# Theory of Statistics II (Fall 2016)

J.P.Kim

Dept. of Statistics

Finally modified at November 23, 2016

# Preface & Disclaimer

This note is a summary of the lecture Theory of Statistics II (326.522) held at Seoul National University, Fall 2016. Lecturer was B.U.Park, and the note was summarized by J.P.Kim, who is a Ph.D student. There are few textbooks and references in this course. Contents and corresponding references are following.

- Asymptotic Approximations. Reference: *Mathematical Statistics: Basic ideas and selected topics, Vol. I., 2nd edition, P.Bickel & K.Doksum, 2007.*
- Weak Convergence. Reference: *Convergence of Probability Measures, P.Billingsley, 1999.*
- Nonparametric Density Estimation. Reference:

Lecture notes are available at `stat.snu.ac.kr/theostat`. Also I referred to following books when I write this note. The list would be updated continuously.

- *Probability: Theory and Examples, R.Durrett*
- *Mathematical Statistics (in Korean), W.C.Kim*
- *Introduction to Nonparametric Regression, K.Takezawa*

If you want to correct typo or mistakes, please contact to: `joonpyokim@snu.ac.kr`

# Chapter 1

# Asymptotic Approximations

## 1.1 Consistency

### 1.1.1 Preliminary for the chapter

**Definition 1.1.1** (Notations)**.** *Let $\Theta$ be a parameter space. Then we consider a 'random variable' $X$ on the probability space $(\Omega, \mathcal{F}, P_\theta)$ which is a function*

$$X : (\Omega, \mathcal{F}, P_\theta) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_\theta^X),$$

*where $P_\theta^X := P_\theta \circ X^{-1}$. Note that $P_\theta$ is a probability measure from the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$. For the convenience, now we omit the explanation of fundamental setting.*

**Definition 1.1.2** (Convergence)**.** *Let $\{X_n\}$ be a sequence of random variables.*

1. $X_n \xrightarrow[n\to\infty]{a.s} X$ *if* $P\left(\lim_{n\to\infty} X_n = X\right) = 1 \Leftrightarrow P(|X_n - X| > \epsilon \ i.o.) = 0 \ \forall \epsilon > 0$
$$\Leftrightarrow \lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty} (|X_n - X| > \epsilon)\right) = 0 \ \forall \epsilon > 0$$
2. $X_n \xrightarrow[n\to\infty]{P} X$ *if* $\forall \epsilon > 0 \ P(|X_n - X| > \epsilon) \to 0$ *as* $n \to \infty$.

**Proposition 1.1.3.** *$X_n \xrightarrow[n\to\infty]{P} X$ if and only if for any subsequence $\{n_k\} \subseteq \{n\}$ there is a further subsequence $\{n_{k_j}\} \subseteq \{n_k\}$ such that $X_{n_{k_j}} \xrightarrow[j\to\infty]{a.s.} X$.*

*Proof.* Durrett, p.65. □

**Definition 1.1.4** (Consistency)**.** *$\hat{q}_n = q_n(X_1, \cdots, X_n)$ is consistent estimator of $q(\theta)$ if*

$$\hat{q}_n \xrightarrow[n\to\infty]{P_\theta} q(\theta)$$

*for any $\theta \in \Theta$. (We don't know what is the true parameter.)*

**Remark 1.1.5.** There are some tools to obtain consistency.

1. $Var(Z_n) \to 0$, $EZ_n \to \mu$ as $n \to \infty \Rightarrow Z_n \xrightarrow[n\to\infty]{P} \mu$.

$$\begin{aligned} \because P(|Z_n - \mu| > \epsilon) &\le P(|Z_n - EZ_n| + |EZ_n - \mu| > \epsilon) \\ &\le P(|Z_n - EZ_n| > \epsilon/2) + \underbrace{P(|EZ_n - \mu| > \epsilon/2)}_{=0 \text{ for sufficiently large } n} \\ &\le \frac{4}{\epsilon^2} Var(Z_n) \to 0 \end{aligned}$$

2. WLLN: $X_1, \cdots, X_n$: i.i.d. and $E|X_1| < \infty \Rightarrow \overline{X}_n \xrightarrow[n\to\infty]{P} EX_1$.

3. If $Z_n \xrightarrow[n\to\infty]{P} Z$ and $g$ is continuous on the support of $Z$, then $g(Z_n) \xrightarrow[n\to\infty]{P} g(Z)$. Note that uniform convergence of $g$ implies this directly, and for the general case, we can use Proposition 1.1.3.

4. Followings are the corollary of 3. Or, we can prove it directly. Suppose that $Y_n \xrightarrow[n\to\infty]{P} Y$ and $Z_n \xrightarrow[n\to\infty]{P} Z$. Then,

   (a) $Y_n + Z_n \xrightarrow[n\to\infty]{P} Y + Z$.

   (b) $Y_n Z_n \xrightarrow[n\to\infty]{P} YZ$.

   (c) $Y_n/Z_n \xrightarrow[n\to\infty]{P} Y/Z$ provided that $Z \neq 0$ $P$-a.s..

   *Proof.* (b) Note that $|Y_nZ_n - YZ| \le |Y_n||Z_n - Z| + |Z||Y_n - Y| \le |Y_n - Y||Z_n - Z| + |Y||Z_n - Z| + |Z||Y_n - Y|$. Now for any $\eta > 0$ there exists $M > 0$ such that $P(|Y| > M) \le \eta$ and $P(|Z| > M) \le \eta$. Now,

   $$\begin{aligned} P(|Y_nZ_n - YZ| > \epsilon) &\le P(|Y_n||Z_n - Z| > \epsilon/2) + P(|Z||Y_n - Y| > \epsilon/2) \\ &\le P(|Y_n - Y||Z_n - Z| > \epsilon/4) + P(|Y||Z_n - Z| > \epsilon/4) + P(|Z||Y_n - Y| > \epsilon/2) \end{aligned}$$

   and note that $P(|Y||Z_n - Z| > \epsilon/4) = P(|Y||Z_n - Z| > \epsilon/4, |Y| > M) + P(|Y||Z_n - Z| > \epsilon/4, |Y| \le M) \le \eta + P(|Z_n - Z| \ge \epsilon/4M)$. Thus

   $$\limsup_{n\to\infty} P(|Y||Z_n - Z| > \epsilon/4) \le \eta$$

and similarly

$$\limsup_{n\to\infty} P(|Z||Y_n - Y| > \epsilon/2) \le \eta.$$

Now, since

$$\begin{aligned} P(|Y_n - Y||Z_n - Z| > \epsilon/4) &= P(|Y_n - Y||Z_n - Z| > \epsilon/4, |Y_n - Y| > \sqrt{\epsilon/4}) \\ &\quad + P(|Y_n - Y||Z_n - Z| > \epsilon/4, |Y_n - Y| \le \sqrt{\epsilon/4}) \\ &\le P(|Y_n - Y| > \sqrt{\epsilon/4}) + P(|Z_n - Z| \ge \sqrt{\epsilon/4}) \to 0 \end{aligned}$$

as $n \to \infty$, we get

$$\limsup_{n\to\infty} P(|Y_n Z_n - YZ| > \epsilon) \le 2\eta.$$

Finally, since $\eta > 0$ was arbitrary, we get the result.

(c) By (b), it's sufficient to show that $Z_n^{-1} \xrightarrow[n\to\infty]{P} Z^{-1}$. Since $P(Z = 0) = 0$, for any $\eta > 0$ there exists $\delta > 0$ such that $P(|Z| \le \delta) \le \eta$. (If not, $\exists \eta > 0$ such that $\forall \delta > 0$ $P(|Z| \le \delta) > \eta$. Then by continuity of measure, $P(\bigcup_{\delta>0}(|Z| \le \delta)) = P(Z = 0) \ge \eta > 0$. Contradiction.) Thus

$$\begin{aligned} P\left(\left|\frac{1}{Z_n} - \frac{1}{Z}\right| > \epsilon\right) &= P\left(\frac{|Z_n - Z|}{|Z_n Z|} > \epsilon\right) \\ &\le P\left(\frac{|Z_n - Z|}{|Z|(|Z| - |Z_n - Z|)} > \epsilon\right) + P(|Z| < |Z_n - Z|) \\ &\le \underbrace{P\left(\frac{|Z_n - Z|}{|Z|(|Z| - |Z_n - Z|)} > \epsilon, |Z| > \delta, |Z_n - Z| \le \delta/2\right)}_{\le P(|Z_n - Z| > \frac{\delta^2}{2}\epsilon) \xrightarrow[n\to\infty]{} 0} \\ &\quad + \underbrace{P(|Z| \le \delta)}_{\le \eta} + \underbrace{P(|Z_n - Z| > \delta/2)}_{\xrightarrow[n\to\infty]{} 0} \\ &\quad + \underbrace{P(|Z| < |Z_n - Z|, |Z_n - Z| > \delta)}_{\le P(|Z_n - Z| > \delta) \xrightarrow[n\to\infty]{} 0} + \underbrace{P(|Z| < |Z_n - Z|, |Z_n - Z| \le \delta)}_{\le P(|Z| \le \delta) \le \eta} \end{aligned}$$

and hence

$$\limsup_{n\to\infty} P\left(\left|\frac{1}{Z_n} - \frac{1}{Z}\right| > \epsilon\right) \le 2\eta$$

holds. Note that $\eta > 0$ was arbitrary. □

**Definition 1.1.6** (Probabilistic $O$-notation)**.** *Let $X_n$ be a sequence of r.v.'s.*

1. $X_n = O_p(1)$ *if* $\lim_{c\to\infty} \sup_n P(|X_n| > c) = 0 \Leftrightarrow \lim_{c\to\infty} \limsup_{n\to\infty} P(|X_n| > c) = 0$. *("Bounded in probability")*

2. $X_n = o_p(1)$ *if* $X_n \xrightarrow[n\to\infty]{P} 0$.

3. $X_n = O_p(a_n)$ *if* $X_n/a_n = O_p(1)$, *and* $X_n = o_p(a_n)$ *if* $X_n/a_n = o_p(1)$.

**Proposition 1.1.7.** *If* $X_n \xrightarrow[n\to\infty]{d} X$ *for some* $X$, *then* $X_n = O_p(1)$.

*Proof.* For given $\epsilon > 0$, there exists $c$ such that $P(|X| > c) < \epsilon/2$. For such $c$, $P(|X_n| > c) \to P(|X| > c)$, so $\exists N$ s.t.

$$\sup_{n>N} |P(|X_n| > c) - P(|X| > c)| < \frac{\epsilon}{2}.$$

Thus $\sup_{n>N} P(|X_n| > c) < \epsilon$. For $n = 1, 2, \cdots, N$, there exists $c_n$ such that $P(|X_n| > c_n) < \epsilon$, and letting $c^* = \max(c_1, \cdots, c_N, c)$, we get $\sup_n P(|X_n| > c^*) < \epsilon$. □

**Example 1.1.8** (Simple Linear Regression)**.** Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} (0, \sigma^2)$. Also assume that $x_1, \cdots, x_n$ are known and not all equal. Note that

$$\hat{\beta}_{1,n} = \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Since $E(\hat{\beta}_{1,n}) = \beta_1$ and $Var(\hat{\beta}_{1,n}) = \sigma^2/S_{xx}$, we obtain consistency

$$\hat{\beta}_{1,n} \xrightarrow[n\to\infty]{P_{\beta,\sigma^2}} \beta_1$$

provided that $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2 \to \infty$ as $n \to \infty$.

**Example 1.1.9** (Sample correlation coefficient)**.** Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be random sample from the population

$$EX_1 = \mu_1,\ EY_1 = \mu_2,\ Var(X_1) = \sigma_1^2 > 0,\ Var(Y_1) = \sigma_2^2 > 0,\ \text{and } Corr(X_1, Y_1) = \rho.$$

Then by WLLN we get

$$(\overline{X}, \overline{Y}, \overline{X^2}, \overline{Y^2}, \overline{XY}) \xrightarrow[n\to\infty]{P} (EX_1, EY_1, EX_1^2, EY_1^2, EX_1Y_1).$$

Since the function

$$g(u_1, u_2, u_3, u_4, u_5) = \frac{u_5 - u_1 u_2}{\sqrt{u_3 - u_1^2}\sqrt{u_4 - u_2^2}}$$

is continuous *at* $(EX_1, EY_1, EX_1^2, EY_1^2, EX_1Y_1)$, we get

$$\hat{\rho}_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\overline{X^2} - \overline{X}^2}\sqrt{\overline{Y^2} - \overline{Y}^2}} \xrightarrow[n\to\infty]{P} \rho.$$

**Remark 1.1.10.** Note that, if $X_n \xrightarrow[n\to\infty]{P} c$ where $c$ is a constant, then continuity of $g(x)$ at $x = c$ is sufficient for consistency $g(X_n) \xrightarrow[n\to\infty]{P} g(c)$. It is trivial from the definition of continuity.

**Example 1.1.11.** Let $X_1, \cdots, X_n$ be a random sample from a population with cdf $F$. Then we use an *empirical distribution function*

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

for estimation of $F$. Then by WLLN, for each $x$, $\hat{F}_n(x)$ is consistent estimator for $F(x)$,

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{P} F(x).$$

**Remark 1.1.12.** Note that in this case, we can say more strong result, which is known as *Glivenko-Cantelli theorem*:

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow[n\to\infty]{P} 0.$$

Sketch of proof is given here. Since $\hat{F}_n$ and $F$ are nondecreasing and bounded, we can partition $[0, 1]$, which is a range of both functions, into finite number of intervals, and then each interval has a well-defined inverse image which is an interval. For whole proof, see Durrett, p.76.

### 1.1.2 FSE and MLE in Exponential Families

**FSE**

Recall that FSE of $\nu(F)$ is defined as $\nu(\hat{F}_n)$. One example of FSE is MME: to estimate $EX^j =: \nu_j(F) =: \int x^j dF(x)$, we use

$$\hat{m}_j = \nu_j(\hat{F}_n) = \int x^j d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

By WLLN we have $(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)^\top \xrightarrow[n\to\infty]{P} (m_1, m_2, \cdots, m_k)^\top$ where $m_j = EX^j$, so we can obtain consistency of MME easily.

**Proposition 1.1.13.** *Let $q = h(m_1, m_2, \cdots, m_k)$ be a parameter of interest where $m_j$'s are*

*population moments. Then for MME*

$$\hat{q}_n = h(\hat{m}_1, \cdots, \hat{m}_k)$$

*based on a random sample* $X_1, \cdots, X_n$,

$$\hat{q}_n \xrightarrow[n\to\infty]{P} q$$

*holds, provided that* $h$ *is continuous at* $(m_1, \cdots, m_k)^\top$.

We can do similar work in FSE $\nu(F)$. Note that in here, $\nu$ is a functional, so we may define a continuity of functional. We may use sup norm as a metric in the space of distribution functions.

**Definition 1.1.14.** *Let* $\mathcal{F}$ *be a space of distribution functions. In this space, we give the norm* $\|\cdot\|$ *as a sup norm*

$$\|F\| = \sup_x |F(x)|.$$

*Then metric is given as*

$$\|F - G\| = \sup_x |F(x) - G(x)|.$$

*Also, we say that a functional* $\nu : \mathcal{F} \to \mathbb{R}$ *is continuous at* $F$ *if for any* $\epsilon > 0$ *there exists* $\delta > 0$ *such that*

$$\|G - F\| < \delta \Rightarrow |\nu(G) - \nu(F)| < \epsilon.$$

**Remark 1.1.15.** Note that since $\|\hat{F}_n - F\| \to 0$ as $n \to \infty$ from Glivenko-Cantelli theorem, we get consistency of FSE

$$\nu(\hat{F}_n) \xrightarrow[n\to\infty]{P} \nu(F)$$

provided that $\nu$ is continuous at $F$. In many cases, showing continuity may be difficult problem.

**Example 1.1.16** (Best Linear Predictor)**.** Let $X_1, \cdots, X_n$ be $k$-dimensional i.i.d. r.v.'s, and $Y_1, \cdots, Y_n$ be i.i.d. 1-dim random variable. Then we know that

$$BLP(x) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1),$$

where

$$E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } Var\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Thus for sample variance

$$S_{11} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\top$$

$$S_{12} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})^\top = S_{21}^\top$$

$$S_{22} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

we obtain FSE for BLP,

$$\widehat{BLP}^{FSE}(x) = \overline{Y} + S_{21}S_{11}^{-1}(x - \overline{X}).$$

Note that it is the same as simple linear regression line. Detail is given in next proposition.

**Proposition 1.1.17.**

*(a) Solution of minimizing problem*

$$(\beta_0^*, \boldsymbol{\beta}_1^*)^\top = \underset{\beta_0, \boldsymbol{\beta}_1}{\arg\min} E(Y - \beta_0 - \boldsymbol{\beta}_1^\top X)^2$$

*is*

$$BLP(x) := \beta_0^* + \boldsymbol{\beta}_1^{*\top} x = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1).$$

*(b) For $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^\top$ and design matrix $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1)$ where $\boldsymbol{X}_1 = (X_1, \cdots, X_n)^\top$, LSE is*

$$\hat{\boldsymbol{\beta}}_1 = S_{11}^{-1}S_{12} \text{ and } \hat{\beta}_0 = \overline{Y} - \overline{X}^\top \hat{\boldsymbol{\beta}}_1.$$

*Proof.* (a) Two approaches are given. First one is direct proof: It is clear because of

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}_1^\top X)^2 &= E[(Y - \mu_2) - \boldsymbol{\beta}_1^\top (X - \mu_1)]^2 + [\mu_2 - \beta_0 - \boldsymbol{\beta}_1^\top \mu_1]^2 \\ &= \Sigma_{22} - 2\boldsymbol{\beta}_1^\top \Sigma_{12} + \boldsymbol{\beta}_1^\top \Sigma_{11} \boldsymbol{\beta}_1 + [\beta_0 - (\mu_2 - \boldsymbol{\beta}_1^\top \mu_1)]^2. \end{aligned}$$

Second approach uses projection in $\mathcal{L}^2$ space. For convenience, suppose $EX = 0$ and $EY = 0$. Then $(\beta_0^*, \boldsymbol{\beta}_1^*)^\top$ should satisfy

$$\langle \beta_0 + \boldsymbol{\beta}_1^\top X, Y - \beta_0^* - \boldsymbol{\beta}_1^{*\top} X \rangle = 0 \ \forall \beta_0, \beta_1.$$

It yields that

$$\beta_0^* = 0, \ \boldsymbol{\beta}_1^* = \left(E(XX^\top)\right)^{-1} E(XY).$$

(b) $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1$ should satisfy $\mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y}|\mathcal{C}(\boldsymbol{X}))$. For $\mathcal{X}_1 = \boldsymbol{X}_1 - \Pi(\boldsymbol{X}_1|\mathcal{C}(\mathbf{1})) = \boldsymbol{X}_1 - \mathbf{1}\overline{X}^\top$,

$$\mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \mathbf{1}\left(\hat{\beta}_0 + \frac{\mathbf{1}^\top \boldsymbol{X}_1}{n}\hat{\boldsymbol{\beta}}_1\right) + \mathcal{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y}|\mathcal{C}(\mathbf{1})) + \Pi(\boldsymbol{Y}|\mathcal{C}(\mathbf{X_1}))$$

we get

$$\hat{\beta}_0 = \overline{Y} - \overline{X}^\top\hat{\boldsymbol{\beta}}_1 \text{ and } \hat{\boldsymbol{\beta}}_1 = (\mathcal{X}_1^\top\mathcal{X}_1)^{-1}\mathcal{X}_1^\top\boldsymbol{Y}.$$

Now $\mathcal{X}_1^\top\mathcal{X}_1/n = S_{11}$ and $\mathcal{X}_1^\top\boldsymbol{Y}/n = S_{12}$ ends the proof. □

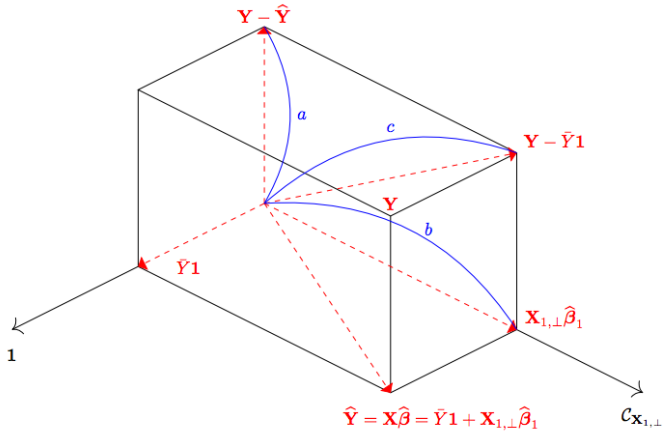


Figure 1.1: Regression with intercept. Image from Lecture Note.

