Class Notes of Stat 8112

1 Bayes estimators

Here are three methods of estimating parameters:

(1) MLE; (2) Moment Method; (3) Bayes Method.

An example of Bayes argument: Let \( X \sim F(x|\theta), \ \theta \in \Theta \). We want to estimate \( g(\theta) \in \mathbb{R}^1 \).

Suppose \( t(X) \) is an estimator and look at

\[
\text{MSE}_\theta(t) = E\theta(t(X) - g(\theta))^2.
\]

The problem is \( \text{MSE}_\theta(t) \) depends on \( \theta \). So minimizing one point may costs at other points.

Bayes idea is to average \( \text{MSE}_\theta(t) \) over \( \theta \) and then minimize over \( t \)’s. Thus we pretend to have a distribution for \( \theta \), say, \( \pi \), and look at

\[
H(t) = E(t(X) - g(\theta))^2
\]

where \( E \) now refers to the joint distribution of \( X \) and \( \theta \), that is

\[
E(t(X) - g(\theta))^2 = \int (t(x) - g(\theta))^2 F(dx|\theta)\pi(d\theta).
\]

Next pick \( t(X) \) to minimize \( H(t) \). The minimizer is called the Bayes estimator.

**Lemma 1.1** Suppose \( Z \) and \( W \) are real random variables defined on the same probability space and \( \mathcal{H} \) be the set of functions from \( \mathbb{R}^1 \) to \( \mathbb{R}^1 \). Then

\[
\min_{h \in \mathcal{H}} E(Z - h(W))^2 = E(Z - E(Z|W))^2.
\]

That is, the minimizer above is \( E(Z|W) \).

**Proof.** Note that

\[
E(Z - h(W))^2 = E(Z - E(Z|W) + E(Z|W) - h(W))^2
= E(Z - E(Z|W))^2 + E(E(Z|W) - h(W))^2
+ 2E\{(Z - E(Z|W))(E(Z|W) - h(W))\}.
\]

Conditioning on \( W \), we have that the cross term is zero. Thus

\[
E(Z - h(W))^2 = E(Z - E(Z|W))^2 + E(E(Z|W) - h(W))^2.
\]
Now $E(Z - E(Z|W))^2$ is fixed. The second term $E(E(Z|W) - h(W))^2$ depends on $h$. So choose $h(W) = E(Z|W)$ to minimize $E(Z - h(W))^2$. ■

Now choose $W = X$, $t = h$ and $Z = g(\theta)$. Then

**Corollary 1.1** The Bayes estimator in (1.1) is $\hat{g}(\theta) = E(g(\theta)|X)$, which is the posterior mean.

At a later occasion, we call $E(g(\theta)|X)$ the posterior mean of $g(\theta)$. Similarly, we have the following multivariate analogue of the above corollary. The proof is in the same spirit of Lemma 1.1. We omit it.

**Theorem 1** Let $X \in \mathbb{R}^m$ and $X \sim F(\cdot|\theta)$, $\theta \in \mathbb{R}^k$. Let $g(\theta) = (g_1(\theta), \cdots, g_k(\theta)) \in \mathbb{R}^k$. The MSE between an estimator $t(X)$ and $g(\theta)$ is

$$E||t(X) - g(\theta)||^2 = \int ||t(x) - g(\theta)||^2 F(dx|\theta) \pi(d\theta).$$

Then Bayes estimator $\hat{g}(\theta) = E(g(\theta)|X) := \langle E(g_1(\theta)|X), \cdots, E(g_k(\theta)|X) \rangle$.

**Definition.** The distribution of $\theta$ (we thought there is such) , $\pi(\theta)$, is called a prior distribution of $\theta$. The conditional distribution of $\theta$ given $X$ is called the posterior distribution.

**Example.** Let $X_1, \cdots, X_n$ be i.i.d. Ber($p$) and $p \sim$ Beta($\alpha$, $\beta$). We know that the density function of $p$ is

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

and $E(p) = \alpha/(\alpha + \beta)$. The joint distribution of $X_1, \cdots, X_n$ and $p$ is

$$f(x_1, \cdots, x_n, p) = f(x_1, \cdots, x_n|p) \cdot \pi(p) = p^n(1-p)^{n-x} \cdot C(\alpha, \beta)p^{\alpha-1}(1-p)^{\beta-1} = C(\alpha, \beta)p^{x+\alpha-1}(1-p)^{n+\beta-x-1}$$

where $x = \sum_i x_i$. The marginal density of $(X_1, \cdots, X_n)$ is

$$f(x_1, \cdots, x_n) = C(\alpha, \beta) \int_0^1 p^{x+\alpha-1}(1-p)^{n+\beta-x-1} dp = C(\alpha, \beta) \cdot B(x + \alpha, n + \beta - x).$$

Therefore,

$$f(p|x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n, p)}{f(x_1, \cdots, x_n)} = D(x, \alpha, \beta) p^{x+\alpha-1}(1-p)^{n+\beta-x-1}.$$
This says that \( p(x) \sim \text{Beta}(x + \alpha, n + \beta - x) \). Thus the Bayes estimator is
\[
\hat{p} = E(p|X) = \frac{X + \alpha}{n + \alpha + \beta},
\]
where \( X = \sum_{i=1}^{n} X_i \). One can easily check that
\[
E(\hat{p}) = \frac{np + \alpha}{n + \alpha + \beta}.
\]
So \( \hat{p} \) is biased unless \( \alpha = \beta = 0 \), which is not allowed in the prior distribution.

The next lemma is useful

**Lemma 1.2** The posterior distribution depends only on sufficient statistics. Let \( t(X) \) be a sufficient statistic. Then \( E(g(\theta)|X) = E(g(\theta)|t(X)) \).

**Proof.** The joint density function of \( X \) and \( \theta \) is \( p(x|\theta)\pi(\theta) \) where \( \pi(\theta) \) is the prior distribution of \( \theta \). Let \( t(X) \) be a sufficient statistic. Then by the factorization theorem
\[
p(x|\theta) = q(t(x)|\theta)h(x).
\]
So the joint distribution function is \( q(t(x)|\theta)h(x)\pi(\theta) \). It follows that the conditional distribution of \( \theta \) given \( X \) is
\[
p(\theta|x) = \frac{q(t(x)|\theta)h(x)\pi(\theta)}{\int q(t(x)|\theta)h(x)\pi(\theta) d\xi} = \frac{q(t(x)|\theta)\pi(\theta)}{\int q(t(x)|\xi)\pi(\xi) d\xi}
\]
which is a function of \( t(x) \) and \( \theta \).

By the above conclusion, \( E(g(\theta)|X) = l(t(X)) \) for some function \( l(\cdot) \). Take conditional expectation for both sides with respect to the algebra \( \sigma(t(X)) \). Note that \( \sigma(t(X)) \subset \sigma(X) \). By the Tower theorem, \( l(t(X)) = E(g(\theta)|t(X)) \). This proves the second conclusion. \( \square \)

**Example.** Let \( X_1, \ldots, X_n \) be iid from \( N(\mu, 1) \) and \( \mu \sim \pi(\mu) = N(\mu_0, \tau_0) \). We know that \( \bar{X} \sim N(\mu, 1/n) \) is a sufficient statistic. The Bayes estimator is \( \hat{\mu} = E(\mu|\bar{X}) \). We need to calculate the joint distribution of \( (\mu, \bar{X})^T \) first.

It is not difficult to see that \( (\mu, \bar{X})^T \) is bivariate normal. We know that \( E\mu = \mu_0, \ Var(\mu) = \tau_0, \ E(\bar{X}) = E(E(\bar{X}|\mu)) = E(\mu) = \mu_0 \). Now
\[
Var(\bar{X}) = E(\bar{X} - \mu + \mu - \mu_0)^2 = E((1/n)^2 + (\mu - \mu_0)^2) = \frac{1}{n} + \tau_0;
\]
\[
Cov(\bar{X}, \mu) = E(E((\bar{X} - \mu_0)(\mu - \mu_0)|\mu)) = E(\mu - \mu_0)^2 = \tau_0.
\]
Thus,
\[
\begin{pmatrix} \bar{X} \\ \mu \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_0 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} (1/n) + \tau_0 & \tau_0 \\ \tau_0 & \tau_0 \end{pmatrix} \right).
\]
Recall the following fact: if
\[
\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),
\]
then
\[
Y_1 | Y_2 \sim N \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (Y_2 - \mu_2), \Sigma_{11} \cdot 2 \right)
\]
where \( \Sigma_{11} \cdot 2 = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \).

It follows that
\[
\mu | \bar{X} \sim N \left( \mu_0 + \frac{\tau_0}{1/n + \tau_0} (\bar{X} - \mu_0), \tau_0 \frac{\tau_0^2}{1/n + \tau_0} \right).
\]

Thus the Bayes estimator is
\[
\hat{\mu} = \mu_0 + \frac{\tau_0}{1/n + \tau_0} (\bar{X} - \mu_0) = \frac{\tau_0}{1/n + \tau_0} \bar{X} + \frac{\mu_0}{1 + n\tau_0}.
\]

One can see that \( \hat{\mu} \to \mu \) as \( \tau_0 \to 0 \) and \( \hat{\mu} \to \bar{X} \) as \( \tau_0 \to \infty \).

**Example.** Let \( X \sim N_p(\theta, I_p) \) and \( \pi(\theta) \sim N_p(0, \tau I_p), \tau > 0 \). The Bayes estimator is the posterior mean \( E(\theta|X) \). We claim that
\[
\begin{pmatrix} X \\ \theta \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (1 + \tau)I_p & \tau I_p \\ \tau I_p & \tau I_p \end{pmatrix} \right).
\]

Indeed, by the prior we know that \( E\theta = 0 \) and \( \text{Cov}(\theta) = \tau I_p \). By conditioning argument one can verify that \( E(X\theta') = \tau I_p \) and \( E(XX') = (1 + \tau)I_p \). So by conditional distribution of normal random variables,
\[
\theta | X \sim N \left( \frac{\tau}{1 + \tau} X, \frac{\tau}{1 + \tau} I_p \right).
\]

So the Bayes estimator is \( t_0(X) = \tau X / (1 + \tau) \).

The usual MVUE is \( t_1(X) = X \). It follows that
\[
\text{MSE}_{t_1}(\theta) = E ||X - \theta||^2 = p
\]
which doesn’t depend on \( \theta \). Now
\[
\text{MSE}_{t_0}(\theta) = E_{\theta} ||\frac{\tau}{1 + \tau} X - \theta||^2 = E_{\theta} ||\frac{\tau}{1 + \tau} (X - \theta) + \frac{-1}{1 + \tau} \theta||^2
\]
\[
= \left( \frac{\tau}{1 + \tau} \right)^2 p + \frac{1}{(1 + \tau)^2} ||\theta||^2
\]
(1.2)
which goes to $p$ as $\tau \to \infty$. What happens to the prior when $\tau \to \infty$? It “converges to” the Lebesgue measure, or “uniform distribution over the real line” which generates no information (recall the information of $X \sim N_p(0, (1 + \tau)I_p)$ is $p/(1 + \tau)$ and $\bar{\theta}_i$ uniform over $[-\tau, \tau]$ with $\bar{\theta}_i, 1 \leq i \leq p$ i.i.d. is $p/(2\tau^2)$. Both go to zero and the later is faster than the former). This explains why the MSE of the former and the limiting MSE are identical.

Now let

$$M = \inf_t \sup_{\theta} E_\theta \|t(X) - \theta\|^2$$

called the Minimax risk and any estimator $t^*$ with

$$M = \sup_{\theta} E_\theta \|t^*(X) - \theta\|^2$$

is called a Minimax estimator.

**THEOREM 2** For any $p \geq 1$, $t_0(X) = X$ is a minimax.

**Proof.** First,

$$M = \inf_t \sup_{\theta} E_\theta \|t(X) - \theta\|^2 \leq \sup_{\theta} E_\theta \|X - \theta\|^2 = p.$$

Second

$$M = \inf_t \sup_{\theta} E_\theta \|t(X) - \theta\|^2 \geq \inf_t \int E_\theta \|t(X) - \theta\|^2 \pi_\tau (d\theta) \geq \left( \frac{\tau}{1 + \tau} \right)^2 p$$

for any $\tau > 0$ where $\pi_\tau (\theta) = N(0, \tau I_p)$, where the last step is from (1.2). Then $M \geq p$ by letting $p \to \infty$. The above says that $M = \sup_{\theta} E_\theta \|X - \theta\|^2 = p$. 

**Question.** Does there exist another estimator $t_1(X)$ such that

$$E_\theta \|t_1(X) - \theta\|^2 \leq E_\theta \|X - \theta\|^2$$

for all $\theta$, and the strict inequality holds for some $\theta$? If so, we say $t_0(X) = X$ is inadmissible (because it can be beaten for any $\theta$ by some estimator $t$, and strictly beaten at some $\theta$).

Here is the answer:

(i) For $p = 1$, there is not. This is proved by Blyth in 1951.

(ii) For $p = 2$, there is not. This result was shown by Stein in 1961.

(iii) When $p \geq 3$, Stein (1956) shows there is such estimator, which is called James-Stein estimator.

Recall the density function of $N(\mu, \sigma^2)$ is $\phi(x) = (\sqrt{2\pi}\sigma)^{-1} \exp(-(x - \mu)^2/(2\sigma^2))$. 5
**Lemma 1.3 (Stein’s lemma).** Let \( Y \sim N(\mu, \sigma^2) \) and \( g(y) \) be a function such that \( g(b) - g(a) = \int_a^b g'(y) \, dy \) for any \( a \) and \( b \) and some function \( g'(y) \). If \( E|g'(Y)| < \infty \) then

\[
Eg(Y)(Y - \mu) = \sigma^2 Eg'(Y).
\]

**Proof.** Let \( \phi(y) = (1/\sqrt{2\pi}\sigma) \exp(-(y - \mu)^2/(2\sigma^2)) \), the density of \( N(\mu, \sigma^2) \). Then \( \phi'(y) = -\phi(y)(y - \mu)/\sigma^2 \). Thus, \( \phi(y) = \int_{-\infty}^{y} \frac{z - \mu}{\sigma^2} \phi(z) \, dz = -\int_{y}^{\infty} \frac{z - \mu}{\sigma^2} \phi(z) \, dz \). We then have that

\[
Eg'(Y) = \int_{-\infty}^{\infty} g'(y)\phi(y) \, dy
= \int_{0}^{\infty} g'(y) \left\{ \int_{y}^{\infty} \frac{z - \mu}{\sigma^2} \phi(z) \, dz \right\} \, dy - \int_{-\infty}^{0} g'(y) \left\{ \int_{-\infty}^{y} \frac{z - \mu}{\sigma^2} \phi(z) \, dz \right\} \, dy
= \left( \int_{0}^{\infty} + \int_{-\infty}^{0} \right) \frac{z - \mu}{\sigma^2} \phi(z) (g(z) - g(0)) \, dz
= \frac{1}{\sigma^2} \int_{-\infty}^{\infty} (z - \mu)g(z)\phi(z) \, dz = \frac{1}{\sigma^2} E(Y - \mu)g(Y).
\]

The Fubini’s theorem is used in the third step; the mean of a centered normal random variable is zero is used in the fifth step.

**Remark 1.** Suppose two functions \( g(x) \) and \( h(x) \) are given such that \( g(b) - g(a) = \int_a^b h(x) \, dx \) for any \( a < b \). This does not mean that \( g'(x) = h(x) \) for all \( x \). Actually, in this case, \( g(x) \) is differentiable almost everywhere and \( g'(x) = h(x) \) a.s. under Lebesgue measure. For example, let \( h(x) = 1 \) if \( x \) is an irrational number, and \( h(x) = 0 \) if \( x \) is rational. Then \( g(x) = x = \int_0^x h(t) \, dt \) for any \( x \in \mathbb{R} \). We know in this case that \( g'(x) = h(x) \) a.s. The following are true from real analysis:

*Fundamental Theorem of Calculus.* If \( f(x) \) is absolutely continuous on \([a, b]\), then \( f(x) \) is differentiable almost everywhere and

\[
f(x) - f(a) = \int_a^x f'(t) \, dt, \quad x \in [a, b].
\]

*Another fact.* Let \( f(x) \) be differentiable everywhere on \([a, b]\) and \( f'(x) \) is integrable over \([a, b]\). Then

\[
f(x) - f(a) = \int_a^x f'(t) \, dt \quad x \in [a, b].
\]

**Remark 2.** What drives the Stein’s lemma is the formula of integration by parts:
suppose $g(x)$ is differentiable then

$$Eg(Y)(Y - \mu) = \int_{\mathbb{R}} g(y)(y - \mu)\phi(y) dy = -\sigma^2 \int_{-\infty}^{\infty} g(y)\phi'(y) dy$$

$$= \sigma^2 \lim_{y \to -\infty} g(y)\phi(y) - \sigma^2 \lim_{y \to +\infty} g(y)\phi(y) + \sigma^2 \int_{-\infty}^{\infty} g'(y)\phi(y) dy.$$  

It is reasonable to assume the two limits are zero. The last term is exactly $\sigma^2 Eg'(Y)$.

**Theorem 3** Let $X \sim N_p(\theta, I_p)$ for some $p \geq 3$. Define

$$\delta_c(X) = \left(1 - \frac{p - 2}{\|X\|^2}\right) X.$$  

Then

$$E\|\delta_c(X) - \theta\|^2 = p - (p - 2)^2 E \left[\frac{c(2 - c)}{\|X\|^2}\right].$$

**Proof.** Let $g_i(x) = c(p - 2)x_i/\|x\|^2$ and $g(x) = (g_1(x), \ldots, g_p(x))$ for $x = (x_1, x_2, \ldots, x_p)$. Then $g(x) = c(p - 2)x/\|x\|^2$. It follows that

$$E\|\delta_c(X) - \theta\|^2 = E\|(X - \theta) - g(X)\|^2 = E\|X - \theta\|^2 + E\|g(X)\|^2 - 2E\langle X - \theta, g(X)\rangle$$

$$= p + E\|g(X)\|^2 - 2\sum_{i=1}^{p} E(X_i - \theta_i)g_i(X).$$

Since $X_1, \ldots, X_p$ are independent, by conditioning and Stein’s lemma,

$$E(X_i - \theta_i)g_i(X) = E\left[ E\{(X_i - \theta_i)g_i(X)|(X_j, j \neq i)\}\right] = E\left[\frac{\partial g_i}{\partial x_i}(X)\right].$$

It is easy to calculate that

$$\frac{\partial g_i}{\partial x_i}(x) = c(p - 2)\frac{\sum_{k=1}^{p} x_k^2 - 2x_i^2}{(\sum_{k=1}^{p} x_k^2)^2} = c(p - 2)\frac{\|x\|^2 - 2x_i^2}{\|x\|^4}.$$  

Thus,

$$\sum_{i=1}^{p} E(X_i - \theta_i)g_i(X) = c(p - 2)E\sum_{i=1}^{p} \frac{\|X\|^2 - 2X_i^2}{\|X\|^4} = c(p - 2)^2 E\left[\frac{1}{\|X\|^2}\right].$$

On the other hand, $E\|g(X)\|^2 = c^2(p - 2)^2E(1/\|X\|^2)$. Combine all above together, the conclusion follows.  

We have to show that $E\|X\|^{-2} < \infty$ for $p \geq 3$. Indeed,

$$E\left[\frac{1}{\|X\|^2}\right] \leq 1 + E\left[\frac{1}{\|X\|^2}\right] I(\|X\| \leq 1) \leq 1 + \int \cdots \int_{\|x\| \leq 1} \frac{1}{\|x\|^2} dx_1 \cdots dx_p.$$
since the density of a normal distribution is bounded by one. By the polar-transformation, the last integral is equal to
\[
\int_0^1 dr \int_0^\pi d\theta_1 \cdots \int_0^\pi d\theta_{p-1} \frac{r^{p-1} J(\theta_1, \cdots, \theta_p)}{r^2} dr \leq C \pi^{p-1} \int_0^1 r^{p-3} dr < \infty
\]
if the integer \( p \geq 3 \), where \( J(\theta_1, \cdots, \theta_p) \) is a multivariate polynomial of \( \sin \theta_i \) and \( \cos \theta_i \), \( i = 1, 2, \cdots, p \), bounded by a constant \( C \).

When \( 0 \leq c \leq 2 \), the term \( c(2 - c) \geq 0 \) and attains the maximum at \( c = 1 \). It follows that

**Corollary 1.2** The estimator \( \delta_c(X) \) dominates \( X \) provided \( 0 < c < 2 \) and \( p \geq 3 \).

When \( c = 1 \), the estimator \( \delta_1(X) \) is called the James-Stein estimator.

**Corollary 1.3** The James-Stein estimator \( \delta_1 \) dominates all estimators \( \delta_c \) with \( c \neq 1 \).

## 2 Likelihood Ratio Test

We consider test \( H_0 : \theta \in \Theta_0 \) vs \( H_a : \theta \in \Theta_a \), where \( \Theta_0 \) and \( \Theta_a \) are disjoint. The probability density or probability mass function of a random observation is \( f(x|\theta) \). The **likelihood ratio test statistic** is

\[
\lambda(x) = \frac{\sup_{H_0} f(x|\theta)}{\sup_{H_0 \cup H_a} f(x|\theta)}
\]

When \( H_0 \) is true, the value of \( \lambda(x) \) tends to be large. So the rejection region is

\[ \{\lambda(x) \leq c\} \]

**Example.** The \( t \)-test as a likelihood ratio test.

Let \( X_1, \cdots, X_n \) be i.i.d. \( N(\mu, \sigma^2) \) with both parameters unknown. Test \( H_0 : \mu = \mu_0 \) vs \( H_a : \mu \neq \mu_0 \). The joint density function of \( X_i \)'s is

\[
f(x_1, \cdots, x_n|\mu, \sigma) = \left( \frac{1}{\sqrt{2\pi}} \right)^n (\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right). \tag{2.1}
\]

This one has maximum at \( (\bar{X}, V) \) where

\[
V = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.
\]
The maximum is
\[
\max_{\mu, \sigma} f(x_1, \ldots, x_n|\mu, \sigma) = CV^{-n/2} \exp\left(-\frac{1}{2V} \sum (X_i - \bar{X})^2\right) = CV^{-n/2} e^{-n/2}
\]
Where \(C = (2\pi)^{-n/2}\). When \(\mu = \mu_0\), \(f(x_1, \ldots, x_n|\mu, \sigma)\) has maximum at
\[
V_0 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.
\]
And the corresponding maximum value is
\[
\max_{\sigma} f(x_1, \ldots, x_n|\mu_0, \sigma) = CV_0^{-n/2} \exp\left(-\frac{1}{2V_0} \sum (X_i - \mu_0)^2\right) = CV_0^{-n/2} e^{-n/2}.
\]
So the likelihood ratio is
\[
\Lambda = \left(\frac{V_0}{V}\right)^{-n/2} = \left(\frac{V + (\bar{X} - \mu_0)^2}{V}\right)^{-n/2}.
\]
So the rejection region is
\[
\{\Lambda \leq a\} = \left\{\left|\frac{\bar{X} - \mu_0}{S}\right| \geq b\right\}
\]
for some constant \(b\), where \(S = \left(1/(n - 1)\right) \sum_{i=1}^{n} (X_i - \bar{X})^2\). This yields exactly a student \(t\) test.

**Example.** Let \(X_1, \ldots, X_n\) be a random sample from
\[
f(x|\theta) = \begin{cases} 
  e^{-(x-\theta)}, & \text{if } x \geq \theta; \\
  0, & \text{otherwise.}
\end{cases}
\]

The joint density function is
\[
f(x|\theta) = \begin{cases} 
  e^{-\sum x_i + n\theta}, & \text{if } \theta \leq x_{(1)}; \\
  0, & \text{otherwise.}
\end{cases}
\]
Consider the test \(H_0 : \theta \leq \theta_0\) versus \(H_a : \theta > \theta_0\). It is easy to see
\[
\max_{\theta \in \mathbb{R}} f(x|\theta) = e^{-\sum x_i + nx_{(1)}},
\]
which is achieved at \(\theta = x_{(1)}\). Now consider \(\max_{\theta \leq \theta_0} f(x|\theta)\). If \(\theta_0 \geq x_{(1)}\), the two maxima are identical; If \(\theta_0 < x_{(1)}\), it is achieved at \(\theta = \theta_0\) that yields the maximum value
\[
\max_{\theta \leq \theta_0} f(x|\theta) = \begin{cases} 
  e^{-\sum x_i + nx_{(1)}}, & \text{if } \theta_0 \geq x_{(1)}; \\
  e^{-\sum x_i + n\theta_0}, & \text{if } \theta_0 < x_{(1)}.
\end{cases}
\]
This gives the likelihood ratio

\[
\lambda(x) = \begin{cases} 
1, & \text{if } x_{(1)} \leq \theta_0; \\
e^{n(\theta_0 - x_{(1)})}, & \text{if } x_{(1)} > \theta_0.
\end{cases}
\]

The rejection region is \(\{\lambda(x) \leq a\}\) which is equivalent to \(\{x_{(1)} \geq b\}\) for some \(b\).

Again, the general form of a hypothesis test is \(H_0 : \theta \in \Theta_0\) vs \(H_a : \theta \in \Theta_a\).

**Theorem 4** Suppose \(X\) has pdf or pmf \(f(x|\theta)\) and \(T(X)\) is sufficient for \(\theta\). Let \(\lambda(x)\) and \(\lambda^*(y)\) be the LRT statistics obtained from \(X\) and \(Y = T(X)\). Then \(\lambda(x) = \lambda^*(T(x))\) for any \(x\).

**Proof.** Suppose the pdf or pmf of \(T(X)\) is \(g(t|\theta)\). To be simple, suppose \(X\) is discrete, that is, \(P_\theta(X = x, T(X) = T(x)) = g(T(x)|\theta)\).

\[
f(x|\theta) = P_\theta(X = x, T(X) = T(x)) = P_\theta(T(X) = T(x))P_\theta(X = x|T(X) = T(x)) = g(T(x)|\theta)h(x).
\]

by the definition of sufficiency, where \(h(x)\) is a function depending only on \(x\) (not \(\theta\)).

Evidently,

\[
\lambda(x) = \frac{\sup_{\theta \in H_0} f(x|\theta)}{\sup_{\theta \in H_0 \cup H_a} f(x|\theta)} \quad \text{and} \quad \lambda^*(t) = \frac{\sup_{\theta \in H_a} g(t|\theta)}{\sup_{\theta \in H_0 \cup H_a} g(t|\theta)}.
\]

By (2.3) and (2.4)

\[
\lambda(x) = \frac{\sup_{\theta \in H_0} f(T(x)|\theta)}{\sup_{\theta \in H_0 \cup H_a} f(T(x)|\theta)} = \lambda^*(T(x)).
\]

This theorem says that the simplification of \(\lambda(x)\) is eventually relevant to \(x\) through a sufficient statistic for the parameter.

### 3 Evaluating tests

Consider the test \(H_0 : \theta \in \Theta_0\) versus \(H_a : \theta \in \Theta_a\), where \(\Theta_0\) and \(\Theta_a\) are disjoint. When making decisions, there are two types of mistakes:

**Type I error:** Reject \(H_0\) when it is true.

**Type II error:** Do not reject \(H_0\) when it is false.
<table>
<thead>
<tr>
<th></th>
<th>Not reject $H_0$</th>
<th>reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>correct</td>
<td>type I error</td>
</tr>
<tr>
<td>$H_1$</td>
<td>Type II error</td>
<td>correct</td>
</tr>
</tbody>
</table>

Suppose $R$ denotes the rejection region for the test. The chances of making the two types of mistakes are $P_\theta(X \in R)$ for $\theta \in \Theta_0$ and $P_\theta(X \in R^c)$, $\theta \in \Theta_a$. So

$$P_\theta(X \in R) = \begin{cases} 
\text{probability of a Type I error, if } \theta \in \Theta_0; \\
1 - \text{probability of a Type II error, if } \theta \in \Theta_a.
\end{cases}$$

**Definition 3.1** The power function of the test above with the rejection region $R$ is a function of $\theta$ defined by $\beta(\theta) = P_\theta(X \in R)$.

**Definition 3.2** For $\alpha \in [0,1]$, a test with power function $\beta(\theta)$ is a size $\alpha$ test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

**Definition 3.3** For $\alpha \in [0,1]$, a test with power function $\beta(\theta)$ is a level $\alpha$ test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

There are some differences between the above two definitions when dealing with complicated models.

