

High Dimensional Probability Notes

Nong Minh Hieu¹

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

Contents

1	Random variables	2
1.1	Basic Inequalities	2
1.2	Limit Theorems	4
1.2.1	Weak Law of Large Numbers	4
1.2.2	Strong Law of Large Numbers	6
1.2.3	Uniform Law of Large Numbers	8
1.2.4	Central Limit Theorem	10
1.3	Convergence of Random Variables	11
1.3.1	Convergence in Distribution	11
1.3.2	Convergence in Probability	12
1.3.3	Convergence in L^p norm	13
1.3.4	Almost-sure Convergence	13
2	Statistical Inference	15
2.1	Sufficiency & Likelihood Principles	15
2.1.1	Sufficiency	15
2.1.2	Likelihood	16
2.2	Point Estimation	18
2.2.1	Bias, Variance, Consistency and MSE	18
2.2.2	Sufficient Statistics & Rao-Blackwell Theorem	18
2.2.3	Estimator Variance & Cramer-Rao Lower Bound	19
2.2.4	Maximum Likelihood Estimation (MLE)	19
A	List of Definitions	20
B	Important Theorems	20
C	Important Propositions	20
D	References	21

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} :

$$X : \Omega \rightarrow \mathbb{R} \quad (1)$$

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:

$$\mathbb{P} : \Sigma \rightarrow [0, 1] \quad (2)$$

For a given probability space $(\Omega, \Sigma, \mathbb{P})$ and a random variable $X : \Omega \rightarrow \mathbb{R}$, 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 .

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}, \quad p \in (0, \infty) \quad (3)$$

$$\|X\|_{L^\infty} = \text{ess sup } |X| \quad (4)$$

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

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

Some basic inequalities and identities:

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

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X) \quad (6)$$

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

$$\|X\|_{L^p} \leq \|X\|_{L^q}, \quad 0 \leq p \leq q \leq \infty. \quad (7)$$

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

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}. \quad (8)$$

- **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:

$$|\mathbb{E}XY| \leq \|X\|_{L^p} \cdot \|Y\|_{L^q}. \quad (9)$$

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

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}. \quad (10)$$

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

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}[|X|^p]}{t^p}, \quad \forall t > 0, p \in [2, \infty). \quad (11)$$

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

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}. \quad (12)$$

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

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

Exercises

Exercise 1.1.1: Generalized Integral Identity

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

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

Solution (Exercise 1.1.1). _____

For $x \in \mathbb{R}$, using the basic integral identity, 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 $x \geq 0 \implies x = |x|$:

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

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:

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

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 $0 \rightarrow \infty$, we also integrate t from $0 \rightarrow \infty$ 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, \dots, X_N be *i.i.d* random variables with mean μ . Consider the sum:

$$S_N = X_1 + \dots + X_N$$

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

$$\lim_{N \rightarrow \infty} \mathbb{P}\left(|S_N/N - \mu| > \epsilon\right) = 0, \quad \forall \epsilon > 0 \quad (16)$$

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

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

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

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

Therefore, we have:

$$\lim_{N \rightarrow \infty} \mathbb{P}\left(|S_N/N - \mu| > \epsilon\right) \leq \lim_{N \rightarrow \infty} \frac{\sigma^2}{N\epsilon^2} = 0.$$

Hence, we have $\lim_{N \rightarrow \infty} \mathbb{P}\left(|S_N/N - \mu| > \epsilon\right) = 0$ 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 $n \geq 1$, 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 $X = X_1 = \dots = X_n$. 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:

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

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, \dots be *i.i.d* random variables. Then:

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

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

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

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

1.2.2 Strong Law of Large Numbers

Theorem 1.3: Strong Law of Large Numbers (SLLN)

Let X_1, \dots, X_N be *i.i.d* random variables with mean $\mu < \infty$. Consider the sum:

$$S_N = X_1 + \dots + X_N$$

Then, the sample mean **converges to μ almost surely** ($S_N/N \xrightarrow{a.s} \mu$):

$$\mathbb{P}\left(\limsup_{N \rightarrow \infty} |S_N/N - \mu| > \epsilon\right) = 0, \quad \forall \epsilon > 0 \quad (21)$$

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

For the sake of simplicity, we will present the proof for (SLLN) with an additional assumption that $\mathbb{E}[|X_n|^4] < \infty, \forall n \geq 1$. 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 $\mathbb{P}\left(\limsup_{N \rightarrow \infty} |S_N/N| > \epsilon\right) = 0$ 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:

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

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

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

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 $1 \leq i \leq n$. Furthermore, we have $\mathbb{E}[X_i^2] = \text{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2$. 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:

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

Therefore, we have:

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

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

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

□.

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:

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

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

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

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:

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

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

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

Furthermore, we have:

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

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

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

□.

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_{θ} over \mathcal{X} that depends on the true parameters θ . Let $\Theta \subset \mathbb{R}^m$ be the parameters space and $f_{\theta} : \mathcal{X} \rightarrow \mathbb{R}$ be a function indexed by $\theta \in \Theta$ that satisfies the following conditions:

1. $\theta \in \Theta$ and Θ is compact.
2. $\mathbb{E}_{\theta}[\sup_{\theta \in \Theta} |f_{\theta}(X)|] < \infty^a$.
3. $f_{\theta}(x)$ is continuous in θ for all $x \in \mathcal{X}$.