**Example 1.1.18** (Multiple Correlation Coefficient)**.** We define a *multiple correlation coefficient (MCC)* as

$$\rho = \max_{\beta_0,\boldsymbol{\beta}_1} \mathrm{Corr}(Y, \beta_0 + \boldsymbol{\beta}_1^\top X)$$

and sample MCC is

$$\hat{\rho}_n = \max_{\beta_0,\boldsymbol{\beta}_1} \widehat{\mathrm{Corr}}(Y, \beta_0 + \boldsymbol{\beta}_1^\top X).$$

Note that,

$$\begin{aligned}\mathrm{Corr}(Y, \beta_0 + \boldsymbol{\beta}_1^\top X) &= \mathrm{Corr}(Y - \mu_2, \boldsymbol{\beta}_1^\top(X - \mu_1)) \\ &= \frac{\Sigma_{21}\boldsymbol{\beta}_1}{\sqrt{\Sigma_{22}}\sqrt{\boldsymbol{\beta}_1^\top\Sigma_{11}\boldsymbol{\beta}_1}} \\ &= \frac{(\Sigma_{11}^{-1/2}\Sigma_{12})^\top(\Sigma_{11}^{1/2}\boldsymbol{\beta}_1)}{\sqrt{\Sigma_{22}}\sqrt{\boldsymbol{\beta}_1^\top\Sigma_{11}\boldsymbol{\beta}_1}}\end{aligned}$$

$$\leq \sqrt{\frac{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}{\Sigma_{22}}}$$

holds by Cauchy-Schwarz inequality, and equality holds when $\boldsymbol{\beta}_1 = \Sigma_{11}^{-1}\Sigma_{12}$. Thus population MCC is obtained as

$$\rho = \sqrt{\frac{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}{\Sigma_{22}}}.$$

Meanwhile, sample correlation is obtained as

$$\widehat{\text{Corr}}(\boldsymbol{Y}, \beta_0 + \boldsymbol{\beta}_1^\top \boldsymbol{X}) = \frac{\langle \boldsymbol{Y} - \overline{Y}\mathbf{1}, (\boldsymbol{X} - \mathbf{1}\overline{X}^\top)\boldsymbol{\beta}_1 \rangle}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\| \|(\boldsymbol{X} - \mathbf{1}\overline{X}^\top)\boldsymbol{\beta}_1\|}$$

so it is the cosine of the angle between the two rays, $\boldsymbol{Y} - \overline{Y}\mathbf{1}$ and $\mathcal{X}_1\boldsymbol{\beta}_1$. Its maximal value is attaiend by $\mathcal{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y} - \overline{Y}\mathbf{1}|\mathcal{C}(\mathcal{X}_1))$. (Or, one may use Cauchy-Schwarz,

$$\widehat{\text{Corr}}(\boldsymbol{Y}, \beta_0 + \boldsymbol{\beta}_1^\top \boldsymbol{X}) = \frac{y^\top \mathcal{X}_1(\mathcal{X}_1^\top\mathcal{X}_1)^{-1}\mathcal{X}_1^\top\mathcal{X}_1\beta_1}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|\|\mathcal{X}_1\boldsymbol{\beta}_1\|} \leq \frac{\sqrt{y^\top \Pi_{\mathcal{X}_1} y}}{\sqrt{y^\top(1 - \Pi_{\mathbf{1}})y}},$$

where equality holds iff $\mathcal{X}_1(\mathcal{X}_1^\top\mathcal{X}_1)^{-1}\mathcal{X}_1^\top y = \mathcal{X}_1\beta_1$, i.e., $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_1$) Thus,

$$\hat{\rho}^2 = \frac{SSR}{SST} = \frac{\hat{\boldsymbol{\beta}}_1^\top \mathcal{X}_1^\top\mathcal{X}_1\hat{\boldsymbol{\beta}}_1}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|^2} = \frac{S_{21}S_{11}^{-1}S_{12}}{S_{22}}.$$

**Example 1.1.19** (Sample Proportions)**.** Let $(X_1, \cdots, X_k)^\top \sim Multi(n, p)$, where $p \in \Theta := \{(p_1, \cdots, p_k)^\top : \sum_{i=1}^k p_i = 1,\ p_i \geq 0\ (i = 1, 2, \cdots, k)\}$. We estimate $p$ with sample proportion

$$\hat{p}_n = \left(\frac{X_1}{n}, \cdots, \frac{X_k}{n}\right)^\top.$$

Then,

(a) $\hat{p}_n$ is consistent estimator of $p$, i.e.,

$$\forall \epsilon > 0,\ \sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \epsilon) \xrightarrow[n\to\infty]{} 0.$$

(b) $q(\hat{p}_n)$ is consistent estimator of $q(p)$ provided that $q$ is (uniformly) continuous on $\Theta$.

*Proof.* (a) Note that there exists a constant $C > 0$ such that

$$\sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \epsilon) \leq \sup_{p\in\Theta} \frac{E|\hat{p}_n - p|^2}{\epsilon^2}$$

$$= \sup_{p\in\Theta} \sum_{i=1}^{k} \frac{p_i(1-p_i)}{n\epsilon^2}$$
$$\leq \frac{C}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0$$

so we get the desired result. Note that first inequality is from Chebyshev's inequality.

(b) Note that $q$ is uniformly continuous on $\Theta$, since $\Theta$ is closed and bounded. Thus the assertion holds. More precisely, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|p' - p| < \delta,\ p, p' \in \Theta \Rightarrow |q(p') - q(p)| < \epsilon.$$

Therefore, we get

$$\sup_{p\in\Theta} P_p(|q(\hat{p}_n) - q(p)| \geq \epsilon) \leq \sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \delta) \xrightarrow[n\to\infty]{} 0.$$

□

**MLE in exponential families**

Consider a random variable $X$ with pdf in canonical exponential family

$$q_\eta(x) = h(x)\exp(\eta^\top T(x) - A(\eta))I_{\mathcal{X}}(x),\ \eta \in \mathcal{E},$$

where $\mathcal{E}$ is natural parameter space in $\mathbb{R}^k$. Our goal is to show consistency of MLE in canonical exponential family.

**Theorem 1.1.20.** *Let*

$$q_\eta(x) = h(x)\exp(\eta^\top T(x) - A(\eta))I_{\mathcal{X}}(x),\ \eta \in \mathcal{E}$$

*be a canonical exponential family with natural parameter space $\mathcal{E} \subseteq \mathbb{R}^k$. Further assume*

*(i) $\mathcal{E}$ is open.*

*(ii) The family is of rank $k$.*

*(iii) $t_0 := T(x) \in C^0$, where $C$ denotes the smallest convex set containing the support of $T(X)$, and $C^0$ be its interior.*

*Then the unique ML estimate $\hat{\eta}(x)$ exists and is the solution of the likelihood equation*

$$\dot{l}_x(\eta) = T(x) - \dot{A}(\eta) = 0.$$

**Remark 1.1.21.** Note that in (iii), $x$ is the observation of $X$, so $t_0$ is the observation of $T(X)$. It is reasonable to consider $t_0$ because ML estimate only depends on $t_0$. Also, recall that (ii) means

$$\begin{aligned} &\nexists a \neq 0 \ s.t. \ \left[P_\eta(a^\top(T(x) - \mu) = 0) = 1 \text{ for some } \eta \in \mathcal{E}\right] \\ \Leftrightarrow &\nexists a \neq 0 \ s.t. \ \left[Var_\eta(a^\top T(x)) = 0 \text{ for some } \eta \in \mathcal{E}\right] \\ \Leftrightarrow &\ddot{A}(\eta) \text{ is positive definite } \forall \eta \in \mathcal{E}. \end{aligned}$$

To prove this, we need some preparation.

**Lemma 1.1.22.**

*(a) ("Supporting Hyperplane Theorem") Let $C \subseteq \mathbb{R}^k$ be a convex set, and $C^0$ be its interior. Then for $t_0 \notin C$ or $t_0 \in \partial C$,*

$$\exists a \neq 0 \ s.t. \ [a^\top t \geq a^\top t_0 \ \forall t \in C].$$

*Conversely, for $t_0 \in C^0$,*

$$\nexists a \neq 0 \ s.t. \ [a^\top t \geq a^\top t_0 \ \forall t \in C].$$

*(b) Let $P(T \in \mathcal{T}) = 1$ and $E(\max_i |T_i|) < \infty$. (i.e., $\mathcal{T}$ is support of $T$.) Then for a convex hull $C$ of $\mathcal{T}$, we get $ET \in C^0$.*

*(c) Assume the above exponential family model with open $\mathcal{E}$. Then the ML estimate exists if the log-likelihood approaches $-\infty$ on the boundary.*

*Proof.* (a) Only second part will be given. (For the first part, see supplementary note.) Let $t_0 \in C^0$. Then $\exists \delta > 0$ such that $B(t_0, \delta) \subseteq C^0$, since $C^0$ is open. Note that for any $u$ s.t. $\|u\| = 1$, we get

$$t_0 - \frac{\delta}{2}u, \ t_0 + \frac{\delta}{2}u \in B(t_0, \delta) \subseteq C.$$

If $\exists a \neq 0$ such that $a^\top t \geq a^\top t_0 \ \forall t \in C$, then

$$a^\top \left(t_0 - \frac{\delta}{2}u\right) \geq a^\top t_0, \ a^\top \left(t_0 + \frac{\delta}{2}u\right) \geq a^\top t_0$$

holds for $u = a/|a|$, which yields contradiction. (Note that convexity condition is not used)

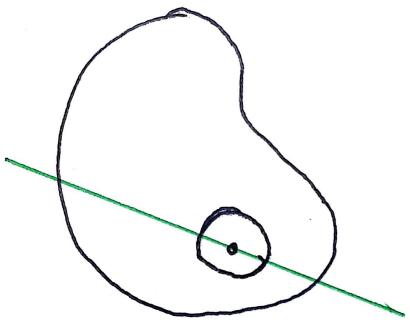

Figure 1.2: Proof of (a)

(b) Note that since $C$ is a convex set, $\mu := ET \in C$ holds. (Convex set contains average of itself) Assume $\mu \notin C^0$. Then $\mu \in \partial C$. Then by (a), $\exists a \neq 0$ such that $a^\top t \geq a^\top \mu$ for any $t \in C$. It implies that, $\exists a \neq 0$ such that $P(a^\top (T - \mu) \geq 0) = 1$, since $\mathcal{T} \subseteq C$. It implies that

$$P(a^\top (T - \mu) = 0) = 1,$$

by the fact that

$$f \geq 0, \ \int f d\mu = 0 \Rightarrow f = 0 \ \mu - a.e..$$

It is contradictory to (ii), which is full rank condition of the exponential family.

(c) Done in TheoStat I.

*Proof of theorem.* By lemma, it's sufficient to show that:

(1) $l(\theta)$ diverges to $-\infty$ at the boundary. (Existence)

(2) Uniqueness

Note that Uniqueness is clear since $l_x(\eta)$ is $\mathcal{C}^2$ function and strictly concave from $\ddot{A}(\eta) > 0$. Thus, our claim is

**Claim.** $l(\theta)$ approches $-\infty$ on the boundary $\partial\mathcal{E}$.

Let $\eta^0 \in \partial\mathcal{E}$. Then there is $\eta_n \xrightarrow[n\to\infty]{} \eta^0$ such that $\eta_n \in \mathcal{E}$. Now our claim is, for any such sequence $\eta_n$, we get $l_x(\eta_n) \xrightarrow[n\to\infty]{} -\infty$. Note that $|\eta_n| \xrightarrow[n\to\infty]{} \infty$ or $\sup |\eta_n| < \infty$. Also note that,

for both cases, from $l_x(\eta) = \log h(x) + \eta^\top T(x) - A(\eta)$ and $e^{A(\eta)} = \int_{\mathcal{X}} h(x) e^{\eta^\top T(x)} d\mu(x)$ , we get

$$\begin{aligned} -l_x(\eta_n) + \log h(x) &= A(\eta_n) - \eta_n^\top t_0 \\ &= \log \int_{\mathcal{X}} \exp\left(\eta_n^\top (T(y) - t_0)\right) h(y) d\mu(y). \end{aligned}$$

Case 1. $|\eta_n| \to \infty$.

Then since

$$\begin{aligned} \int_{\mathcal{X}} e^{\eta_n^\top (T(y)-t_0)} h(y) d\mu(y) &\geq \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} e^{|\eta_n| \cdot \frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0)} h(y) d\mu(y) \\ &\geq \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} e^{|\eta_n|/k} h(y) d\mu(y) \\ &= \exp(|\eta_n|/k) \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y), \end{aligned}$$

if we can conclude

$$\inf_n \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) > 0,$$

by the assumption $|\eta_n| \to \infty$, we get $l_x(\eta_n) \to -\infty$. Note that if

$$\inf_{u:\|u\|=1} \int_{u^\top (T(y)-t_0) > 0} h(y) d\mu(y) > 0,$$

then

$$\inf_n \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > 0} h(y) d\mu(y) > 0,$$

and from

$$\inf_n \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) \xrightarrow[k\to\infty]{} \inf_n \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > 0} h(y) d\mu(y),$$

we get $\exists \epsilon > 0$ & $k$ s.t.

$$\inf_n \int_{\frac{\eta_n^\top}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) > \epsilon$$

and the assertion holds. So our claim is:

**Claim.** $\inf_{u:\|u\|=1} \int_{u^\top (T(y)-t_0) > 0} h(y) d\mu(y) > 0.$

Assume not. If

$$\inf_{u:\|u\|=1} \int_{u^\top (T(y)-t_0)>0} h(y)d\mu(y) = 0,$$

then since $\{u : \|u\| = 1\}$ is compact, there exists $u_0 \in \{u : \|u\| = 1\}$ such that

$$\int_{u_0^\top (T(y)-t_0)>0} h(y)d\mu(y) = 0.$$

It implies $h(y) = 0$ on the set $\{y : u_0^\top (T(y) - t_0) > 0\}$ $\mu$-a.e., and hence

$$\int_{u_0^\top (T(y)-t_0)>0} h(y)e^{\eta^\top T(y)-A(\eta)}d\mu(y) = 0,$$

which implies that

$$P_\eta(u_0^\top (T(X) - t_0) > 0) = 0.$$

Thus, we get

$$P_\eta(u_0^\top (T(X) - t_0) \leq 0) = 1,$$

which is equivalent to

$$u_0^\top (t - t_0) \leq 0 \ \forall t \in \mathcal{T}.$$

Since $C$ is convex hull of $\mathcal{T}$, it implies

$$u_0^\top (t - t_0) \leq 0 \ \forall t \in C,$$

however, this yields contradiction to

$$\nexists a \neq 0 \ s.t. \ a^\top (t - t_0) \leq 0 \ \forall t \in C,$$

from $t_0 \in C^0$.

Case 2. $\sup |\eta_n| < \infty$

In this case, we get

$$\liminf_{n\to\infty} \int_{\mathcal{X}} e^{\eta_n^\top (T(y)-t_0)} h(y)d\mu(y) \geq \int_{\mathcal{X}} e^{\eta^{0\top} (T(y)-t_0)} h(y)d\mu(y) \overset{(*)}{=} \infty$$

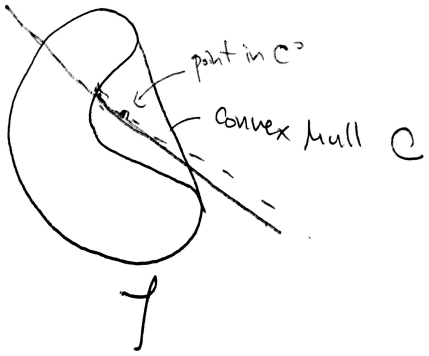


Figure 1.3: Convex hull of $\mathcal{T}$

by Fatou's lemma. $(*)$ holds because $\mathcal{E}$ is natural parameter space, and $\eta^0 \in \partial\mathcal{E}$ implies $\eta^0 \notin \mathcal{E}$, since $\mathcal{E}$ is open. Thus $-l_x(\eta_n) \xrightarrow[n\to\infty]{} \infty$.

□

Now we are ready to prove consistency.

**Theorem 1.1.23.** *Let $X_1, \cdots, X_n$ be a random sample from a population with pdf*

$$p_\eta(x) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}I_{\mathcal{X}}(x),\ \eta \in \mathcal{E}$$

*where $\mathcal{E}$ is the natural parameter space in $\mathbb{R}^k$. Further, assume that*

*(i) $\mathcal{E}$ is open.*

*(ii) The family is of rank $k$.*

*Then, the followings hold:*

*(a) $P_\eta\left(\hat{\eta}_n^{MLE}\ exists\right) \xrightarrow[n\to\infty]{} 1$*

*(b) $\hat{\eta}_n^{MLE}$ is consistent.*

*Proof.* (a) Let $\overline{T}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} T(X_i)$. Then by WLLN, we get

$$\lim_{n\to\infty} P_\eta(|\overline{T}_n - E_\eta T(X_1)| < \epsilon) = 1\ \forall\epsilon > 0.$$

Also note that $E_\eta T(X_1) \in C^0$, where $C^0$ is the interior of the convex hull of the support of $T(X_1)$. Then since $C^0$ is open, open ball $(|\overline{T}_n - E_\eta T(X_1)| < \epsilon)$ is contained in $C^0$ for sufficiently

small $\epsilon > 0$, which implies

$$\lim_{n\to\infty} P_\eta(\overline{T}_n \in C^0) = 1.$$

Now consider $\overline{T}_n$ instead of $T(X_1)$ in previous theorems, and note that (convex hull of support of $\overline{T}_n$) = (convex hull of support of $T(X_1)$). Then we can find that

$$(\overline{T}_n \in C^0) \subseteq (\hat{\eta}_n^{MLE} \text{ exists})$$

and therefore

$$\lim_{n\to\infty} P_\eta(\hat{\eta}_n^{MLE} \text{ exists}) = 1.$$

(b) From $\ddot{A} > 0$, we get $\dot{A}(\eta)$ is one-to-one and continuous for any $\eta$. Then we get

$$(\overline{T}_n \in C^0) \subseteq (\hat{\eta}_n^{MLE} \text{ exists}) = (\dot{A}(\hat{\eta}_n^{MLE}) = \overline{T}_n)$$

and hence

$$\lim_{n\to\infty} P_\eta(\hat{\eta}_n^{MLE} = (\dot{A})^{-1}(\overline{T}_n)) = 1 \ \forall \eta \in \mathcal{E}. \tag{1.1}$$

Further, by inverse function theorem, and $C^2$ property of $A$, we have that $(\dot{A})^{-1}$ is continuous. Thus by WLLN and continuous mapping theorem,

$$(\dot{A})^{-1}(\overline{T}_n) \xrightarrow[n\to\infty]{P_\eta} (\dot{A})^{-1}(E_\eta T(X_1)) = (\dot{A})^{-1}(\dot{A}(\eta)) = \eta$$

and since $(\dot{A})^{-1}(\overline{T}_n) \approx \hat{\eta}_n^{MLE}$ in the sense of (1.1), we get

$$\lim_{n\to\infty} P_\eta(|\hat{\eta}_n^{MLE} - \eta| < \epsilon) = 1 \ \forall \epsilon > 0,$$

i.e., $\hat{\eta}_n^{MLE} \xrightarrow[n\to\infty]{P_\eta} \eta$. □

Now let's see some general results. Suppose we have $\lim_{n\to\infty} \Psi_n(\theta) = \Psi_0(\theta)$ and

$$\theta_n : \text{solution of } \Psi_n(\theta) = 0, \ \theta \in C \ (n = 1, 2, \cdots)$$
$$\theta_0 : \text{solution of } \Psi_0(\theta) = 0, \ \theta \in C.$$

Under what conditions, $\lim_{n\to\infty} \theta_n = \theta_0$? We need following four conditions:

Uniform convergence of $\Psi_n$, Continuity of $\Psi_0$, Uniqueness of $\theta_0$, and Compactness of $C$.