**Example.** Let $X_1, \cdots, X_n$ be from $N(\mu, \sigma^2)$ with $\mu$ unknown but known $\sigma$. Look at the test $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$. It is easy to check that by likelihood ratio test, the rejection region is

$$\left\{ \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > c \right\}.$$

So the power function is

$$\beta(\mu) = P_\theta \left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > c \right) = P_\theta \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) = 1 - \Phi \left( c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right).$$

It is easy to see that

$$\lim_{\mu \to -\infty} \beta(\mu) = 1 \quad \lim_{\mu \to -\infty} \beta(\mu) = 0 \quad \text{and} \quad \beta(\mu_0) = a \text{ if } P(Z > c) = a,$$

where $Z \sim N(0,1)$.
Example. In general, a size \( \alpha \) LRT is constructed by choosing \( c \) such that
\[
\sup_{\theta \in \Theta_0} P_\theta(\lambda(X) \leq c) = \alpha.
\]
Let \( X_1, \ldots, X_n \) be a random sample from \( N(\theta, 1) \). Test \( H_0 : \theta = \theta_0 \) vs \( H_a : \theta \neq \theta_0 \).

Here one can easily see that
\[
\lambda(x) = \exp \left( -\frac{n}{2}(\bar{x} - \theta_0)^2 \right)
\]
So \( \{\lambda(X) \leq c\} = \{|\bar{X} - \theta_0| \geq d\} \). For \( \alpha \) given, \( d \) can be chosen such that
\[
P_{\theta_0}(|Z| > d\sqrt{n}) = \alpha.
\]
For the example in (2.2), the test is \( H_0 : \theta \leq \theta_0 \) vs \( H_a : \theta > \theta_0 \) and the rejection region
\[
\{\lambda(X) \leq c\} = \{X_{(1)} > d\}.
\]
For size \( \alpha \), first choose \( d \) such that
\[
P_{\theta_0}(X_{(1)} > d) = e^{-n(d-\theta_0)}.
\]
Thus, \( d = \theta_0 - (\log \alpha)/n \). Now
\[
\sup_{\theta \in \Theta_0} P_\theta(X_{(1)} > d) = \sup_{\theta \leq \theta_0} P_\theta(X_{(1)} > d) = \sup_{\theta \leq \theta_0} e^{-n(d-\theta)} = e^{-n(d-\theta_0)} = \alpha.
\]

4 Most Powerful Tests

Consider a hypothesis test \( H_0 : \theta \in \Theta_0 \) vs \( H_1 : \theta \in \Theta_a \). There are a lot of decision rules. We plan to select the best one. Since we are not able to reduce the two types of errors at same time (the sum of chances of two errors and two correct decisions is one). We fix the the chance of the first type of error and maximize the powers among all decision rules.

**Definition 4.1** Let \( \mathcal{C} \) be a class of tests for testing \( H_0 : \theta \in \Theta_0 \) vs \( H_1 : \theta \in \Theta_a \). A test in class \( \mathcal{C} \), with power function \( \beta(\theta) \), is a uniformly most powerful (UMP) class \( \mathcal{C} \) test if
\[
\beta(\theta) \geq \beta'(\theta) \quad \text{for every } \theta \in \Theta_a \text{ and every } \beta'(\theta) \text{ that is a power function of a test in class } \mathcal{C}.
\]

In this section we only consider \( \mathcal{C} \) as the set of tests such that the level \( \alpha \) is fixed, i.e.,
\[
\{\beta(\theta); \sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha\}.
\]

The following gives a way to obtain UMP tests.

**Theorem 5** (Neyman-Pearson Lemma). Let \( X \) be a sample with pdf or pmf \( f(x|\theta) \). Consider testing \( H_0 : \theta = \theta_0 \) vs \( H_1 : \theta = \theta_1 \), using a test with rejection region \( R \) that satisfies
\[
x \in R \quad \text{if } f(x|\theta_1) > kf(x|\theta_0) \quad \text{and} \quad x \in R^c \quad \text{if } f(x|\theta_1) < kf(x|\theta_0) \tag{4.1}
\]
for some \( k \geq 0 \), and
\[
\alpha = P_{\theta_0}(X \in R). \tag{4.2}
\]
Then

(i) (Sufficiency). Any test satisfies (4.1) and (4.2) is a UMP level \( \alpha \) test.

(ii) (Necessity). Suppose there exists a test satisfying (4.1) and (4.2) with \( k > 0 \). Then every UMP level \( \alpha \) test is a size \( \alpha \) test (satisfies (4.2)). Every UMP level \( \alpha \) test satisfies (4.1) except on a set \( A \) satisfying \( P_\theta(X \in A) = 0 \) for \( \theta = \theta_0 \) and \( \theta_1 \).

Proof. We will prove the theorem only for the case that \( f(x|\theta) \) is continuous.

Since \( \mathfrak{B}_0 \) consists of only one point, the level \( \alpha \) test and size \( \alpha \) test are equivalent.

(i) Let \( R \) satisfy (4.1) and (4.2) and \( \phi(x) = I(x \in R) \) with power function \( \beta(\theta) \). Now take another rejection region \( R' \) with \( \beta'(\theta) = P_\theta(X \in R') \) and \( \beta'(\theta_0) \leq \alpha \). Set \( \phi'(x) = I(x \in R') \).

It suffices to show that

\[
\beta(\theta_1) \geq \beta'(\theta_1). \tag{4.3}
\]

Actually, \( (\phi(x) - \phi'(x))(f(x|\theta_1) - k f(x|\theta_0)) \geq 0 \) (check it based on if \( f(x|\theta_1) > k f(x|\theta_0) \) or not). Then

\[
0 \leq \int (\phi(x) - \phi'(x))(f(x|\theta_1) - k f(x|\theta_0)) \, dx = \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)). \tag{4.4}
\]

Thus the assertion (4.3) follows from (4.2) and that \( k \geq 0 \).

(ii) Suppose the test satisfying (4.1) and (4.2) has power function \( \beta(\theta) \). By (i) the test is UMP. For the second UMP level \( \alpha \) test, say, it has power function \( \beta'(\theta) \). Then \( \beta(\theta_1) = \beta'(\theta_1) \).

Since \( k > 0 \), (4.4) implies \( \beta'(\theta_0) \geq \beta(\theta_0) = \alpha \). So \( \beta'(\theta_0) = \alpha \).

Given a UMP level \( \alpha \) test with power function \( \beta'(\theta) \). By the proved (ii), \( \beta(\theta_1) = \beta'(\theta_1) \) and \( \beta(\theta_0) = \beta'(\theta_0) = \alpha \). Then (4.4) implies that the expectation in there is zero. Being always nonnegative, \( (\phi(x) - \phi'(x))(f(x|\theta_1) - k f(x|\theta_0)) = 0 \) except on a set \( A \) of Lebesgue measure zero, which leads to (4.1) (by considering \( f(x|\theta_1) > k f(x|\theta_0) \) or not). Since \( X \) has density, \( P_\theta(X \in A) = 0 \) for \( \theta = \theta_0 \) and \( \theta = \theta_1 \).

Example. Let \( X \sim Bin(2, \theta) \). Consider \( H_0 : \theta = 1/2 \) vs \( H_0 : \theta = 3/4 \). Note that

\[
f(x|\theta_1) = \binom{2}{x} \left( \frac{3}{4} \right)^x \left( \frac{1}{4} \right)^{2-x} > k \binom{2}{x} \left( \frac{1}{2} \right)^x \left( \frac{1}{2} \right)^{2-x} = k f(x|\theta_0)
\]

is equivalent to \( \{3^x > 4k\} \).

(i) The case \( k \geq 9/4 \) and the case \( k < 1/4 \) correspond to \( R = \emptyset \) and \( \Omega \), respectively. The according UMP level \( \alpha = 0 \) and \( \alpha = 1 \).

(ii) If \( 1/4 \leq k < 3/4 \), then \( R = \{1, 2\} \), and \( \alpha = P(Bin(2, 1/2) = 1 \text{ or } 2) = 3/4 \).

(iii) If \( 3/4 \leq k < 9/4 \), then \( R = \{2\} \). The level \( \alpha = P(Bin(2, 1/2) = 2) = 1/4 \).
This example says that if we firmly want a size $\alpha$ test, then there are only two such $\alpha$’s: $\alpha = 1/4$ and $\alpha = 3/4$.

**Example.** Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$ with $\sigma$ known. Look at the test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$. The density function $f(x|\theta, \sigma) = (1/\sqrt{2\pi} \sigma) \exp(-(x - \theta)^2/(2\sigma^2))$. Now $f(x|\theta_1) > kf(\theta_0)$ is equivalent to that $R = \{\bar{X} < c\}$. Set $\alpha = P_{\theta_0}(\bar{X} < c)$. One can check easily that $c = \theta_0 + (\sigma z_{1/4}/\sqrt{n})$. So the UMP rejection region is

$$R = \{\bar{X} < \theta_0 + (\sigma z_{1/4}/\sqrt{n})\} \quad \blacksquare$$

**Lemma 4.1** Let $X$ be a random variable. Let also $f(x)$ and $g(x)$ be non-decreasing real functions. Then

$$E(f(X)g(X)) \geq Ef(X) \cdot Eg(X)$$

provided all the above three expectations are finite.

**Proof.** Let the $\mu$ be the distribution of $X$. Noting that $(f(x) - f(y))(g(x) - g(y)) \geq 0$ for any $x$ and $y$. Then

$$0 \leq \int \int (f(x) - f(y))(g(x) - g(y)) \mu(dx) \mu(dy)$$

$$= 2 \int f(x)g(x) \mu(dx) - 2 \int f(x)g(y) \mu(dx) \mu(dy)$$

$$= 2[E(f(X)g(X)) - Ef(X) \cdot Eg(X)].$$

The desired result follows. \hfill \blacksquare

Let $f(x|\theta)$, $\theta \in \mathbb{R}$ be a pmf or pdf with common support, that is, the set $\{x : f(x|\theta) > 0\}$ is identical for every $\theta$. This family of distributions is said to have *monotone likelihood ratio* (MLR) with respect to a statistic $T(x)$ if

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$$

on the support is a non-decreasing function in $T(x)$ for any $\theta_2 > \theta_1$. \hfill (4.5)

**Lemma 4.2** Let $X$ be a random vector with $X \sim f(x|\theta)$, $\theta \in \mathbb{R}$, where $f(x|\theta)$ is a pmf or pdf with a common support. Suppose $f(x|\theta)$ has MLR with a statistic $T(x)$. Let

$$\phi(x) = \begin{cases} 1 & \text{when } T(x) > C, \\ \gamma & \text{when } T(x) = C, \\ 0 & \text{when } T(x) < C \end{cases}$$

\hfill (4.6)
for some $C \in \mathbb{R}$, where $\gamma \in [0,1]$. Then

$$E_{\theta_1} \phi(X) \leq E_{\theta_2} \phi(X)$$

for any $\theta_1 < \theta_2$.

**Proof.** Thanks to MLT, for any $\theta_1 < \theta_2$ there exists a nondecreasing function $g(t)$ such that $f(x|\theta_2)/f(x|\theta_1) = g(T(x))$, where $g(t)$ is a nondecreasing function. Thus

$$E_{\theta_2} \phi(X) = \int \phi(x) \frac{f(x|\theta_2)}{f(x|\theta_1)} f(x|\theta_1) \, dx = E_{\theta_1} (g(T(X)) \cdot h(T(X))),$$

where

$$h(x) = \begin{cases} 1 & \text{when } x > C, \\ \gamma & \text{when } x = C, \\ 0 & \text{when } x < C \end{cases}$$

Since both $g(x)$ and $h(x)$ are non-decreasing, by the positive correlation inequality,

$$E_{\theta_1} (g(T(X)) \cdot h(T(X))) \geq E_{\theta_1} g(T(X)) \cdot E_{\theta_1} h(T(X)) = E_{\theta_1} \phi(X)$$

since $E_{\theta_1} g(T(X)) = 1$. \hfill \Box

In (4.6), if $T(X) > C$ reject $H_0$; if $T(X) < C$, don’t reject $H_0$; if $T(X) = C$, with chance $\gamma$ reject $H_0$ and with chance $1 - \gamma$ don’t reject $H_0$. The last case is equivalent to flipping a coin with chance $\gamma$ for head and $1 - \gamma$ for tail. If the head occurs, reject $H_0$; otherwise, don’t reject $H_0$. This is a randomization. The purpose is to make the test UMP. The randomization occurs only when $T(X)$ is discrete.

**Theorem 6** Consider testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. Let $X \sim f(x|\theta)$, $\theta \in \mathbb{R}$, where $f(x|\theta)$ is a pmf or pdf with a common support. Also let $T$ be a sufficient statistic for $\theta$. If $f(x|\theta)$ has MLR w.r.t. $T(x)$, then $\phi(X)$ in (4.6) is a UMP level $\alpha$ test, where $\alpha = E_{\theta_0} \phi(X)$.

**Proof.** Let $\beta(\theta) = E_{\theta} \phi(X)$. Then by the previous lemma, $\sup_{\theta \leq \theta_0} \beta(\theta) = \sup_{\theta \geq \theta_0} \beta(\theta) = E_{\theta_0} \phi(X) = \alpha$.

We need to show $\beta(\theta_1) \geq \beta'(\theta_1)$ for any $\theta_1 > \theta_0$ and $\beta'$ such that $\sup_{\theta \leq \theta_0} \beta'(\theta) \leq \alpha$. Consider test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. We know that

$$\alpha = P_{\theta_0}(X \in R) = E_{\theta_0} \phi(X) \text{ and } \beta'(\theta_1) \leq \alpha. \quad (4.7)$$

Again, we use the same notation as in Lemma 4.2, $f(x|\theta_1)/f(x|\theta_0) = g(T(x))$, where $g(t)$ is a non-decreasing function in $t$. Take $k = g(C)$ in the Neyman-Pearson Lemma 5.
If $f(x|\theta_1)/f(x|\theta_0) > k$, since $g(t)$ is non-decreasing, then $T(X) > C$, which is the same as saying to reject $H_0$ by the functional form of $\phi(x)$ as in (4.6). If $f(x|\theta_1)/f(x|\theta_0) < k$ then $T(X) < C$, that is, not to reject $H_0$. By Neyman-Pearson lemma and (4.7), $\phi(x)$ corresponds to a UMP level-$\alpha$ test. Thus, $\beta(\theta_1) \geq \beta'(\theta_1)$. ■
5 Probability convergence

Let \((\Omega, \mathcal{F}, P)\) be a probability space. Let Also \(\{X, X_n; n \geq 1\}\) be a sequence of random variables defined on this probability space. We say \(X_n \to X\) in probability or \(X_n\) is consistent with \(X\) if

\[
P(|X_n - X| \geq \epsilon) \to 0 \text{ as } n \to \infty
\]

for any \(\epsilon > 0\).

**Example.** Let \(\{X_n; n \geq 1\}\) be a sequence of independent random variables with mean zero and variance one. Set \(S_n = \sum_{i=1}^{n} X_i, n \geq 1\). Then \(S_n/n \to 0\) in probability as \(n \to \infty\).

In fact,

\[
P\left(\left| \frac{S_n}{n} \right| \geq \epsilon \right) \leq \frac{E(S_n)^2}{n\epsilon^2} = \frac{1}{n\epsilon^2} \to 0
\]

by Chebyshev’s inequality.

Recall \(X\) and \(X_n, n \geq 1\) are real measurable functions defined on \((\Omega, \mathcal{F})\). We say \(X_n\) converges to \(X\) almost surely if

\[
P(\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)) = 1.
\]

For a sequence of events \(E_n; n \geq 1\) in \(\mathcal{F}\). Define

\[
\{E_n; \text{i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} E_k.
\]

Here “i.o.” means “infinitely often”. If \(\omega \in \{E_n; \text{i.o.}\}\), then there are infinitely many \(n\) such that \(\omega \in E_n\). Conversely, if \(\omega \notin \{E_n; \text{i.o.}\}\), or \(\omega \in \{E_n; \text{i.o.}\}^c\), then only finitely many \(E_n\) contain \(\omega\).

In proving almost convergence, the following so-called Borel-Cantelli lemma is very powerful.

**Lemma 5.1 (Borel-Cantelli Lemma).** Let \(E_n, n \geq 1\) be a sequence of events in \(\mathcal{F}\). If

\[
\sum_{n=1}^{\infty} P(E_n) < \infty
\]

then \(P(\text{\(E_n\) occurs only finite times}) = P(\{E_n, \text{i.o.}\}^c) = 1\).

**Proof.** Recall \(P(E_n) = E(I_{E_n})\). Let \(X(\omega) = \sum_{n=1}^{\infty} I_{E_n}(\omega)\). Then

\[
EX = E \sum_{i=1}^{\infty} I_{E_n} = \sum_{i=1}^{\infty} P(E_n) < \infty.
\]
Thus, $X < \infty$ a.s. ■

**Lemma 5.2 (Second Borel-Cantelli Lemma).** Let $E_n; n \geq 1$ be a sequence of independent events in $\mathcal{F}$. If

$$\sum_{n=1}^{\infty} P(E_n) = \infty$$

then $P(E_n, \ i.o.) = 1$.

**Proof.** First,

$$\{E_n, \ i.o.\}^c = \bigcup_n \bigcap_{k \geq n} E_k^c.$$

With $p_k$ denoting $P(E_k)$, we have that

$$P\left(\bigcap_{k \geq n} E_k^c\right) = \prod_{k \geq n} (1 - p_k) \leq \exp\left(-\sum_{k \geq n} p_k\right) = 0$$
as $n \to \infty$, where we used the inequality that $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$. Thus,

$$P(\{E_n, \ i.o.\}^c) \leq \sum_{n \geq 1} P\left(\bigcap_{k \geq n} E_k^c\right) = 0. \quad \blacksquare$$

As an application, we have that

**Example.** Let $\{X_n; n \geq 1\}$ be a sequence of i.i.d. random variables with mean $\mu$ and finite fourth moment. Then

$$\frac{S_n}{n} \to \mu \text{ a.s.}$$
as $n \to \infty$.

**Proof.** W.L.O.G., assume $\mu = 0$. Let $\epsilon > 0$. Then by Chebyshev’s inequality

$$P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) \leq \frac{E(S_n)^4}{n^4\epsilon^4}.$$ 

Now, by independence,

$$E(S_n^4) = E\left(\sum_k X_k^4 + \sum_{1 \leq i \neq j \leq n} X_i^2 X_j^2 + \sum_{1 \leq i \neq j = n} X_i X_j^3\right) = nE(X_1^4) + 6n(n-1)E(X_1^2).$$
Thus, $P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) = O(n^{-2})$. Consequently, $\sum_{n \geq 1} P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) < \infty$ for any $\epsilon > 0$. This means that, with probability one, $|S_n/n| \leq \epsilon$ for sufficiently large $n$. Therefore,

$$\limsup_n \frac{|S_n|}{n} \leq \epsilon \text{ a.s.}$$

for any $\epsilon > 0$. Let $\epsilon \downarrow 0$. The desired conclusion follows. ■

By using a truncation method, we have the following Kolmogorov’s Strong Law of Large Numbers.

**Theorem 7** Let $\{X_n; n \geq 1\}$ be a sequence of i.i.d. random variables with mean $\mu$. Then

$$\frac{S_n}{n} \to \mu \text{ a.s.}$$

as $n \to \infty$.

There are many variations of the above strong law of large numbers. Among them, Etemadi showed that Theorem 7 holds if $\{X_n; n \geq 1\}$ is a sequence of pairwise independent and identically distributed random variables with mean $\mu$.

The next proposition tells the relationship between convergence in probability and convergence almost surely.

**Proposition 5.1** Let $X_n$ and $X$ be random variables taking values in $\mathbb{R}^k$. If $X_n \to X$ a.s. as $n \to \infty$, then $X_n$ converges to $X$ in probability; second, $X_n$ converges to $X$ in probability iff for any subsequence of $\{X_n\}$ there exists a further subsequence, say, $\{X_{n_k}; k \geq 1\}$ such that $X_{n_k} \to X$ a.s. as $k \to \infty$.

**Proof.** The joint almost sure convergence and joint convergence in probability are equivalent to the corresponding pointwise convergences. So W.L.O.G., assume $X_n$ and $X$ are one-dimensional. Given $\epsilon > 0$. The almost sure convergence implies that

$$P(|X_n - X| \geq \epsilon \text{ occurs only for finitely many } n) = 1.$$ 

This says that $P(|X_n - X| \geq \epsilon, i.o.) = 0$. But

$$\lim_n P(|X_n - X| \geq \epsilon) \leq \lim_n P \left( \bigcup_{k \geq n} \{ |X_k - X| \geq \epsilon \} \right) = P(|X_n - X| \geq \epsilon, i.o.) = 0$$

19
where we use the fact that
\[ \bigcup_{k \geq n} \{ |X_k - X| \geq \epsilon \} \uparrow \{ |X_n - X| \geq \epsilon, \text{i.o.} \}. \]

Now suppose \( X_n \to X \) in probability. W.L.O.G., we assume the subsequence is \( X_n \) itself. Then for any \( \epsilon > 0 \), there exists \( n_\epsilon > 0 \) such that \( \sup_{n \geq n_\epsilon} P(|X_n - X| \geq \epsilon) < \epsilon \). We then have a subsequence \( n_k \) such that \( P(|X_{n_k} - X| \geq 2^{-k}) < 2^{-k} \) for each \( k \geq 1 \). By the Borel-Cantelli lemma
\[ P\left( |X_{n_k} - X| < 2^{-k} \text{ for sufficiently large } k \right) = 1 \]
which implies that \( X_{n_k} \to X \) a.s. as \( k \to \infty \).

Now suppose \( X_n \) doesn’t converge to \( X \) in probability. That is, there exists \( \epsilon_0 > 0 \) and subsequence \( \{X_k; \ l \geq 1\} \) such that
\[ P(|X_{k_l} - X| \geq \epsilon_0) \geq \epsilon_0 \]
for all \( l \geq 1 \). Therefore any subsequence of \( X_{k_l} \) can not converge to \( X \) in probability, hence it can not converge to \( X \) almost surely. ■

**Example** Let \( X_i, \ i \geq 1 \) be a sequence of i.i.d. random variables with mean zero and variance one. Let \( S_n = X_1 + \cdots + X_n, \ n \geq 1 \). Then \( S_n/\sqrt{2n \log \log n} \to 0 \) in probability but it does not converge to zero almost surely. Actually \( \limsup_{n \to \infty} S_n/\sqrt{2 \log \log n} = 1 \) a.s.

**Proposition 5.2** Let \( \{X_k; k \geq 1\} \) be a sequence of i.i.d. random variables with \( X_1 \sim N(0, 1) \). Let \( W_n = \max_{1 \leq i \leq n} X_i \). Then
\[ \lim_{n \to \infty} \frac{W_n}{\sqrt{\log n}} = \sqrt{2} \text{ a.s.} \]

**Proof.** Recall
\[ \frac{x}{\sqrt{2\pi (1 + x^2)}} e^{-x^2/2} \leq P(X_1 \geq x) \leq \frac{1}{\sqrt{2\pi x}} e^{-x^2/2} \tag{5.3} \]
for any \( x > 0 \). Given \( \epsilon > 0 \), let \( b_n = (\sqrt{2} + \epsilon) \sqrt{\log n} \). Then
\[ P(W_n \geq b_n) \leq n P(X_1 \geq b_n) \leq \frac{n}{b_n} e^{-b_n^2/2} \leq \frac{1}{n \epsilon^2/2} \]
for \( n \geq 3 \). Let \( n_k = \lfloor k^{1/\epsilon^2} \rfloor + 1 \) for \( k \geq 1 \). Then,
\[ P(W_{n_k} \geq b_{n_k}) \leq \frac{1}{k^2} \]
for large $k$. By the Borel-Cantelli lemma, $P(W_{n_k} \geq b_{n_k}, \text{i.o.}) = 0$. This implies that

$$\limsup_k \frac{W_{n_k}}{b_{n_k}} \leq \sqrt{2} + \epsilon \ a.s.$$  

For any $n \geq n_1$ there exists $k$ such that $n_k \leq n < n_{k+1}$. Note that $n_{k+1}/n_k \to 1$ as $n \to \infty$ and $W_{n_k} \leq W_n \leq W_{n_{k+1}}$ we have that

$$\limsup_n \frac{W_n}{\sqrt{\log n}} \leq \sqrt{2} + \epsilon \ a.s.$$  

Let $\epsilon \downarrow 0$. We obtain that

$$\limsup_n \frac{W_n}{\sqrt{\log n}} \leq \sqrt{2} \ a.s. \quad (5.4)$$  

This proves the upper bound. Now we turn to prove the lower bound. Choose $\epsilon \in (0, \sqrt{2})$. Set $c_n = (\sqrt{2} - \epsilon)\sqrt{\log n}$. Then

$$P(W_n \leq c_n) = P(X_1 \leq c_n)^n = (1 - P(X_1 > c_n))^n \leq e^{-nP(X_1 > c_n)}.$$  

By (5.3),

$$nP(X_1 > c_n) \geq \frac{nc_n}{\sqrt{2\pi} (1 + c_n^2)} e^{-c_n^2/2} \sim \frac{C}{\sqrt{2\pi \log n}} n^{1-(\sqrt{2}-\epsilon)^2/2}$$  

as $n$ is sufficiently large, where $C$ is a constant. Since $1 - (\sqrt{2} - \epsilon)^2/2 > 0$,

$$\sum_{n \geq 2} P(W_n \leq c_n) < \infty.$$  

By the Borel-Cantelli again, $P(W_n \leq c_n, \text{i.o.}) = 0$. This implies that

$$\liminf_n \frac{W_n}{\sqrt{\log n}} \geq \sqrt{2} - \epsilon \ a.s.$$  

Let $\epsilon \downarrow 0$. We obtain

$$\liminf_n \frac{W_n}{\sqrt{\log n}} \geq \sqrt{2} \ a.s.$$  

This together with (5.4) implies the desired result.  

The above results also hold for random variables taking abstract values. Now let’s introduce such random variables.
A metric space is a set $D$ equipped with a metric. A metric is a map $d : D \times D \to [0, \infty)$ with the properties

(i) $d(x, y) = 0$ if and only if $x = y$;
(ii) $d(x, y) = d(y, x)$;
(iii) $d(x, z) \leq d(x, y) + d(y, z)$

for any $x, y$ and $z$. A semi-metric satisfies (ii) and (iii), but not necessarily (i). An open ball is a set of the form $\{ y : d(x, y) < r \}$. A subset of a metric space is open if and only if it is the union of open balls; it is closed if and only if the complement is open. A sequence $x_n$ converges to $x$ if and only if $d(x_n, x) \to 0$, and denoted by $x_n \to x$. The closure $\bar{A}$ of a set $A$ is the set of all points that are the limit of a sequence in $A$; it is the smallest closed set containing $A$. The interior $\mathring{A}$ is the collection of all points $x$ such that $x \in G \subset A$ for some open set $G$; it is the largest open set contained in $A$. A function $f : D \to E$ between metric spaces is continuous at a point $x$ if and only if $f(x_n) \to f(x)$ for every sequence $x_n \to x$; it is continuous at every $x$ if and only if $f^{-1}(G)$ is also open for every open set $G \subset E$. A subset of a metric space is dense if and only if it has a countable dense subset. A subset $K$ of a metric space is compact if and only if it is closed and every sequence in $K$ has a convergent subsequence. A set $K$ is totally bounded if and only if for every $\epsilon > 0$ it can be covered by finitely many balls of radius $\epsilon > 0$. The space is complete if and only if every Cauchy sequence has a limit. A subset of a complete metric space is compact if and only if it is totally bounded and closed.

A norm space $D$ is a vector space equipped with a norm. A norm is a map $\| \cdot \| : D \to [0, \infty)$ such that for every $x, y$ in $D$, and $\alpha \in \mathbb{R}$,

(i) $\|x + y\| \leq \|x\| + \|y\|$;
(ii) $\|\alpha x\| = |\alpha|\|x\|$;
(iii) $\|x\| = 0$ if and only if $x = 0$.

Definition. The Borel $\sigma$-field on a metric space $D$ is the smallest $\sigma$-field that contains the open sets. A function taking values in a metric space is called Borel-measurable if it is measurable relative to the Borel $\sigma$-field. A Borel-measurable map $X : \Omega \to D$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is referred to as a random element with values in $D$.

The following lemma is obvious.

**Lemma 5.3** A continuous map between two metric spaces is Borel-measurable.
**Example.** Let \( C[a, b] = \{ f(x); \ f \) is a continuous function on \([a, b]\}\). Define 
\[
\| f \| = \sup_{a \leq x \leq b} |f(x)|.
\]
This is a complete normed linear space which is also called a Banach space. This space is separable. Any separable Banach space is isomorphic to a subset of \( C[a, b] \).

**Remark.** Proposition 5.1 still holds when the random variables taking values in a Polish space.

Let \( \{ X, X_n; n \geq 1 \} \) be a sequence of random variables. We say \( X_n \) converges to \( X \) weakly or in distribution if
\[
P(X_n \leq x) \to P(X \leq x)
\]
as \( n \to \infty \) for every continuous point \( x \), that is, \( P(X \leq x) = P(X < x) \).