Then, we have:

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n f_{\theta}(X_i) - \mathbb{E}_{\theta}[f_{\theta}(X)] \right| \xrightarrow{p} 0. \quad (25)$$

^a \mathbb{E}_{θ} denotes the expectation taken over the distribution described by p_{θ} .

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:

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

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

$$\begin{aligned} \mathbb{P}\left(\sup_{\theta \in \Theta} |\phi_{\theta}(\mathbf{X})| \geq \epsilon\right) &= \mathbb{P}\left(\left\{\sup_{\theta \in \Theta} \phi_{\theta}(\mathbf{X}) \geq \epsilon\right\} \cup \left\{\sup_{\theta \in \Theta} (-\phi_{\theta}(\mathbf{X})) \geq \epsilon\right\}\right) \\ &\leq \mathbb{P}\left(\sup_{\theta \in \Theta} \phi_{\theta}(\mathbf{X}) \geq \epsilon\right) + \mathbb{P}\left(\sup_{\theta \in \Theta} (-\phi_{\theta}(\mathbf{X})) \geq \epsilon\right) \\ &\leq \frac{1}{\epsilon} \left(\mathbb{E}\left[\sup_{\theta \in \Theta} \phi_{\theta}(\mathbf{X})\right] + \mathbb{E}\left[\sup_{\theta \in \Theta} (-\phi_{\theta}(\mathbf{X}))\right] \right) \quad (\text{Markov's Inequality}) \end{aligned}$$

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:

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

Using the same argument, we can also have:

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

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

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

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

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

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

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

where $M < \infty$ is a constant such that $|f_\theta(x)| \leq M$ 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}_Θ . 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 $\|\theta - \theta'\|_2 < \delta_x \implies |f_\theta(x) - f_{\theta'}(x)| < \epsilon$. Define $\delta > 0$ as follows:

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

¹M 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}_Θ 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 $1 \leq j \leq N$ such that:

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

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

$$\log \mathcal{N}(\mathcal{B}(y_j, r_j), \delta, \|\cdot\|_2) \leq m \log \left(\frac{3r_j}{\delta} \right). \quad (32)$$

Therefore, we have:

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

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

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:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty, \quad (34)$$

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 $1 \leq i \leq n$ and suppose $Z = Z_1 = \dots = Z_n$, we have:

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

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

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

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

1.3 Convergence of Random Variables

In this section, we revise the modes of convergence in 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, \dots with CDFs F_1, F_2, \dots . We say that the sequence converges in distribution to a random variable X with CDF F , denoted $X_n \xrightarrow{d} X$ if:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x), \quad (35)$$

for all $x \in \mathbb{R}$ at which F is continuous. Convergence in distribution can also be referred to as weak convergence 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 $c \neq 0$.

^aIn 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:

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

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

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

Similarly, we have:

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

Taking limits of both inequalities, we have:

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

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

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)).

□.

1.3.2 Convergence in Probability

Definition 1.3 (Convergence in Probability). _____

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

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0, \quad \forall \epsilon > 0. \quad (36)$$

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, \dots 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 $\mathbb{P}(Y \leq a) \leq \mathbb{P}(X \leq a + \epsilon) + \mathbb{P}(|Y - X| \geq \epsilon)$ holds. We have:

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

Using the above inequality, we have:

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

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

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

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

□.

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, \dots 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:

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

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

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

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

1.3.3 Convergence in L^p norm

Definition 1.4 (Convergence in L^p norm).

Given a sequence of random variables X_1, X_2, \dots 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:

$$\lim_{n \rightarrow \infty} \mathbb{E}|X_n - X|^p = 0. \quad (37)$$

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

Let $p \geq 1$ and X_1, X_2, \dots 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:

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

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

1.3.4 Almost-sure Convergence

Definition 1.5 (Convergence almost-surely).

Let X_1, X_2, \dots 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:

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

Remark 1.2 (Consequence of **(BCL)**).

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \implies X_n \xrightarrow{a.s.} X. \quad (38)$$

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

Let X_1, X_2, \dots 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 $f_n : \Omega \rightarrow \mathbb{R}_+$ 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:

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

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

Theorem 1.8: Continuous Mapping Theorem (CMT)

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

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)**).

Since almost-sure convergence implies the other two modes of convergence, we only have to handle the almost-sure convergence case. Since f is continuous, for any $\omega \in \Omega$ such that $X_n(\omega) \rightarrow X(\omega)$, we have $f(X_n(\omega)) \rightarrow f(X(\omega))$. Therefore, we have:

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

Therefore, we have:

$$\begin{aligned} &\mathbb{P}\left(\limsup_{n \rightarrow \infty} \left\{\omega \in \Omega : |f(X_n(\omega)) - f(X(\omega))| \leq \epsilon\right\}\right) \\ &\geq \mathbb{P}\left(\limsup_{n \rightarrow \infty} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \leq \epsilon\right\}\right) = 1, \end{aligned}$$

for all $\epsilon > 0$. Therefore, 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 (FactT)

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample with joint density function p_{θ} over \mathcal{X}^n . The statistic $\mathbf{U} = T(\mathbf{X})$ is sufficient for the parameters θ if and only if we can find functions h, g such that:

$$p_{\theta}(\mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x}),$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $\theta \in \Theta$.