Note that these are sufficient conditions *simultaneously*. Our goal is to obtain similar result for optimization.

**Theorem 1.1.24.** *Suppose that we have* $\lim_{n\to\infty} D_n(\theta) = D_0(\theta)$ *and*

$$\theta_n = \arg\min_{\theta\in C} D_n(\theta) \ (n = 1, 2, \cdots)$$
$$\theta_0 = \arg\min_{\theta\in C} D_0(\theta)$$

*where* $D_n$ *and* $D_0$ *are deterministic functions. Also assume that*

*(i)* $D_n$ *converges to* $D_0$ *uniformly.*

*(ii)* $D_0$ *is continuous on* $C$*.*

*(iii) Minimizer* $\theta_0$ *is unique.*

*(iv)* $C$ *is compact.*

*Then* $\lim_{n\to\infty} \theta_n = \theta_0$*.*

*Proof.* Assume not. In other words, $\theta_n \not\to \theta_0$. Then $\exists\epsilon > 0$ such that $|\theta_n - \theta_0| > \epsilon$ i.o.. It means that there is a subsequence $\{n'\} \subseteq \{n\}$ s.t. $|\theta_{n'} - \theta_0| > \epsilon \ \forall n'$. Now define

$$\Delta_n = \sup_{\theta\in C} |D_n(\theta) - D_0(\theta)|.$$

Then by **uniform convergence** of $D_n$, we get $\Delta_n \xrightarrow[n\to\infty]{} 0$. Now note that

$$\begin{aligned}\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) &= \inf_{|\theta-\theta_0|>\epsilon} \{D_0(\theta) - D_{n'}(\theta) + D_{n'}(\theta)\} \\ &\le \inf_{|\theta-\theta_0|>\epsilon} \{|D_0(\theta) - D_{n'}(\theta)| + D_{n'}(\theta)\} \\ &\le \Delta_{n'} + \inf_{|\theta-\theta_0|>\epsilon} D_{n'}(\theta)\end{aligned}$$

holds. Because minimization of $D_{n'}$ is achieved at $\theta_{n'} \in \{\theta : |\theta - \theta_0| > \epsilon\}$, we get

$$\begin{aligned}\Delta_{n'} + \inf_{|\theta-\theta_0|>\epsilon} D_{n'}(\theta) &\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} D_{n'}(\theta) \\ &\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} \{|D_{n'}(\theta) - D_0(\theta)| + D_0(\theta)\}\end{aligned}$$

$$\leq 2\Delta_{n'} + \inf_{|\theta-\theta_0|\leq\epsilon} D_0(\theta)$$
$$= 2\Delta_{n'} + D_0(\theta_0).$$

The last equality holds from $\theta_0 = \arg\min D_0(\theta)$ and $\theta_0 \in \{\theta : |\theta - \theta_0| \leq \epsilon\}$. Thus

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \leq 2\Delta_{n'} + D_0(\theta_0)$$

holds, which implies

$$\frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0)\right) \leq \Delta_{n'}.$$

Letting $n' \to \infty$, we get

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) = 0.$$

It is contradictory due to our claim that will be shown:

**Claim.** $\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) > 0.$

Intuitively, since $\theta_0$ is unique minimizer, our claim seems trivial, but we also need continuity and compactness condition to guarantee this. (For this see next remark.)

Note that, by definition of infimum, there is a sequence $\{\theta_k\} \subseteq \{\theta : |\theta - \theta_0| > \epsilon\} \cap C$ such that

$$\lim_{k\to\infty} D_0(\theta_k) = \inf_{|\theta-\theta_0|>\epsilon} D_0(\theta).$$

Now, by **compactness of** $C$, there is a subsequence $\{k'\} \subseteq \{k\}$ that makes $\theta_{k'}$ converge to some $\theta_0^*$ ("Bolzano-Weierstrass"), so with the abuse of notation, let $\theta_k \to \theta_0^*$ as $k \to \infty$. Then note that $\theta_0^*$ should belong to $\{\theta : |\theta - \theta_0| \geq \epsilon\} \cap C$, so $\theta_0^* \neq \theta_0$. Now, **continuity of** $D_0$ makes

$$\lim_{k\to\infty} D_0(\theta_k) = D_0(\theta_0^*),$$

which implies

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) = D_0(\theta_0^*).$$

Therefore, by **uniqueness of minimizer**, $D_0(\theta_0^*) > D_0(\theta_0)$, and combining to above result we can obtain

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) > D_0(\theta_0).$$

□

**Remark 1.1.25.** See next figures. Each example tells that we need continuity and compactness, respectively.

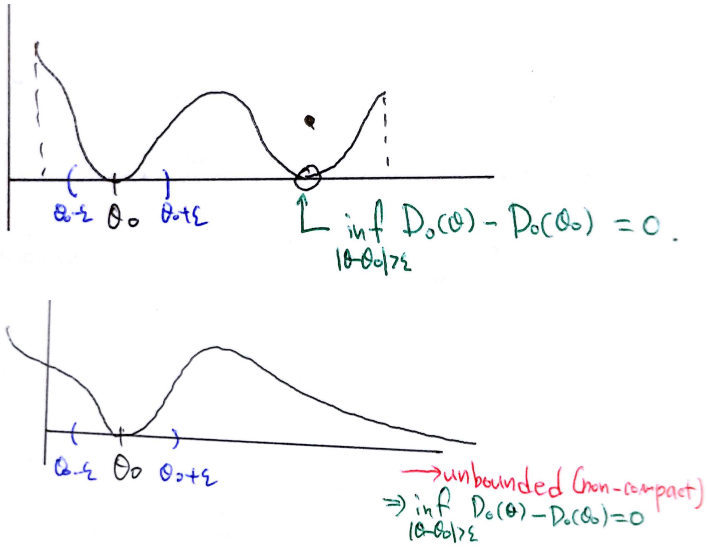


Figure 1.4: Continuity and Compactness are needed.

**Remark 1.1.26.** For deterministic case, one can give an alternative proof. Suppose $\theta_n \not\to \theta_0$. Then since $C$ is compact, we can find a subsequence $\{\theta_{nk}\}$ such that $\theta_{nk} \to \theta_0^*$, $\theta_0^* \neq \theta_0$. (If any convergent subsequence converges to $\theta_0$, then origin sequence should converge to $\theta_0$.) Now for sufficiently large $n_k$,

$$\sup_{\theta \in C} |D_{nk}(\theta) - D_0(\theta)| < \frac{\epsilon}{3}$$

holds, so

$$\begin{aligned} D_0(\theta_0) &\geq D_{nk}(\theta_0) - \frac{\epsilon}{3} \ (\because \text{uniform convergence}) \\ &\geq D_{nk}(\theta_{nk}) - \frac{\epsilon}{3} \ (\because \text{minimizer}) \\ &\geq D_0(\theta_{nk}) - \frac{2}{3}\epsilon \ (\because \text{uniform convergence}) \\ &\geq D_0(\theta_0^*) - \epsilon \ (\because D_0(\theta_{nk}) \to D_0(\theta_0^*) \text{ from continuity of } D_0) \end{aligned}$$

and hence taking $\epsilon \searrow 0$ gives $D_0(\theta_0) \geq D_0(\theta_0^*)$, which is contradictory to uniqueness of $\theta_0$.

In fact, our real goal was, to get the similar result for *random* $D_n$.

**Theorem 1.1.27.** *Let $D_n$ be a sequence of random functions, and $D_0$ be deterministic. Simi-*

*larly, define*

$$\hat{\theta}_n = \arg\min_{\theta \in C} D_n(\theta) \ (n = 1, 2, \cdots)$$

$$\theta_0 = \arg\min_{\theta \in C} D_0(\theta).$$

*Now suppose that*

*(i)* $D_n$ *converges in probability to* $D_0$ ***uniformly****. It means that,*

$$\sup_{\theta \in C} |D_n(\theta) - D_0(\theta)| \xrightarrow[n\to\infty]{P} 0.$$

*(ii)* $D_0$ *is continuous on* $C$*.*

*(iii) Minimizer* $\theta_0$ *is unique.*

*(iv)* $C$ *is compact.*

*Then* $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_0$*.*

*Proof.* Note that in the proof of theorem 1.1.24, we did not used convergence in deriving

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \le 2\Delta_{n'} + D_0(\theta_0).$$

Rather, we only used $|\theta_{n'} - \theta_0| > \epsilon$. (Convergence is used when deriving $\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0)$)
Thus,

$$|\hat{\theta}_n - \theta_0| > \epsilon \Rightarrow \inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \le 2\Delta_n + D_0(\theta_0) \Rightarrow \Delta_n \ge \frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0)\right)$$

holds. Define

$$\frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0)\right) =: \delta(\epsilon).$$

Then, we get

$$\left(|\hat{\theta}_n - \theta_0| > \epsilon\right) \subseteq \left(\Delta_n \ge \delta(\epsilon)\right),$$

and therefore, by uniform P-convergence, $\Delta_n \xrightarrow[n\to\infty]{P} 0$ and hence

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \le P(\Delta_n \ge \delta(\epsilon)) \xrightarrow[n\to\infty]{} 0.$$

□

**Example 1.1.28** (Consistency of MLE when $\Theta$ is finite)**.** Let $X_1, \cdots, X_n$ be a random sample from a population with pdf $f_\theta(\cdot)$, $\theta \in \Theta$. Assume that the parametrization is identifiable and $\Theta = \{\theta_1, \cdots, \theta_k\}$. Then

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0,$$

provided that

(0) (Identifiability) $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$

(1) (Kullback-Leibler divergence) $E_{\theta_0}\left|\log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right| < \infty$.

*Proof.* Note that, we defined

$$\hat{\theta}_n^{MLE} = \arg\min_{\theta\in\Theta} D_n(\theta) \text{ for } D_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)},$$

and by Kullback-Leibler divergence,

$$\theta_0 = \arg\min_{\theta\in\Theta} D_0(\theta) \text{ for } D_0(\theta) = -E_{\theta_0} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}.$$

Then,

(i) $\Theta = \{\theta_1, \cdots, \theta_k\}$ is compact.

(ii) $\theta_0$ is unique minimizer of $D_0$. (For this, see next remark.)

(iii) Uniform convergence is achieved from

$$\begin{aligned} P_{\theta_0}\left\{\max_{1\le j\le k} |D_n(\theta_j) - D_0(\theta_j)| > \epsilon\right\} &= P_{\theta_0}\left\{\bigcup_{1\le j\le k} (|D_n(\theta_j) - D_0(\theta_j)| > \epsilon)\right\} \\ &\le \sum_{j=1}^{k} P_{\theta_0}\left(|D_n(\theta_j) - D_0(\theta_j)| > \epsilon\right) \\ &= o(1) \text{ by WLLN.} \end{aligned}$$

so we can derive the result similarly. In precise, it's sufficient to show

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) > 0$$

for $\epsilon$ s.t. $|\theta_n - \theta_0| > \epsilon$ i.o.. Uniqueness of $\theta_0$ implies it clearly, because $\Theta$ is finite in here. Note that continuity of $D_0$ is not considered. □

**Remark 1.1.29.** *Kullback-Leibler divergence.* Since $1 + \log z \leq z$, we get

$$\begin{aligned} -E_{\theta_0} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} &= -\int \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} dP_{\theta_0} \\ &\geq 1 - \int_{S(\theta_0)} \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x) d\mu(x) \\ &\geq 0, \end{aligned}$$

and hence $D_0(\theta) \geq 0$. In here $S(\theta_0) = \{x : f_{\theta_0}(x) > 0\}$ and $S(\theta) = \{x : f_\theta(x) > 0\}$. Note that $1 + \log z \leq z \Leftrightarrow z = 1$. Thus equality of $D_0(\theta) = 0$ holds if and only if

$$\begin{aligned} \frac{f_\theta(x)}{f_{\theta_0}(x)} &= 1 \ \mu - \text{a.e. on } S(\theta_0) \\ \text{and } \int_{S(\theta_0)} f_\theta(x) d\mu(x) &= 1. \end{aligned}$$

Since

$$\begin{aligned} 1 = \int_{S(\theta)} f_\theta(x) d\mu(x) &= \int_{S(\theta_0) \cup S(\theta)} f_\theta(x) d\mu(x) \\ &= \int_{S(\theta_0)} f_\theta(x) d\mu(x) + \int_{S(\theta) \backslash S(\theta_0)} f_\theta(x) d\mu(x) \end{aligned}$$

we get

$$\int_{S(\theta_0)} f_\theta(x) d\mu(x) = 1 \Leftrightarrow \int_{S(\theta) \backslash S(\theta_0)} f_\theta(x) d\mu(x) = 0.$$

However, by definition of the support, $f_\theta(x) > 0$ on $S(\theta) \backslash S(\theta_0)$, and hence

$$\int_{S(\theta) \backslash S(\theta_0)} f_\theta(x) d\mu(x) = 0 \Leftrightarrow \mu(S(\theta) \backslash S(\theta_0)) = 0.$$

Thus $D_0(\theta)$ holds if and only if

$$\begin{aligned} f_\theta(x) = f_{\theta_0}(x) \ \mu - \text{a.e. on } S(\theta_0) \\ \text{and } \mu(S(\theta) \backslash S(\theta_0)) = 0. \end{aligned}$$

However, note that

$$f_\theta(x) = f_{\theta_0}(x) \ \mu - \text{a.e. on } S(\theta_0) \text{ implies } \mu(S(\theta) \backslash S(\theta_0)) = 0,$$

because

$$\begin{aligned}1 = \int_{S(\theta)} f_\theta(x)d\mu(x) &= \int_{S(\theta_0)} f_\theta(x)d\mu(x) + \int_{S(\theta)\backslash S(\theta_0)} f_\theta(x)d\mu(x) \\ &= \int_{S(\theta_0)} f_{\theta_0}(x)d\mu(x) + \int_{S(\theta)\backslash S(\theta_0)} f_\theta(x)d\mu(x) \\ &= 1 + \int_{S(\theta)\backslash S(\theta_0)} f_\theta(x)d\mu(x).\end{aligned}$$

Therefore we get,

$$D_0(\theta) = 0 \Leftrightarrow f_\theta(x) = f_{\theta_0}(x) \ \mu - \text{a.e. on } S(\theta_0).$$

Now $\mu(S(\theta)\backslash S(\theta_0)) = 0$ implies $f_\theta(x) = f_{\theta_0}(x)$ $\mu -$ a.e. on $S(\theta)\backslash S(\theta_0)$, and therefore $f_\theta(x) = f_{\theta_0}(x)$ $\mu -$ a.e., if $f_\theta(x) = f_{\theta_0}(x)$ $\mu -$ a.e. on $S(\theta_0)$.Therefore we get

$$D_0(\theta) = 0 \Leftrightarrow f_\theta(x) = f_{\theta_0}(x) \ \mu - \text{a.e.} \Leftrightarrow \theta = \theta_0 \ (\because \text{identifiability}).$$

It means that $\theta_0$ is unique minimizer of $D_0(\theta)$.

**Example 1.1.30** (Consistency of MCE)**.** Let $X_1, \cdots, X_n$ be a random sample from $P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$, and

$$\hat{\theta}_n^{MCE} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(X_i, \theta).$$

Assume the following along with $E_{\theta_0}|\rho(X_1, \theta)| < \infty \ \forall \theta_0, \theta \in \Theta$:

For a *fixed* $\theta_0 \in \Theta$, $\exists$a compact set $K \subseteq \Theta$ containing $\theta_0$ such that

(i) (Unique minimizer) $\theta_0 = \arg\min_{\theta\in K} E_{\theta_0}\rho(X_1, \theta)$, and $\theta_0$ is the unique minimizer.

(ii) (Uniform convergence) $\sup_{\theta\in K}|\overline{\rho}_n(\theta) - E_{\theta_0}\rho(X_1, \theta)| \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$

(iii) ($K$ instead of $\Theta$) $P_{\theta_0}(\hat{\theta}_n^{MCE} \in K) \xrightarrow[n\to\infty]{} 1.$

(iv) (Continuous $D_0$) A function $\theta \mapsto E_{\theta_0}\rho(X_1, \theta)$ is continuous on $K$.

In here,

$$\overline{\rho}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho(X_i, \theta).$$

Then $\hat{\theta}_n^{MCE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$.

*Proof.* Note that $\Theta$ need not be compact. Thus, we may use $K$ instead of $\Theta$. By (the proof of)

theorem 1.1.24, we get

$$P_{\theta_0}\left[|\hat{\theta}_n^{MCE} - \theta_0| > \epsilon,\ \hat{\theta}_n^{MCE} \in K\right] \xrightarrow[n\to\infty]{} 0.$$

Thus, we get

$$P_{\theta_0}\left[|\hat{\theta}_n^{MCE} - \theta_0| > \epsilon\right] \leq P_{\theta_0}\left[|\hat{\theta}_n^{MCE} - \theta_0| > \epsilon,\ \hat{\theta}_n^{MCE} \in K\right] + P_{\theta_0}\left[\hat{\theta}_n^{MCE} \notin K\right] \xrightarrow[n\to\infty]{} 0.$$

**Remark 1.1.31.** Indeed, we did not see consistency of MCE yet, but we only verified for fixed $\theta_0 \in \Theta$. For the consistency of MCE, we need that *for any* $\theta_0 \in \Theta$ $\exists K \subseteq \Theta$ *containing* $\theta_0$ *such that the conditions (i)-(iv) are fulfilled.* Suppose that

(a) *for all* compact $K \subseteq \Theta$ and for all $\theta_0 \in \Theta$,

$$\sup_{\theta\in K} |\overline{\rho}_n(\theta) - E_{\theta_0}\rho(X_1, \theta)| \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

(b) *for any* $\theta_0 \in \Theta$ there exists a compact subset $K$ of $\Theta$ containing $\theta_0$ such that

$$P_{\theta_0}\left(\inf_{\theta\in K^c}(\overline{\rho}_n(\theta) - \overline{\rho}_n(\theta_0)) > 0\right) \xrightarrow[n\to\infty]{} 1.$$

(c) $\theta \mapsto E_{\theta_0}\rho(X_1, \theta)$ is continuous on $\Theta$.

Then *for any* $\theta_0 \in \Theta$ there exists a compact subset $K$ of $\Theta$ containing $\theta_0$ such that (ii)-(iv) hold. Note that, (b) implies (iii) with (i) and (c).

Also note that, MLE is a special case for MCE, $\rho(x, \theta) = -\log f(x, \theta)$.

**Remark 1.1.32.** In many cases, it's difficult to verify uniform convergence condition. For this, following **convexity lemma** is useful: *If $K$ is convex,*

$$\overline{\rho}_n(\theta) \xrightarrow[n\to\infty]{P_{\theta_0}} E_{\theta_0}\rho(X_1, \theta)\ \forall\theta \in K, \qquad \text{(“pointwise convergence”)}$$

*and $\overline{\rho}_n$ is a convex function on $K$ with probability 1 under $P_{\theta_0}$, then we get “uniform convergence”*

$$\sup_{\theta\in K} |\overline{\rho}_n(\theta) - E_{\theta_0}\rho(X_1, \theta)| \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

See D. Pollard (1991), Econometric Theory, 7, 186-199.

**Remark 1.1.33.** The condition (b) in remark 1.1.31 is satisfied if the empirical contrast

$$\overline{\rho}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho(X_i, \theta)$$

is convex on a convex open parameter space $\Theta \subseteq \mathbb{R}^k$, and approaches $+\infty$ on the boundary with probability tending to 1.

## 1.2 The Delta Method

Basic intuition of the Delta Method is Taylor expansion.

**Theorem 1.2.1.** *Suppose* $\sqrt{n}(X_n - a) \xrightarrow[n\to\infty]{d} X$. *Then*

$$\sqrt{n}(g(X_n) - g(a)) = \dot{g}(a)\sqrt{n}(X_n - a) + o_P(1)$$

*and hence*

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow[n\to\infty]{d} \dot{g}(a)X,$$

*provided* $g$ *is differentiable at* $a$.