**Remark.** The set of discontinuous points of any random variable is countable.

**Lemma 5.4** Let \( X_n \) and \( X \) be random variables. The following are equivalent:

(i) \( X_n \) converges weakly to \( X \);

(ii) \( Ef(X_n) \to Ef(X) \) for any bounded continuous function \( f(x) \) defined on \( \mathbb{R} \);

(iii) \( \liminf_n P(X_n \in G) \geq P(X \in G) \) for any open set \( G \);

(iv) \( \limsup_n P(X_n \in F) \leq P(X \in F) \) for any closed set \( F \);

**Remark.** The above lemma is actually true for any finite dimensional space. Further, it is true for random variables taking values in separable, complete metric spaces, which are also called Polish spaces.

**Proof.** (i) \( \implies \) (ii). Given \( \epsilon > 0 \), choose \( K > 0 \) such that \( K \) and \( -K \) are continuous point of all \( X_n \)'s and \( X \) and \( P(-K \leq X \leq K) \geq 1 - \epsilon \). By (i)
\[
\lim_n P(|X_n| \leq K) = \lim_n P(X_n \leq K) - \lim_n P(X_n \leq -K) = P(X \leq K) - P(X \leq -K) \geq 1 - \epsilon
\]
In other words
\[
\lim_n P(|X_n| > K) \leq \epsilon.
\]
(5.5)

Since \( f(x) \) is continuous on \([-K, K]\), we can cut the interval into, say, \( m \) subintervals \([a_i, a_{i+1}]\), \( 1 \leq i \leq m \) such that each \( a_i \) is a continuous point of \( X \) and \( \sup_{a_i \leq x \leq a_{i+1}} |f(x) - f(a_i)| \leq \epsilon \). Set
\[
g(x) = \sum_{i=1}^{m} f(a_i) I(a_i \leq x < a_{i+1}).
\]
It is easy to see that
\[ |f(x) - g(x)| \leq \epsilon \text{ for } |x| \leq K \quad \text{and} \quad |f(x) - g(x)| \leq \|f\| \text{ for } |x| > K, \]
where \( \|f\| = \sup_{x \in \mathbb{R}} |f(x)| \). It follows that
\[ E|f(Y) - Eg(Y)| \leq \epsilon + \|f\| \cdot P(|Y| > K) \]
for any random variable \( Y \). Apply this to \( X_n \) and \( X \) we obtain from (5.5) that
\[ \limsup_n |Ef(X_n) - Eg(X_n)| \leq 2(1 + \|f\|)\epsilon. \]
Easily \( Eg(X_n) \to Eg(X) \). Therefore by triangle inequality,
\[ \limsup_n |Ef(X_n) - Ef(X)| \leq 2(1 + \|f\|)\epsilon. \]
Then (ii) follows by letting \( \epsilon \downarrow 0 \).

(ii) \( \implies \) (iii). Define \( f_m(x) = (md(x, G^c)) \land 1 \) for \( m \geq 1 \). Since \( G \) is open, \( f_m(x) \) is a continuous function bounded by one and \( f_m(x) \uparrow 1_G(x) \) for each \( x \) as \( m \to \infty \). It follows that
\[ \liminf_n P(X_n \in G) \geq \liminf_n Ef_m(X_n) \to Ef_m(X) \]
for each \( m \). (iii) follows by letting \( m \to \infty \).

(iii) and (iv) are equivalent because \( F^c \) is open.

(iv) \( \implies \) (i). Let \( x \) be a continuous point. Then by (iii) and (iv)
\[ \limsup_n P(X_n \leq x) \leq P(X \leq x) = P(X < x) \quad \text{and} \]
\[ \liminf_n P(X_n \leq x) \geq \liminf_n P(X_n < x) \geq P(X < x). \]
Then (i) follows. \( \blacksquare \)

**Example.** Let \( X_n \) be uniformly distributed over \( \{1/n, 2/n, \ldots, n/n\} \). Prove that \( X_n \) converges weakly to \( U[0, 1] \).

Actually, for any bounded continuous function \( f(x) \),
\[ Ef(X_n) = \frac{1}{n} \sum_{i=1}^{n} f \left( \frac{i}{n} \right) \to \int_0^1 f(t) \, dt = Ef(Y) \]
where \( Y \sim U[0, 1] \).
Example. Let $X_i$, $i \geq 1$ be i.i.d. random variables. Define

$$
\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \quad \text{through} \quad \mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A) = \frac{\# \{i : X_i \in A \}}{n}
$$

for any set $A$. Then, for any measurable function $f(x)$, by the strong law of large numbers,

$$
\int f(x) \mu_n(dx) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \to Ef(X_1) = \int f(x) \mu(dx)
$$

where $\mu = \mathcal{L}(X_1)$. This concludes that $\mu_n \to \mu$ weakly as $n \to \infty$.

Corollary 5.1 Suppose $X_n$ converges to $X$ weakly and $g(x)$ is a continuous map defined on $\mathbb{R}^1$ (not necessarily one-dimensional). Then $g(X_n)$ converges to $g(X)$ weakly.

The following two results are called Delta methods.

**Theorem 8** Let $Y_n$ be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \Rightarrow N(0, \sigma^2)$. Let $g$ be a real-valued function such that $g'(\theta) \neq 0$. Then

$$
\sqrt{n}(g(Y_n) - g(\theta)) \Rightarrow N(0, \sigma^2(g'(\theta))^2).
$$

**Proof.** By Taylor’s expansion $g(x) = g(\theta) + g'(\theta)(x - \theta) + a(x - \theta)$ for some real function $a(x)$ such that $a(t)/t \to 0$ as $t \to 0$. Then,

$$
\sqrt{n}(g(Y_n) - g(\theta)) = g'(\theta)\sqrt{n}(Y_n - \theta) + \sqrt{n} \cdot a(Y_n - \theta).
$$

We only need to show that $\sqrt{n} \cdot a(Y_n - \theta) \to 0$ in probability as $n \to \infty$ (latter we will see this easily by Slusky’s lemma. But why now?) Indeed, for any $m \geq 1$ there exists a $\delta > 0$ such that $|a(x)| \leq |x|/m$ for all $|x| \leq \delta$. Thus,

$$
P(|\sqrt{n} \cdot a(Y_n - \theta)| \geq b) \leq P(|Y_n - \theta| > \delta) + P(|\sqrt{n}(Y_n - \theta)| \geq mb) \to P(|N(0, \sigma^2)| \geq mb)
$$

as $n \to \infty$. Now let $m \uparrow \infty$, we obtain that $\sqrt{n} \cdot a(Y_n - \theta) \to 0$ in probability. \hfill \blacksquare

Similarly, one can prove that

**Theorem 9** Let $Y_n$ be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \Rightarrow N(0, \sigma^2)$. Let $g$ be a real-valued function such that $g''(x)$ exists around $x = \theta$ with $g'(\theta) = 0$ and $g''(\theta) \neq 0$. Then

$$
n(g(Y_n) - g(\theta)) \Rightarrow \frac{1}{2} \sigma^2 g''(\theta) \chi_1^2.
$$
Proof. By Taylor’s expansion again,
\[ n(g(Y_n) - g(\theta)) = \frac{1}{2}\sigma^2 g''(\theta) \left[ n \cdot \left( \frac{Y_n - \theta}{\sigma} \right)^2 \right] + o(n|Y_n - \theta|^2). \]

This leads to the desired conclusion by the same arguments as in Theorem 8. ■

**Proposition 5.3** Let \( \{X, X_n; n \geq 1\} \) be a sequence of random variables on \( \mathbb{R}^k \). Look at the following three statements:

(i) \( X_n \) converges to \( X \) almost surely;
(ii) \( X_n \) converges to \( X \) in probability;
(iii) \( X_n \) converges to \( X \) weakly.

Then (i) \( \implies \) (ii) \( \implies \) (iii).

Proof. The implication “(i) \( \implies \) (ii)” is proved earlier. Now we prove that (ii) \( \implies \) (iii).

For any closed set \( F \) let
\[ F^\epsilon = \{ x : \inf_{y \in F} d(x, y) \leq \epsilon \}. \]

Then \( F^\epsilon \) is a closed set. When \( X \) takes values in \( \mathbb{R} \), the metric \( d(x, y) = |x - y| \). It is easy to see that
\[ \{X_n \in F\} \subset \{X \in F^{1/k}\} \cup \{|X_n - X| \geq 1/k\} \]
for any integer \( k \geq 1 \). Therefore
\[ P(X_n \in F) \leq P(X \in F^{1/k}) + P(|X_n - X| \geq 1/k). \]

Let \( n \to \infty \) we have that
\[ \limsup_n P(X_n \in F) \leq P(X \in F^{1/k}) \]

Note that \( F^{1/k} \downarrow F \) as \( k \to \infty \). (iii) follows by letting \( k \to \infty \). ■

Any random vector \( X \) taking values in \( \mathbb{R}^k \) is tight: For every \( \epsilon > 0 \) there exists a number \( M > 0 \) such that \( P(\|X\| > M) < \epsilon \). A set of random vectors \( \{X_\alpha; \alpha \in A\} \) is called uniformly tight if there exists a constant \( M > 0 \) such that
\[ \sup_{\alpha \in A} P(\|X_\alpha\| > M) < \epsilon. \]
A random vector $X$ taking values in a polish space $(\mathbb{D}, d)$ is also tight: For every $\epsilon > 0$ there exists a compact set $K \subset \mathbb{D}$ such that $P(X \notin K) < \epsilon$. This will be proved later.

**Remark.** It is easy to see that a finite number of random variables is uniformly tight if every individual is tight. Also, if $\{X_\alpha; \alpha \in A\}$ and $\{X_\alpha; \alpha \in B\}$ are uniformly tight, respectively, then $\{X_\alpha, \alpha \in A \cup B\}$ is also tight. Further, if $X_n \Longrightarrow X$, then $\{X, X_n; n \geq 1\}$ is uniformly tight. Actually, for any $\epsilon > 0$, choose $M > 0$ such that $P(||X|| \geq M) \leq 1 - \epsilon$. Since the set $\{x \in \mathbb{R}^k, ||x|| > M\}$ is an open set, $\liminf_{n \to \infty} P(||X_n|| > M) \geq P(||X|| > M) \geq 1 - \epsilon$. The tightness of $\{X, X_n\}$ follows.

**Lemma 5.5** Let $\mu$ be a probability measure on $(\mathbb{D}, \mathcal{B})$, where $(\mathbb{D}, d)$ is a Polish space and $\mathcal{B}$ is the Borel $\sigma$-algebra. Then $\mu$ is tight.

**Proof.** Since $(\mathbb{D}, d)$ is Polish, it is separable, there exists a countable set $\{x_1, x_2, \cdots\} \subset \mathbb{D}$ such that it is dense in $\mathbb{D}$.

Given $\epsilon > 0$. For any integer $k \geq 1$, since $\bigcup_{n \geq 1} B(x_n, 1/k) = \mathbb{D}$, there exists $n_k$ such that $\mu(\bigcup_{n=1}^{n_k} B(x_n, 1/k)) > 1 - \epsilon/2^k$. Set $\mathbb{D}_k = \bigcup_{n=1}^{n_k} B(x_n, 1/k)$ and $M = \bigcap_{k \geq 1} \mathbb{D}_k$. Then $M$ is totally bounded ($M$ is always covered by a finite number of balls with any prescribed radius). Also

$$
\mu(M^c) \leq \sum_{k=1}^{\infty} \mu(\mathbb{D}_k^c) \leq \sum_{k=1}^{\infty} \frac{\epsilon}{2^k} = \epsilon.
$$

Let $K$ be the closure of $M$. Since $\mathbb{D}$ is complete, $M \subset K \subset \mathbb{D}$ and $K$ is compact. Also by the above fact, $\mu(K^c) < \epsilon$. ■

**Lemma 5.6** Let $\{X_n; n \geq 1\}$ be a sequence of random variables. Let $F_n$ be the cdf of $X_n$. Then there exists a subsequence $F_{n_j}$ with the property that $F_{n_j}(x) \to F(x)$ at each continuity point of $x$ of a possibly defective distribution function $F$. Further, if $\{X_n\}$ is tight, then $F(x)$ is a cumulative distribution function.

**Proof.** Put all rational numbers $q_1, q_2, \cdots$ in a sequence. Because the sequence $F_n(q_1)$ is contained in the interval $[0, 1]$, it has a converging subsequence. Call the indexing subsequence $\{n_j^1\}_{j=1}^{\infty}$ and the limit $G(q_1)$. Next, extract a further subsequence $\{n_j^2\} \subset \{n_j^1\}$ along which $F_n(q_2)$ converges to a limit $G(q_2)$, a further subsequence $\{n_j^3\} \subset \{n_j^2\}$ along which $F_n(q_3)$ converges to a limit $G(q_3)$, and so forth. The tail of the diagonal sequence $n_j := n_j^1 \in \{n_j^i\}$ belongs to every sequence $n_j^i$. Hence $F_{n_j}(q_i) \to G(q_i)$ for every $i = 1, 2, \cdots$. Since each $F_n$ is nondecreasing, $G(q) \leq G(q')$ if $q \leq q'$. Define

$$
F(x) = \inf_{q > x} G(q).
$$

27
It is not difficult to verify that $F(x)$ is decreasing and right-continuous at every point $x$. It is also easy to show that $F_{n_j}(x) \to F(x)$ at each continuity point of $x$ of $F$ by monotonicity of $F(x)$.

Now assume $\{X_n\}$ is tight. For given $\epsilon \in (0,1)$, there exists a rational number $M > 0$ such that $P(-M < X_n \leq M) \geq 1 - \epsilon$. Equivalently, $F_n(M) - F_n(-M) \geq 1 - \epsilon$ for all $n$. Passing the subsequence $n_j$ to $\infty$. We have that $G(M) - G(-M) \geq 1 - \epsilon$. By definition, $F(M) - F(-M) \geq 1 - \epsilon$. Therefore $F(x)$ is a cdf. 

**Theorem 10** (Almost-sure representations). If $X_n \Rightarrow X_0$ then there exist random variables $Y_n$ on $([0,1], \mathcal{B}[0,1])$ such that $\mathcal{L}(Y_n) = \mathcal{L}(X_n)$ for each $n \geq 0$ and $Y_n \to Y_0$ almost surely.

**Proof.** Let $F_n(x)$ be the cdf of $X_n$. Let $U$ be a random variable uniformly distributed on $([0,1], \mathcal{B}[0,1])$. Define $Y_n = F_n^{-1}(U)$ where the inverse here is the generalized inverse of a non-decreasing function defined by

$$F_n^{-1}(t) = \inf\{x : F_n(x) \geq t\}$$

for $0 \leq t \leq 1$. One can verify that $F_n^{-1}(t) \to F_0^{-1}(t)$ for every $t$ such that $F_0(x)$ is continuous at $t$. We know that the set of the discontinuous points is countable. Therefore the Lebesgue measure is zero. This shows that $Y_n = F_n^{-1}(U) \to F_0^{-1}(U) = Y_0$ almost surely.

It is easy to check that $\mathcal{L}(X_n) = \mathcal{L}(Y_n)$ for all $n = 0, 1, 2, \cdots$. 

**Lemma 5.7** (Slusky). Let $\{a_n\}$, $\{b_n\}$ and $\{X_n\}$ are sequences of random variables such that $a_n \to a$ in probability and $b_n \to b$ in probability and $X_n \Rightarrow X$ for constants $a$ and $b$ and some random variable $X$. Then $a_nX_n + b_n \Rightarrow aX + b$ as $n \to \infty$.

**Proof.** We will show that $(a_nX_n + b_n) - (aX_n + b) \to 0$ in probability. If this is the case, then the desired conclusion follows from the weak convergence of $X_n$ (why?). Given $\epsilon > 0$. Since $\{X_n\}$ is convergent, it is tight. So there exists $M > 0$ such that $P(|X_n| \geq M) \leq \epsilon$. Moreover, by the given condition,

$$P\left(|a_n - a| \geq \frac{\epsilon}{2M}\right) < \epsilon \text{ and } P(|b_n - b| \geq \epsilon) < \epsilon$$

as $n$ is sufficiently large. Note that

$$|(a_nX_n + b_n) - (aX_n + b)| \leq |a_n - a||X_n| + |b_n - b|.$$
Thus,

\[ P\left( |(a_nX_n + b_n) - (aX_n + b)| \geq 2\epsilon \right) \leq P\left( |a_n - a| \geq \frac{\epsilon}{2M} \right) + P(|X_n| \geq M) + P(|b_n - b| \geq \epsilon) < 3\epsilon \]

as \( n \) is sufficiently large. □

**Remark.** All the above statements of weak convergence of random variables \( X_n \) can be replaced by their distributions \( \mu_n := \mathcal{L}(X_n) \). So \( \lim_{n \to \infty} X_n = X \) is equivalent to \( \lim_{n \to \infty} \mu_n = \mu \).

**Theorem 11** (Lévy’s continuity theorem). Let \( \mu_n; n = 0, 1, 2, \ldots \) be a sequence of probability measures on \( \mathbb{R}^k \) with characteristic function \( \phi_n \). We have that

(i) If \( \mu_n \Rightarrow \mu_0 \) then \( \phi_n(t) \to \phi_0(t) \) for every \( t \in \mathbb{R}^k \);

(ii) If \( \phi_n(t) \) converges pointwise to a limit \( \phi_0(t) \) that is continuous at 0, then the associated sequence of distributions \( \mu_n \) is tight and converges weakly to the measure \( \mu_0 \) with characteristic function \( \phi_0 \).

**Remark.** The condition that \( \phi_0(t) \) is continuous at 0 is essential. Let \( X_n \sim N(0, n) \), \( n \geq 1 \). So \( P(X_n \leq x) \to 1/2 \) for any \( x \), which means that \( X_n \) doesn’t converges weakly. Note that \( \phi_n(t) = e^{-nt^2/2} \to 0 \) as \( n \to \infty \) and \( \phi_n(0) = 1 \). So \( \phi_0(t) \) is not continuous. If one ignores the continuity condition at zero, the conclusion that \( X_n \) converges weakly may wrongly follow.

**Proof of Theorem 11.** (i) is trivial.

(ii) Because marginal tightness implies joint tightness. W.L.O.G., assume that \( \mu_n \) is a probability measure on \( \mathbb{R} \) and generated by \( X_n \). Note that for every \( x \) and \( \delta > 0 \),

\[ 1\{|\delta x| > 2\} \leq 2 \left( 1 - \frac{\sin \delta x}{\delta x} \right) = \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \cos tx) \, dt. \]

Replace \( x \) by \( X_n \), take expectations, and use Fubini’s theorem to obtain that

\[ P\left( |X_n| > \frac{2}{\delta} \right) \leq \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - Ee^{itX_n}) \, dt = \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \phi_n(t)) \, dt. \]

By the given condition,

\[ \limsup_n P\left( |X_n| > \frac{2}{\delta} \right) \leq \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \phi_0(t)) \, dt. \]

The right hand side above goes to zero as \( \delta \downarrow 0 \). So tightness follows.
To prove the weak convergence, we have to show that \( \int f(x) \, du_n \to \int f(x) \, d\mu \) for every bounded continuous function \( f(x) \). It is equivalent to that for any subsequence, there is a further subsequence, say \( n_k \), such that \( \int f(x) \, du_{n_k} \to \int f(x) \, d\mu_0 \).

For any subsequence, by Lemma 5.6, the Helly selection principle, there is a further subsequence, say \( \mu_{n_k} \), such that it converges to \( \mu_0 \) weakly, by the earlier proof, the c.f. of \( \mu_0 \) is \( \phi_0 \). We know that a c.f. uniquely determines a distribution. This says that every limit is identical. So \( \int f(x) \, du_{n_k} \to \int f(x) \, d\mu_0 \).

**Theorem 12** (Central Limit Theorem). Let \( X_1, X_2, \ldots, X_n \) be a sequence of i.i.d. random variables with mean zero and variance one. Let \( \bar{X}_n = (X_1 + \cdots + X_n)/n \). Then \( \sqrt{n} \bar{X}_n \Rightarrow N(0, 1) \).

**Proof.** Let \( \phi(t) \) be the c.f. of \( X_1 \). Since \( E X_1 = 0 \) and \( E X_1^2 = 1 \) we have that \( \phi'(0) = iEX_1 = 0 \) and \( \phi''(0) = t^2EX_1^2 = -1 \). By Taylor’s expansion
\[
\phi \left( \frac{t}{\sqrt{n}} \right) = 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right)
\]
as \( n \) is sufficiently large. Hence
\[
E e^{it\sqrt{n} \bar{X}_n} = \phi \left( \frac{t}{\sqrt{n}} \right)^n = \left( 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right) \right)^n \to e^{-t^2/2}
\]
as \( n \to \infty \). By Levy’s continuity theorem, the desired conclusion follows.

To deal with multivariate random variables, we have the next tool.

**Theorem 13** (Cramér-Wold device). Let \( X_n \) and \( X \) be random variables taking values in \( \mathbb{R}^k \). Then
\[
X_n \Longrightarrow X \text{ if and only if } t^T X_n \Longrightarrow t^T X \text{ for all } t \in \mathbb{R}^k.
\]

**Proof.** By Levy’s continuity Theorem 11, \( X_n \Longrightarrow X \) if and only if \( E \exp(it^T X_n) \to E \exp(it^T X) \) for each \( t \in \mathbb{R}^k \), which is equivalent to \( E \exp(iu(t^T X_n)) \to E \exp(iu(t^T X)) \) for any real number \( u \). This is same as saying that the c.f. of \( t^T X_n \) converges to that of \( t^T X \) for each \( t \in \mathbb{R}^k \). It is equivalent to that \( t^T X_n \Longrightarrow t^T X \) for all \( t \in \mathbb{R}^k \).

As an application, we have the multivariate analogue of the one-dimensional CLT.

**Theorem 14** Let \( X_1, X_2, \ldots \) be a sequence of i.i.d. random vectors in \( \mathbb{R}^k \) with mean vector \( \mu = EX_1 \) and covariance matrix \( \Sigma = E( X_1 - \mu)(X_1 - \mu)^T \). Then
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) = \sqrt{n} (\bar{X}_n - \mu) \Longrightarrow N_k(0, \Sigma).
\]
Proof. By the Cramér-Wold device, we need to show that
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t^T X_i - t^T \mu) \implies N(0, t^T \Sigma t)
\]
for any \( t \in \mathbb{R}^k \). Note that \( \{t^T X_i - t^T \mu, \ i = 1, 2, \ldots \} \) are i.i.d. real random variables. By the one-dimensional CLT,
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t^T X_i - t^T \mu) \implies N(0, \text{Var}(t^T X_1)).
\]
This leads to the desired conclusion since \( \text{Var}(t^T X_1) = t^T \Sigma t \).

We state the following theorem without giving the proof. The spirit of the proof is similar to that of Theorem 12. It is called Lindeberg-Feller Theorem.

**Theorem 15** Let \( \{k_n; \ n \geq 1\} \) be a sequence of positive integers and \( k_n \to +\infty \) as \( n \to \infty \). For each \( n \) let \( Y_{n,1}, \ldots, Y_{n,k_n} \) be independent random vectors with finite variances such that
\[
\sum_{i=1}^{k_n} \text{Cov}(Y_{n,i}) \to \Sigma \quad \text{and} \quad \sum_{i=1}^{k_n} E\|Y_{n,i}\|^2 I\{\|Y_{n,i}\| > \epsilon\} \to 0
\]
for every \( \epsilon > 0 \). Then the sequence \( \Sigma_{i=1}^{k_n} (Y_{n,i} - EY_{n,i}) \implies N(0, \Sigma) \).

The Central Limit Theorems hold not only for independent random variables, it is also true for some other dependent random variables of special structures.

Let \( Y_1, Y_2, \ldots \) be a sequence of random variables. A sequence of \( \sigma \)-algebra \( \mathcal{F}_n \) satisfies that \( \mathcal{F}_n \supset \sigma(Y_1, \ldots, Y_n) \) and is non-decreasing \( \mathcal{F}_n \subset \mathcal{F}_{n+1} \) for all \( n \geq 0 \), where \( \mathcal{F}_0 = \{\emptyset, \Omega\} \). When \( E(Y_{n+1} | \mathcal{F}_n) = 0 \) for any \( n \geq 0 \), we say \( \{Y_n; \ n \geq 1\} \) are martingale differences.

**Theorem 16** For each \( n \geq 1 \), let \( \{X_{ni}; 1 \leq i \leq k_n\} \) be a sequence of martingale differences relative to nested \( \sigma \)-algebras \( \{\mathcal{F}_{ni}; 1 \leq i \leq k_n\} \), that is, \( \mathcal{F}_{ni} \subset \mathcal{F}_{nj} \) for \( i < j \), such that
1. \( \{\max_i |X_{ni}|; \ n \geq 1\} \) is uniformly integrable;
2. \( E(\max_i |X_{ni}|) \to 0 \);
3. \( \sum_i X_{ni}^2 \to 1 \) in probability.

Then \( S_n = \sum_i X_{ni} \to N(0,1) \) in distribution.

**Example.** Let \( X_i; \ i \geq 1 \) be i.i.d. with mean zero and variance one. Then one can verify that the required conditions in Theorem 16 hold for \( X_{ni} = X_i/\sqrt{n} \) for \( i = 1, 2, \ldots, n \).
Therefore we recollect the classical CLT that \( \sum_{i=1}^{n} X_i / \sqrt{n} \to N(0, 1) \). Actually,

\[
P\left( \max_i |X_{ni}| \geq \epsilon \right) \leq n P(|X_1| \geq \sqrt{n} \epsilon) \leq \frac{n}{n \epsilon^2} E|X_1|^2 I\{|X_1| \geq \sqrt{n} \epsilon\} \to 0 \quad \text{and} \quad \frac{E(\max_i |X_{ni}|)^2}{n} \leq \frac{E \sum_{i=1}^{n} X_i^2}{n} = 1
\]

This shows that \( \max_i |X_{ni}| \to 0 \) in probability and \( \{\max_i |X_{ni}|, n \geq 1\} \) is uniformly integrable. So condition 2 holds.

**Lemma 5.8** Let \( \{U_n; n \geq 1\} \) and \( T_n; n \geq 1\) be two sequences of random variables such that

1. \( U_n \to a \) in probability, where \( a \) is a constant;
2. \( \{T_n\} \) is uniformly integrable;
3. \( \{U_nT_n\} \) is uniformly integrable;
4. \( ET_n \to 1 \).

Then \( E(T_nU_n) \to a \).

**Proof.** Write \( T_n U_n = T_n(U_n - a) + aT_n \). Then \( E(T_nU_n) = E\{T_n(U_n - a)\} + aET_n \). Evidently, \( \{T_n(U_n - a)\} \) is uniformly integrable by the given condition. Since \( \{T_n\} \) is uniformly integrable, it is not difficult to verify that \( T_n(U_n - a) \to 0 \) in probability. Then the result follows. ■

By the Taylor expansion, we have the following equality

\[
e^{ix} = (1 + ix)e^{-x^2/2 + r(x)}
\]

for some function \( r(x) = O(x^3) \) as \( x \to \infty \). Look at the set-up in Theorem 16, we have that \( e^{it\sum_j X_{nj}} = T_n U_n \), where

\[
T_n = \prod_j (1 + itX_{nj}) \quad \text{and} \quad U_n = \exp\left(-\left(t^2/2\right) \sum_j X_{nj}^2 + \sum_j r(tX_{nj})\right).
\]

We next obtain a general result on CLT by using Lemma 5.8.

**Theorem 17** Let \( \{X_{nj}, 1 \leq i \leq k_n, n \geq 1\} \) be a triangular array of random variables. Suppose

1. \( ET_n \to 1 \);
2. \( \{T_n\} \) is uniformly integrable;
3. $\sum_j X_{nj}^2 \to 1$ in probability;
4. $E(\max_j |X_{nj}|) \to 0$.

Then $S_n = \sum_j X_{nj}$ converges to $N(0, 1)$ in distribution.