Proof (Factorisation Theorem (FactT)).

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:

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

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

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

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

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

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

From the above, we have:

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

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

2.1.2 Likelihood

Definition 2.3 (Likelihood Function).

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

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n p_{\theta}(x_i), \quad \theta \in \Theta. \quad (39)$$

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

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log p_{\theta}(x_i), \quad \theta \in \Theta. \quad (40)$$

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 $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample whose distribution belongs to a family of distributions $\mathcal{P} = \{p_{\theta} : \theta \in \Theta\}$ and let $\mathbf{x} = (x_1, \dots, x_n)$ 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:

$$\theta_{MLE} = \arg \max_{\theta \in \Theta} L(\theta; \mathbf{x}). \quad (41)$$

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

Proposition 2.1: Consistency of MLE

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample from a distribution p_{θ} over \mathcal{X} dependent on a true set of parameters θ . 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 $\Theta_{MLE} \xrightarrow{p} \theta$, provided that the following conditions are met:

1. $\theta \in \Theta$ and Θ is a compact space.
2. $\log p_{\theta}(x)$ is continuous in θ for almost all $x \in \mathcal{X}$.
3. $\mathbb{E}_{\theta}[\sup_{\theta \in \Theta} |\log p_{\theta}(X)|] < \infty$.
4. The mapping $\xi \mapsto p_{\xi}$, $\xi \in \Theta$ is one-to-one (Identifiability).^a

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

^aIn 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 hypothesis testing” 1994, Theorem 2.5 but we attempt our own proof anyway. The general proof strategy is listed below:

1. First, prove that $\theta = \arg \max_{\xi \in \Theta} \mathbb{E}_{\theta}[\log p_{\xi}(X)]$.
2. Then, by (ULLN): $\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \xrightarrow{p} \mathbb{E}_{\theta}[\log p_{\theta}(X)]$, $\forall \theta \in \Theta$.
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:

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

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

$$M_n = \xi \mapsto \frac{1}{n} \sum_{i=1}^n \log p_{\xi}(X_i),$$

and $M = \xi \mapsto \mathbb{E}_{\theta}[\log p_{\xi}(X)]$.

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 $n \geq N$, with probability of at least $1 - \delta$, we have:

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

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

$$M(\hat{\theta}_n) \geq M_n(\hat{\theta}_n) - \frac{\epsilon}{2} \geq M_n(\theta) - \frac{\epsilon}{2} \geq M(\theta) - \epsilon.$$

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

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

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

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\theta - \hat{\theta}_n| < \xi) = 1 \implies \hat{\theta}_n \xrightarrow{P} \theta.$$

□.

Proposition 2.2: Asymptotic Normality of MLE

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample from a distribution p_{θ} dependent on a true set of parameters θ . 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:

$$\frac{\Theta_{MLE} - \theta}{\sqrt{\text{Var}(\Theta_{MLE})/n}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (42)$$

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 $\mathbf{X} = (X_1, \dots, X_n) \sim p(\cdot; \boldsymbol{\theta})$ 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:

$$\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}} \left[\left(\frac{\partial}{\partial \theta} \log L(\theta; \mathbf{X}) \right)^2 \mid \boldsymbol{\theta} \right]. \quad (43)$$

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

1.1	Definition (Random variable)	2
1.2	Definition (Convergence in Distribution)	11
1.3	Definition (Convergence in Probability)	12
1.4	Definition (Convergence in L^p norm)	13
1.5	Definition (Convergence almost-surely)	13
2.1	Definition (Sufficient Statistics)	15
2.2	Definition (Sufficiency Principle)	15
2.3	Definition (Likelihood Function)	16
2.4	Definition (Maximum Likelihood Estimator)	17
2.5	Definition (Fisher Information)	19

B Important Theorems

1.1	Weak Law of Large Numbers (WLLN)	4
1.2	Levy Continuity Theorem (LCT)	5
1.3	Strong Law of Large Numbers (SLLN)	6
1.4	Borel-Cantelli Lemma (BCL)	7
1.5	Uniform Law of Large Numbers (ULLN)	8
1.6	Central Limit Theorem (CLT)	10
1.7	Slutsky's Theorem (SLUTSKY)	11
1.8	Continuous Mapping Theorem (CMT)	14
2.1	Factorisation Theorem (FacT)	16
2.2	Rao-Blackwell Theorem (RB)	18
2.3	Cramer-Rao Lower Bound (CRLB)	19

C Important Propositions

1.1	$X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$	12
1.2	$X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$	12
1.3	$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$	13
1.4	$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X$	14
2.1	Consistency of MLE	17
2.2	Asymptotic Normality of MLE	18

D 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](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.
- undefinedinar, 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].