*Proof.* By Taylor theorem, $\exists R(x, a)$ s.t.

$$g(x) = g(a) + (\dot{g}(a) + R(x, a))(x - a)$$

where $R(x, a) \to 0$ as $x \to a$. Note that if $X_n \xrightarrow[n\to\infty]{P} a$ and $R(x, a) \to 0$ as $x \to a$ then $R(X_n, a) \xrightarrow[n\to\infty]{P} 0$ ($\because \forall\epsilon > 0\ \exists\delta > 0\ s.t. |x - a| < \delta \Rightarrow |R(x, a)| < \epsilon$ implies

$$P(|R(X_n, a)| > \epsilon) \le P(|X_n - a| \ge \delta) \xrightarrow[n\to\infty]{} 0$$

and then $R(X_n, a) = o_P(1)$). Thus

$$g(X_n) = g(a) + (\dot{g}(a) + R(X_n, a))(X_n - a)$$

and hence

$$\sqrt{n}(g(X_n) - g(a)) = \dot{g}(a)\sqrt{n}(X_n - a) + \underbrace{R(X_n, a)}_{=o_P(1)}\underbrace{\sqrt{n}(X_n - a)}_{=O_P(1)} = \dot{g}(a)\sqrt{n}(X_n - a) + o_P(1).$$

In multivariate case, statement becomes $\dot{g}(a)^\top (X_n - a)$. □

**Remark 1.2.2.** When $g$ is a function of several variables, the differentiability means the total differentiability, which is implied by the existence of "continuous partial derivatives."

**Example 1.2.3.** $(X_1, Y_1), \cdots, (X_n, Y_n)$ : iid from $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Let

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

(i) As far as the distribution of $\hat{\rho}_n$ is concerned, we may assume $\mu_1 = \mu_2 = 0,\ \sigma_1^2 = \sigma_2^2 = 1$. Let

$$W_i = (X_i, Y_i, X_i^2, Y_i^2, X_i Y_i)^\top \overset{i.i.d.}{\sim} (0, 0, 1, 1, \rho),$$

and $Z_n = \sqrt{n}(\overline{W}_n - (0, 0, 1, 1, \rho)^\top)$, i.e.,

$$Z_{n1} = \sqrt{n}\overline{X},\ Z_{n2} = \sqrt{n}\overline{Y},\ Z_{n3} = \sqrt{n}(\overline{X^2} - 1),\ Z_{n4} = \sqrt{n}(\overline{Y^2} - 1),\ Z_{n5} = \sqrt{n}(\overline{XY} - \rho).$$

Note that $Z_n = O_P(1)$. Then

$$\begin{aligned}
\hat{\rho}_n &= \frac{\frac{1}{\sqrt{n}} Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}} Z_{n1}\right)\left(\frac{1}{\sqrt{n}} Z_{n2}\right)}{\sqrt{1 + \frac{1}{\sqrt{n}} Z_{n3} - \left(\frac{1}{\sqrt{n}} Z_{n1}\right)^2} \sqrt{1 + \frac{1}{\sqrt{n}} Z_{n4} - \left(\frac{1}{\sqrt{n}} Z_{n2}\right)^2}} \\
&= \left(\frac{1}{\sqrt{n}} Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}} Z_{n1}\right)\left(\frac{1}{\sqrt{n}} Z_{n2}\right)\right) \\
&\cdot \left(1 + \frac{1}{\sqrt{n}} Z_{n3} - \left(\frac{1}{\sqrt{n}} Z_{n1}\right)^2\right)^{-1/2} \left(1 + \frac{1}{\sqrt{n}} Z_{n4} - \left(\frac{1}{\sqrt{n}} Z_{n2}\right)^2\right)^{-1/2} \\
&= \left(\frac{1}{\sqrt{n}} Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&\cdot \left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}} Z_{n3} - \left(\frac{1}{\sqrt{n}} Z_{n1}\right)^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}} Z_{n4} - \left(\frac{1}{\sqrt{n}} Z_{n2}\right)^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \left(\frac{1}{\sqrt{n}} Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}} Z_{n3} + o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}} Z_{n4} + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \left(\frac{1}{\sqrt{n}} Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}} Z_{n3} - \frac{1}{2}\frac{1}{\sqrt{n}} Z_{n4} + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \rho + \frac{1}{\sqrt{n}} Z_{n5} - \frac{\rho}{2}\left(\frac{1}{\sqrt{n}} Z_{n3} + \frac{1}{\sqrt{n}} Z_{n4}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

holds, so we get

$$\begin{aligned}
\sqrt{n}(\hat{\rho}_n - \rho) &= Z_{n5} - \frac{\rho}{2}Z_{n3} - \frac{\rho}{2}Z_{n4} + o_P(1) \\
&= \sqrt{n}\left((\overline{XY} - \rho) - \frac{\rho}{2}(\overline{X^2} - 1) - \frac{\rho}{2}(\overline{Y^2} - 1)\right) + o_P(1) \\
&= \sqrt{n}\left(\overline{XY} - \frac{\rho}{2}\overline{X^2} - \frac{\rho}{2}\overline{Y^2}\right) + o_P(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_iY_i - \frac{\rho}{2}X_i^2 - \frac{\rho}{2}Y_i^2\right) + o_P(1) \\
&\xrightarrow[n\to\infty]{d} N(0, Var\left(X_1Y_1 - \frac{\rho}{2}X_1^2 - \frac{\rho}{2}Y_1^2\right)).
\end{aligned}$$

(ii) Now additionally suppose that $(X_i, Y_i)$'s are from bivariate normal distribution. Then $Y_1 - \rho X_1$ is independent of $X_1$. Letting $Z_1 = Y_1 - \rho X_1$, we get $Var(Z_1) = 1-\rho^2$, $Var(Z_1^2) = 2(1-\rho^2)^2$ and hence

$$\begin{aligned}
Var\left(X_1Y_1 - \frac{\rho}{2}X_1^2 - \frac{\rho}{2}Y_1^2\right) &= Var\left((1-\rho^2)X_1Z_1 - \frac{\rho}{2}Z_1^2 + \frac{\rho}{2}(1-\rho^2)X_1^2\right) \\
&= Var\left(\frac{\rho}{2}(1-\rho^2)X_1^2 - \frac{\rho}{2}Z_1^2\right) + Var\left((1-\rho^2)X_1Z_1\right) \\
&\quad + \underbrace{2\,Cov\left(\frac{\rho}{2}(1-\rho^2)X_1^2 - \frac{\rho}{2}Z_1^2, (1-\rho^2)X_1Z_1\right)}_{=0} \\
&= \frac{\rho^2}{4}\left((1-\rho^2)^2 Var(X_1^2) - 2(1-\rho^2)Cov(X_1^2, Z_1^2) + Var(Z_1^2)\right) \\
&\quad + (1-\rho^2)^2 Var(X_1Z_1) \\
&= \frac{\rho^2}{4}\left(2(1-\rho^2)^2 + 2(1-\rho^2)^2\right) + (1-\rho^2)^2(1-\rho^2) \\
&= \rho^2(1-\rho^2)^2 + (1-\rho^2)^2(1-\rho^2) \\
&= (1-\rho^2)^2
\end{aligned}$$

holds. It implies that

$$\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow[n\to\infty]{d} N(0, (1-\rho^2)^2).$$

Therefore, if we define $h(\rho)$ as $h'(\rho) = (1-\rho^2)^{-1}$, i.e.,

$$h(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}, \qquad \text{("Fisher's } z\text{-transform")}$$

then we get

$$\sqrt{n}(h(\hat{\rho}_n) - h(\rho)) \xrightarrow[n\to\infty]{d} N(0, 1),$$

and with this, we can find a confidence region "with stabilized variance."

We can also expand with higher order terms.

**Theorem 1.2.4** (Higher order stochastic expansion)**.** *Let $X_1, \cdots, X_n$ be a random sample with $EX_1 = \mu$ and finite $Var(X_1) = \Sigma$.*

*(a) (1-dim case) For $g$ with $\exists \ddot{g}$,*

$$g(\overline{X}_n) = g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2 + o_P\left(\frac{1}{n}\right),$$

*where $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$.*

*(b) (general case) For $g$ with $\exists \ddot{g}$,*

$$g(\overline{X}_n) = g(\mu) + \dot{g}(\mu)^\top(\overline{X}_n - \mu) + \frac{1}{2}(\overline{X}_n - \mu)^\top\ddot{g}(\mu)(\overline{X}_n - \mu) + o_P\left(\frac{1}{n}\right).$$

*Proof.* Again, use Taylor theorem. Only prove (a). Note that

$$g(x) = g(a) + \dot{g}(a)(x-a) + \frac{1}{2}\left(\ddot{g}(a) + R(x,a)\right)(x-a)^2$$

for $R(x,a) \to 0$ as $x \to a$, so letting $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$, we get

$$\begin{aligned} g(\overline{X}_n) &= g(\mu) + \dot{g}(\mu)(\overline{X}_n - \mu) + \frac{1}{2}\ddot{g}(\mu)(\overline{X}_n - \mu)^2 + \frac{1}{2}\underbrace{R(\overline{X}_n, \mu)}_{=o_P(1)}\underbrace{(\overline{X}_n - \mu)^2}_{=O_P(1/n)} \\ &= g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2 + o_P(1/n) \end{aligned}$$

which implies the conclusion.

**Remark 1.2.5.** For general case, following notation is also frequently used. For $\left(Z_n^i\right)_{i=1}^d = \sqrt{n}(\overline{X}_n - \mu)$,

$$g(\overline{X}_n) = g(\mu) + \frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i + \frac{1}{2n}g_{/ij}(\mu)Z_n^iZ_n^j + o_P\left(\frac{1}{n}\right).$$

In here, we omit the "$\sum$," i.e.,

$$g_{/i}(\mu)Z_n^i := \sum_{i=1}^d g_{/i}(\mu)Z_n^i,\ g_{/ij}(\mu)Z_n^iZ_n^j = \sum_{i=1}^d\sum_{j=1}^d g_{/ij}(\mu)Z_n^iZ_n^j.$$

**Example 1.2.6** (Estimation of Reliability)**.** Let $X_1, \cdots, X_n$ be a random sample from $Exp(\lambda)$, where $\lambda > 0$ is a rate. Consider an estimation problem of "reliability"

$$\eta = P(X_1 > a) = e^{-a\lambda}.$$

(i) Note that

$$\hat{\eta}_n^{MLE} = e^{-a/\overline{X}}.$$

Let $Z_n = \sqrt{n}(\lambda\overline{X} - 1)$. Then $Z_n \xrightarrow[n\to\infty]{d} N(0,1)$ and

$$\overline{X}^{-1} = \lambda\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}.$$

Thus, we get

$$\begin{aligned}
\hat{\eta}_n^{MLE} &= \exp\left(-a\lambda\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}\right) \\
&= \exp\left(-a\lambda\left(1 - \frac{Z_n}{\sqrt{n}} + \frac{Z_n^2}{n} + o_P\left(\frac{1}{n}\right)\right)\right) \\
&= e^{-a\lambda}\left(1 + \left(a\lambda\frac{Z_n}{\sqrt{n}} - a\lambda\frac{Z_n^2}{n}\right) + \frac{1}{2}\left(a\lambda\frac{Z_n}{\sqrt{n}} - a\lambda\frac{Z_n^2}{n}\right)^2 + o_P\left(\frac{1}{n}\right)\right) \\
&= e^{-a\lambda}\left(1 + a\lambda\frac{Z_n}{\sqrt{n}} + \frac{-a\lambda + (a\lambda)^2/2}{n}Z_n^2 + o_P\left(\frac{1}{n}\right)\right) \\
&= \eta + \frac{a\lambda e^{-a\lambda}}{\sqrt{n}}Z_n + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n}Z_n^2 + o_P\left(\frac{1}{n}\right) \qquad (1.2)
\end{aligned}$$

from $(1+x)^{-1} = 1 - x + x^2 + o(x^2)$ and $e^x = 1 + x + x^2/2 + o(x^2)$ when $x \approx 0$.

(ii) Now consider

$$\hat{\eta}^{UMVUE} = \left(1 - \frac{a}{n\overline{X}}\right)^{n-1} I\left(\frac{a}{n\overline{X}} < 1\right).$$

Note that from $P\left(\frac{a}{n\overline{X}} < 1\right) \xrightarrow[n\to\infty]{} 1$, we can let $\hat{\eta}^{UMVUE} = \left(1 - \frac{a}{n\overline{X}}\right)^{n-1}$ (See next remark). Then

$$\begin{aligned}
\log\hat{\eta}^{UMVUE} &= (n-1)\log\left(1 - \frac{a}{n\overline{X}}\right) \\
&= (n-1)\log\left(1 - \frac{a\lambda}{n}\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}\right) \\
&= (n-1)\log\left(1 - \frac{a\lambda}{n}\left(1 - \frac{Z_n}{\sqrt{n}} + \frac{Z_n^2}{n} + o_P\left(\frac{1}{n}\right)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= (n-1)\log\left(1 - \frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2 + o_P\left(\frac{1}{n^2}\right)\right) \\
&= (n-1)\left\{\left(-\frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2\right) - \frac{1}{2}\left(-\frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2\right)^2\right\} + o_P\left(\frac{1}{n}\right) \\
&= -a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n - \frac{a\lambda}{n}Z_n^2 - \frac{(a\lambda)^2}{2n} + \frac{a\lambda}{n} + o_P\left(\frac{1}{n}\right) \\
&= -a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n} + o_P\left(\frac{1}{n}\right)
\end{aligned}
$$

implies

$$
\begin{aligned}
\hat{\eta}^{UMVUE} &= \exp\left(-a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n} + o_P\left(\frac{1}{n}\right)\right) \\
&= e^{-a\lambda}\left(1 + \left(\frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n}\right) + \frac{1}{2}\left(\frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n}\right)^2\right) \\
&\quad + o_P\left(\frac{1}{n}\right) \\
&= e^{-a\lambda}\left(1 + \frac{a\lambda}{\sqrt{n}}Z_n + \left(-a\lambda Z_n^2 + a\lambda - \frac{(a\lambda)^2}{2} + \frac{1}{2}(a\lambda)^2 Z_n^2\right)\frac{1}{n}\right) + o_P\left(\frac{1}{n}\right) \\
&= \eta + \frac{a\lambda e^{-a\lambda}}{\sqrt{n}}Z_n + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n}(Z_n^2 - 1) + o_P\left(\frac{1}{n}\right) \qquad (1.3)
\end{aligned}
$$

from $\log(1+x) = x - x^2/2 + o(x^2),\ x \approx 0$. Comparing (1.3) to (1.2), we can say that UMVUE is "closer" than MLE to $\eta$, since MLE's leading term has a bias

$$\frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n},$$

while UMVUE's leading term has no bias. Like this case, if one suggests a new estimator, then in many cases, one compares 2nd order term to judge its asymptotic behavior.

**Remark 1.2.7.** If there is an event that occurring probability converges to 1, then in an asymptotic sense, we may ignore such event, in the sense that:

(i) If $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$ and $P(X_n \le x,\ \mathcal{E}_n) \xrightarrow[n\to\infty]{} F(x)$, then $X_n \xrightarrow[n\to\infty]{d} F$.

(ii) If $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$ and $X_n = X + O_P(n^{-\alpha})$ on $\mathcal{E}_n$, then $X_n = X + O_P(n^{-\alpha})$ in general. Convergence rate of $P(\mathcal{E}_n)$ does not matter!

($\because P(n^\alpha|X_n - X| \ge C) \le P(n^\alpha|X_n - X| \ge C,\ \mathcal{E}_n) + \underbrace{P(\mathcal{E}_n^c)}_{\xrightarrow[n\to\infty]{}0}$, take $\limsup_n$ and $\lim_C$ on both sides.)

**Example 1.2.8.** Consider a sample correlation coefficient again. Assume $EX_1^4 < \infty$ and $EY_1^4 < \infty$. Then

$$\begin{aligned}
\hat{\rho}_n &= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)\left(\frac{1}{\sqrt{n}}Z_{n2}\right)\right) \\
&\cdot \left(1 + \frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)^{-1/2}\left(1 + \frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)^{-1/2} \\
&= \left(\rho + \frac{1}{\sqrt{n}}Z_{n5} - \frac{1}{n}Z_{n1}Z_{n2} + o_P\left(\frac{1}{n}\right)\right) \\
&\cdot \left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right) + \frac{3}{8}\left(\frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)^2 + o_P\left(\frac{1}{n}\right)\right) \\
&\cdot \left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right) + \frac{3}{8}\left(\frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)^2 + o_P\left(\frac{1}{n}\right)\right) \\
&= \left(\rho + \frac{1}{\sqrt{n}}Z_{n5} - \frac{1}{n}Z_{n1}Z_{n2} + o_P\left(\frac{1}{n}\right)\right) \\
&\cdot \left(1 - \frac{1}{2\sqrt{n}}Z_{n3} + \frac{1}{n}\left(\frac{1}{2}Z_{n1}^2 + \frac{3}{8}Z_{n3}^2\right) + o_P\left(\frac{1}{n}\right)\right) \\
&\cdot \left(1 - \frac{1}{2\sqrt{n}}Z_{n4} + \frac{1}{n}\left(\frac{1}{2}Z_{n2}^2 + \frac{3}{8}Z_{n4}^2\right) + o_P\left(\frac{1}{n}\right)\right) \\
&= \rho + \frac{1}{\sqrt{n}}\left(Z_{n5} - \frac{\rho}{2}Z_{n3} - \frac{\rho}{2}Z_{n4}\right) \\
&+ \frac{1}{n}\left(-Z_{n1}Z_{n2} - \frac{1}{2}Z_{n3}Z_{n5} - \frac{1}{2}Z_{n4}Z_{n5} + \rho\left(\frac{1}{4}Z_{n3}Z_{n4} + \frac{1}{2}Z_{n1}^2 + \frac{1}{2}Z_{n2}^2 + \frac{3}{8}Z_{n3}^2 + \frac{3}{8}Z_{n4}^2\right)\right) + o_P\left(\frac{1}{n}\right)
\end{aligned}$$

holds. The leading term has bias

$$\frac{1}{n}\left\{\frac{\rho}{4}EX_1^2Y_1^2 + \frac{3}{8}\rho(EX_1^4 + EY_1^4) - \frac{1}{2}(EX_1^3Y_1 + EX_1Y_1^3)\right\}$$

from

$$\begin{aligned}
E(Z_{3n}Z_{4n}) &= EX_1^2Y_1^2 - 1 \\
E(Z_{3n}^2 + Z_{4n}^2) &= EX_1^4 + EY_1^4 - 2 \\
E(Z_{1n}^2 + Z_{2n}^2) &= 2 \\
E(Z_{3n}Z_{5n} + Z_{4n}Z_{5n}) &= EX_1^3Y_1 + EX_1Y_1^3 - 2\rho \\
E(Z_{1n}Z_{2n}) &= \rho.
\end{aligned}$$

For bivariate normal case, it becomes

$$-\frac{1}{2n}\rho(1-\rho^2).$$

**Remark 1.2.9.** Note that in using stochastic expansion, we can get the mean, variance, skewness, ... of the leading term and might expect that they become the approximation of the moments of $g(\overline{X}_n)$, but this is not true! For example, if $X_n \sim Ber(1/n)$, then

$$P(nX_n > \epsilon) = P(X_n = 1) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$$

from $X_n = 0$ or 1, so $nX_n = o_P(1)$, but we get $E(nX_n) = 1\ \forall n$. Thus, we need to check the behavior of the remainder.

As we can see in this example, we cannot say that $E(o_p(1)) = o(1)$, i.e., "convergence in probability does not imply convergence in $\mathcal{L}^1$." (Similarly, "bounded in probability does not imply uniform integrability") By Vitalli's theorem, it is known that if $\{X_n\}$ is uniformly integrable, then

$$X_n \xrightarrow[n\to\infty]{P} X \text{ implies } EX_n \xrightarrow[n\to\infty]{} EX.$$

From now on, we compare "moments of leading terms" and "approximation of moments."