Proof. Since $|T_n U_n| = |e^{iS_n}| = 1$, we know that $\{T_n U_n\}$ is uniformly integrable. By Lemma 5.8, it is enough to show that $U_n \to e^{-t^2/2}$ in probability. Review (5.7) and condition 3, it suffices to show that $\sum_j r(tX_{nj}) \to 0$ in probability. The assertion $r(x) = O(x^3)$ say that there exists $A > 0$ and $\delta > 0$ such that $|r(x)| \leq A|x|^3$ as $|x| < \delta$. Then

$$P(|\sum_j r(tX_{nj})| > \epsilon) \leq P(\max_j |X_{nj}| \geq \delta) + P(|\sum_j r(tX_{nj})| > \epsilon, \max_j |X_{nj}| < \delta)$$

$$\leq \delta^{-1} E(\max_j |X_{nj}|) + P(\max_j |X_{nj}|, \sum_j |X_{nj}|^2 > \epsilon) \to 0$$

as $n \to \infty$ by the given conditions. ■

Proof of Theorem 16. Define $Z_{n1} = X_{n1}$ and

$$Z_{nj} = X_{nj}I(\sum_{k=1}^{j-1} X_{nk}^2 \leq 2), \quad 2 \leq j \leq k_n.$$

It is easy to check that $\{Z_{nj}\}$ is still a martingale relative to the original $\sigma$-algebra. Now define $J = \inf\{j \geq 1; \sum_{1 \leq k \leq j} X_{nk}^2 > 2\} \wedge k_n$. Then

$$P(X_{nk} \neq Z_{nk} \text{ for some } k \leq k_n) = P(J \leq k_n - 1)$$

$$\leq P(\sum_{1 \leq k \leq k_n} X_{nk}^2 > 2) \to 0$$

as $n \to \infty$. Therefore, $P(S_n \neq \sum_{j=1}^{k_n} Z_{nj}) \to 0$ as $n \to \infty$. To prove the result, we only need to show

$$\sum_{j=1}^{k_n} Z_{nj} \Rightarrow N(0, 1). \quad (5.8)$$

We now apply Theorem 17 to prove this. Replacing $X_{nj}$ by $Z_{nj}$ in (5.7), we have new $T_n$ and $U_n$. Let’s literally verify the four conditions in Theorem 17. By a martingale property and iteration, $ET_n = 1$. So condition 1 holds. Now $\max_j |Z_{nj}| \leq \max_j |X_{nj}| \to 0$ in probability as $n \to \infty$. So condition 4 holds. It is also easy to check that $\sum_j Z_{nj}^2 \to 1$ in probability. Then it remains to show condition 2.
By definition, \( T_n = \prod_{j=1}^{k_n} (1 + itZ_{nj}) = \prod_{1 \leq j \leq J} (1 + itZ_{nj}). \) Thus,

\[
|T_n| = \prod_{k=1}^{J-1} (1 + t^2 X_{nk}^2)^{1/2} |1 + tX_n| \\
\leq \exp \left( \frac{t^2}{2} \sum_{k=1}^{J-1} X_{nk}^2 \right) (1 + |t| \cdot |X_n|) \\
\leq e^{t^2} (1 + |t| \cdot \max_j |X_{nj}|)
\]

which is uniform integrable by condition 1 in Theorem 16. ■

The next is an extreme limit theorem, which is different than the CLT. The limiting distribution is called an extreme distribution.

**Theorem 18** Let \( X_i; 1 \leq i \leq n \) be i.i.d. \( N(0, 1) \) random variables. Let \( W_n = \max_{1 \leq i \leq n} X_i \).

Then

\[
P \left( W_n \leq \sqrt{2 \log n} - \frac{\log_2 n + x}{2 \sqrt{2 \log n}} \right) \to e^{-\frac{1}{2\sqrt{\pi}} e^{x/2}}
\]

for any \( x \in \mathbb{R} \), where \( \log_2 n = \log(\log n) \).

**Proof.** Let \( t_n \) be the right hand side of “\( \leq \)” in the probability above. Then

\[
P(W_n \leq t_n) = P(X_1 \leq t_n)^n = (1 - P(X_1 > t_n))^n.
\]

Since \( (1 - x_n)^n \sim e^{-a} \) as \( n \to \infty \) if \( x_n \sim a/n \). To prove the theorem, it suffices to show that

\[
P(X_1 > t_n) \sim \frac{1}{2\sqrt{\pi} n} e^{x/2}.
\]  

(5.9)

Actually, we know that

\[
P(X_1 > x) \sim \frac{1}{\sqrt{2\pi x}} e^{-x^2/2}
\]

as \( x \to +\infty \). It is easy to calculate that

\[
\frac{1}{\sqrt{2\pi t_n}} \sim \frac{1}{2\sqrt{\pi \log n}} \quad \text{and} \quad \frac{t_n^2}{2} \sim \log n - \log(\sqrt{\log n}) - \frac{x}{2}
\]

as \( n \to \infty \). This leads to (5.9). ■
As an application of the CLT of i.i.d. random variables, we will be able to derive the classical $\chi^2$-test as follows.

Let’s consider a multinomial distribution with $n$ trials, $k$ classes with parameter $p = (p_1, \cdots, p_k)$. Roll a die twenty times. Let $X_i$ be the number of the occurrences of “$i$ dots”, $1 \leq i \leq 6$. Then $(X_1, X_2, \cdots, X_6)$ follows a multinomial distribution with “success rate” $p = (p_1, \cdots, p_6)$. How to test the die is fair or not? From an introductory course, we know that, the null hypothesis is $H_0 : p_1 = \cdots = p_6 = 1/6$. The test statistic is

$$\chi^2 = \sum_{i=1}^{6} \frac{(X_i - np_i)^2}{np_i}.$$  

We use the fact that $\chi^2$ is roughly $\chi^2(5)$ as $n$ is large. We will prove this next.

In general, $X_n = (X_{n,1}, \cdots, X_{n,k})$ follows a multinomial distribution with $n$ trials and “success rate” $p = (p_1, \cdots, p_k)$. Of course, $\sum_{i=1}^{k} X_{n,i} = n$. To be more precise,

$$P(X_{n,1} = x_{n,1}, \cdots, X_{n,k} = x_{n,k}) = \frac{n!}{x_{n,1}! \cdots x_{n,k}!} p_{1}^{x_{n,1}} \cdots p_{k}^{x_{n,k}},$$  

where $\sum_{i=1}^{k} x_{n,i} = n$. Now we prove the following theorem,

**THEOREM 19** As $n \to \infty$,

$$\chi_n^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \Longrightarrow \chi^2(k - 1).$$

We need a lemma.

**LEMMA 5.9** Let $Y \sim N_k(0, \Sigma)$. Then

$$\|Y\|^2 \overset{d}{=} \sum_{i=1}^{k} \lambda_i Z_i^2,$$

where $\lambda_1, \cdots, \lambda_k$ are eigenvalues of $\Sigma$ and $Z_1, \cdots, Z_k$ are i.i.d. $N(0,1)$.

**Proof.** Decompose $\Sigma = O \text{diag}(\lambda_1, \cdots, \lambda_k)O^T$ for some orthogonal matrix $O$. Then the $k$ coordinates of $OY$ are independent with mean zero and variances $\lambda_i$’s. So $\|Y\|^2 = \|OY\|^2$ is equal to $\sum_{i=1}^{k} \lambda_i Z_i^2$, where $Z_i$’s are i.i.d. $N(0,1)$. ■

Another fact we will use is that $AX_n \Longrightarrow AX$ for any $k \times k$ matrix $A$ if $X_n \Longrightarrow X$, where $X_n$ and $X$ are $R^k$-valued random vectors. This can be shown easily by the Cramer-Wold device.
Proof of the Theorem. Let $Y_1, \cdots, Y_n$ be i.i.d. with a multinomial distribution with 1 trials and “success rate” $p = (p_1, \cdots, p_k)$. Then $X_n = Y_1 + \cdots + Y_n$. Each $Y_i$ is a $k$-dimensional vector. By CLT,

$$\frac{X_n - EX_n}{\sqrt{n}} \rightarrow N_k(0, \text{Cov}(Y_1)),$$

where it is easy to see that $EX_n = np$ and

$$\Sigma_1 := \text{Cov}(Y_1) = \begin{pmatrix}
p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\
-p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\
\vdots & \vdots & \ddots & \vdots \\
-p_1p_k & -p_2p_k & \cdots & p_k(1-p_k)
\end{pmatrix}.$$

Set $D = \text{diag}(1/\sqrt{p_1}, \cdots, 1/\sqrt{p_k})$. Then

$$DS_1D^T = I - \begin{pmatrix}\sqrt{p_1} \\
\sqrt{p_2} \\
\vdots \\
\sqrt{p_k}\end{pmatrix} \begin{pmatrix}\sqrt{p_1} \\
\sqrt{p_2} \\
\vdots \\
\sqrt{p_k}\end{pmatrix}^T.$$

Since $\sum_{i=1}^k p_i = 1$, the above matrix is symmetric and idempotent or a projection. Also, trace of $\Sigma_1$ is $k-1$. So $\Sigma_1$ has eigenvalue 0 with one multiplicity and 1 with $k-1$ multiplicity. Therefore, by the lemma, we have that

$$\frac{D(X_n - np)}{\sqrt{n}} \rightarrow N(0, D\Sigma_1D^T).$$

By the lemma, since $g(x) = \|x\|^2$ is a continuous function on $\mathbb{R}^k$,

$$\sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} = \left\| \frac{D(X_n - np)}{\sqrt{n}} \right\|^2 \Rightarrow \sum_{i=1}^{k-1} Z_i^2 \sim \chi^2(k-1).$$

6 Central Limit Theorems for Stationary Random Variables and Markov Chains

Central limit theorems hold not only for independent random variables but also for dependent random variables. In the next we will study the Central Limit Theorems for $\alpha$-mixing random variables and and martingale (a special class of dependent random variables).

Let $X_1, X_2, \cdots$ be a sequence of random variables. For each $n \geq 1$, define $\alpha_n$ by

$$\alpha(n) := \sup |P(A \cap B) - P(A)P(B)|$$

(6.10)
where the supremum is taken over all \( A \in \sigma(X_1, \cdots, X_k), \ B \in \sigma(X_{k+n}, X_{k+n+1}, \cdots) \), and \( k \geq 1 \). Obviously, \( \alpha(n) \) is decreasing. When \( \alpha(n) \to 0 \), then \( X_k \) and \( X_{k+n} \) are approximately independent. In this case the sequence \( \{X_n\} \) is said to be \( \alpha \)-mixing. If the distribution of the random vector \( (X_n, X_{n+1}, \cdots, X_{n+j}) \) does not depend on \( n \), the sequence is said to be stationary.

The sequence is called \( m \)-dependent if \( (X_1, \cdots, X_k) \) and \( (X_{k+n}, \cdots, X_{k+n+l}) \) are independent for any \( n > m, \ k \geq 1 \) and \( l \geq 1 \). In this case the sequence is \( \alpha \)-mixing with \( \alpha_n = 0 \) for \( n > m \). An independent sequence is 0-dependent.

For stationary process, we have the following the Central Limit Theorems.

**THEOREM 20** Suppose that \( X_1, X_2, \cdots \) is stationary with \( EX_1 = 0 \) and \( |X_1| \leq A \) for some constant \( A > 0 \). Let \( S_n = X_1 + \cdots + X_n \). If \( \sum_{i=1}^{\infty} \alpha(n) < \infty \), then

\[
\frac{Var(S_n)}{n} \to \sigma^2 = E(X_1^2) + 2 \sum_{k=1}^{\infty} E(X_1X_{k+1})
\]

where the series converges absolutely. If \( \sigma > 0 \), then \( S_n/\sqrt{n} \to N(0, \sigma^2) \).

**THEOREM 21** Suppose that \( X_1, X_2, \cdots \) is stationary with \( EX_1 = 0 \) and \( E|X_1|^{2+\delta} < \infty \) for some constant \( \delta > 0 \). Let \( S_n = X_1 + \cdots + X_n \). If \( \sum_{i=1}^{\infty} \alpha(n)^{\delta(2+\delta)} < \infty \), then

\[
\frac{Var(S_n)}{n} \to \sigma^2 = E(X_1^2) + 2 \sum_{k=1}^{\infty} E(X_1X_{k+1})
\]

where the series converges absolutely. If \( \sigma > 0 \), then \( S_n/\sqrt{n} \to N(0, \sigma^2) \).

Easily we have the following corollary.

**COROLLARY 6.1** Let \( X_1, X_2, \cdots \) be \( m \)-dependent random variables with the same distribution. If \( EX_1 = 0 \) and \( EX_1^{2+\delta} < \infty \) for some \( \delta > 0 \). Then \( S_n/\sqrt{n} \to N(0, \sigma^2) \), where \( \sigma^2 = EX_1^2 + 2 \sum_{i=1}^{m+1} E(X_1X_i) \).

**Remark.** If we argue by using the truncation method as in the proof of Theorem 21, we can show that the above corollary still holds only under \( EX_1^2 < \infty \).

**LEMMA 6.1** If \( Y \in \sigma(X_1, \cdots, X_k) \) and \( Z \in \sigma(X_{k+n}, X_{k+n+1}, \cdots, X_{k+m}) \) for \( n \leq m \leq +\infty \). Assume \( |Y| \leq C \) and \( |Z| \leq D \). Then

\[
|E(YZ) - (EY)EZ| \leq 4CD\alpha(n).
\]
Proof. W.l.o.g, assume that $C = D = 1$. Now since expectations can be approximated by finite sums, so w.l.o.g, we further assume that

$$Y = \sum_{i=1}^{m} y_i A_i \text{ and } Z = \sum_{j=1}^{n} z_j B_j$$

where $A_i$’s and $B_j$’s are respectively partitions of the sample space. Then

$$|E(Y Z) - (EY)EZ| = \left| \sum_{i} y_i \sum_{j} z_j (P(A_i \cap B_j) - P(A_i)P(B_j)) \right|$$

$$\leq \sum_{i} \left| \sum_{j} z_j (P(A_i \cap B_j) - P(A_i)P(B_j)) \right|.$$ 

Let $\xi_i = 1$ or $-1$ depends on whether or not $z_j(P(A_i \cap B_j) - P(A_i)P(B_j)) \geq 0$ or not. Then the last sum is equal to $\sum_{i} \sum_{j} \xi_i z_j(P(A_i \cap B_j) - P(A_i)P(B_j))$ which is also identical to $\sum_{j} z_j \sum_{i} \xi_i (P(A_i \cap B_j) - P(A_i)P(B_j))$. By the same argument, there are $\eta_i$ equal to 1 or $-1$ such that the last quantity is bounded by $\sum_{j} \sum_{i} \xi_i \eta_i (P(A_i \cap B_j) - P(A_i)P(B_j))$. Let $E_1$ be the union of $A_i$ such that $\xi_i = 1$ and $E_2 = E_1^c$; Let $F_1$ and $F_2$ be defined similarly for $B_j$ according to the sign of $\eta_j$’s. Then

$$|E(Y Z) - (EY)EZ| \leq \sum_{1 \leq i,j \leq 2} |P(E_i \cap F_j) - P(E_i)P(F_j)| \leq 4\alpha(n)$$

**Lemma 6.2** If $Y \in \sigma(X_1, \ldots, X_k)$ and $Z \in \sigma(X_{k+n}, X_{k+n+1}, \ldots)$. Assume for some $\delta > 0$, $E(Y^{2+\delta}) \leq C$ and $E(Z^{2+\delta}) \leq D$. Then

$$|E(Y Z) - (EY)EZ| \leq 5(1+C+D)\alpha(n)^{\delta/2+\delta}.$$ 

**Proof.** Let $Y_0 = YI(|Y| \leq a)$ and $Y_1 = YI(|Y| > a)$; and $Z_0 = ZI(|Z| \leq a)$ and $Z_1 = ZI(|Z| > a)$. Then

$$|E(Y Z) - (EY)EZ| \leq \sum_{1 \leq i,j \leq 2} |E(Y_i Z_j) - (EY_i)EZ_j|.$$ 

By the previous lemma, the term corresponding to $i = j = 0$ is bounded by $4a^2\alpha(n)$. For $i = j = 1$, $|E(Y_1 Z_1) - (EY_1)EZ_1| = |\text{Cov}(Y_1, Z_1)|$ which is bounded by $\sqrt{E(Y_1^2)E(Z_1^2)} \leq \sqrt{CD}/a^\delta$ since $E(Y_1^2) \leq C/a^\delta$. Now

$$|E(Y_0 Z_1) - (EY_0)EZ_1| \leq E(|Y_0 - EY_0| \cdot |Z_1 - EZ_1|) \leq 2a \cdot 2E|Z_1|$$

which is less than $4D/a^\delta$. The term for $i = 1, j = 0$ is bounded by $4C/a^\delta$. In summary

$$|E(Y Z) - (EY)EZ| \leq 4a^2\alpha(n) + \frac{\sqrt{CD} + 4C + 4D}{a^2} \leq 4a^2\alpha(n) + \frac{4.5(C + D)}{a^\delta}.$$ 

38
Now take \( a = \alpha(n)^{-1/(2+\delta)} \). The conclusion follows.

**Proof of Theorem 20.** By Lemma 6.2, \(|EX_1X_1+k| \leq 4A^2\alpha(k)\), so the series (6.11) converges absolutely. Moreover, let \( \rho_k = E(X_1X_1+k) \). Then \( ES^2_n = nEX^2_1 + 2\sum_{1 \leq i < j \leq n} E(X_iX_j) = n\rho_0 + 2\sum_{k=1}^{n-1} (n-k)\rho_k \). Then it is to see that

\[
\frac{E(S^2_n)}{n} \to \rho_0 + 2\sum_{k=1}^{\infty} \rho_k. \tag{6.12}
\]

We claim that

\[
E(S^4_n) = n^3\delta_n \tag{6.13}
\]

for a sequence of constants \( \delta_n \to 0 \). Actually, \( E(S^4_n) \leq \sum_{1 \leq i,j,k,l \leq n} |E(X_iX_jX_kX_l)| \). Among the terms in the sum there are \( n \) terms of \( X_i^4 \); at most \( n^2 \) of \( X_i^2X_j^2 \) and \( X_i^3X_j \) respectively for \( i \neq j \). So the contribution of these terms is bounded by \( Cn^2 \) for an universal constant \( C \). There are only two type of terms left: \( X_i^2X_jX_k \) and \( X_iX_jX_kX_l \) where these indices are all different. We will deal with the second type of terms only. The first one can be done similarly and more easily. Note that

\[
\sum_{1 \leq i,j,k,l \leq n} |E(X_iX_jX_kX_l)| \leq 4!n \sum_{1 \leq i+j+k \leq n} |E(X_1X_{1+i}X_{1+i+j}X_{1+i+j+k})| \tag{6.14}
\]

where the three indices in the last sum are not zero. By Lemma 6.2, the term inside is bounded by \( 4A^4\alpha_{i} \). By the same argument, it is also bounded by \( 4A^4\alpha_{k} \). Therefore, the sum in (6.14) is bounded by

\[
4 \cdot 4!A^4n^2 \sum_{2 \leq i+k \leq n} \min\{\alpha_i, \alpha_k\} \leq 2Kn^2 \sum_{2 \leq i+k \leq n} \min\{\alpha_i, \alpha_k\} \leq 2Kn^2 \sum_{k=1}^{n} k\alpha_k = n\delta_n \tag{6.15}
\]

for a sequence of constants \( \delta_n \) such that \( 0 < \delta_n \to 0 \). This is because

\[
\sum_{k=1}^{n} k\alpha_k = \sum_{k \leq \sqrt{n}} k\alpha_k + \sum_{\sqrt{n} < k \leq n} k\alpha_k \leq \sqrt{n} \sum_{k=1}^{\infty} \alpha_k + n \sum_{k \geq \sqrt{n}} \alpha_k = o(n).
\]
Therefore (6.13) follows. From this, we know that $ES_n^4 \leq n\delta_n \leq n^3(\delta_n + (1/n)) \leq n^3\delta_n'$ where $\delta_n := \sup_{k \geq n}\{\delta_k + (1/k) ; k \geq n\}$ is decreasing to 0 and is bounded below by 1/n. Thus, w.l.o.g., assume $\delta_n$ has this property.

Define $p_n = [\sqrt{n}\log(1/\delta_n)]$ and $q_n = [n/p_n]$ and $r_n = [n/(p_n + q_n)]$. Then

$$\sqrt{n} \leq p_n \leq \sqrt{n}\log n \quad \text{and} \quad \frac{p_n^2\delta_{p_n}}{n} \leq \delta_{p_n}\log^2\left(\frac{1}{\delta_{p_n}}\right) \leq \delta_{p_n}\log^2\left(\frac{1}{\delta_{p_n}}\right) \to 0 \quad (6.16)$$

as $n \to \infty$. Now for finite $n$, break $X_1, X_2, \ldots, X_n$ into some blocks as follows:

$$X_1, \ldots, X_{p_n} | X_{p_n+1}, \ldots, X_{p_n+q_n} | X_{p_n+q_n+1}, \ldots, X_{2p_n+q_n} | \ldots, X_n$$

For $i = 1, 2, \ldots, r_n$, set

$$U_{n,i} = X_{(i-1)(p_n+q)}+1 + X_{(i-1)(p_n+q)}+2 + \cdots + X_{i(p_n+q)};$$
$$V_{n,i} = X_{ip_n+(i-1)q+1} + X_{ip_n+(i-1)q+2} + \cdots + X_{i(p_n+q)};$$
$$W_n = X_{r_n(p_n+q)+1} + X_{r_n(p_n+q)+2} + \cdots + X_n,$$

and $S_n' = \sum_{i=1}^{r_n} U_{n,i}$ and $S_n'' = \sum_{i=1}^{r_n} V_{n,i}$. In what follows, we will show that $S_n''$ and $W_n$ are negligible. Now, by (6.12) and its reasoning

$$Var(S_n'') = r_n Var(V_{n,1}) + 2 \sum_{k=2}^{r_n} (r_n-k) E(V_{n,1}V_{n,k}) \leq Cr_nq_n + 2r_n \sum_{k=2}^{r_n} |E(V_{n,1}V_{n,k})|.$$

By assumption, $|V_{n,i}| \leq Aq_n$ for any $i \geq 1$. By Lemma 6.1, $|E(V_{n,1}V_{n,k})| \leq 4A^2q_n^2\alpha_{(k-1)p_n}$. Thus, by monotonicity,

$$\sum_{k=2}^{r_n} |E(V_{n,1}V_{n,k})| \leq 4A^2q_n^2 \sum_{k=2}^{r_n} \alpha_{(k-1)p_n} \leq 4A^2q_n^2 \sum_{k=2}^{r_n} \frac{1}{p_n} \sum_{j=(k-2)p_n+1}^{(k-1)p_n} \alpha_j \leq 4A^2q_n^2 \sum_{j=1}^{\infty} \alpha_j.$$

Since $q_n/p_n \to 0$, the above two inequalities imply that

$$\frac{Var(S_n'')}{n} \to 0 \quad (6.17)$$

as $n \to \infty$. This says that $S_n''/\sqrt{n} \to 0$ in probability. By Chebyshev’s inequality and (6.12) we see that $W_n/\sqrt{n} \to 0$ in probability. Thus, to prove the theorem, it suffices to show that $S_n'/\sqrt{n} \Rightarrow N(0, \sigma^2)$. This will be done through the following two steps

$$\frac{\tilde{S}_n}{\sqrt{n}} \Rightarrow N(0, \sigma^2) \quad \text{and} \quad Ee^{itS_n'/\sqrt{n}} - Ee^{it\tilde{S}_n/\sqrt{n}} \to 0 \quad (6.18)$$
for any \( t \in \mathbb{R} \), where \( \tilde{S}_n = \sum_{i=1}^{r_n} U'_{n,i} \) and \( \{U'_{n,i} : 1 = 1, 2, \cdots, r_n\} \) are i.i.d. with the same distribution as that of \( U_{n,1} \). First, by (6.12), \( \text{Var}(\tilde{S}_n) = r_n \text{Var}(U_{n,1}) \sim r_n \sigma_n^2 \). It follows that \( \text{Var}(\tilde{S}_n)/n \to \sigma^2 \). Also, \( EU'_{n,i} = 0 \). By (6.13) and (6.16),
\[
\frac{1}{\text{Var}(\tilde{S}_n)^2} \sum_{i=1}^{r_n} E(U'_{n,i})^4 \approx \frac{r_n^2 \sigma_n^4}{n^2 \sigma^4} \sim \frac{\sigma^2}{r_n^2} \sigma_n^4 \to 0
\]
as \( n \to \infty \). By the Lyapunov central limit theorem, we know the first assertion in (6.18) holds. Last, by Lemma 6.12,
\[
|e^{it\sqrt{n}} - (e^{it\sqrt{U_{n,1}/n}}) \exp(it \sum_{j=2}^{r_n} U_{n,j}/\sqrt{n})| \leq 16\alpha(q_n).
\]
Then repeat this inequality for \( \exp(it \sum_{j=2}^{r_n} U_{n,j}/\sqrt{n}) \) and so on, we obtain that
\[
|e^{it\sqrt{n}} - e^{it\sqrt{\tilde{S}_n}/\sqrt{n}}| = |e^{it\sqrt{S_n}/\sqrt{n}} - \prod_{j=1}^{r_n} e^{it\sqrt{U'_n}/\sqrt{n}}| \leq 16\alpha(q_n)r_n.
\]
Since \( \alpha(n) \) is decreasing and \( \sum_{n \geq 1} \alpha(n) < \infty \), we have \( n\alpha(n)/2 \leq \sum_{\lceil n/2 \rceil \leq i \leq n} \alpha(i) \to 0 \). Thus the above bound is \( o(r_n/q_n) = o(n/(p_nq_n)) \to 0 \) as \( n \to \infty \). The second statement in (6.18) follows. The proof is complete.

**Proof of Theorem 21.** By Lemma 6.2, \( |EX_1X_k| \leq 5(1 + 2E|X_1|^{2+\delta})\alpha(n)^{2/(2+\delta)} \), which is summable by the assumption. Thus the series \( EX_1^2 + 2\sum_{i=2}^\infty E(X_1X_k) \) is absolutely convergent. By the exact same argument as in proving (6.11), we obtain that \( \text{Var}(S_n)/n \to \sigma^2 \).

Now for a given constant \( A > 0 \) and all \( i = 1, 2, \cdots \), define
\[
Y_i = X_i \mathbb{I}\{|X_i| \leq A\} - E(X_i \mathbb{I}\{|X_i| \leq A\}) \quad \text{and} \quad Z_i = X_i \mathbb{I}\{|X_i| > A\} - E(X_i \mathbb{I}\{|X_i| > A\}),
\]
and \( S' = \sum_{i=1}^{r_n} Y_i \) and \( S'' = \sum_{i=1}^{r_n} Z_i \). Obviously, \( S_n = S' + S'' \). By Theorem 20,
\[
\frac{S_n}{\sqrt{n}} \Rightarrow N(0, \sigma(A)^2)
\]
as \( n \to \infty \) for each fixed \( A > 0 \), where \( \sigma(A)^2 := EY_1^2 + 2\sum_{k=2}^\infty EY_k \). By the summable assumption on \( \alpha(n) \), it is easy to check that both the sum in the definition of \( \sigma(A)^2 \) and \( \tilde{\sigma}(A)^2 := EZ_1^2 + 2\sum_{k=2}^\infty E(Z_1Z_k) \) are absolutely convergent, and
\[
\sigma(A)^2 \to \sigma^2 \quad \text{and} \quad \tilde{\sigma}(A)^2 \to 0
\]
as \( A \to +\infty \). Now

\[
\lim_{n \to \infty} \sup \mathcal{P} \left( \frac{|S_n'| \| \sqrt{n} \geq \epsilon \right) \leq \lim_{n \to \infty} \sup \frac{\text{Var}(S_n'/\sqrt{n})}{n\epsilon^2} \to \frac{\tilde{\sigma}(A)}{\epsilon^2}.
\] (6.21)

by using the same argument as in (6.2). Since \( \mathcal{P}(S_n'/\sqrt{n} \in F) \leq \mathcal{P}(S_n'/\sqrt{n} \in \bar{F}^c) + \mathcal{P}(|S_n'|/\sqrt{n} \geq \epsilon) \) for any closed set \( F \subset \mathbb{R} \) and any \( \epsilon > 0 \), where \( \bar{F}^c \) is the closed \( \epsilon \)-neighborhood of \( F \). Thus, from (6.19) and (6.21)

\[
\lim_{n \to \infty} \sup \mathcal{P} \left( \frac{S_n}{\sqrt{n}} \in F \right) \leq \mathcal{P}(N(0, \sigma^2) \in \bar{F}^c) + \lim_{n \to \infty} \sup \mathcal{P} \left( \frac{|S_n'|}{\sqrt{n}} \geq \epsilon \right)
\]

for any \( A > 0 \) and \( \epsilon > 0 \). First let \( A \to +\infty \), and then let \( \epsilon \to 0^+ \) to obtain from (6.20) that

\[
\lim_{n \to \infty} \sup \mathcal{P} \left( \frac{S_n}{\sqrt{n}} \in F \right) \leq \mathcal{P}(N(0, \sigma^2) \in F).
\]

The proof is complete. \( \blacksquare \)

Define by \( \tilde{\alpha}(n) \) the supremum of

\[
|E \left( f(X_1, \cdots, X_k)g(X_{k+n}, \cdots, X_{k+m}) \right) - Ef(X_1, \cdots, X_k) \cdot Eg(X_{k+n}, \cdots, X_{k+m})| \quad (6.22)
\]

over all \( k \geq 1, m \geq 0 \) and measurable functions \( f(x) \) defined on \( \mathbb{R}^m \) and \( g(x) \) defined on \( \mathbb{R}^{m-n+1} \) with \( |f(x)| \leq 1 \) and \( |g(x)| \leq 1 \) for all \( x \in \mathbb{R} \).

