$ and $\mathbf{u} = \frac{1}{n}\sum_{i=1}^n x_i$. We have:

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

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.

□.

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 Maximum Likelihood Estimator $\theta_{MLE} \in \Theta$ 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 **book:newey1994** *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)***:* $\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \xrightarrow{p} \mathbb{E}_{\boldsymbol{\theta}_*}[\log p_\theta(X)], \quad \forall\theta\in\Theta$.
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* $\|M_n - M\|_\infty \xrightarrow{p} 0$*. 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:*

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

where $\hat\theta = \arg\max_{\theta\in\Theta} M_n(\theta)$*. 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:*

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

□.

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 $\boldsymbol{\theta}_*$. Let Θ be the parameters space. Then, the Maximum Likelihood Estimator $\Theta_{MLE} = \arg\max_{\theta\in\Theta} L(\theta;\mathbf{X})$ is asymptotically normal:

$$\frac{\Theta_{MLE} - \boldsymbol{\theta}_*}{\sqrt{\mathrm{Var}(\Theta_{MLE})/n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{47}$$

Proof (Proposition 2.2).

□.

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)

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 parameterized by* $\boldsymbol{\theta}_*$*. The* *(total) Fisher Information* *about* θ *in the random sample* $\mathbf{X}$ *is defined as follows:*

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

The Fisher Informaton is the total information about $\boldsymbol{\theta}_*$ *contained in the sample* $\mathbf{X}$.

Theorem 2.3: Cramer-Rao Lower Bound (CRLB)

2.2.4 Maximum Likelihood Estimation (MLE)

A List of Definitions

B Important Theorems

C Important Propositions

D References