**Example 1.2.10.** Recall mgf

$$mgf_X(t) = Ee^{tX} \ |t| < \epsilon$$

and cgf

$$cgf_X(t) = \log mgf_X(t) = \log Ee^{tX}. \ |t| < \epsilon$$

For $m_r = EX^r$, mgf has a Taylor expansion

$$mgf_X(t) = 1 + m_1 t + \frac{m_2}{2!}t^2 + \frac{m_3}{3!}t^3 + \cdots,$$

and if $X$ and $Y$ are independent,

$$mgf_{X+Y}(t) = mgf_X(t) \cdot mgf_Y(t)$$

and

$$cgf_{aX+b}(t) = cgf_X(at) + bt$$

holds for constants $a$ and $b$. From this we get

$$c_r(aX+b) = a^r c_r(X),$$

where $c_r$ denotes $r$th cumulant. Also recall that, for $A \approx 0$,

$$\log(1+A) = A - \frac{1}{2}A^2 + \frac{1}{3}A^3 - \cdots,$$

and with this, we can obtain

$$cgf_X(t) = \log\left(1 + \underbrace{(mgf_X(t)-1)}_{=A}\right) = c_1 t + \frac{c_2}{2!}t^2 + \frac{c_3}{3!}t^3 + \cdots$$

where

$$c_1 = m_1,\ c_2 = m_2 - m_1^2,\ c_3 = m_3 - 3m_1m_2 + 2m_1^3,\ c_4 = m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - m_1^4, \cdots.$$

If observations are normalized, i.e., $m_1 = 0$ and $m_2 = 1$, then

$$c_1 = 0,\ c_2 = 1,\ c_3 = m_3,\ c_4 = m_4 - 3.$$

**Example 1.2.11.** Let

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{\frac{X_i - \mu}{\sigma}}_{=:\tilde{Z}_i \overset{d}{\equiv} Z_1}.$$

Then from

$$cgf_{Z_n}(t) = cgf_{n^{-1/2}\sum \tilde{Z}_i}(t) = n \cdot cgf_{Z_1}\left(\frac{t}{\sqrt{n}}\right),$$

we obtain

$$c_r(Z_n) = n \cdot \left(\frac{1}{\sqrt{n}}\right)^r c_r(Z_1) = n^{-\frac{r}{2}+1} c_r(Z_1).$$

From this, we obtain

$$EZ_n^3 = c_3(Z_n) = \frac{1}{\sqrt{n}}c_3(Z_1) \tag{1.4}$$

$$EZ_n^4 = c_4(Z_n) + 3 = \frac{1}{n}c_4(Z_1) + 3. \tag{1.5}$$

**Example 1.2.12.** Now see the multivariate case. Let $X = (X_1, \cdots, X_d)^\top$ and $t = (t_1, \cdots, t_d)^\top$.

Then

$$mgf_X(t) = Ee^{t^\top X} = Ee^{t_1X_1+\cdots+t_dX_d}$$

and

$$m_1 = \left[\frac{\partial}{\partial t_i} mgf_X(t)\Big|_{t=0}\right]_i$$

$$m_2 = \left[\frac{\partial^2}{\partial t_i \partial t_j} mgf_X(t)\Big|_{t=0}\right]_{i,j}$$

$$m_3 = \left[\frac{\partial^3}{\partial t_i \partial t_j \partial t_k} mgf_X(t)\Big|_{t=0}\right]_{i,j,k}$$

$$\vdots$$

and we get

$$mgf_X(t) = 1 + \sum_i m_1(i)t_i + \frac{1}{2!}\sum_{i,j} m_2(i,j)t_it_j + \frac{1}{3!}\sum_{i,j,k} m_3(i,j,k)t_it_jt_k + \cdots.$$

If data is centered, i.e., $EX = 0$, then $m_1 = 0$ and so

$$cgf_X(t) = \frac{1}{2!}\sum_{i,j} m_2(i,j)t_it_j + \frac{1}{3!}\sum_{i,j,k} m_3(i,j,k)t_it_jt_k + \frac{1}{4!}\sum_{i,j,k,l}\left(m_4(i,j,k,l) - 3m_2(i,j)m_2(k,l)\right)t_it_jt_kt_l + \cdots.$$

Now let $Z_n = (Z_n^1, \cdots, Z_n^d)^\top = \sqrt{n}(\overline{X}_n - \mu)$. Then

$$cgf_{Z_n}(t) = n \cdot cgf_{Z_1}(t/\sqrt{n})$$

implies

$$EZ_n^iZ_n^j =: \sigma^{i,j},\ (\sigma^{i,j})_{i,j} = Var(X_1)$$

$$EZ_n^iZ_n^jZ_n^k = \frac{1}{\sqrt{n}}c_3(i,j,k)$$

$$EZ_n^iZ_n^jZ_n^kZ_n^l = \frac{1}{n}c_4(i,j,k,l) + 3\sigma^{ij}\sigma^{kl}$$

where

$$c_3(i,j,k) = c_3(Z_1)(i,j,k) = EZ_1^iZ_1^jZ_1^k$$

$$c_4(i,j,k,l) = c_4(Z_1)(i,j,k,l) = EZ_1^iZ_1^jZ_1^kZ_1^l - 3(EZ_1^iZ_1^j)(EZ_1^kZ_1^l).$$

**Proposition 1.2.13** (Moments of the leading terms)**.** *Let $X_1, \cdots, X_n$ be i.i.d. with $E|X_1|^4 <$*

$\infty$[1].

*(a) (Univariate case) For $g$ with $\exists \ddot{g}$ and $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$,*

$$g(\overline{X}_n) = \underbrace{g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2}_{=:W_n} + o_P(n^{-1}),$$

*and for the leading term $W_n$,*

$$\begin{aligned}
E(W_n) &= g(\mu) + \frac{\sigma^2}{2n}\ddot{g}(\mu) \\
Var(W_n) &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{1}{n^2}\left(\sigma^3 c_3(Z_1)\dot{g}(\mu)\ddot{g}(\mu) + \frac{1}{2}\sigma^4\left(\ddot{g}(\mu)\right)^2\right) + O(n^{-3}) \\
E(W_n - EW_n)^3 &= \frac{1}{n^2}\left(\sigma^3 c_3(Z_1)\left(\dot{g}(\mu)\right)^3 + 3\sigma^4\left(\dot{g}(\mu)\right)^2\left(\ddot{g}(\mu)\right)\right) + O(n^{-3}).
\end{aligned}$$

*(b) (Multivariate case) For $g$ with $\exists \ddot{g}$ and $(Z_n^i) = \sqrt{n}(\overline{X}_n - \mu)$,*

$$g(\overline{X}_n) = \underbrace{g(\mu) + \frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i + \frac{1}{2n}g_{/ij}(\mu)Z_n^i Z_n^j}_{=:W_n} + o_P(n^{-1}),$$

*and for the leading term $W_n$,*

$$\begin{aligned}
E(W_n) &= g(\mu) + \frac{1}{2n}g_{/ij}(\mu)\sigma^{ij} \\
Var(W_n) &= \frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij} + O(n^{-2})
\end{aligned}$$

*where $(\sigma^{ij}) = Var(X_1)$.*

*Proof.* (a) Nothing but tedious calculation. First,

$$E(W_n) = g(\mu) + \frac{\sigma^2}{2n}\ddot{g}(\mu)$$

is easily obtained. Next, note that

$$\begin{aligned}
Var(W_n) &= Var\left(\frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2\right) \\
&= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 Var(Z_n) + \frac{\sigma^3}{n\sqrt{n}}\dot{g}(\mu)\ddot{g}(\mu)Cov(Z_n, Z_n^2) + \frac{\sigma^4}{4n^2}\left(\ddot{g}(\mu)\right)^2 Var(Z_n^2).
\end{aligned}$$

First, $Var(Z_n) = 1$. Also, $Var(Z_n^2) = E(Z_n^4) - \left[E(Z_n^2)\right]^2 = 2 + n^{-1}c_4(Z_1)$ from (1.5). Finally,

[1] In fact, stronger condition is needed: mgf of $X_1$ exists

$Cov(Z_n, Z_n^2) = EZ_n^3 - EZ_n \cdot EZ_n^2 = n^{-1/2}c_3(Z_1)$ from (1.4). Now we get

$$\begin{aligned} Var(W_n) &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{\sigma^3}{n^2}\dot{g}(\mu)\ddot{g}(\mu)c_3(Z_1) + \frac{\sigma^4}{2n^2}\left(\ddot{g}(\mu)\right)^2 + \frac{\sigma^4}{4n^3}\left(\ddot{g}(\mu)\right)^2 c_4(Z_1) \\ &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{1}{n^2}\left(\sigma^3 c_3(Z_1)\dot{g}(\mu)\ddot{g}(\mu) + \frac{1}{2}\sigma^4\left(\ddot{g}(\mu)\right)^2\right) + O(n^{-3}). \end{aligned}$$

For $E(W_n - EW_n)^3$, note that $EZ_n^5 = O(n^{-1/2})$ and $EZ_n^6 = O(1)$. (Check!) Then

$$\begin{aligned} E(W_n - EW_n)^3 &= E\left[\frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)(Z_n^2 - 1)\right]^3 \\ &= \frac{\sigma^3}{n\sqrt{n}}\dot{g}(\mu)^3 \underbrace{EZ_n^3}_{=n^{-1/2}c_3(Z_1)} + \frac{3\sigma^4}{2n^2}\dot{g}(\mu)^2\ddot{g}(\mu)\underbrace{E[Z_n^2(Z_n^2-1)]}_{=n^{-1}c_4(Z_1)+2} \\ &\quad + \underbrace{\frac{\sigma^5}{4n^2\sqrt{n}}\dot{g}(\mu)\ddot{g}(\mu)^2 E[Z_n(Z_n^2-1)^2] + \frac{\sigma^6}{8n^3}\ddot{g}(\mu)^3 E[(Z_n^2-1)^3]}_{=O(n^{-3})} \\ &= \frac{1}{n^2}\left(\sigma^3 c_3(Z_1)\left(\dot{g}(\mu)\right)^3 + 3\sigma^4\left(\dot{g}(\mu)\right)^2\left(\ddot{g}(\mu)\right)\right) + O(n^{-3}) \end{aligned}$$

by (1.4) and (1.5).

(b) Note that

$$W_n = g(\mu) + \frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i + \frac{1}{2n}g_{/ij}(\mu)Z_n^iZ_n^j = g(\mu) + \frac{1}{\sqrt{n}}\sum_{i=1}^d g_{/i}(\mu)Z_n^i + \frac{1}{2n}\sum_{i,j=1}^d g_{/ij}(\mu)Z_n^iZ_n^j$$

so

$$\begin{aligned} Var(W_n) &= \frac{1}{n}\sum_{i=1}^d\sum_{j=1}^d g_{/i}(\mu)g_{/j}(\mu)Cov(Z_n^i, Z_n^j) + \frac{1}{n\sqrt{n}}\sum_{i=1}^d\sum_{k,l=1}^d g_{/i}(\mu)g_{/kl}(\mu)Cov(Z_n^i, Z_n^kZ_n^l) \\ &\quad + \frac{1}{4n^2}\sum_{i,j=1}^d\sum_{k,l=1}^d g_{/ij}(\mu)g_{/kl}(\mu)Cov(Z_n^iZ_n^j, Z_n^kZ_n^l) \\ &= \frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\underbrace{Cov(Z_n^i, Z_n^j)}_{=\sigma^{ij}} + \frac{1}{n\sqrt{n}}g_{/i}(\mu)g_{/kl}(\mu)\underbrace{Cov(Z_n^i, Z_n^kZ_n^l)}_{=EZ_n^iZ_n^kZ_n^l} \\ &\quad + \frac{1}{4n^2}g_{/ij}(\mu)g_{/kl}(\mu)\underbrace{Cov(Z_n^iZ_n^j, Z_n^kZ_n^l)}_{EZ_n^iZ_n^jZ_n^kZ_n^l - (EZ_n^iZ_n^j)(EZ_n^kZ_n^l)} \\ &= \frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij} + \frac{1}{n^2}g_{/i}(\mu)g_{/kl}(\mu)c_3(i,j,k) + \frac{1}{4n^2}g_{/ij}(\mu)g_{/kl}(\mu)\left(\frac{1}{n}c_4(i,j,k,l) + 2\sigma^{ij}\sigma^{kl}\right) \\ &= \frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij} + O(n^{-2}). \end{aligned}$$

□

**Proposition 1.2.14** (Approximation of moments)**.** *Let $X_1, \cdots, X_n$ be i.i.d. with $E|X_1|^4 < \infty$.*

*(a) (Univariate case) For $g$ with bounded $g^{(r)}$ $(r = 0, 1, \cdots, 4)$,*

$$E(g(\overline{X}_n)) = g(\mu) + \frac{\sigma^2}{2n} g^{(2)}(\mu) + O(n^{-2})$$
$$Var(g(\overline{X}_n)) = \frac{\sigma^2}{n} g^{(1)}(\mu)^2 + O(n^{-2})$$

*where $\mu = EX_1$, $\sigma^2 = Var(X_1)$.*

*(b) (Multivariate case) For $g$ with bounded and contiuous $g_{/I}$ $(|I| = 0, 1, \cdots, 4)$,*

$$E(g(\overline{X}_n)) = g(\mu) + \frac{1}{2n} g_{/ij}(\mu)\sigma^{ij} + O(n^{-2})$$
$$Var(g(\overline{X}_n)) = \frac{1}{n} g_{/i}(\mu) g_{/j}(\mu)\sigma^{ij} + O(n^{-2})$$

*where $\mu = EX_1$, $(\sigma^{ij}) = Var(X_1)$.*

*Proof.* (a) Note that,

$$g(\overline{X}_n) = g(\mu) + \frac{\sigma}{\sqrt{n}} g^{(1)}(\mu) Z_n + \frac{\sigma^2}{2n} g^{(2)}(\mu) Z_n^2 + \frac{1}{3!} \frac{\sigma^3}{n\sqrt{n}} g^{(3)}(\mu) Z_n^3 + R_n,$$

where

$$R_n = \frac{1}{4!} \frac{\sigma^4}{n^2} g^{(4)}(\xi_n) Z_n^4, \ \xi_n : \text{ a number between } \mu \text{ and } \overline{X}_n.$$

Note that

$$E|R_n| \le \frac{1}{4!} \frac{\sigma^4}{n^2} \sup_x |g^{(4)}(x)| \cdot EZ_n^4 = O(n^{-2})$$

from "boundedness of $g^{(4)}$" and $EZ_n^4 = 3 + n^{-1} c_4(Z_1)$. Also note that

$$\frac{\sigma^3}{n\sqrt{n}} EZ_n^3 = \frac{\sigma^3}{n\sqrt{n}} \frac{1}{\sqrt{n}} c_3(Z_1) = O(n^{-2}).$$

From these, we obtain

$$Eg(\overline{X}_n) = g(\mu) + \frac{\sigma^2}{2n} g^{(2)}(\mu).$$

Next, for the variance, we get

$$Varg(\overline{X}_n) = \frac{\sigma^2}{n} g^{(1)}(\mu)^2 Var(Z_n) + \frac{\sigma^2}{4n^2} g^{(2)}(\mu)^2 Var(Z_n^2)$$
$$+ O(n^{-3}) Var(Z_n^3) + Var(R_n)$$

$$+ O(n^{-3/2})Cov(Z_n, Z_n^2) + O(n^{-2})Cov(Z_n, Z_n^3) + O(n^{-5/2})Cov(Z_n^2, Z_n^3)$$
$$+ O(n^{-1/2})Cov(Z_n, R_n) + O(n^{-1})Cov(Z_n^2, R_n) + O(n^{-3/2})Cov(Z_n^3, R_n).$$

Note that,

$$Var(Z_n) = 1,\ Var(Z_n^2) = 2 + \frac{1}{n}c_4(Z_1) = O(1),\ Var(Z_n^3) = EZ_n^6 - (EZ_n^3)^2 \le O(1),$$

$$Var(R_n) \le ER_n^2 = O(n^{-4}) \cdot EZ_n^8 \le O(n^{-4}),$$

$$|Cov(Z_n, R_n)| \le \sqrt{VarZ_n}\sqrt{VarR_n} \le O(n^{-2}),$$

$$|Cov(Z_n^2, R_n)| \le \sqrt{VarZ_n^2}\sqrt{VarR_n} \le O(n^{-2}),$$

$$|Cov(Z_n^3, R_n)| \le \sqrt{VarZ_n^3}\sqrt{VarR_n} \le O(n^{-2}),$$

$$Cov(Z_n, Z_n^2) = EZ_n^3 - (EZ_n)(EZ_n^2) = O(n^{-1/2}),$$

$$|Cov(Z_n, Z_n^3)| \le \sqrt{VarZ_n}\sqrt{VarZ_n^3} \le O(1),$$

$$|Cov(Z_n^2, Z_n^3)| \le \sqrt{VarZ_n^2}\sqrt{VarZ_n^3} \le O(1).$$

From these, we get

$$\begin{aligned}
Varg(\overline{X}_n) = {} & \frac{\sigma^2}{n}g^{(1)}(\mu)^2 \underbrace{Var(Z_n)}_{=1} + \frac{\sigma^2}{4n^2}g^{(2)}(\mu)^2 \underbrace{Var(Z_n^2)}_{=O(1)} \\
& + O(n^{-3}) \underbrace{Var(Z_n^3)}_{=O(1)} + \underbrace{Var(R_n)}_{\le O(n^{-4})} \\
& + O(n^{-3/2}) \underbrace{Cov(Z_n, Z_n^2)}_{=O(n^{-1/2})} + O(n^{-2}) \underbrace{Cov(Z_n, Z_n^3)}_{\le O(1)} + O(n^{-5/2}) \underbrace{Cov(Z_n^2, Z_n^3)}_{\le O(1)} \\
& + O(n^{-1/2}) \underbrace{Cov(Z_n, R_n)}_{\le O(n^{-2})} + O(n^{-1}) \underbrace{Cov(Z_n^2, R_n)}_{\le O(n^{-2})} + O(n^{-3/2}) \underbrace{Cov(Z_n^3, R_n)}_{\le O(n^{-2})} \\
= {} & \frac{\sigma^2}{n}g^{(1)}(\mu)^2 + O(n^{-2}).
\end{aligned}$$

(b) Recall the Taylor theorem for multivariate function: for $(k+1)$-time continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\mathbf{x}) = \sum_{|\alpha| \le k} \frac{D^\alpha f(\mathbf{a})}{\alpha!}(\mathbf{x} - \mathbf{a})^\alpha + \sum_{|\beta| = k+1} R_\beta(\mathbf{x})(\mathbf{x} - \mathbf{a})^\beta,$$

where for $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$ we define $|\alpha| = \alpha_1 + \cdots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$, and $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$,

$$R_\beta(\mathbf{x}) = \frac{|\beta|}{\beta!} \int_0^1 (1-t)^{|\beta|-1} D^\beta f(\mathbf{a} + t(\mathbf{x}-\mathbf{a}))dt.$$

In here,

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Using this, we obtain

$$g(\overline{X}_n) = g(\mu) + \frac{1}{\sqrt{n}} g_{/i}(\mu) Z_n^i + \frac{1}{2!}\frac{1}{n} g_{/ij} Z_n^i Z_n^j + \frac{1}{3!}\frac{1}{n\sqrt{n}} g_{/ijk} Z_n^i Z_n^j Z_n^k + R_n,$$

where

$$|R_n| \leq \frac{1}{6n^2} \left| \int_0^1 (1-u)^3 g_{/ijkl} \left( \mu + \frac{1}{\sqrt{n}} u Z_n \right) Z_n^i Z_n^j Z_n^k Z_n^l du \right|.$$

By boundedness of $g_{/I}$ and

$$E|Z_n^i Z_n^j Z_n^k Z_n^l| \leq \frac{1}{4} E((Z_n^i)^4 + (Z_n^j)^4 + (Z_n^k)^4 + (Z_n^l)^4) = O(1), \qquad \text{(“AM-GM”)}$$

we obtain the desired result with similar procedure as univariate case. □

**Example 1.2.15** (Estimation of reliability)**.** Let $X_1, \cdots, X_n$ be a random sample from $Exp(\lambda)$, $\lambda > 0$. Define

$$\eta = P_\lambda(X_1 > a) = e^{-a\lambda}.$$

Then

$$\hat{\eta}_n^{MLE} = \exp(-a/\overline{X}).$$

Here, $g(x) = \exp(-a/x)$ is infinitely differentiable with bounded derivatives. Thus

$$E(\hat{\eta}_n^{MLE}) = \eta + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n} + O(n^{-2})$$

$$Var(\hat{\eta}_n^{MLE}) = \frac{(a\lambda)^2 e^{-2a\lambda}}{n} + O(n^{-2}),$$

the leading terms of which agree with the mean and variance of the leading term in its stochastic expansion. On the other hand, for

$$\hat{\eta}_n^{UMVUE} = \left(1 - \frac{a}{n\overline{x}}\right)^{n-1} I\left(\frac{a}{n\overline{X}} < 1\right),$$

we cannot apply the result for the approximation of moments. But, it can be shown that its variance is also approximated by the same approximation formula.