**Lemma 6.3** The following are true:

(i) \( \alpha(n) = \sup_{k \geq 1} \sup_{m \geq 0} \sup_{A \in \sigma(X_1, \cdots, X_k), B \in \sigma(X_{k+n}, \cdots, X_{k+n+m})} |\mathcal{P}(A \cap B) - \mathcal{P}(A)\mathcal{P}(B)|. \)

(ii) For all \( n \geq 1 \), we have that \( \alpha(n) \leq \tilde{\alpha}(n) \leq 4\alpha(n) \).

**Proof.** (i) For two sets \( A \) and \( B \), let \( A \Delta B = (A \setminus B) \cup (B \setminus A) \), which is their symmetric difference. Then \( I_A \Delta B = |I_A - I_B| \). Let \( Z_i; i \geq 1 \) be a sequence of random variables. We claim that

\[
\sigma(Z_1, Z_2, \cdots) = \{B \in \sigma(Z_1, Z_2, \cdots); \text{ for any } \epsilon > 0, \text{ there exists } l \geq 1 \text{ and } C \in \sigma(Z_1, \cdots, Z_l) \text{ such that } \mathcal{P}(B \Delta C) < \epsilon \}. \quad (6.23)
\]

If this is true, then for any \( \epsilon > 0 \), there exists \( l < \infty \) and \( C \in \sigma(Z_1, \cdots, Z_l) \) such that \( \mathcal{P}(B) - \mathcal{P}(C) | < \epsilon. \) Thus (ii) follows.
Now we prove this claim. First, note that
\[ \sigma(Z_1, Z_2, \cdots) = \sigma\{\{Z_i \in E_i\}; E_i \in \mathcal{B}(\mathbb{R}); i \geq 1\}. \]
Denote by \( \mathcal{F} \) the right hand side of (6.23). Then \( \mathcal{F} \) contains all \( \{Z_i \in E_i\}; E_i \in \mathcal{B}(\mathbb{R}); i \geq 1 \). We only need to show that the right hand side is a \( \sigma \)-algebra.

It is easy to see that (i) \( \Omega \in \mathcal{F} \); (ii) \( B^c \in \mathcal{F} \) if \( B \in \mathcal{F} \) since \( A \Delta B = A^c \Delta B^c \). (iii) If \( B_i \in \mathcal{F} \) for \( i \geq 1 \), there exist \( m_i < \infty \) and \( C_i \in \sigma(Z_1, \cdots, Z_{m_i}) \) such that \( P(B_i \Delta C_i) < \epsilon/2^i \) for all \( i \geq 1 \). Evidently, \( \bigcup_{i=1}^\infty B_i \uparrow \bigcup_{i=1}^\infty B_i \) as \( n \to \infty \). Therefore, there exists \( n_0 < \infty \) such that
\[ |P(\bigcup_{i=1}^\infty B_i) - P(\bigcup_{i=1}^{n_0} B_i)| \leq \epsilon/2. \]
It is easy to check that \( \bigcup_{i=1}^\infty B_i \Delta (\bigcup_{i=1}^{n_0} C_i) \subset \bigcup_{i=1}^{n_0} (B_i \Delta C_i) \). Write \( B = \bigcup_{i=1}^\infty B_i \) and \( \tilde{B} = \bigcup_{i=1}^{n_0} B_i \) and \( C = \bigcup_{i=1}^{n_0} C_i \). Note \( B \Delta C \subset (B\setminus\tilde{B}) \cup (\tilde{B} \Delta C) \). These facts show that \( P(B \Delta C) < \epsilon \) and \( C \in \sigma(Z_1, Z_2, \cdots, Z_l) \) for some \( l < \infty \) Thus \( \mathcal{F} \) is a \( \sigma \)-algebra.

(ii) Take \( f \) and \( g \) to be indicator functions, from (i) we get \( \alpha(n) \leq \tilde{\alpha}(n) \). By Lemma 6.1, we have that \( \tilde{\alpha}(n) \leq 4\alpha(n) \). 

Let \( \{X_i; i \geq 1\} \) be a sequence of random variables. We say it is a Markov chain if
\[ P(X_{k+1} \in A|X_1, \cdots, X_k) = P(X_{k+1} \in A|X_k) \]
for any Borel set \( A \) and any \( k \geq 1 \). A Markov chain has the property that **given the present observations, the past and future observations are (conditionally) independent**. This is literally stated in Proposition 10.3.

Let’s look at the strong mixing coefficient \( \tilde{\alpha}(n) \). Recall the definition of \( \tilde{\alpha}(n) \). If \( X_1, X_2, \cdots \) is a Markov chain, for simplicity, we write \( f \) and \( g \), respectively, for \( f(X_1, \cdots, X_k) \) and \( g(X_{k+n}, \cdots, X_{k+n}) \). Let \( \tilde{f} = E(f|X_{k+1}) \) and \( \tilde{g} = E(g|X_{k+1}) \). Then by Proposition 10.3, \( E(fg) = E(\tilde{f}\tilde{g}) \), and \( Ef = E\tilde{f} \) and \( Eg = E\tilde{g} \). Thus,
\[ E(fg) - (Ef)Eg = E(\tilde{f}\tilde{g}) - (E\tilde{f})E\tilde{g}. \]
The point is that \( \tilde{f} \) is \( X_{k+1} \) measurable, and \( \tilde{g} \) is \( X_{k+n-1} \) measurable. Therefore
\[ \alpha(n) \leq \sup_{|f| \leq 1, |g| \leq 1} |E(f(X_{k+1})g(X_{k+n-1}) - (Ef(X_{k+1}))Eg(X_{k+n-1})| \quad (6.25) \]
\[ \leq 4 \sup_{A, B} |P(X_{k+1} \in A, X_{k+n-1} \in B) - P(X_{k+1} \in A)P(X_{k+n-1} \in B)| \quad (6.26) \]
by Lemma 6.1 by taking \( k \) there to be \( k + 1 \) and \( k + n \) to be \( k + n + 1 \) and \( X_i = 0 \) for all positive integer \( i \) excluding \( k + 1 \) and \( k + n + 1 \).

If the Markov process is stationary, then
\[ \alpha(n + 1) \leq 4 \sup_{A, B} |P(X_1 \in A, X_n \in B) - P(X_1 \in A)P(X_n \in B)|. \quad (6.27) \]

43
Let \( P^n(x, B) \) represent the probability of \( X_{n+1} \in B \) given \( X_1 = x \), and \( \pi \) be the marginal probability distribution of \( X_1 \). Then

\[
P(X_{n+1} \in B, X_1 \in A) = \int_A P^n(x, B) \pi(dx)
\]

From this,

\[
\alpha(n+2) \leq 4 \cdot \sup_{A,B} \left| \int_A (P^n(x, B) - P(X_n \in B)) \pi(dx) \right|.
\]

For two probability measures \( \mu \) and \( \nu \), define its total variation distance \( \|\mu - \nu\| = \sup_A |\mu(A) - \nu(A)| \). Then

\[
\alpha(n+2) \leq 4 \int \|P^n(x, \cdot) - \pi\| \pi(dx), \ n \geq 2.
\]

In practice, for example, in the theory of Monte Carlo Markov Chains, we have some conditions to guarantee the convergence rate of \( \|P^n(x, \cdot) - \pi\| \). In other words, we have a function \( M(x) \geq 0 \) and \( \gamma(n) \) such that

\[
\|P^n(x, \cdot) - \pi\| \leq M(x) \gamma(n)
\]

for all \( x \) and \( n \). Here are some examples:

- If a chain is so-called polynomially ergodic of order \( m \), then \( \gamma(n) = n^{-m} \);
- If a chain is geometrically ergodic, then \( \gamma(n) = t^n \) for some \( t \in (0, 1) \);
- If a chain is uniformly geometrically ergodic, then \( \gamma(n) = t^n \) for some \( t \in (0, 1) \) and \( \sup_x M(x) < \infty \).

These three conditions, combining with the Central Limit Theorems of stationary random variables stated in Theorems 20 and 21, imply the following CLT for Markov chains directly.

Let \( E_\pi M = \int M(x) \pi(dx) \).

**Theorem 22** Suppose \( \{X_1, X_2, \ldots\} \) is a stationary Markov chain with \( \mathcal{L}(X_1) = \pi \) and \( E_\pi M < \infty \). If the chain is Geometrically ergodic, in particular, if it is uniformly geometrically ergodic, then

\[
\frac{S_n - nEX_1}{\sqrt{n}} \Rightarrow N(0, \sigma^2)
\]

where \( \sigma^2 = E(X_1^2) + 2 \sum_{i=2}^{\infty} \infty E(X_1X_k) \).
Theorem 23 Suppose \( \{X_1, X_2, \cdots \} \) is a stationary Markov chain with \( \mathcal{L}(X_1) = \pi \) and \( E_\pi M < \infty \).

(i) If \( |X_1| \leq C \) for some constant \( C > 0 \), and the chain is polynomially ergodic of order \( m > 1 \), then (6.28) holds.

(ii) If the chain is polynomially ergodic of order \( m \), and \( E|X_1|^{2+\delta} < \infty \) for some \( \delta > 0 \) satisfying \( m\delta > 2 + \delta \), then (6.28) holds.

Remark. In practice, given an initial distribution \( \lambda \), and a transitional kernel \( P(x, A) \), here is the way to generate a Markov chain: randomly draw \( X_0 = x \) under distribution \( \lambda \), randomly draw \( X_1 \) under distribution \( P(X_0, \cdot) \) and \( X_{n+1} \) under \( P(X_n, \cdot) \) for any \( n \geq 1 \). Through this a Markov chain \( X_1, X_2, \cdots \) with initial distribution \( \lambda \) and transitional kernel \( P(x, \cdot) \) has been generated. Let \( \pi \) be the stationary distribution of this chain, that is, \( \pi(\cdot) = \int P(x, \cdot)\pi(dx) \).

Theorem 17.1.6 (on p. 420 from “Markov Chains and Stochastic Stability” by S.P. Meyn and R.L. Tweedie) says that a Markov CLT holds for a certain initial distribution \( \lambda_0 \), it then holds for any initial distribution \( \lambda \). In particular, if Theorem 22 or Theorem 23 (\( \lambda = \pi \) in this case) holds, then the CLT holds for the Markov chain of the same transitional probability kernel \( P(x, \cdot) \) and stationary distribution \( \pi \) for any initial distribution \( \lambda \).

7 U-Statistics

Let \( X_1, \cdots, X_n \) be a sequence of i.i.d. random variables. Sometimes we are interested in estimating \( \theta = Eh(X_1, \cdots, X_m) \) for \( 1 \leq m \leq n \). To make discussion simple, we assume \( h(x_1, \cdots, x_m) \) is a symmetric function. If \( X(1) \leq X(2) \leq \cdots \leq X(n) \), the order statistic, is sufficient and complete, then

\[
U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} h(X_{i_1}, \cdots, X_{i_m})
\]  

(7.1)

is an unbiased estimator of \( \theta \) and is also a function of \( X(1), \cdots, X(n) \), and hence is uniformly minimum variance unbiased estimator (UMVUE). To see \( U_n \) is a function of \( X(1), \cdots, X(n) \), one can rewrite it as

\[
U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} h(X(i_1), \cdots, X(i_m)).
\]  

(7.2)

We can see that \( U_n \) is a conditional expectation of \( h(x_1, \cdots, X_m) \). In fact, the following is true.
\textbf{PROPOSITION 7.1} The equality \( U_n = E(h(X_1, \ldots, X_m)|X(1), \ldots, X(n)) \) holds for all \( 1 \leq m \leq n \).

\textbf{Proof.} One can see that

\[ E(h(X_1, \ldots, X_m)|X(1), \ldots, X(n)) = E(h(X_{i_1}, \ldots, X_{i_m})|X(1), \ldots, X(n)) \]

for any \( 1 \leq i_1 < \cdots < i_m \leq n \). Summing both sides above over all such indices and then dividing by \( \binom{n}{m} \), we obtain

\[ E(h(X_1, \ldots, X_m)|X(1), \ldots, X(n)) = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}) = U_n \]

by (7.2) \quad \blacksquare

\textbf{Example} Let \( \theta = E(X_1^2) \). We know \( E(X_1^2) = (1/2)E(X_1 - X_2)^2 \). Taking \( h(x_1, x_2) = (1/2)(x_1 - x_2)^2 \), then the corresponding \( U \)-statistic is

\[ U_n = \frac{2}{n(n-1)} \cdot \frac{1}{2} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 = \frac{1}{n-1} \sum_{l=1}^{n} (X_l - \bar{X})^2 = S^2, \]

which is the sample variance.

\textbf{Proof.} One-sample Wilcoxon statistic \( \theta = P(X_1 + X_2 \leq 0) \). The statistic is

\[ U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} I(X_i + X_j \leq 0) \]

which is \( U \)-statistic.

\textbf{Example.} Gini’s mean difference is to measure the concentration of the distribution, the parameter is \( \theta = E|X_1 - X_2| \). The statistic is

\[ U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} |X_i - X_j|. \]

This is again a \( U \)-statistic.

Now we calculate the variance of \( U \)-statistics. Assume \( E[h(X_1, \cdots, X_m)^2] < \infty \). Set

\[ h_k(x_1, \ldots, x_k) = E(h(X_1, \cdots, X_m)|X_1 = x_1, \ldots, X_k = x_k) = Eh(x_1, \cdots, x_k, X_{k+1}, \ldots, X_m). \quad (7.3) \]

Evidently, \( h_m(x_1, \cdots, x_m) = h(x_1, \ldots, x_m) \).
**Theorem 24** *(Hoeffding’s Theorem)* For a $U$-statistic given by (7.1) with $E[h(X_1, \cdots , X_m)^2] < \infty$,

$$\text{Var}(U_n) = \frac{1}{(n^m)^2} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k$$

where $\zeta_k = \text{Var}(h_k(X_1, \cdots , X_k))$.

**Proof.** Consider two sets $\{i_1, \cdots , i_m\}$ and $\{j_1, \cdots , j_m\}$ of $m$ distinct integers from $\{1, 2, \cdots , n\}$ with exactly $k$ integers in common. Then the total number of distinct choices is $\binom{n}{m} \binom{m}{k} \binom{n-m}{m-k}$.

Now

$$\text{Var}(U_n) = \frac{1}{(n^m)^2} \sum \text{Cov}(h(X_{i_1}, \cdots , X_{i_m}), h(X_{j_1}, \cdots , X_{j_m}))$$

where the sum is over all indices $1 \leq i_1 < \cdots < i_m \leq n$ and $1 \leq j_1 < \cdots < j_m \leq n$. If $\{i_1, \cdots , i_m\}$ and $\{j_1, \cdots , j_m\}$ have no integers in common, the covariance is zero by independence. Now we classify the sum over the number of integers in common. Then

$$\text{Var}(U_n) = \frac{1}{(n^m)^2} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \cdot \text{Cov}(h(X_1, \cdots , X_m), h(X_{i_1}, \cdots , X_{i_k}, X_{m+1}, \cdots , X_{2m-k}))$$

$$= \frac{1}{(n^m)^2} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \cdot \text{Var}(\zeta_k)$$

where the fact (from conditioning) that the covariance above equal to $\text{Var}(\zeta_k)$ is used in the last step. 

**Lemma 7.1** Under the condition of Theorem 24,

(i) $0 = \zeta_0 \leq \zeta_1 \leq \cdots \leq \zeta_m = \text{Var}(h)$;

(ii) For fixed $m$ and some $k \geq 1$ such that $\zeta_0 = \cdots = \zeta_{k-1} = 0$ and $\zeta_k > 0$, we have

$$\text{Var}(U_n) = \frac{k! (n^m)^2 \zeta_k}{n^k} + O\left(\frac{1}{n^{k+1}}\right)$$

as $n \to \infty$.

**Proof.** (i) For any $1 \leq k \leq m - 1$,

$$h_k(X_1, \cdots , X_k) = E[h(X_1, \cdots , X_n)|X_1, \cdots , X_k]$$

$$= E[h_{k+1}(X_1, \cdots , X_{k+1})|X_1, \cdots , X_k].$$
Then by Jensen’s inequality, \( E(h_k^2) \leq E(h_{k+1}^2) \). Note that \( Eh_k = Eh_{k+1} \). Then \( \text{Var}(h_k) \leq \text{Var}(h_{k+1}) \). So (i) follows.

(ii) By Theorem 24,

\[
\text{Var}(U_n) = \frac{1}{(m)^n} \binom{m}{k} \binom{n - m}{m - k} \cdot \frac{1}{h_m} \sum_{i=1}^{m} \binom{m}{i} \binom{n - m}{m - i} \cdot \zeta_i.
\]

Since \( m \) is fixed, \( \binom{n}{m} \sim n^m/m! \) and \( \binom{n-m}{m-i} \leq n^{m-i} \) for any \( k+1 \leq i \leq m \). Thus the last term above is bounded by \( O(n^{-(k+1)}) \) as \( n \to \infty \). Now

\[
= \frac{1}{(m)^n} \binom{m}{k} \binom{n - m}{m - k} \cdot \frac{m!}{n(n - 1) \cdots (n - m + 1)} \cdot \frac{(n - m)!}{(m - k)! (n - 2m + k)!} \cdot \frac{1}{h_m} \sum_{i=1}^{m} \binom{m}{i} \binom{n - m}{m - i} \cdot \frac{(n - m) \cdots (n - 2m + k)}{(n - k) \cdots (n - m + 1)}
\]

\[
= \frac{k!}{n^k} \binom{m}{k} \frac{k!}{n^k} \frac{1}{k!} \prod_{i=1}^{k-1} \left( 1 - \frac{i}{n} \right)^{-1} \prod_{i=m}^{m-1} \left( 1 - \frac{i}{n} \right) \prod_{i=k}^{n-m-1} \left( 1 - \frac{i}{n} \right) \cdot \frac{1}{h_m} \sum_{i=1}^{m} \binom{m}{i} \binom{n - m}{m - i} \cdot \frac{(n - m) \cdots (n - 2m + k)}{(n - k) \cdots (n - m + 1)}
\]

Note that since \( m \) is fixed, each of the above product is equal to \( 1 + O(1/n) \). The conclusion then follows.

Let \( X_1, \ldots, X_n \) be a sample, and \( T_n \) be a statistic based on this sample. The projection of \( T_n \) on some random variables \( Y_1, \ldots, Y_p \) is defined by

\[
T_n' = ET_n + \sum_{i=1}^{p} (E(T_n|Y_i) - ET_n).
\]

Suppose \( T_n \) is a symmetric function of \( X_1, \ldots, X_n \). Set \( \psi(X_i) = E(T_n|X_i) \) for \( i = 1, 2, \ldots, n \). Then \( \psi(X_1), \ldots, \psi(X_n) \) are i.i.d. random variables with mean \( ET_n \). If \( \text{Var}(T_n) < \infty \), then

\[
\frac{1}{\sqrt{n\text{Var}(\psi(X_1))}} \sum_{i=1}^{n} (\psi(X_i) - ET_n) \to N(0,1) \quad (7.4)
\]

in distribution as \( n \to \infty \). Now Let \( T_n' \) be the projection of \( T_n \) on \( X_1, \ldots, X_n \). Then

\[
T_n - T_n' = (T_n - ET_n) - \sum_{i=1}^{n} (\psi(X_i) - ET_n).
\]

We will next show that \( T_n - T_n' \) is negligible comparing the order appeared in (7.4). Then the CLT holds for \( T_n \) by (7.4) and the Slusky lemma.
Lemma 7.2 Let $T_n$ be a symmetric statistic with $\text{Var}(T_n) < \infty$ for every $n$, and $T_n'$ be the projection of $T_n$ on $X_1, \cdots, X_n$. Then $ET_n' = ET_n$ and

$$E(T_n - T_n')^2 = \text{Var}(T_n) - \text{Var}(T_n').$$

Proof. Since $ET_n = ET_n'$,

$$E(T_n - T_n')^2 = \text{Var}(T_n) + \text{Var}(T_n') - 2\text{Cov}(T_n, T_n').$$

First, $\text{Var}(T_n') = n\text{Var}(ET_n|X_1)$ by independence. Second, by (7.5)

$$\text{Cov}(T_n, T_n') = \text{Var}(T_n) + n\text{Var}(ET_n|X_1) - 2\sum_{i=1}^n \text{Cov}(T_n - ET_n)(\psi(X_i) - ET_n).$$

(7.6)

Now the $i$-th covariance above is equal to $E(T_n' E(T_n|X_i)) - E(E(T_n|X_i)) \text{Var}(E(T_n|X_1))$. We already see $\text{Var}(T_n') = n\text{Var}(ET_n|X_1)$. So the proof is complete by the two facts and (7.6).

Theorem 25 Let $U_n$ be the statistic given by (7.1) with $E[h(X_1, \cdots, X_m)]^2 < \infty$.

(i) If $\zeta_1 > 0$, then $\sqrt{n}(U_n - ET_n) \rightarrow N(0, m^2 \zeta_1)$ in distribution.

(ii) If $\zeta_1 = 0$ and $\zeta_2 > 0$, then

$$n(U_n - EU_n) - \frac{m(m-1)}{2} \sum_{j=1}^\infty \lambda_j (\chi_j^2 - 1),$$

where $\chi_j^2$’s are i.i.d. $\chi_2(1)$-random variables, and $\lambda_j$’s are constants satisfying $\sum_{j=1}^\infty \lambda_j^2 = \zeta_2$.

Proof. We will only prove (i) here. The proof of (ii) is very technical, it is omitted. Let $U_n'$ be the projection of $U_n$ on $X_1, \cdots, X_n$. Then

$$U_n' = Eh + \frac{1}{m} \sum_{1 \leq i_1 < \cdots < i_m \leq n} E(h(X_{i_1}, \cdots, X_{i_k})|X_i) - Eh.$$

Observe that

$$\frac{1}{m} \sum_{1 \leq i_1 < \cdots < i_m \leq n} E(h(X_{i_1}, \cdots, X_{i_k})|X_1) - Eh$$

$$= \frac{1}{m} \sum_{2 \leq i_2 < \cdots < i_m \leq n} (h_1(X_1) - Eh)$$

$$= \frac{(n-1)}{m} (h_1(X_1) - Eh) = \frac{m}{n} (h_1(X_1) - Eh).$$

49
Thus

\[ U_n' = Eh + \frac{m}{n} \sum_{i=1}^{n} (h_1(X_i) - Eh). \]

This says that \( \text{Var}(U_n') = \frac{m^2}{n} \zeta_1 \), where \( \zeta_1 = \text{Var}(h_1(X_1)) \). By lemma 7.1, \( \text{Var}(U_n) = \frac{m^2}{n} \zeta_1 + O(n^{-2}) \). From Lemma 7.2, \( \text{Var}(U_n - U_n') = O(n^{-2}) \) as \( n \to \infty \). By the Chebyshev inequality, \( \sqrt{n}(U_n - U_n') \to 0 \) in probability. The result follows from (7.4) and the Slusky lemma by noting \( \text{Vart}(\psi(X_1)) = \text{Var}(h_1(X_1)) = \zeta_1 \). ■
8 Empirical Processes

Let \( X_1, X_2, \cdots \) be a random sample, where \( X_i \) takes values in a metric space \( M \). Define a probability measure \( \mu_n \) such that

\[
\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A), \quad A \subset M.
\]

The measure \( \mu_n \) is called a probability measure. If \( X_1 \) takes real values, that is, \( M = \mathbb{R} \), set

\[
F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t), \quad t \in \mathbb{R}, \quad A = (-\infty, t].
\]

The random process \( \{F_n(t), \ t \in \mathbb{R}\} \) is called an empirical process. By the LLN and CLT,

\[
F_n(t) \to F(t) \quad \text{a.s.} \quad \text{and} \quad \sqrt{n}(F_n(t) - F(t)) \Rightarrow N(0, F(t)(1 - F(t))),
\]

where \( F(t) \) is the cdf of \( X_1 \). Among many interesting questions, here we are interested in the uniform convergence of the above almost sure and weak convergences. Specifically, we want to answer the following two questions:

1) Does \( \Delta_n := \sup_t |F_n(t) - F(t)| \to 0 \quad \text{a.s.} \)?

2) Fix \( n \geq 1 \), regard \( \{F_n(t); \ t \in \mathbb{R}\} \) as one point in a big space, does the CLT hold?

The answers are yes for both questions. We first answer the first question.

**Theorem 26** (Glivenko-Cantelli) Let \( X_1, X_2, \cdots \) be a sequence of i.i.d. random variables with cdf \( F(t) \). Then

\[
\Delta_n = \sup_t |F_n(t) - F(t)| \to 0 \quad \text{a.s.}
\]
as \( n \to \infty \).

**Remark.** When \( F(t) \) is continuous and strictly increasing in the support of \( X_1 \), the above distribution of \( \Delta_n \) is independent of the distribution of \( X_1 \). First, \( F(X_i)'s \) are i.i.d. \( U[0,1] \)-distributed random variables. Second,

\[
\Delta_n = \sup_t |\frac{1}{n} \sum_i I(X_i \leq t) - F(t)| = \sup_t |\frac{1}{n} \sum_i I(F(X_i) \leq F(t)) - F(t)|
\]

\[
= \sup_{0 \leq y \leq 1} \left| \frac{1}{n} \sum_i I(Y_i \leq y) - y \right|,
\]
where \( \{Y_i = F(X_i); 1 \leq i \leq n\} \) are i.i.d. random variables with uniform distribution over 
\([0, 1]\).

**Proof.** Set

\[
F_n(t-) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < t) \quad \text{and} \quad F(t-) = P(X_1 < t), \quad t \in \mathbb{R}.
\]

By the strong law of large numbers, \( F_n(t) \to F(t) \) a.s. and \( F_n(t-) \to F(t-) \) a.s. as \( n \to \infty \).

Given \( \epsilon > 0 \), choose \(-\infty = t_0 < t_1 < \cdots < t_k = \infty \) such that \( F(t_i-) - F(t_{i-1}) < \epsilon \) for every \( i \). Now, for \( t_{i-1} \leq t < t_i \),

\[
F_n(t) - F(t) \leq F_n(t_i-) - F(t_i-) + \epsilon
\]
\[
F_n(t) - F(t) \geq F_n(t_i) - F(t_i) - \epsilon.
\]

This says that

\[
\Delta_n \leq \max_{1 \leq i \leq k} \{ |F_n(t_i) - F(t_i)|, |F_n(t_i-) - F(t_i-)| \} + \epsilon \to \epsilon \quad \text{a.s.}
\]

as \( n \to \infty \). The conclusion follows by letting \( \epsilon \downarrow 0 \). \[\Box\]

Given \((t_1, t_2, \cdots, t_k) \in \mathbb{R}^k\), it is easy to check that \( \sqrt{n}(F_n(t_1) - F(t_1), \cdots, F_n(t_k) - F(t_k)) \Rightarrow N_k(0, \Sigma) \) where

\[
\Sigma = (\sigma_{ij}) \quad \text{and} \quad \sigma_{ij} = F(t_i \wedge t_j) - F(t_i)F(t_j), \quad 1 \leq i, j \leq n.
\]

Let \( D[-\infty, \infty] \) be the set of all functions defined on \((-\infty, \infty)\) such that they are right-continuous and the left limits exist everywhere. Equipped with a so-called Skorohod metric, it becomes a Polish space. The random vector \( \sqrt{n}(F_n - F) \), viewed as an element in \( D[-\infty, \infty] \), converges weakly to a continuous Gaussian process \( \xi(t) \), where \( \xi(\pm \infty) = 0 \) and the covariance structure is given in (8.1). This is usually called a Brownian bridge. It has same distribution of \( G_\lambda \circ F \), where \( G_\lambda \) is the limiting point when \( F \) is the cumulative distribution function of the uniform distribution over \([0, 1]\).

**Theorem 27 (Donsker).** If \( X_1, X_2, \cdots \) are i.i.d. random variables with distribution function \( F \), then the sequence of empirical processes \( \sqrt{n}(F_n - F) \) converges in distribution in the space \( D[-\infty, \infty] \) to a random element \( G_F \), whose marginal distributions are zero-mean with covariance function (8.1).