**Remark 1.2.16.** The boundedness of the derivatives for the approximation of moments is rather stronger than needed. Whenever the approximation can be proved, the formulae agree with the moments of the leading term of its stochastic expansion. So only the validity of the order of the remainder needs to be proved. For example, in the bivariate normal case, the mean and variance of the sample correlation coefficient can be approximated as follows;

$$E(\hat{\rho}_n) = \frac{-\rho(1-\rho^2)}{2n} + O(n^{-2})$$

$$Var(\hat{\rho}_n) = \frac{(1-\rho^2)^2}{n} + O(n^{-2}).$$

## 1.3 Asymptotic Theory

### 1.3.1 MLE in Exponential Family

**Proposition 1.3.1.** *Let $X_1, \cdots, X_n$ be a random sample from a population with pdf*

$$p_\eta(x) = h(x)\exp(\eta^\top T(x) - A(\eta))I_{\mathcal{X}}(x),\ \eta \in \mathcal{E},$$

*where $\mathcal{E}$ is a natural parameter space in $\mathbb{R}^k$. Further, assume that*

*(i) $\mathcal{E}$ is open.*

*(ii) The family is of rank $k$.*

*Then*

*(a)* $\hat{\eta}_n^{MLE} = \eta + (\ddot{A}(\eta))^{-1}(\overline{T}_n - \dot{A}(\eta)) + o_{P_\eta}(n^{-1/2}) = \eta + (-\ddot{l}(\eta))^{-1}\dot{l}(\eta) + o_{P_\eta}(n^{-1/2})$.

*(b)* $\sqrt{n}(\hat{\eta}_n^{MLE} - \eta) \xrightarrow[n\to\infty]{d} N\left(0, (\ddot{A}(\eta))^{-1}\right) = N(0, I_1^{-1}(\eta))$.

*Proof.* Recall that, under the same assumptions,

$$(\overline{T}_n \in C^0) \subseteq (\dot{A}(\hat{\eta}_n^{MLE}) = \overline{T}_n) \subseteq \left(\hat{\eta}_n^{MLE} = (\dot{A})^{-1}(\overline{T}_n)\right),$$

with the probabilities of these events tending to 1 (See theorem 1.1.20). Also note that, by full rank condition, $\dot{A}$ is one-to-one and differentiable on $\mathcal{E}$ with $\ddot{A} > 0$.

Now let $t = \dot{A}(\eta)$. Then by **Inverse Function Theorem** (see next Remark), $\dot{A}^{-1}$ is also differentiable and

$$D(\dot{A}^{-1})(t) = \frac{\partial}{\partial t}\dot{A}^{-1}(t) = \left(\ddot{A}(\eta)\right)^{-1}.$$

CLT implies

$$\sqrt{n}(\overline{T}_n - E_\eta T(X_1)) \xrightarrow[n\to\infty]{d} N\left(0, Var_\eta T(X_1)\right),$$

and recall that $E_\eta T(X_1) = \dot{A}(\eta)$, $Var_\eta T(X_1) = \ddot{A}(\eta)$. Therefore, $\Delta$-method implies

$$\begin{aligned}\hat{\eta}_n^{MLE} &= \dot{A}^{-1}(\overline{T}_n) \\ &= \dot{A}^{-1}(t) + \left(\ddot{A}(\eta)\right)^{-1}(\overline{T}_n - t) + o(|\overline{T}_n - t|) \\ &= \eta + \left(\ddot{A}(\eta)\right)^{-1}(\overline{T}_n - t) + o_{P_\eta}(n^{-1/2}),\end{aligned}$$

and hence

$$\begin{aligned}\sqrt{n}(\hat{\eta}_n^{MLE} - \eta) &= \left(\ddot{A}(\eta)\right)^{-1}\sqrt{n}(\overline{T}_n - t) + o_{P_\eta}(1) \\ &\xrightarrow[n\to\infty]{d} N\left(0, Var_\eta T(X_1)^{-1}\right).\end{aligned}$$

Rest part is obtained from $\overline{T}_n - \dot{A}(\eta) = \dot{l}_n(\eta)/n$ and $-\ddot{l}_n(\eta) = n\ddot{A}(\eta)$. □

**Remark 1.3.2** (Inverse Function Theorem)**.** Let $F : U \to \mathbb{R}^d$, where $U \subseteq \mathbb{R}^d$ is an open set. If

(i) $F$ is one-to-one

(ii) $F$ is Fréchet differentiable near $x_0 \in U$

(iii) $DF_{x_0} := \left[\left.\frac{\partial F}{\partial x_j}\right|_{x=x_0}\right]_{i,j}$ is invertible.

Then $F^{-1} : F(U) \to U$ is also Fréchet differentiable at $y_0 = F(x_0)$, and it satisfies $D(F^{-1})(y_0) = (DF_{x_0})^{-1}$.

**Example 1.3.3** (Multinomial case)**.** Let $X_1, \cdots, X_n$ be a random sample from Multi$(1, p(\theta))$, where $p(\theta) = (p_1(\theta), \cdots, p_k(\theta))^\top$, and $\theta \in \Theta \subseteq \mathbb{R}$. For example, consider Hardy-Weinberg proportions $p(\theta) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)^\top$, $0 < \theta < 1$. Assume that

(i) $\Theta$ is open and $0 < p_i(\theta) < 1$, $\sum_{i=1}^{k} p_i(\theta) = 1$.

(ii) $p(\theta) = (p_1(\theta), \cdots, p_k(\theta))^\top$ is twice (totally) differentiable.

Then we can derive the asymptotic distribution of estimator of $\theta$. Let $\theta = h(p(\theta))$ for any $\theta \in \Theta$ for some differentiable function $h$. Let

$$\hat{p}(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i = \left(\frac{N_1}{n}, \cdots, \frac{N_k}{n}\right)^\top,$$

where $N_j = \sum_{i=1}^{n} I(X_{ij} = 1)$. Then we get

$$E_\theta \hat{p}(\theta) = p(\theta)$$

and hence

$$h(\overline{X}_n) = h(p(\theta)) + \dot{h}(p(\theta))^\top(\overline{X}_n - p(\theta)) + o(|\overline{X}_n - p(\theta)|).$$

Note that $Z_n := \sqrt{n}(\overline{X}_n - p(\theta)) = O_P(1)$, and it has an asymptotic distribution

$$Z_n \xrightarrow[n\to\infty]{d} N(0, \Sigma(\theta)),\ \Sigma(\theta) := \mathrm{diag}(p(\theta)) - p(\theta)p(\theta)^\top,$$

and therefore we get

$$\sqrt{n}(h(\overline{X}_n) - h(p(\theta))) = \dot{h}(p(\theta))^\top Z_n + o_P(1),$$

which implies

$$\sqrt{n}(h(\overline{X}_n) - h(p(\theta))) \xrightarrow[n\to\infty]{d} N(0, \sigma_h^2(\theta)),$$

where

$$\sigma_h^2(\theta) = \dot{h}(p(\theta))^\top \Sigma(\theta) \dot{h}(p(\theta)).$$

Furthermore, we can obtain that

$$\sigma_h^2(\theta) \geq I_1^{-1}(\theta),$$

where equality holds iff

$$\dot{h}(p(\theta))^\top(X_1 - p(\theta)) = I_1^{-1}(\theta)\dot{l}_1(\theta).$$

It can be shown as following. First, note that

$$\dot{h}(p(\theta))^\top \dot{p}(\theta) = 1.$$

It implies that

$$1 = \dot{h}(p(\theta))^\top \frac{\partial}{\partial\theta} E_\theta X_1$$

$$
\begin{aligned}
&= \dot{h}(p(\theta))^\top \int \frac{\partial}{\partial\theta} x f(x:\theta) d\mu(x) \\
&= \dot{h}(p(\theta))^\top \int x \frac{\partial}{\partial\theta} f(x:\theta) d\mu(x) \\
&= \dot{h}(p(\theta))^\top Cov_\theta \left( X_1, \frac{\partial}{\partial\theta} l_1(\theta) \right) \ (\because E_\theta \frac{\partial}{\partial\theta} f(X_1:\theta) = E_\theta \frac{\partial}{\partial\theta} l_1(\theta) = 0) \\
&= Cov_\theta \left( \dot{h}(p(\theta))^\top X_1, \frac{\partial}{\partial\theta} l_1(\theta) \right) \\
&\le \sqrt{Var_\theta \left( \dot{h}(p(\theta))^\top X_1 \right) Var_\theta \left( \frac{\partial}{\partial\theta} l_1(\theta) \right)} \\
&= \sqrt{\dot{h}(p(\theta))^\top \Sigma(\theta) \dot{h}(p(\theta)) \cdot I_1(\theta)}
\end{aligned}
$$

holds. In here, "=" holds when $\dot{h}(p(\theta))^\top X_1$ and $\partial l_1(\theta)/\partial\theta$ has a linear relationship, i.e.,

$$\dot{h}(p(\theta))^\top (X_1 - p(\theta)) = I_1^{-1}(\theta)\dot{l}_1(\theta).$$

**Remark 1.3.4.** Actually, previous example shows *how to deal with asymptotic distribution of FSE*, more generally.

### 1.3.2 Asymptotic Normality of MCE

Our real goal of this section is right here.

**Theorem 1.3.5** (Asymptotic Normality of MCE)**.** *Let $X_1, \cdots, X_n$ be a random sample from $P_\theta$, where $\theta \in \Theta$ and parameter space $\Theta$ is open in $\mathbb{R}^k$. Let*

$$\hat{\theta}_n^{MCE} = \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \rho(X_i, \theta)$$

$$\theta_0 = \arg\min_{\theta\in\Theta} E_{\theta_0} \rho(X_1, \theta).$$

*Under the assumption of their existence, let*

$$
\begin{aligned}
&\Psi_1(\theta) = \Psi(X_1, \theta) = \frac{\partial}{\partial\theta}\rho(X_1, \theta), && \overline{\Psi}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \Psi(X_i, \theta) \\
&\dot{\Psi}_1(\theta) = \frac{\partial \Psi_1(\theta)}{\partial\theta}, && \dot{\overline{\Psi}}_n(\theta) = \frac{\partial \overline{\Psi}_n(\theta)}{\partial\theta} \\
&\ddot{\Psi}_1(\theta) = \frac{\partial^2 \Psi_1(\theta)}{\partial\theta^2}, && \ddot{\overline{\Psi}}_n(\theta) = \frac{\partial^2 \overline{\Psi}_n(\theta)}{\partial\theta^2}
\end{aligned}
$$

*Assume that*

*(A0)* $P_{\theta_0}\left(\overline{\Psi}_n(\hat{\theta}_n^{MCE})=0\right) \xrightarrow[n\to\infty]{} 1.$

*(A1)* $E_{\theta_0}\Psi_1(\theta_0)=0.$

*(A2)* $Var_{\theta_0}(\Psi_1(\theta_0))$ *exists.*

*(A3)* $E_{\theta_0}(\dot{\Psi}_1(\theta_0))$ *exists and is nonsingular.*

*(A4)* $\exists\delta>0$ *and* $\exists M(X_1)=M_{\theta_0,\delta}(X_1)$ *s.t.*

$$\max_{\substack{|\theta-\theta_0|\le\delta\\ \theta\in\Theta}}|\ddot{\Psi}_1(\theta)|\le M(X_1),\ \textit{where } E_{\theta_0}M(X_1)<\infty.$$

*(A5)* $\hat{\theta}_n^{MCE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$ *as* $n\to\infty$.

*Then*

$$\hat{\theta}_n^{MCE}=\theta_0+\left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1}\overline{\Psi}_n(\theta_0)+o_{P_{\theta_0}}(n^{-1/2}),$$

*so that*

$$\sqrt{n}(\hat{\theta}_n^{MCE}-\theta_0)\xrightarrow[n\to\infty]{d} N\left(0,\left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1}Var_{\theta_0}(\Psi_1(\theta_0))\left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1}\right)\ \textit{under } P_{\theta_0}.$$

**Remark 1.3.6** (Gradient of vector map)**.** Let $F:\mathbb{R}^k\to\mathbb{R}^d$ be a smooth function, where

$$F(x)=(F_1(x),\cdots,F_d(x))^\top.$$

(1) (1st-order gradient)

$$\frac{\partial F}{\partial x}(x_0):=DF(x_0)=\left(\frac{\partial F}{\partial x_1}(x_0),\frac{\partial F}{\partial x_2}(x_0),\cdots,\frac{\partial F}{\partial x_k}(x_0)\right)\in\mathbb{R}^{d\times k},$$

where $\dfrac{\partial F}{\partial x_j}(x_0)=\left(\dfrac{\partial F_1}{\partial x_j}(x_0),\cdots,\dfrac{\partial F_d}{\partial x_j}(x_0)\right)^\top$ is a column vector. It can be interpreted as "a linear map."

(2) (2nd-order gradient)

$$\frac{\partial^2 F}{\partial x^2}(x_0):=D^2F(x_0)=\left(\frac{\partial}{\partial x_1}\frac{\partial F}{\partial x}(x_0),\cdots,\frac{\partial}{\partial x_k}\frac{\partial F}{\partial x}(x_0)\right)\in\mathbb{R}^{d\times k\times k},$$

where $\dfrac{\partial}{\partial x_i}\dfrac{\partial F}{\partial x}(x_0)$ is $d\times k$ matrix with $\dfrac{\partial^2 F}{\partial x_i x_j}(x_0)$ as the $j$th column vector. Note that it can be interpreted as "a bi-linear map."

(3) (Taylor expansion of vector-valued map)

$$\begin{aligned} F(x) &\approx F(x_0) + \sum_{j=1}^{k} \frac{\partial F}{\partial x_j}(x_0)(x_j - x_{0j}) + \frac{1}{2}\sum_{i,j} \frac{\partial^2 F}{\partial x_i \partial x_j}(x_0)(x_i - x_{0i})(x_j - x_{0j}) \\ &= F(x_0) + \underbrace{\frac{\partial F}{\partial x_j}(x_0)}_{\text{matrix}} \underbrace{(x - x_0)}_{\text{vector}} + \frac{1}{2}(x - x_0)^\top \underbrace{\frac{\partial^2 F}{\partial x_i \partial x_j}(x_0)}_{\text{3-array}}(x - x_0). \end{aligned}$$

In here, "matrix $DF(x_0)$ × vector" becomes a vector, and "quadratic form with 3-array $D^2F(x_0)$" becomes vector-valued. In this view, $DF(x_0)$ and $D^2F(x_0)$ can be interpreted as a linear and bi-linear map, respectively.

*Proof.* By Taylor's theorem, $\exists \theta_n^*$ in line$(\theta_0, \hat{\theta}_n)$ such that $\|\theta_n^* - \theta_0\| \le \|\hat{\theta}_n - \theta_0\|$ and

$$\overline{\Psi}_n(\hat{\theta}_n) = \overline{\Psi}_n(\theta_0) + \dot{\overline{\Psi}}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\overline{\Psi}}_n(\theta_n^*)(\hat{\theta}_n - \theta_0).$$

Also note that, from (A4) and (A5),

$$\lim_{K\to\infty} \sup_n P_{\theta_0}\left(\ddot{\overline{\Psi}}_n(\theta_n^*) \ge K\right) = 0$$

holds, and hence

$$\ddot{\overline{\Psi}}_n(\theta_n^*) = O_{P_{\theta_0}}(1),$$

i.e.,

$$\ddot{\overline{\Psi}}_n(\theta_n^*)(\hat{\theta}_n - \theta_0) = o_{P_{\theta_0}}(1).$$

($\because$ (Motivation: since $\ddot{\overline{\Psi}}_n$ is dominated by integrable majorant, it is "uniformly integrable," hence it is "tight.") From

$$\begin{aligned} P_{\theta_0}\left(\left|\ddot{\overline{\Psi}}_n(\theta_n^*)\right| > K\right) &= P_{\theta_0}\left(\left|\ddot{\overline{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| \le \delta\right) + P_{\theta_0}\left(\left|\ddot{\overline{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| > \delta\right) \\ &\le P_{\theta_0}\left(\left|\ddot{\overline{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| \le \delta\right) + P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| > \delta\right) \\ &\underset{\text{(A4)}}{\le} P_{\theta_0}\left(M(X_1) > K\right) + \underbrace{P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| > \delta\right)}_{\xrightarrow[n\to\infty]{\text{(A5)}} 0}, \end{aligned}$$

for any $\epsilon > 0$ we get for large $N$

$$\sup_{n>N} P_{\theta_0}\left(|\ddot{\overline{\Psi}}_n(\theta_n^*)| > K\right) \le \frac{1}{K} E_{\theta_0} M(X_1) + \epsilon,$$

which implies

$$\lim_{K\to\infty}\sup_{n>N} P_{\theta_0}\left(|\ddot{\overline{\Psi}}_n(\theta_n^*)| > K\right) = 0.$$

(Let $N$ be larger and take $\epsilon \searrow 0$)) Thus, we get

$$\begin{aligned}
\overline{\Psi}_n(\hat{\theta}_n) &= \overline{\Psi}_n(\theta_0) + \dot{\overline{\Psi}}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\overline{\Psi}}_n(\theta_n^*)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(\dot{\overline{\Psi}}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\overline{\Psi}}_n(\theta_n^*)\right)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(\dot{\overline{\Psi}}_n(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\theta}_n - \theta_0).
\end{aligned}$$

Now, note that

(1) $P_{\theta_0}(\overline{\Psi}_n(\hat{\theta}_n) = 0) \xrightarrow[n\to\infty]{} 1$.

(2) $E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)$ is nonsingular with probability 1 (See remark 1.3.7)

(3) If $X_n = Y_n + O_P(a_n)$ on $\mathcal{E}_n$ with $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$, then $X_n = Y_n + O_P(a_n)$ (on whole space), and the same holds for $o_P$ (See remark 1.2.7).

Thus, on the set with probability tending to 1,

$$\begin{aligned}
\hat{\theta}_n - \theta_0 &= \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)\right)^{-1}\overline{\Psi}_n(\theta_0) \\
&= \left[\left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1} + o_{P_{\theta_0}}(1)\right]\overline{\Psi}_n(\theta_0)
\end{aligned}$$

holds, which yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1}\underbrace{\sqrt{n}\overline{\Psi}_n(\theta_0)}_{O_{P_{\theta_0}}(1)} + o_{P_{\theta_0}}(1)$$

and therefore

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{d} N\left(0, \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1} Var_{\theta_0}\Psi_1(\theta_0)\left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1}\right).$$

□

**Remark 1.3.7.** If $A \in \mathbb{R}^{d\times d}$ is symmetric positive definite matrix, then for small perbutation $\Delta$ s.t. $\|\Delta\|_2 < \sigma_{min}(A)$, $\text{rank}(A+\Delta) = d$. Note that $\text{rank}(A) = d$. In here, $\sigma_{min}(A)$ denotes the

smallest eigenvalue of $A$, and $\|\cdot\|_p$ is a matrix norm induced by corresponding $\mathcal{L}^p$ vector norm, i.e.,

$$\|\Delta\|_p = \sup_{x:\|x\|_p=1} \|\Delta x\|_p.$$

*Proof.* (Motivation: for $c \approx 0$ and $x \neq 0$,

$$\frac{1}{x+c} = \frac{x}{c} - \frac{x^2}{c^2} + \frac{x^3}{c^3} - \cdots$$

exists, or for small vectors $u, v$, $A + uv^\top$ is nonsingular from

$$(A+uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+u^\top A^{-1}v},$$

if $A$ is invertible) Let $\text{rank}(A+\Delta) < d$. Then $\exists x_0 \neq 0$ s.t. $(A+\Delta)x_0 = 0$ and $\|x_0\|_2 = 1$. Then by definition of matrix norm,

$$\|\Delta\|_2 \geq \|\Delta x_0\|_2 = \|Ax_0\|_2 \geq \sigma_{min}(A)$$

holds. The last inequality is from spectral theorem. It is contradictory to our assumption that $\Delta$ is small. □

### 1.3.3 Asymptotic Normality and Efficiency of MLE

Note that MLE is just a special case of MCE.