Later we will see that

\[
\sqrt{n}\|F_n - F\|_\infty = \sqrt{n}\sup_t |F_n(t) - F(t)| \Rightarrow \sup_t |G_F(t)|,
\]

52
where
\[ P \left( \sup_t |G_F(t)| \geq x \right) = 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2x^2}, \ x \geq 0. \]

Also, the DKW (Dvoretsky, Kiefer, and Wolfowitz) inequality says that
\[ P(\sqrt{n}\|F_n - F\|_\infty > x) \leq 2e^{-2x^2}, \ x > 0. \]

Let \(X_1, X_2, \cdots\) be i.i.d. random variables taking values in \((\mathcal{X}, \mathcal{A})\) and \(\mathcal{L}(X_1) = P\). We write \(\mu_f = \int f(x) \mu(dx)\). So
\[ P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \text{and} \quad P f = \int_{\mathcal{X}} f(x) P(dx). \]

By the Glivenko-Cantelli theorem
\[ \sup_{x \in \mathbb{R}} |P_n([-\infty, x]) - P([-\infty, x])| \to 0 \quad \text{a.s.} \]

Also, if \(P\) has no point mass, then
\[ \sup_{\text{any measurable set } A} |P_n(A) - P(A)| = 1 \]
because \(P_n(A) = 1\) and \(P(A) = 0\) for \(A = \{X_1, X_2, \cdots, X_n\}\). We are searching for \(F\) such that
\[ \sup_{f \in F} |P_n f - Pf| \to 0 \quad \text{a.s.} \quad (8.2) \]

The previous two examples say (8.2) holds for \(F = \{I(-\infty, x], \ x \in \mathbb{R}\}\) but doesn’t hold for the power set \(2^\mathbb{R}\). The class \(F\) is called \(P\)-Glivenko-Cantelli if (8.2) holds.

Define \(G_n = \sqrt{n}(P_n - P)\). Given \(k\) measurable functions \(f_1, \cdots, f_k\) such that \(P f_i^2 < \infty\) for all \(i\), one can also check by the multivariate Central Limit Theorem (CLT) that
\[ (G_n f_1, \cdots, G_n f_k) \Longrightarrow N_k(0, \Sigma) \]
where \(\Sigma = (\sigma_{ij})_{1 \leq i,j \leq n}\) and \(\sigma_{ij} = P(f_i f_j) - (P f_i) P f_j\). This tells us that \(\{G_n f, \ f \in F\}\) satisfies CLT in \(\mathbb{R}^k\). If \(F\) has infinite many members, how we define weak convergence? We first define a space similar to \(\mathbb{R}^k\). Set
\[ l^\infty(F) = \{\text{bounded function } z : F \to \mathbb{R}\} \]
with norm \(\|z\|_\infty = \sup_{F \in F} |z(F)|\). Then \((l^\infty(F), \| \cdot \|_\infty)\) is a Banach space; it is separable if \(F\) is countable. When \(F\) has finite elements, \((l^\infty(F), \| \cdot \|_\infty) = (\mathbb{R}^k, \| \cdot \|_\infty)\).

So, \(G_n : F \to \mathbb{R}\) is a random element and can be viewed as an element in Banach space \(l^\infty(F)\). We say \(F\) is \(P\)-Donsker if \(G_n \Longrightarrow \mathcal{G}\), where \(\mathcal{G}\) is a tight element in \(l^\infty(F)\).
Lemma 8.1 (Bernstein’s inequality). Let \(X_1, \ldots, X_n\) be independent random variables mean zero, \(|X_j| \leq K\) for some constant \(K\) and all \(j\), and \(\sigma_j^2 = EX_j^2 > 0\). Let \(S_n = \sum_{i=1}^n X_i\) and \(s_n^2 = \sum_{j=1}^n \sigma_j^2\). Then

\[
P(\{|S_n| \geq x\}) \leq 2e^{-x^2/2(s_n^2 + Kx)}, \quad x > 0.
\]

Proof. It suffices to show that

\[
P(S_n \geq x) \leq e^{-x^2/2(s_n^2 + Kx)}, \quad x > 0.
\] (8.3)

First, for any \(\lambda > 0\),

\[
P(S_n > x) \leq e^{-\lambda x} E e^{\lambda S_n} = e^{-\lambda x} \prod_{i=1}^n E e^{\lambda X_i}.
\]

Now

\[
E e^{\lambda X_j} = 1 + \sum_{i=2}^\infty \frac{\lambda^i}{i!} EX_j^i \leq 1 + \frac{\sigma_j^2 \lambda^2}{2} \sum_{i=2}^\infty (\lambda K)^{i-2} = 1 + \frac{\sigma_j^2 \lambda^2}{2(1 - \lambda K)} \leq \exp \left( \frac{\sigma_j^2 \lambda^2}{2(1 - \lambda K)} \right)
\]

if \(\lambda K < 1\). Thus

\[
P(S_n > x) \leq \exp \left( -\lambda x + \frac{\lambda^2 s_n^2}{2(1 - \lambda K)} \right)
\]

Now (8.3) follows by choosing \(\lambda = x/(s_n^2 + Kx)\). ■

Recall \(G_n f = \sum_{i=1}^n (f(X_i) - Pf)/\sqrt{n}\). The random variable \(Y_i := (f(X_i) - Pf)/\sqrt{n}\) here corresponds to \(X_i\) in the above lemma. Note that \(\text{Var}(\sum_{i=1}^n Y_i) = \text{Var}(f(X_1)) \leq Pf^2\) and \(\|Y_i\|_\infty \leq 2\|f\|_\infty / \sqrt{n}\). Apply the above lemma to \(\sum_{i=1}^n Y_i\), we obtain

Corollary 8.1 For any bounded, measurable function \(f\),

\[
P(|G_n f| \geq x) \leq 2 \exp \left( -\frac{1}{4} Pf^2 + x\|f\|_\infty / \sqrt{n} \right)
\]

for any \(x > 0\).

Notice that

\[
P(|G_n f| \geq x) \leq 2e^{-Cx} \quad \text{when } x \text{ is large and}
\]

\[
P(|G_n f| \geq x) \leq 2e^{-C'x^2} \quad \text{when } x \text{ is small.} \tag{8.4}
\]

Let’s estimate \(E \max_{1 \leq i \leq m} |Y_i|\) provided \(m\) is large and \(P(|Y_i| \geq x) \leq e^{-x}\) for all \(i\) and \(x > 0\). Then \(E|Y_i|^k \leq k!\) for \(k = 1, 2, \cdots\). The immediate estimate is

\[
E \max_{1 \leq i \leq m} |Y_i| \leq \sum_{i=1}^m E|Y_i| \leq m.
\]
By Holder’s inequality,
\[ E \max_{1 \leq i \leq m} |Y_i| \leq (E \max_{1 \leq i \leq m} |Y_i|^2)^{1/2} \leq \left( \sum_{i=1}^{m} E|Y_i|^2 \right)^{1/2} \leq \sqrt{2m}. \]

Let \( \psi(x) = e^x - 1, \ x \geq 0. \) Following this logic, by Jensen’s inequality
\[ \psi(E \max_{1 \leq i \leq m} |Y_i|/2) \leq m \cdot \max_{1 \leq i \leq m} E\psi(|Y_i|/2) \leq 2m. \]

Take inverse \( \psi^{-1} \) for both sides. We obtain that
\[ E \max_{1 \leq i \leq m} |Y_i| \leq 2 \log(2m). \]

This together with (8.4) leads to the following lemma.

**Lemma 8.2** For any finite class \( \mathcal{F} \) of bounded, measurable, square-integrable functions with \(|\mathcal{F}|\) elements, there is an universal constant \( C > 0 \) such that
\[ C \cdot E \max_f |G_n f| \leq \max_f \|f\|_\infty \log(1 + |\mathcal{F}|) + \max_f \|f\|_{P,2} \sqrt{\log(1 + |\mathcal{F}|)}, \]
where \( \|f\|_{P,2} = (Pf^2)^{1/2} \).

**Proof.** Define \( a = 24\|f\|_\infty / \sqrt{n} \) and \( b = 24Pf^2 \). Define \( A_f = (G_n f) I\{|G_n f| > b/a\} \) and \( B_f = (G_n f) I\{|G_n f| \leq b/a\} \), then \( G_n f = A_f + B_f \). It follows that
\[ E \max_f |G_n f| \leq E \max_f |A_f| + E \max_f |B_f|. \] (8.5)

For \( x \geq b/a \) and \( x \leq b/a \) the exponent in the Bernstein inequality is bounded above by \(-3x/a\) and \(-3x^2/b\), respectively. By Bernstein’s inequality,
\[ P(|A_f| \geq x) \leq P(|G_n f| \geq x \vee b/a) \leq 2 \exp\left(-\frac{3x}{a}\right), \]
\[ P(|B_f| \geq x) = P(b/a \geq |G_n f| \geq x) \leq P(|G_n f| \geq x \wedge b/a) \leq 2 \exp\left(-\frac{3x^2}{b}\right) \]
for all \( x \geq 0 \). Let \( \psi_p(x) = \exp(x^p) - 1, \ x \geq 0, \ p \geq 1. \)
\[ E\psi_1 \left( \frac{|A_f|}{a} \right) = E \int_0^{|A_f|/a} e^x dx = \int_0^\infty P(|A_f| > ax) e^x dx \leq 1. \]

By a similar argument we find that \( E\psi_2(|B_f|/\sqrt{b}) \leq 1. \) Because \( \psi_p(\cdot) \) is convex for all \( p \geq 1. \)

By Jensen’s inequality
\[ \psi_1 \left( E \max_f \frac{|A_f|}{a} \right) \leq E\psi_1 \left( \frac{\max_f |A_f|}{a} \right) \leq E \sum_f \psi_1 \left( \frac{|A_f|}{a} \right) \leq |\mathcal{F}|. \]
Taking inverse function, we have the first term on the right hand side. Similarly we have
the second term. The conclusion follows from (8.5). ■

We actually use the following lemma above

**LEMMA 8.3** Let $X$ be a real valued random variable and $f : [0, \infty) \to \mathbb{R}$ be differentiable
with $\int_0^s |f'(t)| \, dt < \infty$ for any $s > 0$ and $\int_0^\infty P(X > t)|f'(t)| \, dt < \infty$. Then

$$Ef(X) = \int_0^\infty P(X \geq t)f'(t) \, dt + f(0).$$

**Proof.** First, since $\int_0^t |f'(t)| \, dt < \infty$ for any $t > 0$, we have

$$f(X) - f(0) = \int_0^X f'(t) \, dt = \int_0^\infty I(X \geq t)f'(t) \, dt.$$

Then, by Fubini’s theorem,

$$Ef(X) = E\int_0^\infty I(X \geq t)f'(t) \, dt + f(0) = \int_0^\infty P(X \geq t)f'(t) \, dt + f(0).$$

**Remark.** When $f(x)$ is differentiable, $f'(x)$ is not necessarily Lebesgue integrable even
though it is Riemann-integrable. The following is an example: Let

$$f(x) = \begin{cases} x^2 \cos(1/x^2), & \text{if } x \neq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$f'(x) = \begin{cases} 2x \cos(1/x^2) + (2/x) \sin(1/x^2), & \text{if } x \neq 0; \\ 0, & \text{otherwise} \end{cases}$$

is not Lebesgue-integrable on $[0,1]$, but obviously Riemann-integrable. We only need to
show $g(x) := (2/x) \sin(1/x^2)$ is not integrable. Suppose it is,

$$\int_0^1 \frac{1}{x} \left| \sin \left( \frac{1}{x^2} \right) \right| \, dx = \frac{1}{2} \int_0^1 \left| \frac{\sin t}{t} \right| \, dt = +\infty,$$

yields a contradiction.

Now we describe the size of $\mathcal{F}$. For any $f \in \mathcal{F}$, define

$$\|f\|_{P,r} = (P|f'|^r)^{1/r}.$$

Let $l$ and $u$ be two functions, the *bracket* $[l, u]$ is set of all functions $f$ such that $l \leq f \leq u$.

An $\epsilon$-*bracket* in $L_r(P)$ is a bracket $[l, u]$ such that $\|u - l\|_{P,r} < \epsilon$. The bracketing number
$N(\epsilon, \mathcal{F}, L_r(P))$ is the minimum numbers of $\epsilon$-brackets needed to cover $\mathcal{F}$. The *entropy with bracketing*

is the logarithm of the bracketing number.
8.1 Outer Measures and Expectations

Recall $X$ is a random variable from $(\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ if $X^{-1}(B) \in \mathcal{G}$ for any set $B \in \mathcal{B}(\mathbb{R})$.

For an arbitrary map $X$ the inverse may not be in $\mathcal{G}$, particularly for small $\mathcal{G}$. For example, when $\mathcal{G} = \{\emptyset, \Omega\}$, many maps are not random variables. In empirical processes, we will deal with $Z = \sup_{t \in T} X_t$ for some index set $T$. If $T$ is big, then $Z$ may not be measurable. It does not make sense to study expectations and probabilities related to such a random variable. But we have another way to get around this.

**Definition** Let $X$ be an arbitrary map from $(\Omega, \mathcal{G}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define

$$E^* X = \inf \{E Y; \ Y \geq X \text{ and } Y \text{ is a measurable map : } (\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))\};$$

$$P^*(X \in A) = \inf \{P(X \in B); \ \{X \in B\} \in \mathcal{G}, \ B \supset A\}, \ A, B \in \mathcal{B}(\mathbb{R}).$$

One can easily show that $E^* X$, as an infimum, can be achieved, i.e., there exists a random variable $X^* : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $EX^* = E^* X$. Further, $X^*$ is $\mathbb{P}$-almost surely unique, i.e., if there exist two such random variables $X^*_1$ and $X^*_2$, then $P(X^*_1 = X^*_2) = 1$. We call $X^*$ the **measurable cover function**. Obviously

$$(X_1 + X_2)^* \leq X_1^* + X_2^* \text{ and } X_1^* \leq X_2^* \text{ if } X_1 \leq X_2.$$ 

One can define $E_*$ and $P_*$ similarly.

Let $(M, d)$ be a metric space. A sequence of arbitrary maps $X_n : (\Omega_n, \mathcal{G}_n) \to (M, d)$ converges in distribution to a **random vector** $X$ if

$$E^* f(X_n) \to Ef(X)$$

for any bounded, continuous function $f$ defined on $(M, d)$. We still have an analogue of the Portmanteau theorem.

**Theorem 28** The following are equivalent:

(i) $E^* f(X_n) \to Ef(X)$ for every bounded, continuous function $f$ defined on $(M, d)$;

(ii) $E^* f(X_n) \to Ef(X)$ for every bounded, Lipschitz function $f$ defined on $(M, d)$, that is, there is a constant $C > 0$ such that $|f(x) - f(y)| \leq Cd(x,y)$ for any $x, y \in M$;

(iii) $\liminf_n P_n(X_n \in G) \geq P(X \in G)$ for any open set $G$;

(iv) $\limsup_n P^*(X_n \in F) \leq P(X \in F)$ for any closed set $F$;

(v) $\limsup_n P^*(X_n \in H) = P(X \in H)$ for any set $H$ such that $P(X \in \partial H) = 0$, where $\partial H$ is the boundary of $H$.  

57
Let
\[ J_1(\delta, F, L_2(P)) = \int_0^\delta \sqrt{\log N_1(\epsilon, F, L_2(P))} \, d\epsilon. \]

The proof of the following theorem is easy. It is omitted.

**Theorem 29** Every class \( F \) of measurable functions such that \( N_1(\epsilon, F, L_1(P)) < \infty \) for every \( \epsilon > 0 \) is \( P \)-Glivenko-Cantelli.

**Theorem 30** Every class \( F \) of measurable functions with \( J_1(1, F, L_2(P)) < \infty \) is \( P \)-Donsker.

To prove the theorem, we need some preparation.

**Theorem 31** A sequence of arbitrary maps \( X_n = (X_{n,t}, t \in T) : (\Omega_n, \mathcal{G}_n) \to l^\infty(T) \) converges weakly to a tight random element if and only if both of the following conditions hold:

(i) The sequence \( (X_{n,t_1}, \ldots, X_{n,t_k}) \) converges in distribution in \( \mathbb{R}^k \) for every finite set of points \( t_1, \ldots, t_k \) in \( T \);

(ii) for every \( \epsilon, \eta > 0 \) there exists a partition of \( T \) into finitely many sets \( T_1, \ldots, T_k \) such that
\[
\limsup_{n \to \infty} P^* \left( \sup_i \sup_{s,t \in T_i} |X_{n,s} - X_{n,t}| \geq \epsilon \right) \leq \eta.
\]

**Proof.** We only prove sufficiency.

**Step 1: A preparation.** For each integer \( m \geq 1 \), let \( T_1^m, \ldots, T_k^m \) be a partition of \( T \) such that
\[
\limsup_{n \to \infty} P^* \left( \sup_j \sup_{s,t \in T_j^m} |X_{n,s} - X_{n,t}| \geq \frac{1}{2^m} \right) \leq \frac{1}{2^m}. \tag{8.1}
\]

Since the supremum above becomes smaller when a partition becomes more refined, w.l.o.g., assume the partitions are successive refinements as \( m \) increases. Define a semi-metric
\[
\rho_m(s, t) = \begin{cases} 
0, & \text{if } s, t \text{ belong to the same partitioning set } T_j^m \text{ for some } j; \\
1, & \text{otherwise.}
\end{cases}
\]

Easily, by the nesting of the partitions, \( \rho_1 \leq \rho_2 \leq \cdots \). Define
\[
\rho(s, t) = \sum_{m=1}^\infty \frac{\rho_m(s, t)}{2^m} \quad \text{for } s, t \in T.
\]
Obviously, $\rho(s, t) \leq \sum_{k=m+1}^{\infty} 2^{-k} \leq 2^{-m}$ when $s, t \in T_j^n$ for some $j$. So $(T, \rho)$ is totally bounded. Let $T_0$ be the countable $\rho$-dense subset constructed by choosing an arbitrary point $t_j^m$ from every $T_j^m$.

**Step 2: Construct the limit of $X_n$.** For two finite subsets of $T$, $S = (s_1, \cdots, s_p)$ and $U = (s_1, \cdots, s_p, \cdots, s_q)$, by assumption (i), there are two probability measures $\mu_p$ on $\mathbb{R}^p$ and $\mu_q$ on $\mathbb{R}^q$ such that

$$
(X_{n,s_1}, \cdots, X_{n,s_p}) \Rightarrow \mu_p; \quad (X_{n,s_1}, \cdots, X_{n,s_p}, \cdots, X_{n,s_q}) \Rightarrow \mu_q
$$

where $\mu_q(A \times \mathbb{R}^{q-p}) = \mu_p(A)$ for any Borel set $A \subset \mathbb{R}^p$.

By the Kolmogorov consistency theorem there exists a stochastic process $\{X_t; \ t \in T_0\}$ on some probability space such that $(X_{n,t_1}, \cdots, X_{n,t_k}) \Rightarrow (X_{t_1}, \cdots, X_{t_k})$ for every finite set of points $t_1, \cdots, t_k$ in $T_0$. It follows that, for a finite set $S \subset T$,

$$
\sup_i \sup_{s,t \in T_j^m} |X_{n,s} - X_{n,t}| \rightarrow \sup_i \sup_{s,t \in S} |X_s - X_t|
$$

as $n \rightarrow \infty$ for fixed $m \geq 1$. By the Portmanteau theorem and (8.1)

$$
P \left( \sup_i \sup_{s,t \in T_j^m} |X_s - X_t| > \frac{1}{2^m} \right) \leq \limsup_{n \rightarrow \infty} P^* \left( \sup_i \sup_{s,t \in S} |X_{n,s} - X_{n,t}| \geq \frac{1}{2^m} \right) \leq \frac{1}{2^m}
$$

for each $m \geq 1$. By Monotone Convergence Theorem, since $T$ is countable, let $S \uparrow T$, we obtain that

$$
P \left( \sup_i \sup_{s,t \in T_j^m} |X_s - X_t| > \frac{1}{2^m} \right) \leq \frac{1}{2^m}.
$$

If $\rho(s, t) < 2^{-m}$, then $\rho_m(s, t) < 1$ (otherwise, $\rho_m(s, t) = 1$ implies $\rho(s, t) \geq 2^{-m}$). This says that $s, t \in T_j^m$ for the same $j$. So the above inequality implies that

$$
P \left( \sup_{\rho(s,t) < 2^{-m}} |X_s - X_t| > \frac{1}{2^m} \right) \leq \frac{1}{2^m}
$$

for each $m \geq 1$. By the Borel-Cantelli lemma, when $m$ is large, with probability one

$$
\sup_{\rho(s,t) < 2^{-m}} |X_s - X_t| \leq \frac{1}{2^m}
$$
as \( m \) is large enough. This says that, with probability one, \( X_t \), as a function of \( t \) is uniformly continuous on \( T_0 \). Extending this from \( T_0 \) to \( T \), we obtain a \( \rho \)-continuous process \( X_t, t \in T \). We then have that with probability one
\[
|X_s - X_t| \leq \frac{1}{2^m} \quad \text{for any } s, t \in T \text{ with } \rho(s, t) < \frac{1}{2^m}
\] (8.2)
for sufficiently large \( m \).

**Step 3: Show** \( X_n \Longrightarrow X \). Define \( \pi_m : T \rightarrow T \) as the map that maps every point in the partitioning set \( T_i^m \) onto the point \( t_i^m \in T_i^m \). For any bounded, Lipschitz continuous function \( f : l^\infty(T) \rightarrow \mathbb{R} \),
\[
\lim_{n \rightarrow \infty} |E^* f(X_n) - E f(X)| \leq \lim_{n \rightarrow \infty} |E^* f(X_n) - E^* f(X_n \circ \pi_m)| + \lim_{n \rightarrow \infty} |E^* f(X_n \circ \pi_m) - E f(X \circ \pi_m)| + |E f(X \circ \pi_m) - f(X)|.
\]

\( X_n \circ \pi_m \) and \( X \circ \pi_m \) are essentially finite dimensional random vectors, by assumption (i), the first one converges to the second in distribution. So the middle term above is zero.

Now, if \( s, t \in T_i^m, \rho(s, t) \leq 2^{-m} \). It follows that
\[
\|X \circ \pi_m - X\|_\infty = \sup_{t \in T_i^m} |X_t - X_{t_i^m}| < \frac{1}{2^{m-1}}.
\] (8.3)
as \( m \) is large enough. Thus, \( E|f(X \circ \pi_m) - f(X)| \leq \|f\|_\infty 2^{-m+1} \). Since \( f \) is Lipschitz,
\[
|E^* f(X_n) - E^* f(X_n \circ \pi_m)| \leq \|f\|_\infty 2^{-m} + P(\|X_n - X_n \circ \pi_m\| \geq 2^{-m})
\]
By the same argument as in (8.3), using (8.1) to obtain
\[
\lim_{n \rightarrow \infty} |E^* f(X_n) - E f(X)| \leq \|f\|_\infty 2^{-m} + 2^{-m} + 2^{-m+1}\|f\|_\infty.
\]

Letting \( m \rightarrow \infty \), we have that \( E^* f(X_n) \rightarrow E f(X) \) as \( n \rightarrow \infty \). ■

Let \( \mathcal{F} \) be a class of measurable functions \( f : \mathcal{X} \rightarrow \mathbb{R} \). Review
\[
a(\delta) = \frac{\delta}{\sqrt{\log N(\delta, \mathcal{F}, L_2(P))}}
\]
\[
J_1(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P))} \, d\epsilon.
\]

**Lemma 8.4** Suppose there is \( \delta > 0 \) and a measurable function \( F > 0 \) such that \( P^* f^2 \leq \delta^2 \) and \( |f| \leq F \) for every \( f \in \mathcal{F} \). Then there exists a constant \( C > 0 \) such that
\[
C \cdot E^*_P \|G_n\|_F \leq J_1(\delta, \mathcal{F}, L_2(P)) + \sqrt{n} P^* F\{F > \sqrt{n} a(\delta)\}.
\]

60
Recall \( \mathbb{G}_n f = (1/\sqrt{n}) \sum_{i=1}^{n} (f(X_i) - Pf) \). If \(|f| \leq g\), then
\[
|\mathbb{G}_n f| \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(X_i) + P g).
\]

It follows that
\[
E^* \| \mathbb{G}_n f I \{ F > \sqrt{n}a(\delta) \} \|_F \leq 2\sqrt{n}PF \{ F > \sqrt{n}a(\delta) \}.
\]

We will bound \( E^* \| \mathbb{G}_n f I \{ F \leq \sqrt{n}a(\delta) \} \|_F \) next. The bracketing number of the class of functions \( fI\{ F > \sqrt{n}a(\delta) \} \) if \( f \) ranges over \( \mathcal{F} \) is smaller than the bracketing number of the class \( \mathcal{F} \). To simplify notation, we assume, w.l.o.g., \(|f| \leq \sqrt{n}a(\delta)\) for every \( f \in \mathcal{F} \).

**Step 1: Truncation.** Choose an integer \( q_0 \) such that \( 4\delta \leq 2^{-q_0} \leq 8\delta \).

We claim there exists a nested sequence of \( \mathcal{F} \)-partitions \( \{ \mathcal{F}_i; 1 \leq i \leq N_q \} \), indexed by the integers \( q \geq q_0 \), into \( N_q \) disjoint subsets and measurable functions \( \Delta_{qi} \leq 2F \) such that
\[
C \sum_{q=q_0}^{\infty} \frac{1}{2^q} \log N_q \leq \int_0^\delta \sqrt{\log N(\epsilon)} \, d\epsilon, \tag{8.4}
\]

and
\[
\sup_{f,g \in \mathcal{F}_i} |f - g| \leq \Delta_{qi}, \text{ and } P \Delta_{qi}^2 \leq \frac{1}{2^q} \tag{8.5}
\]

for an universal constant \( C > 0 \). For convenience, write \( N(\epsilon) = N(\epsilon; \mathcal{F}, L_2(P)) \). Thus,
\[
\int_0^\delta \sqrt{\log N(\epsilon)} \, d\epsilon \geq \int_0^{1/2^{q_0}+3} \sqrt{\log N(\epsilon)} \, d\epsilon = \sum_{q=q_0+3}^{\infty} \int_{1/2^{q+1}}^{1/2^{q+2}} \sqrt{\log N(\epsilon)} \, d\epsilon \geq \sum_{q=q_0+3}^{\infty} \frac{1}{2^{q+1}} \sqrt{\log N \left( \frac{1}{2^{q+3}} \right)}.
\]

Re-indexing the sum, we have that
\[
\frac{1}{16} \sum_{q=q_0}^{\infty} \frac{1}{2^q} \sqrt{\log N_q} \leq \int_0^\delta \sqrt{\log N(\epsilon)} \, d\epsilon,
\]

where \( N_q = N(2^{-q}) \). By definition of \( N_q \), there exists a partition \( \{ \mathcal{F}_i; l_{qi}, u_{qi}; 1 \leq i \leq N_q \} \), where \( l_{qi} \) and \( u_{qi} \) are functions such that \( E(u_{qi} - l_{qi})^2 < 2^{-q} \). Set \( \Delta_{qi} = u_{qi} - l_{qi} \). Then \( \Delta_{qi} \leq 2F \) and \( \sup_{f,g \in \mathcal{F}_i} |f - g| \leq \Delta_{qi} \), and \( P \Delta_{qi}^2 \leq 2^{-2q} \).

We can also assume, w.l.o.g., that the partitions \( \{ \mathcal{F}_i; 1 \leq i \leq N_q \} \) are successive refined as \( q \) increases. Actually, at \( q \)-th level, we can make intersections of elements from \( \{ \mathcal{F}_i; 1 \leq i \leq N_q \} \) as \( k \) goes from \( q_0 \) to \( q \) : \( \cap_{k=q_0}^{q} \mathcal{F}_{ki} \). The total possible number of such intersections is no more than \( \tilde{N}_q := N_{q_0} \cdots N_{q_0+1} \cdots N_q \). Since the current \( \mathcal{F}_q \)'s become
smaller, all requirements on $\mathcal{F}_{qi}$ still hold obviously except possibly (8.4). Now we verify (8.4). Actually, noticing $\sqrt{\log N_q} \leq \sum_{k=q_0}^q \sqrt{\log N_k}$. Then

$$
\sum_{q \geq q_0} \frac{1}{2q} \sqrt{\log N_q} \leq \sum_{q \geq q_0} \sum_{k=q_0}^q \frac{1}{2q} \sqrt{\log N_k} = \sum_{k=q_0}^q \sum_{q \geq k} \frac{1}{2q} \sqrt{\log N_k} = 2 \sum_{k=q_0}^q \frac{1}{2k} \sqrt{\log N_k}.
$$

So (8.4) still holds when replacing $N_q$ by $\bar{N}_q$.

**Step 3: Chaining-a skill.** For each fixed $q \geq q_0$, choose a fixed element $f_{qi}$ from each partition set $\mathcal{F}_{qi}$, and set

$$
\pi_q f = f_{qi}, \quad \Delta_q f = \Delta_{qi}, \quad \text{if } f \in \mathcal{F}_{qi}.
$$

Thus, $\pi_q f$ and $\Delta_q f$ runs through a set of $\bar{N}_q$ functions if $f$ run through $\mathcal{F}$. Without loss of generality, we can assume

$$
\Delta_q f \leq \Delta_{q-1} f \quad \text{for any } f \in \mathcal{F} \text{ and } q \geq q_0 + 1.
$$

Actually, let $\bar{\Delta}_{qi}$ be the measurable cover function of $\sup_{f,g \in \mathcal{F}_{qi}} |f-g|$. Then (8.5) also holds. Let $\mathcal{F}_{q-1,j}$ be a partition set at the $q-1$ level such that $\mathcal{F}_{qi} \subset \mathcal{F}_{q-1,j}$. Then $\sup_{f,g \in \mathcal{F}_{qi}} |f-g| \leq \sup_{f,g \in \mathcal{F}_{q-1,j}} |f-g|$. Thus, $\bar{\Delta}_{qi} \leq \bar{\Delta}_{q-1,j}$. The assertion (8.7) follows by replacing $\Delta_{qi}$ with $\bar{\Delta}_{qi}$.

By (8.6), $P(\pi_q f - f)^2 \leq \max_q P\Delta_{qi}^2 < 2^{-2q}$. Thus, $P \sum_q (\pi_q f - f)^2 = \sum_q P(\pi_q f - f)^2 < \infty$. So $\sum_q (\pi_q f - f)^2 < \infty$ a.s. This implies

$$
\pi_q f \to f \quad \text{a.s. on } P
$$

as $q \to \infty$ for any $f \in \mathcal{F}$. Define

$$
a_q = 2^{-q}/\sqrt{\log N_{q+1}},
$$

$$
\tau = \tau(n, f) = \inf\{q \geq q_0 : \Delta_q f > \sqrt{n a_q}\}.
$$

The value of $\tau$ is thought to be $+\infty$ if the above set is empty. This is the first time $\Delta_q f > \sqrt{n a_q}$. By construction, $2a(\delta) = 2\delta(\log N_{1|}(\delta, \mathcal{F}, L_2(P)))^{-1/2} \leq a_{q_0}$ since the denominator is decreasing in $\delta$. By Step 1, we know that $|\Delta_q f| \leq 2\sqrt{n a(\delta)} \leq \sqrt{n a_{q_0}}$. This says that $\tau > q_0$.

We claim that

$$f - \pi_{q_0} f = \sum_{q_0+1}^{\infty} (f - \pi_q f) I\{\tau = q\} + \sum_{q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f) I\{\tau \geq q\}, \quad \text{a.s on } P.
$$

In fact, write $f - \pi_{q_0} f = (f - \pi_{q_1} f) + \sum_{q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f)$ for $q_1 > q_0$. Now,
bounded by 2

Lemma 8.2 again to obtain $P$

By Holder’s inequality,$\sqrt{2}$

for some universal constant $C > 0$.

Step 4. Bound terms in the chain. Apply $G_n$ to the both sides of (8.9). For $|f| \leq g$, note that $|G_n f| \leq |G_n g| + 2\sqrt{n} P g$. By (8.6), $|f - \pi_q f| \leq \Delta_q f$. One obtains that

$$E^*\left\| \sum_{q_0+1}^{\infty} G_n (f - \pi_q f) I\{\tau = q\} \right\|_F$$

$$\leq \sum_{q_0+1}^{\infty} E^*\left\| G_n (\Delta_q f) I\{\tau = q\} \right\|_F + 2\sqrt{n} \sum_{q_0+1}^{\infty} \left\| P(\Delta_q f I\{\tau = q\}) \right\|_F. \quad (8.10)$$

Now, by (8.7), $\Delta_q f I\{\tau = q\} \leq \Delta_q - 1 f I\{\tau = q\} \leq \sqrt{n} a_{q-1}$. Moreover, $P(\Delta_q f I\{\tau = q\})^2 \leq 2^{-2q}$. By Lemma 8.2, the middle term above is bounded by $\sum_{q_0+1}^{\infty} (a_{q-1} \log N_q + 2^{-q} \sqrt{\log N_q})$.

By Holder’s inequality, $P(\Delta_q f I\{\tau = q\}) \leq (P(\Delta_q f)^2)^{1/2} P(\tau = q)^{1/2} \leq 2^{-q} P(\Delta_q f > \sqrt{n} a_q)^{1/2} \leq 2^{-q} (\sqrt{n} a_q)^{-1} (P(\Delta_q f)^2)^{1/2} \leq 2^{-2q} (\sqrt{n} a_q)^{-1}$. So the last term in (8.10) is bounded by $2 \cdot 2^{-2q}/a_q$. In summary,

$$E^*\left\| \sum_{q_0+1}^{\infty} G_n (f - \pi_q f) I\{\tau = q\} \right\|_F \leq C \sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\log N_q} \quad (8.11)$$

for some universal constant $C > 0$.

Second, there are at most $N_q$ functions $\pi_q f - \pi_{q-1} f$ and two values the indicator functions $I(\tau \geq q)$ takes. Because the partitions are nested, the function $|\pi_q f - \pi_{q-1} f| I\{\tau \geq q\} \leq \Delta_q - 1 f I\{\tau \geq q\} \leq \sqrt{n} a_{q-1}$. The $L_2(P)$-norm of $\pi_q f - \pi_{q-1} f$ is bounded by $2^{-q+1}$. Applying Lemma 8.2 again to obtain

$$E^*\left\| \sum_{q_0+1}^{\infty} G_n (\pi_q f - \pi_{q-1} f) I\{\tau \geq q\} \right\|_F \leq \sum_{q_0+1}^{\infty} (a_{q-1} \log N_q + 2^{-q} \sqrt{\log N_q})$$

$$\leq C \sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\log N_q} \quad (8.12)$$

for some universal constant $C > 0$. 

63
At last, we consider $\pi q_0 f$. Because $|\pi q_0 f| \leq F \leq a(\delta) \sqrt{n} \leq \sqrt{n} a(q_0)$ and $P(\pi q_0 f)^2 \leq \delta^2$ by assumption, another application of Lemma 8.2 leads to

$$E^* \|G_n \pi q_0 f\|_F \leq a(q_0) \log N_{q_0} + \delta \sqrt{\log N_{q_0}}.$$ 

In view of the choice of $q_0$, this is no more than the bound in (8.12). All above inequalities together with (8.4) yield the desired result. ■

**Corollary 8.2** For any class $\mathcal{F}$ of measurable functions with envelope function $F$, there exists an universal constant $C$ such that

$$E^* \|G_n \|_F \leq C \cdot J_1(\|F\|_{P,2}, \mathcal{F}, L_2(P)).$$

**Proof.** Since $\mathcal{F}$ has a single bracket $[-F, F]$, we have that $N_1(\delta, \mathcal{F}, L_2(P)) = 1$ for $\delta = 2\|F\|_{P,2}$. Review the definition in Lemma 8.4. Choose $\delta = \|F\|_{P,2}$. It follows that

$$a(\delta) = \frac{\|F\|_{P,2}}{\sqrt{\log N_1(\|F\|_{P,2}, \mathcal{F}, L_2(P))}}.$$

Now $\sqrt{n} P^* FI(F > \sqrt{n} a(\delta)) \leq \|F\|_{P,2}^2 / a(\delta) = \|F\|_{P,2} / \sqrt{\log N_1(\|F\|_{P,2}, \mathcal{F}, L_2(P))} \log N_{q_0} \|F\|_{P,2,2}$ by Markov’s inequality, which is bounded by $J_1(\|F\|_{P,2}, \mathcal{F}, L_2(P))$ since the integrand is non-decreasing and hence

$$\int_0^{\|F\|_{P,2}} \sqrt{\log N_1(\epsilon, \mathcal{F}, L_2(P))} d\epsilon \geq \|F\|_{P,2} \sqrt{\log N_{q_0}(\|F\|_{P,2}, \mathcal{F}, L_2(P))}. \quad \blacksquare$$

**Proof of Theorem 30.** We will use Theorem 31 to prove this theorem. The part (i) is easily satisfied. Now we verify (ii).

Note there is no cover on $\mathcal{F}$ and we don’t know if $P f^2 < \delta^2$. Let $\mathcal{G} = \{f - g, f, g \in \mathcal{F}\}$.

With a given set of $\epsilon$-brackets $[l_i, u_i]$ over $\mathcal{F}$ we can construct $2\epsilon$-brackets over $\mathcal{G}$ by taking differences $[l_i - u_j, u_i - l_j]$ of upper and lower bounds. Therefore, the bracketing number $N_1(2\epsilon, \mathcal{G}, L_2(P))$ are bounded by the square of the bracketing number $N_1(\epsilon, \mathcal{F}, L_2(P))$. Easily, $\|(u_i - l_j) - (l_i - u_j)\| \leq 2\epsilon$. Hence

$$N_1(2\epsilon, \mathcal{G}, L_2(P)) \leq N_1(\epsilon, \mathcal{F}, L_2(P))^2.$$

This says that

$$J_1(\epsilon, \mathcal{G}, L_2(P)) < \infty. \quad (8.13)$$
For a given, small \( \delta > 0 \), by the definition of \( N(\delta, \mathcal{F}, L_2(P)) \), choose a minimal number of brackets of size \( \delta \) that cover \( \mathcal{F} \), and use them to form a partition of \( \mathcal{F} = \cup_i \mathcal{F}_i \). The subsets \( \mathcal{G} \) consisting of differences \( f - g \) of functions \( f \) and \( g \) belonging to the same partitioning set consists of functions of \( L_2(P) \)-norm smaller than \( \delta \). Hence, by Lemma 8.4, there exists a finite number \( a(\delta) := a(\delta)_F < a(\delta)_G \) and a universal constant \( C \) such that

\[
C \cdot E^* \sup_{i} \sup_{f,g \in \mathcal{F}_i} |G_n(f-g)| = C E^* \sup_{h \in \mathcal{H}} |G_n h|
\leq J_{\mathcal{H}}(\mathcal{G}, L_2(P)) + 2\sqrt{n} \mathcal{PF}_1(F > a(\delta)\sqrt{n})
\leq J_{\mathcal{G}}(\mathcal{G}, L_2(P)) + 2\sqrt{n} \mathcal{PF}_1(F > a(\delta)\sqrt{n}).
\]

since \( \mathcal{H} := \{ f - g; f, g \in \mathcal{F}_i, \text{ for all } i \} \subset \mathcal{G} \), where the envelope function \( F \) can be taken equal to the supremum of the absolute values of the upper and lower bounds of finitely many brackets that cover \( \mathcal{F} \), for instance a minimal set of brackets of size 1. This \( F \) is square integrable. The second term above is bounded by \( a(\delta)^{-1} \mathcal{PF}_2(\mathcal{F} > a(\delta)\sqrt{n}) \to 0 \) as \( n \to \infty \) for fixed \( \delta \). First let \( n \to \infty \), then let \( \delta \downarrow 0 \) the left hand side goes to 0. So part (ii) in (31) is valid by using Markov’s inequality.

\[ \blacksquare \]

**Example.** Let \( \mathcal{F} = \{ f_t = I(-\infty, t], \ t \in \mathbb{R} \} \). Then

\[
\|f_t - f_s\|_{2,P} = (F(t) - F(s))^{1/2} \text{ for } s < t.
\]

Cut \([0, 1]\) into many pieces with length less than \( \epsilon^2 \). Since \( F(t) \) is increasing, right continuous and of left limits, there exist a partition, say, \( -\infty = t_0 < t_1 < \cdots < t_k = \infty \) with \( k = [1/\epsilon^2] + 1 \) such that

\[
F(t_{i-1}) - F(t_{i-1}) < \epsilon^2 \text{ for } i = 1, 2, \cdots, k.
\]

Since \( \|I(-\infty, t_i) - I(-\infty, t_{i-1})\|_{2,2} = (F(t_i) - F(t_{i-1}))^{1/2} < \epsilon \). Namely, \((I(-\infty, t_{i-1}), I(-\infty, t_i))\) is an \( \epsilon \)-bracket for \( i = 1, 2, \cdots, k \). It follows that

\[
N(\epsilon, \mathcal{F}, L_2(P)) \leq \frac{2}{\epsilon^2} \text{ for } \epsilon \in (0, 1).
\]

But \( \int_0^1 \sqrt{\log(1/\epsilon)} \ d\epsilon < \infty \). The previous classical Gliivenko-Cantelli’s Lemma 26 and Donsker’s Theorem 27 follow from Theorems 29 and 30.

Since \( \|f\| = \sup_{t \in \mathbb{R}} |f(t)| \) is the norm in \( l^\infty \), it is continuous. By the Delta method, we have

\[
\sup_{t \in \mathbb{R}} \sqrt{n}|F_n(t) - F(t)| \to \sup_{t \in \mathbb{R}} G(t)
\]
weakly, i.e., in distribution, as \( n \to \infty \), where \( G(t) \) is the Gaussian process we mentioned before.

**Example.** Let \( \mathcal{F} = \{ f_\theta; \, \theta \in \Theta \} \) be a collection of measurable functions with \( \Theta \subset \mathbb{R}^d \) bounded. Suppose \( Pm^r < \infty \) and \( |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)||\theta_1 - \theta_2|| \) for any \( \theta_1 \) and \( \theta_2 \). Then

\[
N_{|[\epsilon]}(\epsilon||m||_{P,r}, \mathcal{F}, L^r(P)) \leq K \left( \frac{\text{diam}(\Theta)}{\epsilon} \right)^d \tag{8.14}
\]

for any \( 0 < \epsilon < \text{diam}(\Theta) \). As long as \( ||\theta_1 - \theta_2|| < \epsilon \), we have that \( l(x) := f_{\theta_1}(x) - m(x)\epsilon(x) \leq f_{\theta_2}(x) \leq f_{\theta_1}(x) + m(x)\epsilon(x) =: u(x) \). Also, \( ||u - l||_{P,r} = \epsilon||m||_{P,r} \). Naturally, \([l,u]\) is an \( \epsilon||m||_{P,r}\)-bracket. It suffices to calculate the minimal number of balls of radius \( \epsilon \) need to cover \( \Theta \).

Note that \( \Theta \subset \mathbb{R}^d \) is in a cube with side length no bigger than \( \text{diam}(\Theta) \). We can cover \( \Theta \) with fewer than \( (2\text{diam}(\Theta)/\epsilon)^d \) cubes of size \( \epsilon \). The circumscribed balls have radius a multiple of \( \epsilon \) and also cover \( \Theta \). The intersection of these balls with \( \Theta \) cover \( \Theta \). So the claim is true. This says that \( \mathcal{F} \) is a \( P \)-Donsker class.

**Example (Sobolev classes).** For \( k \geq 1 \), let

\[
\mathcal{F} = \left\{ f : [0,1] \to \mathbb{R}; \, ||f||_\infty \leq 1 \text{ and } \int_0^1 (f^{(k)}(x))^2 \, dx \leq 1 \right\}.
\]

Then there exists an universal constant \( K \) such that

\[
\log N_{|[\epsilon]}(\epsilon, \mathcal{F}, || \cdot ||_\infty) \leq K \left( \frac{1}{\epsilon} \right)^{1/k}
\]

Since \( ||f||_{P,2} \leq ||f||_\infty \) for any \( P \), it is easy to check that \( N_{|[\epsilon]}(\epsilon, \mathcal{F}, L_2(P)) \leq N_{|[\epsilon]}(\epsilon, \mathcal{F}, || \cdot ||_\infty) \) for any \( P \). So \( \mathcal{F} \) is \( P \)-Donsker class for any \( P \).

**Example (Bounded Variation).** Let

\[
\mathcal{F} = \{ f : \mathbb{R} \to [-1,1] \text{ of bounded variation} \}.
\]

Any function of bounded variation is the difference of two monotone increasing functions. Then for any \( r \geq 1 \) and probability measure \( P \),

\[
\log N_{|[\epsilon]}(\epsilon, \mathcal{F}, L_r(P)) \leq K \left( \frac{1}{\epsilon} \right).
\]

Therefore \( \mathcal{F} \) is \( P \)-Donsker for every \( P \).
9 Consistency and Asymptotic Normality of Maximum Likelihood Estimators

9.1 Consistency

Let \( X_1, X_2, \cdots, X_n \) be a random sample from a population distribution with pdf or pmf \( f(x|\theta) \), where \( \theta \) is an unknown parameter. The MLE \( \hat{\theta} \) under certain conditions on \( f(x|\theta) \) will be consistent and satisfy Central Limit Theorems. But for some cases those will not be true. Let’s see a good example and a pathological example.

Example. Let \( X_1, X_2, \cdots, X_n \) be i.i.d. from \( \text{Exp}(\theta) \) with pdf

\[
f(x|\theta) = \begin{cases} 
\theta e^{-\theta x}, & \text{if } x > 0; \\
0, & \text{otherwise}.
\end{cases}
\]

So \( EX_1 = 1/\theta \) and \( \text{Var}(X_1) = 1/\theta^2 \). It is easy to check that \( \hat{\theta} = 1/\bar{X} \). By the CLT,

\[
\sqrt{n} \left( \bar{X} - \frac{1}{\theta} \right) \rightarrow N(0, \theta^{-2})
\]

as \( n \to \infty \). Let \( g(x) = 1/x, x > 0 \). Then \( g(EX_1) = \theta \) and \( g'(EX_1) = -\theta^2 \). By the Delta method,

\[
\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, g'(\mu)^2 \sigma^2) = N(0, \theta^2)
\]

as \( n \to \infty \). Of course, \( \hat{\theta} \to \theta \) in probability.

Example. Let \( X_1, X_2, \cdots, X_n \) be i.i.d. from \( U[0, \theta] \), where \( \theta \) is unknown. The MLE estimator \( \hat{\theta} = \max X_i \). First, it is easy to check that \( \hat{\theta} \to \theta \) in probability. But \( \hat{\theta} \) doesn’t follow the CLT. In fact,

\[
P \left( \frac{n(\theta - \hat{\theta})}{\theta} \leq x \right) \rightarrow \begin{cases} 
1 - e^{-x}, & \text{if } x > 0; \\
0, & \text{otherwise}
\end{cases}
\]

as \( n \to \infty \).

For some cases even the consistency, i.e., \( \hat{\theta} \to \theta \) in probability, doesn’t hold. Next, we will study some sufficient conditions for consistency. Later we provide sufficient conditions for CLT.

Let \( X_1, \cdots, X_n \) be a random sample from a density \( p_{\theta} \) with reference measure \( \mu \), that is, \( P_\theta(X_1 \in A) = \int_A p_{\theta}(x)\mu(dx) \), where \( \theta \in \Theta \). The maximum likelihood estimator \( \hat{\theta}_n \) maximizes the function \( h(\theta) := \sum \log p_{\theta}(X_i) \) over \( \Theta \), or equivalently, the function

\[
M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\theta}}{p_{\theta_0}}(X_i),
\]

67
where \( \theta_0 \) is the true parameter. Under suitable conditions, by the weak law of large numbers,

\[
M_n(\theta) \to M(\theta) := E_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}}(X_1) = \int_{\mathbb{R}} p_{\theta_0}(x) \log \frac{p_\theta}{p_{\theta_0}}(x) \mu(dx)
\]

(9.1)

in probability as \( n \to \infty \). The number \(-M(\theta)\) is called the Kullback-Leibler divergence of \( p_\theta \) and \( p_{\theta_0} \). Let \( Y = p_{\theta_0}(Z)/p_\theta(Z) \), where \( Z \sim p_\theta \). Then \( E_\theta Y = 1 \) and

\[
M(\theta) = -E_\theta(Y \log Y) \leq -(E_\theta Y) \log(E_\theta Y) = 0
\]

by Jensen’s inequality. Of course \( M(\theta_0) = 0 \), that is, \( \theta_0 \) attains the maximum of \( M(\theta) \).

Obviously, \( M(\theta_0) = M_n(\theta_0) = 0 \). The following gives the consistency of \( \hat{\theta}_n \) with \( \theta_0 \).

**Theorem 32** Suppose \( \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \overset{P}{\to} 0 \) and \( \sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0) \) for any \( \epsilon > 0 \). Then \( \hat{\theta}_n \to \theta_0 \) in probability.

**Proof.** For any \( \epsilon > 0 \), we need to show that

\[
P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \to 0
\]

(9.2)
as \( n \to \infty \). By the condition, there exists \( \delta > 0 \) such that

\[
\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0) - \delta.
\]

Thus, if \( d(\hat{\theta}_n, \theta_0) \geq \epsilon \), then \( M(\hat{\theta}_n) < M(\theta_0) - \delta \). Note that

\[
M(\hat{\theta}_n) - M(\theta_0) = M_n(\hat{\theta}_n) - M_n(\theta_0) + (M_n(\theta_0) - M(\theta_0)) + (M(\hat{\theta}_n) - M_n(\hat{\theta}_n))
\]

\[
\geq -2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|
\]
since \( M_n(\hat{\theta}_n) \geq M_n(\theta_0) \). Consequently,

\[
P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \leq P(M(\hat{\theta}_n) - M(\theta_0) < -\delta) = P(\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \delta/2) \to 0.
\]

by the first assumption. \( \blacksquare \)

Let \( X_1, \cdots, X_n \) be a random sample from a density \( p_\theta \) with reference measure \( \mu \), that is, \( P_\theta(X_1 \in A) = \int_A p_\theta(x) \mu(dx) \), where \( \theta \in \Theta \). The maximum likelihood estimator \( \hat{\theta}_n \) maximizes the function \( \sum \log p_\theta(X_i) \) over \( \Theta \), or equivalently, the function

\[
M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_\theta}{p_{\theta_0}}(X_i),
\]

68
where $\theta_0$ is the true parameter. This is usually practiced by solving the equation $\partial M_n(\theta)/\partial \theta = 0$ for $\theta$. Let

$$\psi \theta(x) = \frac{\partial \log(p_\theta(x)/p_\theta(x))}{\partial \theta}.$$ 

Then the maximum likelihood estimator $\hat{\theta}_n$ is the zeroes of the function

$$\Psi_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \psi \theta(X_i). \quad (9.3)$$

Though we still use the notation $\Psi$ in the following theorem, it is not necessarily the one above. It can be any function of properties stated below.

**Theorem 33** Let $\Theta$ be a subset of the real line and let $\Psi_n$ be random functions and $\Psi$ a fixed function of $\theta$ such that $\Psi_n(\theta) \to \Psi(\theta)$ in probability for every $\theta$. Assume that $\Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$ or is nondecreasing satisfying $\Psi_n(\hat{\theta}_n) = o_P(1)$. Let $\theta_0$ be a point such that $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$ for every $\epsilon > 0$. Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

**Proof.** **Case 1.** Suppose $\Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$. Then

$$P(\Psi_n(\theta_0 - \epsilon) < 0, \Psi_n(\theta_0 + \epsilon) > 0) \leq P(\hat{\theta}_n - \epsilon \leq \theta_0 \leq \hat{\theta}_n + \epsilon).$$

Since $\Psi_n(\theta_0 \pm \epsilon) \to \Psi(\theta_0 \pm \epsilon)$ in probability, and $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$, the left hand side goes to one.

**Case 2.** Suppose $\Psi_n(\theta)$ is nondecreasing satisfying $\Psi_n(\hat{\theta}_n) = o_P(1)$. Then

$$P(|\hat{\theta}_n - \theta_0| \geq \epsilon) \leq P(\hat{\theta}_n > \theta_0 + \epsilon) + P(\hat{\theta}_n < \theta_0 - \epsilon) \leq P(\Psi_n(\hat{\theta}_n) \geq \Psi_n(\theta_0 + \epsilon)) + P(\Psi_n(\hat{\theta}_n) \leq \Psi_n(\theta_0 - \epsilon)).$$

Now, $\Psi_n(\theta_0 \pm \epsilon) \to \Psi(\theta_0 \pm \epsilon)$ in probability. This together with $\Psi_n(\hat{\theta}_n) = o_P(1)$ shows that $P(|\hat{\theta}_n - \theta_0| \geq \epsilon) \to 0$ as $n \to \infty$. \hfill \blacksquare

**Example.** Let $X_1, \cdots, X_n$ be a random sample from $\text{Exp}(\theta)$, that is, the density function is $p_\theta(x) = \theta^{-1} \exp(-x/\theta) I(x \geq 0)$ for $\theta > 0$. We know that, under the true model, the MLE $\hat{\theta}_n = 1/\bar{X}_n \to \theta_0$ in probability. Let’s verify that this conclusion indeed can be deduced from Theorem 33.

Actually, $\psi \theta(x) = (\log p_\theta(x))' = -\theta^{-1} + x \theta^{-2}$ for $x \geq 0$. Thus $\Psi_n(\theta) = -\theta^{-1} + X_n \theta^{-2}$, which goes to $\Psi(\theta) = E_{\theta_0}(-\theta^{-1} + X_1 \theta^{-2}) = \theta^{-2}(\theta_0 - \theta)$, that is positive or negative depending on if $\theta$ is smaller or bigger than $\theta_0$. Also, $\hat{\theta}_n = 1/\bar{X}_n$. Applying Theorem 33 to $-\Psi_n$ and $-\Psi$, we obtain the consistency result.
9.2 Asymptotic Normality

Now we study the central limit theorems for the MLE.

Now we illustrate the idea of showing the normality of MLE. Recalling (9.3). Do the Taylor’s expansion for \( \psi_{\theta}(X_i) \).

\[
\psi_{\hat{\theta}}(X_i) = \psi_{\theta_{0}}(X_i) + (\hat{\theta}_n - \theta_{0})\dot{\psi}_{\theta_{0}}(X_i) + \frac{1}{2}(\hat{\theta}_n - \theta_{0})^2\ddot{\psi}_{\theta_{0}}(X_i)
\]

We will use the following notation:

\[
Pf = \int f(x)\mu(dx) \quad \text{for any real function } f(x) \quad \text{and} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.
\]

Then \( P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \). Thus,

\[
0 = \Psi_n(\hat{\theta}_n) = P_n\psi_{\theta_{0}} + (\hat{\theta}_n - \theta_{0})P_n\dot{\psi}_{\theta_{0}} + \frac{1}{2}(\hat{\theta}_n - \theta_{0})^2P_n\ddot{\psi}_{\theta_{0}}.
\]

Reorganize it in the following form

\[
\sqrt{n}(\hat{\theta}_n - \theta_{0}) = \frac{-\sqrt{n}(P_n\psi_{\theta_{0}})}{P_n\psi_{\theta_{0}} + \frac{1}{2}(\hat{\theta}_n - \theta_{0})P_n\dot{\psi}_{\theta_{0}}}.
\]

Recall the Fisher information

\[
I(\theta_{0}) = E_{\theta_{0}}\left(\frac{\partial \log p_{\theta}(X_1)}{\partial \theta} \big|_{\theta=\theta_{0}}\right)^2 = E_{\theta_{0}}(\psi_{\theta_{0}}(X_1))^2 = -E_{\theta_{0}}(\frac{\partial^2 \log p_{\theta}(X_1)}{\partial^2 \theta} \big|_{\theta=\theta_{0}}) = -E_{\theta_{0}}(\ddot{\psi}_{\theta_{0}}(X_1)).
\]

By CLT and LLN, \( \sqrt{n}(P_n\psi_{\theta_{0}}) \Rightarrow N(0, I(\theta_{0})) \), \( P_n\dot{\psi}_{\theta_{0}} \rightarrow -I(\theta_{0}) \) and \( P_n\ddot{\psi}_{\theta_{0}} \rightarrow p_{\theta_{0}}\dot{\psi}_{\theta_{0}}(X_1) \) in probability. This illustrates that

\[
\sqrt{n}(\hat{\theta}_n - \theta_{0}) \Rightarrow N(0, I(\theta_{0})^{-1})
\]

as \( n \rightarrow \infty \). The next two theorems will make these steps rigorous.

Let \( g(\theta) = E_{\theta_{0}}(m_{\theta}(X_1)) = \int m_{\theta}(x)p_{\theta_{0}}(x)\mu(dx) \). We need the following condition:

\[
g(\theta) = g(\theta_{0}) + \frac{1}{2}(\theta - \theta_{0})^T V_{\theta_{0}}(\theta - \theta_{0}) + o(||\theta - \theta_{0}||^2), \quad (9.1)
\]

where

\[
V_{\theta_{0}} = \left(E_{\theta_{0}}\left(\frac{\partial m_{\theta}}{\partial \theta_i \partial \theta_j} \big|_{\theta=\theta_{0}}\right)\right)_{1 \leq i,j \leq d}, \quad \theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d.
\]
THEOREM 34  For each \( \theta \) in an open subset of Euclidean space let \( x \mapsto m_\theta(x) \) be a measurable function such that \( \theta \mapsto m_\theta(x) \) is differentiable at \( \theta_0 \) for \( P \)-almost every \( x \) with derivative \( \dot{m}_{\theta_0}(x) \) and such that, for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \) and a measurable function \( n(x) \) with \( \text{En}(X_1)^2 < \infty \)

\[
|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq n(x)||\theta_1 - \theta_2||.
\]

Furthermore, assume the map \( \theta \mapsto \text{Em}_\theta(X_1) \) has expression as in (9.1). If \( P_n\dot{m}_{\hat{\theta}_n} \geq \sup_\theta P_n m_\theta - o_P(n^{-1}) \) and \( \hat{\theta}_n \xrightarrow{P} \theta_0 \), then

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{m}_{\theta_0}(X_i) + o_P(1).
\]

In particular, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, V_{\theta_0}^{-1} E(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T) V_{\theta_0}^{-1}) \) as \( n \to \infty \).