**Theorem 1.3.8.** *Let $X_1, \cdots, X_n$ be a random sample from $P_\theta$, where $\theta \in \Theta$ and parameter space $\Theta$ is open in $\mathbb{R}^k$. Recall that MLE is an MCE with*

$$\rho(x,\theta) = -\log p_\theta(x),\ p_\theta : \textit{pdf of } P_\theta.$$

*Under the assumption of their existence, denote*

$$\dot{l}_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\log p_\theta(X_i), \qquad \bar{\dot{l}}_n(\theta) = \frac{1}{n}\dot{l}_n(\theta)$$
$$\ddot{l}_n(\theta) = \frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i), \qquad \bar{\ddot{l}}_n(\theta) = \frac{1}{n}\ddot{l}_n(\theta)$$
$$\dddot{l}_n(\theta) = \frac{\partial^3}{\partial\theta^3}\log p_\theta(X_i), \qquad \bar{\dddot{l}}_n(\theta) = \frac{1}{n}\dddot{l}_n(\theta).$$

*Also assume that*

*(M0)* $P_{\theta_0}\left(\dot{l}_n(\hat{\theta}_n^{MLE})=0\right) \xrightarrow[n\to\infty]{} 1.$

*(M1)* $E_{\theta_0}\dot{l}_1(\theta_0)=0.$

*(M2)* $I(\theta_0)=Var_{\theta_0}(\dot{l}_1(\theta_0))$ *exists.*

*(M3)* $E_{\theta_0}(\ddot{l}_1(\theta_0))$ *exists and is nonsingular.*

*(M4)* $\exists\delta_0>0$ *and* $\exists M(X_1)=M_{\theta_0,\delta}(X_1)$ *s.t.*

$$\max_{\substack{|\theta-\theta_0|\le\delta_0\\ \theta\in\Theta}} |\dddot{l}_1(\theta)| \le M(X_1), \text{ where } E_{\theta_0}M(X_1)<\infty.$$

*(M5)* $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$ *as* $n\to\infty$.

*(M6)* $I(\theta_0)=E_{\theta_0}(-\ddot{l}_1(\theta_0))$.

*Under (M0)* $\sim$ *(M6),*

$$\hat{\theta}_n^{MLE}=\theta_0+I(\theta_0)^{-1}\bar{\dot{l}}_n(\theta_0)+o_{P_{\theta_0}}(n^{-1/2}),$$

*so that*

$$\sqrt{n}(\hat{\theta}_n^{MLE}-\theta_0) \xrightarrow[n\to\infty]{d} N\left(0, I(\theta_0)^{-1}\right).$$

Even though MCE is more general version of MLE, we frequently use MLE to estimate the parameter. Following theorem says that, "MLE is more (asymptotically) efficient than MCE," i.e., log contrast function makes MCE the most efficient.

**Theorem 1.3.9** (Asymptotic efficiency of MLE)**.** *Assume (A0)* $\sim$ *(A6), and (M0)* $\sim$ *(M6) hold, where*

$$\textit{(A6)}: E_{\theta_0}\dot{\Psi}_1(\theta_0) = -E_{\theta_0}\dot{l}_1(\theta_0)\Psi_1^{\top}(\theta_0).$$

*Then*

$$\sqrt{n}(\hat{\theta}_n^{MLE}-\theta_0) \xrightarrow[n\to\infty]{d} N(0, I(\theta_0)^{-1})$$

$$\sqrt{n}(\hat{\theta}_n^{MCE}-\theta_0) \xrightarrow[n\to\infty]{d} N(0, \Sigma_\Psi(\theta_0))$$

*with* $\Sigma_\Psi(\theta_0)-I(\theta_0)^{-1}$ *being nonnegative definite (i.e., "* $\Sigma_\Psi(\theta_0)\ge I(\theta_0)^{-1}$ *"), and "=" holds if and only if*

$$\Psi_1(\theta_0)=(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1(\theta_0)^{-1}\dot{l}_1(\theta_0). \qquad \text{("Determining contrast function")}$$

*Proof.* By Cauchy-Schwarz inequality,

$$\begin{aligned} Corr(\lambda^\top \dot{l}_1(\theta_0), \gamma^\top \Psi_1(\theta_0)) &= \frac{\lambda^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}} \\ &= \frac{\lambda^\top I_1^{1/2}(\theta_0) I_1^{-1/2}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}} \\ &\leq \frac{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}} \\ &= \frac{\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}} \end{aligned}$$

holds, and hence

$$\max_{\lambda\neq 0} Corr(\lambda^\top \dot{l}_1(\theta_0), \gamma^\top \Psi_1(\theta_0)) = \frac{\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}$$

is obtained (we can find $\lambda$ that achieves maximum). Since correlation coefficient is less than 1, we get

$$\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma \leq \gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma$$

for any $\gamma$. Using $(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}\gamma$ instead of $\gamma$, we obtain that

$$\gamma^\top I_1^{-1}(\theta_0)\gamma \leq \gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} Var_{\theta_0}\Psi_1(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}\gamma,$$

i.e.,

$$I_1^{-1}(\theta_0) \leq (-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} Var_{\theta_0}\Psi_1(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} = \Sigma_\Psi(\theta_0).$$

Equality holds iff

$$\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} I_1^{-1}(\theta_0)\dot{l}_1(\theta_0) = \gamma^\top\Psi_1(\theta_0),$$

i.e.,

$$\Psi_1(\theta_0) = (-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} I_1^{-1}(\theta_0)\dot{l}_1(\theta_0).$$

□

**Remark 1.3.10.** The claim that the MLE has smaller variance than other asymptotically normal estimators was known as *Fisher's conjecture*. This is true for a certain class of estimators in a *regular* parametric model. Essential property for such a comparison is the "*uniform*" convergence to the normal distribution as it can be seen in the following example.

**Example 1.3.11** (Hodge's example: "Superefficient estimator")**.** Let $\overline{X}$ be the sample mean in $N(\theta, 1)$. Let

$$\hat{\theta}_n^s = \overline{X} I(|\overline{X}| > n^{-1/4}).$$

(Actually, not only 1/4, but any positive number less than 1/2 is OK.)

If $\theta \neq 0$, then

$$\begin{aligned} P_\theta\left(\hat{\theta}_n^s = \overline{X}\right) &= 1 - P_\theta(|\overline{X}| \leq n^{-1/4}) \\ &= 1 - P_\theta(-n^{-1/4} \leq \overline{X} \leq n^{-1/4}) \\ &= 1 - P_\theta(-\sqrt{n}\theta - n^{1/4} \leq \sqrt{n}(\overline{X} - \theta) \leq -\sqrt{n}\theta + n^{1/4}) \\ &= 1 - \underbrace{\Phi\left(-\sqrt{n}\theta + n^{1/4}\right)}_{\xrightarrow[n\to\infty]{}0 \text{ if } \theta>0;1 \text{ if } \theta<0} + \underbrace{\Phi\left(-\sqrt{n}\theta - n^{1/4}\right)}_{\xrightarrow[n\to\infty]{}0 \text{ if } \theta>0;1 \text{ if } \theta<0} \\ &\xrightarrow[n\to\infty]{} 1 \end{aligned}$$

holds. Intuitively, if $\theta \neq 0$, $\overline{X}$ converges to nonzero value with $\sqrt{n}$-rate, so $I(\overline{X} > n^{-1/4})$ becomes 1. Thus,

$$\begin{aligned} P_\theta\left(\sqrt{n}(\hat{\theta}_n^s - \theta) \leq x\right) &= P_\theta\left(\sqrt{n}(\hat{\theta}_n^s - \theta) \leq x, \hat{\theta}_n^s = \overline{X}\right) + P_\theta\left(\sqrt{n}(\hat{\theta}_n^s - \theta) \leq x, \hat{\theta}_n^s \neq \overline{X}\right) \\ &= P_\theta\left(\sqrt{n}(\overline{X} - \theta) \leq x, \hat{\theta}_n^s = \overline{X}\right) + P_\theta\left(\sqrt{n}(\hat{\theta}_n^s - \theta) \leq x, \hat{\theta}_n^s \neq \overline{X}\right) \\ &\xrightarrow[n\to\infty]{} \Phi(x) \end{aligned}$$

holds, i.e.,

$$\sqrt{n}(\hat{\theta}_n^s - \theta) \xrightarrow[n\to\infty]{d} N(0,1) \text{ under } P_\theta.$$

Now, assume that $\theta = 0$. Then $\sqrt{n}\overline{X} \sim N(0,1)$, so

$$P_\theta(\hat{\theta}_n^s = 0) = P_\theta(|\overline{X}| \leq n^{-1/4}) = \Phi(n^{-1/4}) - \Phi(-n^{-1/4}) \xrightarrow[n\to\infty]{} 1.$$

Then we obtain $\lim_{n\to\infty} P_\theta\left(\hat{\theta}_n^s = 0\right) = 1$, and therefore,

$$\begin{aligned} \lim_{n\to\infty} P_\theta\left(\sqrt{n}(\hat{\theta}_n^s - \theta) \leq x\right) &= \lim_{n\to\infty} P_\theta(0 \leq x) \\ &= \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \end{aligned}$$

$$= I_{[0,\infty)}(x),$$

which yields

$$\sqrt{n}(\hat{\theta}_n^s - \theta) \xrightarrow[n\to\infty]{d} N(0,0) \text{ under } P_\theta.$$

Thus, we get

$$\sqrt{n}(\hat{\theta}_n^s - \theta) \xrightarrow[n\to\infty]{d} N(0, \sigma_s^2(\theta)) \text{ under } P_\theta,$$

where

$$\sigma_s^2(\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}.$$

It implies that, there is a "superefficient" estimator than "optimal" one with variance $I(\theta)^{-1}$, i.e., *Fisher's conjecture* is wrong!

**Remark 1.3.12.** Note that Fisher's conjecture is wrong in general, but it is "partially correct" in "regular" model. It means "uniform convegence" of CLT. First note that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d\ (P_\theta)} N(0, \nu(\theta))$$

only means "pointwise" convergence *for each* $\theta$. It is equivalent to, for some metric d, and law (distribution function) $L_\theta$ under $P_\theta$,

$$d\left(L_\theta(\sqrt{n}(\hat{\theta}_n - \theta)), \Phi\left(\frac{\cdot}{\nu(\theta)}\right)\right) \xrightarrow[n\to\infty]{} 0.$$

"Regular estimator" means, on the other hand, that

$$\sup_{\theta:|\theta-\theta_0|<Mn^{-1/2}} d\left(L_\theta(\sqrt{n}(\hat{\theta}_n - \theta)), \Phi\left(\frac{\cdot}{\nu(\theta_0)}\right)\right) \xrightarrow[n\to\infty]{} 0$$

holds *for any* $\theta_0 \in \Theta$. Essentially it means that *for any sequence* $\{\theta_n\}$ s.t. $\sqrt{n}|\theta_n - \theta| \leq M$, we get

$$P_{\theta_n}\left(\sqrt{n}(\hat{\theta}_n - \theta_n) \leq x\right) \xrightarrow[n\to\infty]{} \Phi\left(\frac{x}{\nu(\theta)}\right).$$

In Hodge's example, $\hat{\theta}_n^s$ is not regular, because cases '$\theta = 0$' and '$\theta \approx 0$' are different. Take $\{\theta_n\}$ s.t. $\theta_n \to 0$ at $\sqrt{n}$-rate, e.g.,

$$\theta_n = \frac{a_0}{\sqrt{n}}.$$

Then

$$\sqrt{n}\left(\hat{\theta}_n^s - \frac{a_0}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{d\ (P_{a_0/\sqrt{n}})} -a_0,$$

and "limiting distribution depends on $a_0$," i.e., "the sequence $\{\theta_n\}$."

**Example 1.3.13** (Linear model with stochastic covariates)**.** Consider the model

$$Y|Z = z \sim N(\alpha + z^\top\beta, \sigma^2),\ \alpha \in \mathbb{R},\ \beta \in \mathbb{R}^k,\ \sigma^2 > 0,$$

where $E(Z) = 0$ and $Var(Z)$ is non-singular. Assume that there are $n$ independent copies of $Y$ and $Z$, and denote $Z_{(n)} = (Z_1, \cdots, Z_n)^\top$. Then as long as the distribution of $Z$ does not depend on $\theta = (\alpha, \beta, \sigma^2)$, we get MLE of $\beta$ is equivalent to least square procedure, from

$$l_n(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \alpha - z_i^\top\beta)^2 - \frac{n}{2}\log 2\pi\sigma^2 + \sum_{i=1}^n \log pdf_Z(z_i).$$

In other words,

$$\begin{aligned}
\hat{\beta}^{MLE} &= (\tilde{Z}_{(n)}^\top \tilde{Z}_{(n)})^{-1}\tilde{Z}_{(n)}^\top Y,\ \tilde{Z}_{(n)} = (Z_1 - \overline{Z}, \cdots, Z_n - \overline{Z})^\top \\
\hat{\alpha}^{MLE} &= \overline{Y} - \overline{Z}^\top\hat{\beta}^{MLE} \\
\hat{\sigma^2}^{MLE} &= \frac{1}{n}\|Y - (\mathbf{1}\hat{\alpha}^{MLE} + Z_{(n)}\hat{\beta}^{MLE})\|^2.
\end{aligned}$$

It also says that,

$$l_1(\theta) = -\frac{1}{2\sigma^2}(Y_1 - \alpha - z_1^\top\beta)^2 - \frac{1}{2}\log 2\pi\sigma^2 + \log pdf_Z(z_1)$$

and hence

$$\dot{l}_1(\theta) = \left(\frac{\epsilon_1}{\sigma^2}, z_1^\top\frac{\epsilon_1}{\sigma^2}, \frac{\epsilon_1^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right)^\top,\ \text{where } \epsilon_1 = Y_1 - \alpha - z_1^\top\beta.$$

Thus

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & & \\ & \frac{1}{\sigma^2}E(Z_1Z_1^\top) & \\ & & \frac{1}{2\sigma^4} \end{pmatrix}.$$

(It can be easily obtained from $\epsilon_1|z_1 \sim N(0, \sigma^2)$. Note that

$$Var(\epsilon_1) = EVar(\epsilon_1|z_1) + VarE(\epsilon_1|z_1) = \sigma^2,$$

$$Var(z_1\epsilon_1) = EVar(z_1\epsilon_1|z_1) + VarE(z_1\epsilon_1|z_1) = E(z_1\sigma^2 z_1^\top) + Varz_1 \cdot 0,$$

and similarly we can obtain $Var(\epsilon_1^2)$.) It implies that,

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta) \xrightarrow[n\to\infty]{d} N(0, \Sigma),$$

where

$$\Sigma = I(\theta)^{-1} = \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 \left(E(Z_1 Z_1^\top)\right)^{-1} & \\ & & 2\sigma^4 \end{pmatrix},$$

*provided that* $E(Z_1 Z_1^\top)$, or equivalently $Var(Z_1)$, is non-singular.

### 1.3.4 Asymptotic Null distribution of LRT

Consider a random sample $X_1, \cdots, X_n$ from $P_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. Denote $\theta = (\xi^\top, \eta^\top)^\top$, where $\eta \in \mathbb{R}^{k_0}$. We wish to test

$$H_0 : \xi = \xi_0 \; vs \; H_1 : \xi \neq \xi_0.$$

(Note that it is composite null!) Let $\Theta$ be a $k$-dimensional open set, and

$$\Theta_0 = \{(\xi_0^\top, \eta^\top)^\top : (\xi_0^\top, \eta^\top)^\top \in \Theta\}$$

be a $k_0$-dimensional open set. Now denote

$$\dot{l}(\theta) = \begin{pmatrix} \dot{l}_1(\theta) \\ \dot{l}_2(\theta) \end{pmatrix}, \; I(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix},$$

where

$$\dot{l}_1(\theta) = \frac{\partial}{\partial \xi} l(\theta), \; \dot{l}_2(\theta) = \frac{\partial}{\partial \eta} l(\theta).$$

**Theorem 1.3.14** (Asymptotic null distribution of LRT)**.** *Assume (M0) $\sim$ (M6). Then under* $H_0 : \xi = \xi_0$

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) \xrightarrow[n\to\infty]{d} \chi^2(k - k_0)$$

*where*

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} l(\theta), \; \hat{\theta}_n^0 = \arg\max_{\theta\in\Theta_0} l(\theta).$$

*Proof.* Recall that,

$$l(\theta_0) = l(\hat{\theta}_n) + \frac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta_0)^\top \left[\frac{1}{n}\ddot{l}(\theta_0) + o_{P_{\theta_0}}(1)\right]\sqrt{n}(\hat{\theta}_n - \theta_0)$$

holds. Since $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_{P_{\theta_0}}(1)$, and $O_{P_{\theta_0}}(1) \cdot o_{P_{\theta_0}}(1) = o_{P_{\theta_0}}(1)$, we get

$$2(l(\hat{\theta}_n) - l(\theta_0)) = \sqrt{n}(\hat{\theta}_n - \theta_0)^\top I(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_{P_{\theta_0}}(1),$$

from

$$-\frac{1}{n}\ddot{l}(\theta_0) \xrightarrow[n\to\infty]{P_{\theta_0}} I(\theta_0).$$

Also recall that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}(\theta_0) + o_{P_{\theta_0}}(1).$$

It implies that

$$2(l(\hat{\theta}_n) - l(\theta_0)) = \sqrt{n}\bar{\dot{l}}(\theta_0)^\top I(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}(\theta_0) + o_{P_{\theta_0}}(1).$$

Similarly, we get

$$2(l(\hat{\theta}_n^0) - l(\theta_0)) = \sqrt{n}(\hat{\eta}_n^0 - \eta_0)^\top I_{22}(\theta_0)\sqrt{n}(\hat{\eta}_n^0 - \eta_0) + o_{P_{\theta_0}}(1)$$

and

$$\sqrt{n}(\hat{\eta}_n^0 - \eta_0) = I_{22}(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(1)$$

under $H_0$, so we get

$$2(l(\hat{\theta}_n^0) - l(\theta_0)) = \sqrt{n}\bar{\dot{l}}_2(\theta_0)^\top I_{22}(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(1)$$

under $H_0$. Thus, we get

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) = \sqrt{n}\bar{\dot{l}}(\theta_0)^\top I(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}(\theta_0) - \sqrt{n}\bar{\dot{l}}_2(\theta_0)^\top I_{22}(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(1)$$

under $H_0$. Now note that

$$I(\theta_0)^{-1} = \begin{pmatrix} I_{11}(\theta_0) & I_{12}(\theta_0) \\ I_{21}(\theta_0) & I_{22}(\theta_0) \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} J_1 & 0 \\ -I_{22}^{-1}(\theta_0)I_{21}(\theta_0) & J_2 \end{pmatrix} \begin{pmatrix} I_{11\cdot 2}^{-1}(\theta_0) & 0 \\ 0 & I_{22}^{-1}(\theta_0) \end{pmatrix} \begin{pmatrix} J_1 & -I_{12}(\theta_0)I_{22}^{-1}(\theta_0) \\ 0 & J_2 \end{pmatrix}$$

holds, for identity matrices $J_1$ and $J_2$ with suitable sizes. Then

$$\begin{aligned} \bar{\dot{l}}(\theta_0)^\top I(\theta_0)^{-1}\bar{\dot{l}}(\theta_0) &= \bar{\dot{l}}(\theta_0)^\top \begin{pmatrix} J_1 & 0 \\ -I_{22}^{-1}(\theta_0)I_{21}(\theta_0) & J_2 \end{pmatrix} \begin{pmatrix} I_{11\cdot 2}^{-1}(\theta_0) & 0 \\ 0 & I_{22}^{-1}(\theta_0) \end{pmatrix} \begin{pmatrix} J_1 & -I_{12}(\theta_0)I_{22}^{-1}(\theta_0) \\ 0 & J_2 \end{pmatrix} \bar{\dot{l}}(\theta_0) \\ &= \begin{pmatrix} \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0) \\ \bar{\dot{l}}_2(\theta_0) \end{pmatrix}^\top \begin{pmatrix} I_{11\cdot 2}^{-1}(\theta_0) & 0 \\ 0 & I_{22}^{-1}(\theta_0) \end{pmatrix} \begin{pmatrix} \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0) \\ \bar{\dot{l}}_2(\theta_0) \end{pmatrix} \end{aligned}$$

holds, so we get

$$\begin{aligned} 2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) = \sqrt{n}\left(\bar{\dot{l}}_1(\theta_0) - I_{21}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0)\right)^\top I_{11\cdot 2}^{-1}(\theta_0)\sqrt{n}\left(\bar{\dot{l}}_1(\theta_0) - I_{21}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0)\right) \\ + o_{P_{\theta_0}}(1). \qquad (1.6) \end{aligned}$$

Now by CLT,

$$\begin{aligned} \sqrt{n}(\bar{\dot{l}}_1(\theta_0) - I_{21}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0)) &\xrightarrow[n\to\infty]{d} \left(J_1 \quad -I_{12}(\theta_0)I_{22}^{-1}(\theta_0)\right) N(0, I(\theta_0)) \\ &= N(0, I_{11}(\theta_0) - I_{12}(\theta_0)I_{22}^{-1}(\theta_0)I_{21}(\theta_0)) \\ &= N(0, I_{11\cdot 2}(\theta_0)), \end{aligned}$$

and hence,

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) \xrightarrow[n\to\infty]{d} \chi^2(k - k_0).$$

□

**Remark 1.3.15.** (i) The extension to a null hypothesis

$$H_0 : g_1(\theta) = 0, \cdots, g_{k_1}(\theta) = 0$$

is rather trivial by considering a smooth reparametrization

$$\xi = g_1(\theta), \cdots, \xi_{k_1}(\theta) = g_{k_1}(\theta),\ \eta_1 = g_{k_1+1}(\theta), \cdots, \eta_{k_0} = g_k(\theta)$$

whenever such $g_j(\theta)$ $(j = k_1 + 1, \cdots, k)$ can be found.