A statistical model \((p_\theta, \theta \in \Theta)\) is called differentiable in quadratic mean if there exists a measurable vector-valued function \( \dot{i}_{\theta_0} \) such that, as \( \theta \to \theta_0 \),

\[
\int \left[ \sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{i}_{\theta_0} \sqrt{p_{\theta_0}} \right]^2 du = o(||\theta - \theta_0||^2).
\]

THEOREM 35  Suppose that the model \((P_0 : \theta \in \Theta)\) is differentiable in quadratic mean at an inner point \( \theta_0 \) of \( \Theta \subset \mathbb{R}^k \). Furthermore, suppose that there exists a measurable function \( l(x) \) with \( E_{\theta_0} l^2(X_1) < \infty \) such that, for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \),

\[
| \log p_{\theta_1}(x) - \log p_{\theta_2}(x) | \leq l(x)||\theta_1 - \theta_2||.
\]

If the Fisher information matrix \( I_{\theta_0} \) is non-singular and \( \hat{\theta}_n \) is consistent, then

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{i}_{\theta_0}(X_i) + o_P(1)
\]

In particular, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, I_{\theta_0}^{-1}) \) as \( n \to \infty \), where

\[
I(\theta_0) = \left( E \frac{\partial^2 \log p_{\theta_0}(X_1)}{\partial \theta_i \partial \theta_j} \right)_{1 \leq i,j \leq k}.
\]

We need some preparations in proving the above theorems.

Given functions \( x \mapsto m_\theta(x) \), \( \theta \in \mathbb{R}^d \), we need conditions that ensure that, for a given sequence \( r_n \to \infty \) and any sequence \( \tilde{h}_n = O_P(1) \),

\[
G_n \left( r_n (m_{\theta_0} + \tilde{h}_n/r_n - m_{\theta_0}) - \tilde{h}_n^T \dot{m}_{\theta_0} \right) \xrightarrow{P} 0.
\]

(9.2)
**Lemma 9.1** For each \( \theta \) in an open subset of Euclidean space let \( m_\theta(x) \), as a function of \( x \), be measurable for each \( \theta \); and as a function of \( \theta \), is differentiable for almost every \( x \) (w.r.t. \( P \)) with derivative \( \dot{m}_\theta(x) \) and such that for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \) and a measurable function \( \dot{m} \) such that \( \dot{m}^2 < \infty \),

\[
\|m_{\theta_1}(x) - m_{\theta_2}(x)\| \leq \dot{m}(x)\|\theta_1 - \theta_2\|.
\]

Then (9.2) holds for every random sequence \( \tilde{h}_n \) that is bounded in probability.

**Proof.** Because \( \tilde{h}_n \) is bounded in probability, to show (9.2), it is enough to show, w.l.o.g.

\[
\sup_{|\theta| \leq 1} |G_n(r_n(m_{\theta_0 + \theta/r_n} - m_{\theta_0}) - \theta^T \dot{m}_{\theta_0})| \xrightarrow{P} 0
\]
as \( n \to \infty \). Define

\[
F_n = \{ f_\theta := r_n(m_{\theta_0 + \theta/r_n} - m_{\theta_0}) - \theta^T \dot{m}_{\theta_0}, \ |\theta| \leq 1 \}.
\]

Then \( |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq 2\dot{m}_{\theta_0}(x)\|\theta_1 - \theta_2\| \)
for any \( \theta_1 \) and \( \theta_2 \) in the unit ball in \( \mathbb{R}^d \) by the Lipschitz condition. Further, set

\[
H_n = \sup_{|\theta| \leq 1} |r_n(m_{\theta_0 + \theta/r_n} - m_{\theta_0}) - \theta^T \dot{m}_{\theta_0}|.
\]

Then \( H_n \) is a cover and \( H_n \to 0 \) as \( n \to \infty \) by the definition of partial derivative. By Bounded Convergence Theorem, \( \delta_n := (PH_n^2)^{1/2} \to 0 \). Thus, by Corollary 8.2 and Example 8.14,

\[
E^* \sup_{|\theta| \leq 1} |G_n(m_\theta - m_{\theta_0})| \leq C \int_0^{\delta_n} \sqrt{\log(K\epsilon^{-d})} \, d\epsilon \to 0
\]
as \( n \to \infty \). The desired conclusion follows. \( \blacksquare \)

We need some preparation before proving Theorem 34.

Let \( P_n \) be the empirical distribution of a random sample of size \( n \) from a distribution \( P \), and, for every \( \theta \) in a metric space \( (\Theta, d) \), \( m_\theta(x) \) be a measurable function. Let \( \hat{\theta}_n \) (nearly) maximize the criterion function \( P_n m_\theta \). The number \( \theta_0 \) is the truth, that is, the maximizer of \( m_\theta \) over \( \theta \in \Theta \). Recall \( G_n = \sqrt{n}(P_n - P) \).

**Theorem 36 (Rate of Convergence)** Assume that for fixed constants \( C \) and \( \alpha > \beta \), for every \( n \) and every sufficiently small \( \delta > 0 \),

\[
\sup_{d(\theta, \theta_0) > \delta} P(m_\theta - m_{\theta_0}) \leq -C\delta^\alpha,
\]

\[
E^* \sup_{d(\theta, \theta_0) < \delta} |G_n(m_\theta - m_{\theta_0})| \leq C\delta^\beta.
\]

72
If the sequence \( \hat{\theta}_n \) satisfies \( \mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - \mathcal{O}(n^{\alpha/(2^3-2\alpha)}) \) and converges in outer probability to \( \theta_0 \), then \( n^{1/(2\alpha-2\beta)}d(\hat{\theta}_n, \theta_0) = \mathcal{O}^*_n(1) \).

**Proof.** Set \( r_n = n^{1/(2\alpha-2\beta)} \) and \( \mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - R_n \) with \( 0 \leq R_n = \mathcal{O}(n^{\alpha/(2^3-2\alpha)}) \). Partition \((0, \infty)\) by \( S_{j,n} = \{ \theta : 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j \} \) for all integers \( j \). If \( r_n d(\hat{\theta}_n, \theta_0) \geq 2^j \) for a given \( M \), then \( \hat{\theta}_n \) is in one of the shells \( S_{j,n} \) with \( j \geq M \). In that case, \( \sup_{\theta \in S_{j,n}}(\mathbb{P}_n m_{\theta} - \mathbb{P}_n m_{\theta_0}) \geq -R_n \). It follows that

\[
P^*(r_n d(\hat{\theta}_n, \theta_0) \geq 2^M) \leq \sum_{j \geq M, 2^j \leq r_n} P^* \left( \sup_{\theta \in S_{j,n}}(\mathbb{P}_n m_{\theta} - \mathbb{P}_n m_{\theta_0}) \geq -\frac{K}{r_n^{\alpha}} \right) \tag{9.3}
\]

The middle term on right goes to zero; the last term can be arbitrarily small by choosing large \( K \). We only need to show the sum is arbitrarily small when \( M \) is large enough as \( n \to \infty \). By the given condition

\[
\sup_{\theta \in S_{j,n}} P(m_{\theta} - m_{\theta_0}) \leq -C\frac{2^{j-1} \alpha}{r_n^{\alpha}}.
\]

For \( M \) such that \((1/2)C2^{(M-1)\alpha} \geq K\), by the fact that \( \sup_s (f_s + g_s) \leq \sup_s f_s + \sup_s g_s \),

\[
P^* \left( \sup_{\theta \in S_{j,n}}(\mathbb{P}_n m_{\theta} - \mathbb{P}_n m_{\theta_0}) \geq -\frac{K}{r_n^{\alpha}} \right) \leq P^* \left( \sup_{\theta \in S_{j,n}}\mathbb{P}_n (m_{\theta} - m_{\theta_0}) \geq C\sqrt{n}\frac{2^{(j-1)\alpha}}{2r_n^{\alpha}} \right).
\]

Therefore, the sum in (9.3) is bounded by

\[
\sum_{j \geq M, 2^j \leq r_n} P^* \left( \sup_{\theta \in S_{j,n}}|\mathbb{P}_n (m_{\theta} - m_{\theta_0})| \geq C\sqrt{n}\frac{2^{(j-1)\alpha}}{2r_n^{\alpha}} \right) \leq \sum_{j \geq M} \frac{(2^j/r_n^{\alpha})^2 2r_n^{\alpha}}{\sqrt{n}2^{(j-1)\alpha}}
\]

by Markov’s inequality and the definition of \( r_n \). The right hand side goes to zero as \( M \to \infty \).

\( \blacksquare \)

**Corollary 9.1** For each \( \theta \) in an open subset of Euclidean space, \( m_\theta(x) \) is a measurable function such that, there exists \( \breve{m}(x) \) with \( \breve{m}^{\alpha/\beta} < \infty \) satisfying

\[
|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \breve{m}(x) \|\theta_1 - \theta_2\|.
\]

Furthermore, assume

\[
P m_\theta = P m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0} (\theta - \theta_0) + o(\|\theta - \theta_0\|^2) \tag{9.4}
\]

with \( V_{\theta_0} \) nonsingular. If \( P_n m_{\hat{\theta}_n} \geq P_n m_{\theta_0} - \mathcal{O}(n^{-1}) \) and \( \hat{\theta}_n \to \theta_0 \), then \( \sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{O}(1) \).
Proof. (9.4) implies the first condition of Theorem 36 holds with $\alpha = 2$ since $V_{\theta_0}$ is nonsingular. Now we use Corollary 8.2 to the class of functions $\mathcal{F} = \{m_\theta - m_{\theta_0}; \|\theta - \theta_0\| < \delta\}$ to see the second condition is valid with $\beta = 1$. This class has envelope function $F = \delta \hat{m}$, while

$$E^2 \sup_{\|\theta - \theta_0\| < \delta} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \leq C \cdot \int_0^\delta \frac{\log K(1)}{d\epsilon} \left( \frac{\delta}{\epsilon} \right)^d d\epsilon = C_1 \delta$$

by (8.14) with $\mathrm{Diam}(\Theta) = \text{const} \cdot \delta^d$, where $C_1$ depends on $\|\hat{m}\|_{P,2}$.

Proof of Theorem 34. By Lemma 9.1,

$$\mathbb{G}_n \left( r_n (m_{\theta_0 + \hat{h}_n/r_n} - m_{\theta_0} - \hat{h}_n^T \hat{m}_{\theta_0}) \right) \overset{P}{\to} 0.$$  \hfill (9.5)

Expand $P(r_n (m_{\theta_0 + \hat{h}_n/r_n} - m_{\theta_0})$ by condition 9.1, we obtain

$$n \mathbb{P}_n (m_{\theta_0 + \hat{h}_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2} \hat{h}_n^T V_{\theta_0} \hat{h}_n + \hat{h}_n^T \mathbb{G}_n \hat{m}_{\theta_0} + o_P(1).$$

By Corollary 9.1, $\sqrt{n}(\hat{\theta}_n - \theta)$ is bounded in probability (same as tightness). Take $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta)$ and $\hat{h}_n = -V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0}$, we then obtain the Taylor’s expansions of $\mathbb{P}_n m_{\theta_0}$ and $\mathbb{P}_n m_{\theta_0} - V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0}$ as follows

$$n \mathbb{P}_n (m_{\theta_0} - m_{\theta_0}) = \frac{1}{2} \hat{h}_n^T V_{\theta_0} \hat{h}_n + \hat{h}_n^T \mathbb{G}_n \hat{m}_{\theta_0} + o_P(1),$$

$$n \mathbb{P}_n (m_{\theta_0} - V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0}/\sqrt{n} - m_{\theta_0}) = \frac{1}{2} \mathbb{G}_n \hat{m}_{\theta_0}^T V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0} + o_P(1),$$

where the second one is obtained through a bit algebra. By the definition of $\hat{\theta}_n$, the left hand side of the first equation is greater than that of the second one. So are he right hand sides. Take the difference and make it to a complete square, we then have

$$\frac{1}{2} (\hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0})^T V_{\theta_0} (\hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0}) + o_P(1) \geq 0.$$  

We know $V_{\theta_0}$ is strictly negative-definite, the quadratic form must converge to zero in probability. This is also the case $\|\hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0}\|$, that is, $\hat{h}_n = -V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_{\theta_0} + o_P(1)$.

■
10 Appendix

Let $\mathcal{A}$ be a collection of subsets of $\Omega$ and $\mathcal{B}$ is generated by $\mathcal{A}$, that is, $\mathcal{B} = \sigma(\mathcal{A})$. Let $P$ be a probability measure on $(\Omega, \mathcal{B})$.

**Lemma 10.1** Suppose $\mathcal{A}$ has the following property: (i) $\Omega \in \mathcal{A}$, (ii) $A^c \in \mathcal{A}$ if $A \in \mathcal{A}$, and (iii) $\bigcup_{i=1}^n A_i \in \mathcal{A}$ if $A_i \in \mathcal{A}$ for all $1 \leq i \leq m$. Then, for any $B \in \mathcal{B}$ and $\epsilon > 0$, there exists $A \in \mathcal{A}$ such that $P(B \Delta A) < \epsilon$.

**Proof.** Let $B'$ be the set of $B \in \mathcal{B}$ satisfying the conclusion. Obviously, $\mathcal{A} \subset B' \subset \mathcal{B}$. It is enough to verify that $B'$ is a $\sigma$-algebra.

It is easy to see that (i) $\Omega \in B'$; (ii) $B^c \in B'$ if $B \in B'$ since $A \Delta B = A^c \Delta B^c$. (iii) If $B_i \in B'$ for $i \geq 1$, there exist $A_i \in \mathcal{A}$ such that $P(B_i \Delta A_i) < \epsilon/2^i$ for all $i \geq 1$. Evidently, $\bigcup_{i=1}^n B_i \uparrow \bigcup_{i=1}^\infty B_i$ as $n \to \infty$. Therefore, there exists $n_0 < \infty$ such that $|P(\bigcup_{i=1}^\infty B_i) - P(\bigcup_{i=1}^{n_0} B_i)| < \epsilon/2$. It is easy to check that $(\bigcup_{i=1}^m B_i) \Delta (\bigcup_{i=1}^m A_i) \subset \bigcup_{i=1}^n (B_i \Delta A_i)$. Write $B = \bigcup_{i=1}^m B_i$ and $\tilde{B} = \bigcup_{i=1}^m B_i$ and $A = \bigcup_{i=1}^m A_i$. Then $A \in \mathcal{A}$. Note $B \Delta A \subset (B \setminus \tilde{B}) \cup (\tilde{B} \Delta A)$.

The above facts show that $P(B \Delta A) < \epsilon$. Thus $B'$ is a $\sigma$-algebra. $\blacksquare$

**Lemma 10.2** Let $X_1, X_2, \ldots, X_m, m \geq 1$ be random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$. Let $f(x_1, \ldots, x_m)$ be a real measurable function with $E|f(X_1, \ldots, X_m)|^p < \infty$ for some $p \geq 1$. Then there exists $\{f_n(X_1, \ldots, X_m); n \geq 1\}$ such that

(i) $f_n(X_1, \ldots, X_m) \to f(X_1, \ldots, X_m)$ a.s.

(ii) $f_n(X_1, \ldots, X_m) \to f(X_1, \ldots, X_m)$ in $L^p(\Omega, \mathcal{F}, \mathcal{P})$.

(iii) For each $n \geq 1$, $f_n(X_1, \ldots, X_m) = \sum_{i=1}^{k_m} c_i g_i(X_1) \cdots g_m(X_m)$ for some $k_m < \infty$, constants $c_i$, and $g_{i,j}(X_j) = I_{A_{i,j}}(X_j)$ for some sets $A_{i,j} \in \mathcal{B}(\mathbb{R})$, and all $1 \leq i \leq k_m$ and $1 \leq j \leq m$.

**Proof.** To save notation, we write $f = f(x_1, \ldots, x_m)$. Since $E|f I(|f| \leq C) - f|^p \to 0$ as $C \to \infty$, choose $C_k$ such that $E|f I(|f| \leq C_k) - f|^p \leq 1/k^2$ for $k = 1, 2, \ldots$. We will show that there exists function $g_k = g_k(x_1, \ldots, x_k)$ of the form in (iii) such that $E|f I(|f| \leq C_k) - g_k|^p \leq 1/k^2$ for all $k \geq 1$. Thus, $E|f - g_k|^p < 2/k^2$ for all $k \geq 1$. The assertion (ii) then follows. Also, it implies $E(\sum_{k \geq 1} |f - g_k|^p) < \infty$, then $\sum_{k \geq 1} |f - g_k|^p < \infty$ a.s. we obtain (i). Therefore, to prove this lemma, we assume w.l.o.g. that $f$ is bounded and we need to show that there exist $f_n = f_n(x_1, \ldots, x_m); n \geq 1$ of the form in (iii) such that

$$E|f - f_n|^p \leq \frac{1}{n^2} \quad (10.6)$$

for all $n \geq 1$. 75
Since $f$ is bounded, for any $n \geq 1$, there exists $h_n$ such that
\[
\sup_{x \in \mathbb{R}^m} |f(x) - h_n(x)| < \frac{1}{2n^2},
\] where $h_n$ is a simple function, i.e., $h_n(x) = \sum_{i=1}^{k_n} c_i I\{x \in B_i\}$ for some $k_n < \infty$, constants $c_i$'s and $B_i \in \mathcal{B}(\mathbb{R}^n)$.

Now set $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ and $\mu$ be the probability measure of $X$ under probability $P$. Let $\mathcal{A}$ be the set of all finite unions of sets in $\mathcal{A}_1 := \{\prod_{i=1}^{m} A_i \in \mathcal{B}(\mathbb{R}^n); \ A_i \in \mathcal{B}(\mathbb{R})\}$. By the construction of $\mathcal{B}(\mathbb{R}^n)$, we know that $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{A})$. It is not difficult to verify that $\mathcal{A}$ satisfies the conditions in Lemma 10.1. Thus there exist $E_i \in \mathcal{A}$ such that
\[
\int |I_{E_i}(x) - I_{E}(x)|^p \, d\mu = \mu(B_i \Delta E_i) < \frac{1}{(2cn^2)^p} \text{ and } E_i = \bigcup_{j=1}^{k_i} \prod_{m} A_{i,j,l}
\] where $c = 1 + \sum_{i=1}^{k_n} |c_i|$ and $A_{i,j,l} \in \mathcal{B}(\mathbb{R})$ for all $i, j$ and $l$. Now, since $\| \cdot \|_p$ is a norm, we have
\[
\|h_n(X) - \sum_{i=1}^{k_n} c_i I_{E_i}(X)\|_p \leq \sum_{i=1}^{k_n} |c_i| \cdot \|I_{E_i}(X) - I_{B_i}(X)\|_p \leq \frac{1}{2n^2}. \tag{10.8}
\]
Observe that $\nu(E) := I_{E}(X)$ is a measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Note that the intersection of any finite product sets $\prod_{i=1}^{m} A_{i,j,l}$ is still in $\mathcal{A}_1$. By the inclusion-exclusion formula, $I_{E_i}(X)$ is a finite linear combination of $I_{F}(X)$ where $F \in \mathcal{A}_1$. Thus, $f_n(X) := \sum_{i=1}^{k_n} c_i I_{E_i}(X)$ is of the form in (iii). Now, by (10.7) and (10.8)
\[
\|f(X) - f_n(X)\|_p \leq \|f(X) - h_n(X)\|_p + \|h_n(X) - f_n(X)\|_p \leq \frac{1}{n^2}.
\]
Thus, (10.6) follows. $\blacksquare$

**Lemma 10.3** Let $\xi_1, \ldots, \xi_k$ be random variables, and $\phi(x), x \in \mathbb{R}^n$ be a measurable function with $E[\phi(\xi_1, \ldots, \xi_k)] < \infty$. Let $\mathcal{F}$ and $\mathcal{G}$ be two $\sigma$-algebras. Suppose
\[
E(f_1(\xi_1) \cdots f_k(\xi_k)|\mathcal{F}) = E(f_1(\xi_1) \cdots f_k(\xi_k)|\mathcal{G}) \text{ a.s.} \tag{10.9}
\]
holds for all bounded, measurable functions $f_i : \mathbb{R} \to \mathbb{R}$, $1 \leq i \leq n$. Then $E(\phi(\xi_1, \ldots, \xi_k)|\mathcal{F}) = E(\phi(\xi_1, \ldots, \xi_k)|\mathcal{G})$ almost surely.

**Proof.** From (10.9), we know that the desired conclusion holds for $\phi$ that is a finite linear combination of $f_1 \cdots f_k$'s. By Lemma 10.2, there exist such functions $\phi_n; n \geq 1$, such that
\[ E[\phi_n(\xi_1, \cdots, \xi_k) - \phi(\xi_1, \cdots, \xi_k)] \rightarrow 0 \text{ as } n \rightarrow \infty. \] Therefore,
\[
E[E(\phi_n(\xi_1, \cdots, \xi_k)|\mathcal{F}) - E(\phi(\xi_1, \cdots, \xi_k)|\mathcal{F})] \\
\leq E[E(\phi_n(\xi_1, \cdots, \xi_k) - \phi(\xi_1, \cdots, \xi_k)] ightarrow 0.
\]

We know that \( E(\phi_n(\xi_1, \cdots, \xi_k)|\mathcal{F}) = E(\phi_n(\xi_1, \cdots, \xi_k)|\mathcal{G}) \) for all \( n \geq 1 \), the conclusion follows since the \( L^1 \)-limit is unique up to a set of measure zero. \( \square \)

**Proposition 10.1** Let \( X_1, X_2, \cdots \) be a Markov chain. Let \( 1 \leq l \leq m \leq n - 1 \). Assume \( \phi(x), x \in \mathbb{R}^{n-m} \) is a measurable function such that \( E[\phi(X_{m+1}, \cdots, X_n)] < \infty \). Then
\[
E(\phi(X_{m+1}, \cdots, X_n)|X_m, \cdots, X_l) = E(\phi(X_{m+1}, \cdots, X_n)|X_m).
\]

**Proof.** Let \( p = n - m \). By Lemma 10.3, we only need to prove the lemma for \( \phi(x) = \phi_1(x_1) \cdots \phi(x_p) \) with \( x = (x_1, \cdots, x_p) \) and \( \phi_i \)'s being bounded functions. We do this by induction.

If \( n = m+1 \), by the Markov property, \( Y := E(\phi_1(X_{m+1})|X_m, \cdots, X_1) = E(\phi_1(X_{m+1})|X_m). \) Review the property that \( E(E(Z|\mathcal{F})|\mathcal{G}) = E(E(Z|\mathcal{G})|\mathcal{F}) = E(Z|\mathcal{F}) \) if \( \mathcal{F} \subset \mathcal{G} \). We know that
\[
E(\phi_1(X_{m+1})|X_m, \cdots, X_l) = E(Y|X_m, \cdots, X_l) = E(\phi_1(X_{m+1})|X_m)
\]
since \( Y = E(\phi_1(X_{m+1})|X_m) \) is \( \sigma(X_m, \cdots, X_l) \)-measurable.

Now assume \( n \geq m + 2 \). Using the same argument to obtain,
\[
E\{\phi_1(X_{m+1}) \cdots \phi_{p+1}(X_{n+1})|X_m, \cdots, X_l)\}
\]
\[
= E\{\phi_1(X_{m+1}) \cdots \phi_{p-1}(X_{n-1})\bar{\phi}_p(X_n)|X_m, \cdots, X_l\}
\]
\[
= E\{\phi_1(X_{m+1}) \cdots \phi_{p-1}(X_{n-1})\tilde{\phi}_p(X_n)|X_m\}
\]
where \( \bar{\phi}_p(X_n) = \phi_p(X_n) \cdot E(\phi_{p+1}(X_{n+1})|X_n) \), and the last step is by induction. Since \( \phi_1(X_{m+1}) \cdots \phi_{p-1}(X_{n-1})\tilde{\phi}_p(X_n) \) is equal to \( E(\phi_1(X_{m+1}) \cdots \phi_{p+1}(X_{n+1})|X_m, \cdots, X_1) \). The conclusion follows. \( \square \)

The following property says that a Markov chain has a reverse property.

**Proposition 10.2** Let \( X_1, X_2, \cdots \) be a Markov chain. Let \( 1 \leq m < n \). Assume \( \phi(x), x \in \mathbb{R}^m \) is a measurable function such that \( E[\phi(X_1, \cdots, X_m)] < \infty \). Then
\[
E(\phi(X_1, \cdots, X_m)|X_{m+1}, \cdots, X_n) = E(\phi(X_1, \cdots, X_m)|X_{m+1}).
\]
Proof. By lemma 10.3, we assume, without loss of generality, \( \phi(x) \) is a bounded function. From the definition of conditional expectation, to prove the proposition, it suffices to show that

\[
E\{\phi(X_1, \cdots, X_m)g(X_{m+1}, \cdots, X_n)\} \\
= E\{E(\phi(X_1, \cdots, X_m)|X_{m+1}) \cdot g(X_{m+1}, \cdots, X_n)\} \\
\tag{10.10}
\]

for any bounded, measurable function \( g(x), x \in \mathbb{R}^n \). By Lemma 10.3 again, we only need to prove (10.10) for \( g(x) = g_1(x_1) \cdots g_p(x_p) \) where \( p = n - m \). Now

\[
E\{E(\phi(X_1, \cdots, X_m)|X_{m+1}) \cdot g_1(X_{m+1}) \cdots g_p(X_n)\} \\
= E\{Z \cdot g_2(X_{m+2}) \cdots g_p(X_n)\}
\]

where \( Z = E(\phi(X_1, \cdots, X_m)g_1(X_{m+1})|X_{m+1}) \). Use the fact \( E(E(\xi|\mathcal{F}) \cdot \eta) = E(\xi \cdot E(\eta|\mathcal{F})) \) for any \( \xi, \eta \) and \( \sigma \)-algebra \( \mathcal{F} \) to obtain that

\[
E\{Z \cdot g_2(X_{m+2}) \cdots g_p(X_n)\} = E(\phi(X_1, \cdots, X_m)g_1(X_{m+1}) \cdot E(\eta|X_{m+1}))
\]

where \( \eta = g_2(X_{m+2}) \cdots g_p(X_n) \). By Proposition 10.1, \( E(\eta|X_{m+1}) = E(\eta|X_{m+1}, \cdots, X_1) \). Therefore the above is identical to

\[
E\{E(\phi(X_1, \cdots, X_m)|X_{m+1}) \cdot \eta|X_{m+1}, \cdots, X_1)\} \\
= E\{\phi(X_1, \cdots, X_m)\eta|X_{m+1}, \cdots, X_1)\}
\]

since \( g(x) = g_1(x_1) \cdots g_p(x_p) \). \( \blacksquare \)

The following proposition says that a Markov chain has the property: given the present, the past and future are independent.

**Proposition 10.3** Let \( 1 \leq k < l \leq m \) be such that \( l - k \geq 2 \). Let \( \phi(x), x \in \mathbb{R}^k \) and \( \psi(x), x \in \mathbb{R}^{m-l+1} \) be two measurable functions satisfying \( E(\phi(X_1, \cdots, X_l)^2) < \infty \) and \( E(\psi(X_1, \cdots, X_m)^2) < \infty \). Then

\[
E\{\phi(X_1, \cdots, X_k) \cdot \psi(X_l, \cdots, X_m)|X_{k+1}, X_{l-1}\} \\
= E\{\phi(X_1, \cdots, X_k)|X_{k+1}, X_{l-1}\} \cdot E\{\psi(X_l, \cdots, X_m)|X_{k+1}, X_{l-1}\}.
\]

**Proof.** For saving notation, we write \( \phi = \phi(X_1, \cdots, X_k) \) and \( \psi = \psi(X_l, \cdots, X_m) \). Then

\[
E(\phi \psi|X_{k+1}, X_{l-1}) = E\{E(\phi \psi|X_1, \cdots X_{l-1})|X_{k+1}, X_{l-1}\} \\
= E\{\phi \cdot E(\psi|X_{l-1})|X_{k+1}, X_{l-1}\}
\]

78
by Proposition 10.1 and that $\phi$ is $\sigma(X_1, \ldots, X_{l-1})$ measurable. The last term is also equal to $E(\psi|X_{l-1}) \cdot E\{\phi|X_{k+1}, X_{l-1}\}$. The conclusion follows since $E(\psi|X_{l-1}) = E(\psi|X_k, X_{l-1})$. 
\[\blacksquare\]