(ii) Large sample confidence set based on the maximum likelihood is obtained by duality.

**Remark 1.3.16.** Note that to perform LRT we should find MLE on $\Theta$ and $\Theta_0$, however, it may not be so easy. Thus our interest is to find *asymptotic equivalent tests* of LRT.

### 1.3.5 Asymptotic Equivalents of LRT

Let $X_1, \cdots, X_n$ be a random sample from $P_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. For $\theta = (\xi^\top, \eta^\top)^\top \in \Theta$, we again wish to test

$$H_0 : \xi = \xi_0 \; vs \; H_1 : \xi \neq \xi_0.$$

We assume the regularity conditions, and in addition, that $I(\theta)$ is continuous. Also, we denotethe true parameter under $H_0$ as $\theta_0 = (\xi_0^\top, \eta_0^\top)^\top$.

**Lemma 1.3.17.** *(a) From*

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) = I(\theta_0)^{-1}\sqrt{n}\bar{\dot{l}}(\theta_0) + o_{P_{\theta_0}}(1),$$

*we get*

$$I_{11}(\theta_0)(\hat{\xi}_n - \xi_0) + I_{12}(\theta_0)(\hat{\eta}_n - \eta_0) = \bar{\dot{l}}_1(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}),$$

$$I_{21}(\theta_0)(\hat{\xi}_n - \xi_0) + I_{22}(\theta_0)(\hat{\eta}_n - \eta_0) = \bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}).$$

*(b)*

$$\sqrt{n}(\hat{\eta}_n^0 - \eta_0) = I_{22}^{-1}(\theta_0)\sqrt{n}\bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(1)$$

$$\hat{\eta}_n^0 - \eta_0 = I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2})$$

*(c)* $2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) = \sqrt{n}(\bar{\dot{l}}_1(\theta_0) - I_{21}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0))^\top I_{11\cdot 2}^{-1}(\theta_0)\sqrt{n}(\bar{\dot{l}}_1(\theta_0) - I_{21}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0)) + o_{P_{\theta_0}}(1)$.

**Theorem 1.3.18** (Wald's test)**.** *Let*

$$W_n = \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)^\top I(\hat{\theta}_n^0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0).$$

*Then,*

$$\begin{aligned} W_n &= \sqrt{n}(\hat{\xi}_n - \xi_0)^\top I_{11\cdot 2}(\theta_0)\sqrt{n}(\hat{\xi}_n - \xi_0) + o_{P_{\theta_0}}(1) \\ &= 2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) + o_{P_{\theta_0}}(1). \end{aligned}$$

*Proof.* First note that, by continuity of $I$,

$$\begin{aligned} W_n &= n(\hat{\theta}_n - \hat{\theta}_n^0)^\top \left( I(\theta_0) + o_{P_{\theta_0}}(1) \right) (\hat{\theta}_n - \hat{\theta}_n^0) \\ &= n(\hat{\theta}_n - \hat{\theta}_n^0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) + o_{P_{\theta_0}}(1). \end{aligned}$$

Then, we get

$$\begin{aligned} W_n &= n(\hat{\theta}_n - \hat{\theta}_n^0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) + o_{P_{\theta_0}}(1) \\ &= n \Big[ (\hat{\xi}_n - \xi_0)^\top I_{11}(\theta_0)(\hat{\xi}_n - \xi_0) + 2(\hat{\xi}_n - \xi_0)^\top I_{12}(\theta_0)(\hat{\eta}_n - \hat{\eta}_n^0) \\ &\quad + (\hat{\eta}_n - \hat{\eta}_n^0)^\top I_{22}(\theta_0)(\hat{\eta}_n - \hat{\eta}_n^0) \Big] + o_{P_{\theta_0}}(1) \\ &= n \Big[ (\hat{\xi}_n - \xi_0)^\top \left( I_{11}(\theta_0)(\hat{\xi}_n - \xi_0) + I_{12}(\theta_0)(\hat{\eta}_n - \hat{\eta}_n^0) \right) \\ &\quad + \left( (\hat{\xi}_n - \xi_0)^\top I_{12}(\theta_0) + (\hat{\eta}_n - \hat{\eta}_n^0)^\top I_{22}(\theta_0) \right) (\hat{\eta}_n - \hat{\eta}_n^0) \Big] + o_{P_{\theta_0}}(1), \end{aligned}$$

and from

$$\begin{aligned} I_{11}(\theta_0)(\hat{\xi}_n - \xi_0) + I_{12}(\theta_0)(\hat{\eta}_n - \hat{\eta}_n^0) &= I_{11}(\theta_0)(\hat{\xi}_n - \xi_0) + I_{12}(\theta_0)(\hat{\eta}_n - \eta_0) - I_{12}(\theta_0)(\hat{\eta}_n^0 - \eta_0) \\ &= \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0)(\hat{\eta}_n^0 - \eta_0) + o_{P_{\theta_0}}(n^{-1/2}) \\ &= \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}) \end{aligned}$$

and

$$\begin{aligned} (\hat{\xi}_n - \xi_0)^\top I_{12}(\theta_0) + (\hat{\eta}_n - \hat{\eta}_n^0)^\top I_{22}(\theta_0) &= (\hat{\xi}_n - \xi_0)^\top I_{12}(\theta_0) + (\hat{\eta}_n - \eta_0)^\top I_{22}(\theta_0) - (\hat{\eta}_n^0 - \eta_0)^\top I_{22}(\theta_0) \\ &= \bar{\dot{l}}_2(\theta_0) - I_{22}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}) \end{aligned}$$

we get

$$\begin{aligned} W_n &= n \Bigg[ \underbrace{(\hat{\xi}_n - \xi_0)^\top}_{=O_{P_{\theta_0}}(n^{-1/2})} \left( \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}) \right) \\ &\quad + \left( \underbrace{\bar{\dot{l}}_2(\theta_0) - I_{22}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0)}_{=0} + o_{P_{\theta_0}}(n^{-1/2}) \right) \underbrace{(\hat{\eta}_n - \hat{\eta}_n^0)}_{=O_{P_{\theta_0}}(n^{-1/2})} \Bigg] + o_{P_{\theta_0}}(1) \\ &= n(\hat{\xi}_n - \xi_0)^\top \left( \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0) \right) + o_{P_{\theta_0}}(1). \end{aligned}$$

Now by (a) of lemma, we get

$$\hat{\xi}_n - \xi_0 = I_{11\cdot 2}^{-1}(\theta_0)\left(\bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0)I_{22}^{-1}(\theta_0)\bar{\dot{l}}_2(\theta_0)\right) + o_{P_{\theta_0}}(n^{-1/2}),$$

and hence,

$$W_n = n(\hat{\xi}_n - \xi_0)^\top I_{11\cdot 2}(\theta_0)(\hat{\xi}_n - \xi_0) + o_{P_{\theta_0}}(1).$$

Use (a) part again, and then we can obtain

$$W_n = 2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) + o_{P_{\theta_0}}(1)$$

by (1.6). □

**Remark 1.3.19.** The leading term in previous theorem depends on unknown $\eta_0$ through $I_{11\cdot 2}(\theta_0)$, so in practice, it should be estimated as

$$n(\hat{\xi}_n - \xi_0)^\top I_{11\cdot 2}(\hat{\theta}_n^0)(\hat{\xi}_n - \xi_0), \qquad \text{(“estimated Wald statistics”)}$$

which is also the leading term of LRT.

**Theorem 1.3.20** (Rao's test; score test statistic)**.** *Let*

$$R_n = n\bar{\dot{l}}(\hat{\theta}_n^0)^\top \widehat{I(\theta^0)}^{-1}\bar{\dot{l}}(\hat{\theta}_n^0),$$

*where*

$$\widehat{I(\theta^0)} = -\bar{\ddot{l}}(\hat{\theta}_n^0)/n \text{ or } I(\hat{\theta}_n^0).$$

*i.e., “observed information.” Define*

$$\widehat{I_{11\cdot 2}(\theta^0)} = \widehat{I_{11}(\theta^0)} - \widehat{I_{12}(\theta^0)}\widehat{I_{22}(\theta^0)}^{-1}\widehat{I_{21}(\theta^0)},$$

*just as*

$$I_{11\cdot 2}(\theta^0) = I_{11}(\theta^0) - I_{12}(\theta^0)I_{22}(\theta^0)^{-1}I_{21}(\theta^0).$$

*Then,*

$$\begin{aligned} R_n &= n\bar{\dot{l}}_1(\hat{\theta}_n^0)^\top \widehat{I_{11\cdot 2}(\theta^0)}^{-1}\bar{\dot{l}}_1(\hat{\theta}_n^0) \\ &= 2(l(\hat{\theta}_n) - l(\hat{\theta}_n^0)) + o_{P_{\theta_0}}(1). \end{aligned}$$

**Remark 1.3.21.** Note that, unlike Wald statistics or LRT statistics, to obtain $R_n$ we only need MLE on the null, $\hat{\theta}_n^0$.

*Proof.* Note that, under $H_0$, $\dot{l}_2(\hat{\theta}_n^0) = 0$ holds,

$$\dot{l}(\hat{\theta}_n^0) = \begin{pmatrix} \dot{l}_1(\hat{\theta}_n^0) \\ 0 \end{pmatrix}$$

so we get

$$R_n = n\bar{\dot{l}}_1(\hat{\theta}_n^0)^\top \widehat{I_{11\cdot 2}(\theta^0)}^{-1} \bar{\dot{l}}_1(\hat{\theta}_n^0)$$

directly. Now with Taylor expansion, $\exists \hat{\theta}_n^{0,*} \in \text{line}\left(\begin{pmatrix} \xi_0 \\ \hat{\eta}_n^0 \end{pmatrix}, \begin{pmatrix} \xi_0 \\ \eta_0 \end{pmatrix}\right)$ s.t.

$$\bar{\dot{l}}_1(\hat{\theta}_n^0) - \bar{\dot{l}}_1(\theta_0) = \frac{\partial}{\partial \eta} \bar{\dot{l}}_1(\theta)\Big|_{\theta=\hat{\theta}_n^{0,*}} (\hat{\eta}_n^0 - \eta_0)$$

(by chain rule, and $\xi_0 - \xi_0 = 0$) Thus by continuity of $I$,

$$\bar{\dot{l}}_1(\hat{\theta}_n^0) - \bar{\dot{l}}_1(\theta_0) = \left(-I_{12}(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\eta}_n^0 - \eta_0)$$

holds. In the same way, we get

$$\bar{\dot{l}}_2(\hat{\theta}_n^0) - \bar{\dot{l}}_2(\theta_0) = \left(-I_{22}(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\eta}_n^0 - \eta_0),$$

but from $\bar{\dot{l}}_2(\hat{\theta}_n^0) = 0$,

$$\hat{\eta}_n^0 - \eta_0 = \left(I_{22}(\theta_0) + o_{P_{\theta_0}}(1)\right)^{-1} \bar{\dot{l}}_2(\theta_0) = \left[I_{22}^{-1}(\theta_0) + o_{P_{\theta_0}}(1)\right] \bar{\dot{l}}_2(\theta_0)$$

is obtained. Therefore,

$$\begin{aligned} \bar{\dot{l}}_1(\hat{\theta}_n^0) &= \bar{\dot{l}}_1(\theta_0) - \left[I_{12}(\theta_0) + o_{P_{\theta_0}}(1)\right](\hat{\eta}_n^0 - \eta_0) \\ &= \bar{\dot{l}}_1(\theta_0) - \left[I_{12}(\theta_0) + o_{P_{\theta_0}}(1)\right]\left[I_{22}^{-1}(\theta_0) + o_{P_{\theta_0}}(1)\right] \bar{\dot{l}}_2(\theta_0) \\ &= \bar{\dot{l}}_1(\theta_0) - I_{12}(\theta_0) I_{22}^{-1}(\theta_0) \bar{\dot{l}}_2(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}) \end{aligned}$$

holds, and the proof completes if one uses

$$\widehat{I_{11\cdot 2}(\theta^0)} = I_{11\cdot 2}(\theta_0) + o_{P_{\theta_0}}(1)$$

and (1.6). □

**Remark 1.3.22.** Note that, for a simple null $H_0 : \theta = \theta_0$,

$$W_n = n(\hat{\theta}_n - \theta_0)I(\theta_0)(\hat{\theta}_n - \theta_0)$$

$$R_n = n\bar{\dot{l}}_1(\theta_0)I(\theta_0)^{-1}\bar{\dot{l}}_2(\theta_0).$$

In this case, $k_0 = 0$, so $W_n$ and $R_n$ converges (in distribution) to $\chi^2(k)$ under $H_0$. For the composite, $\theta_0$ is estimated by $\hat{\theta}_n^0$.

**Example 1.3.23** (GoF test in a multinomial model)**.** Let $X_1, \cdots, X_n$ be a random sample from Multi$(1, (p_1, \cdots, p_r)^\top)$. Let $\theta = (p_1, \cdots, p_k)^\top$, where $k = r - 1$. Recall that

$$\Sigma(\theta) = \text{diag}(\theta_i) - \theta\theta^\top$$

$$I(\theta) = \Sigma(\theta)^{-1} = \text{diag}(\theta_i^{-1}) + \theta_r^{-1}\mathbf{1}\mathbf{1}^\top$$

$$\hat{\theta}^{MLE} = (\hat{p}_1, \cdots, \hat{p}_k)^\top,\ \hat{p}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ji} = \frac{O_i}{n}$$

hold.

(a) Under simple null $H_0 : p = p_0$, we get (for convenience, denote $\hat{\theta}^{MLE}$ as $\hat{\theta}$)

$$\begin{aligned} W_n &= n(\hat{\theta} - \theta_0)^\top I(\theta_0)(\hat{\theta} - \theta_0) \\ &= n(\hat{\theta} - \theta_0)^\top \left(\text{diag}(p_{0i}^{-1}) + p_{0r}^{-1}\mathbf{1}\mathbf{1}^\top\right)(\hat{\theta} - \theta_0) \\ &= n\left(\sum_{i=1}^{k} \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} + \frac{1}{p_{0r}}\left(\sum_{i=1}^{k}\hat{p}_i - \sum_{i=1}^{k} p_{0i}\right)^2\right) \\ &= n\sum_{i=1}^{r} \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}}, \end{aligned}$$

if we define

$$\hat{p}_r = 1 - \hat{\theta}_1 - \cdots - \hat{\theta}_k.$$

Then letting $n\hat{p}_i = \sum_{j=1}^{n} X_{ji} = O_i$, and $E_i^0 = np_{0i}$ be "expected occurrence" under the null, we get

$$W_n = \sum_{i=1}^{r} \frac{(O_i - E_i^0)^2}{E_i^0},$$

and $W_n \xrightarrow[n\to\infty]{d} \chi^2_k = \chi^2_{r-1}$ under $H_0$.

(b) Under the composite null $H_0 : p \in \Theta_0$, let $\Theta_0 = \{p(\eta) : \eta \in \mathcal{E}_0\}$, where $p(\cdot)$ is known and $\mathcal{E}_0$ is a $k_0$-dimensional space. Then $\hat{\theta}_0 = p(\hat{\eta}^0)$, and so similarly as (a) we get

$$W_n = \sum_{i=1}^{r} \frac{(O_i - \hat{E}_i^0)^2}{\hat{E}_i^0},$$

where $\hat{E}_i^0 = np_i(\hat{\eta}^0)$, and

$$W_n \xrightarrow[n\to\infty]{d} \chi^2_{k-k_0}$$

under $H_0$.

**Example 1.3.24** (Testing independence in a contingency table)**.** Consider the model

$$O_{ij} \overset{indep}{\sim} \text{Multi}(n, (p_{ij})),\ i = 1, \cdots, r,\ j = 1, \cdots, c.$$

In here, we wish to test

$$H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j}. \qquad \text{(“Independence”)}$$

Let

| $i \backslash j$ | 1 | 2 | $\cdots$ | $c$ | |
|---|---|---|---|---|---|
| 1 | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1c}$ | $p_{1\cdot}$ |
| 2 | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2c}$ | $p_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| r | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rc}$ | $p_{r\cdot}$ |
| | $p_{\cdot 1}$ | $p_{\cdot 2}$ | $\cdots$ | $p_{\cdot c}$ | 1 |

Table 1.1: Contingency table.

$$\theta = (p_{11}, p_{12}, \cdots, p_{1c}, p_{21}, \cdots, p_{2c}, \cdots, p_{r1}, \cdots, p_{r,c-1})^\top$$

be a $rc - 1$ dimensional parameter, and

$$\Theta = \{(p_{ij}) : p_{\cdot\cdot} = 1,\ p_{ij} > 0\}$$

be a parameter space. From

$$l(\theta) = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log p_{ij} + C,$$

and

$$p_{rc} = 1 - \sum_{(i,j)\neq(r,c)} p_{ij},$$

we get

$$\frac{\partial}{\partial\theta} l(\theta) = \begin{bmatrix} \vdots \\ \dfrac{O_{ij}}{p_{ij}} - \dfrac{O_{rc}}{p_{rc}} \\ \vdots \end{bmatrix}.$$

Note that it is $rc-1$ dimensional vector. Thus likelihood equation becomes

$$\frac{O_{ij}}{p_{ij}} = \frac{O_{rc}}{p_{rc}} \quad \forall i,j,$$

and hence

$$\hat{p}_{ij} = \frac{O_{ij}}{n}.$$

Further, under $H_0$, parameter space is

$$\Theta_0 = \{(p_{i\cdot}p_{\cdot j}) : p_{1\cdot} + \cdots + p_{r\cdot} = 1,\ p_{i\cdot} > 0,\ p_{\cdot 1} + \cdots + p_{\cdot c} = 1,\ p_{\cdot j} > 0\},$$

and likelihood becomes

$$l(\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log p_{i\cdot}p_{\cdot j} + C$$

so

$$\hat{p}^0_{ij} = \hat{p}^0_{i\cdot}\hat{p}^0_{\cdot j} = \frac{O_{i\cdot}O_{\cdot j}}{n^2}.$$

Further, $(i,j)$-th diagonal term of $\partial^2 l(\theta)/\partial\theta^2$ is

$$-\frac{O_{ij}}{p_{ij}^2} - \frac{O_{rc}}{p_{rc}^2},$$

and other elements are $-O_{rc}/p_{rc}^2$, and so

$$I(\theta) = \frac{1}{n} E_\theta \left[ -\frac{\partial^2}{\partial\theta^2} l(\theta) \right] = \mathrm{diag}(\theta_i^{-1}) + \frac{1}{p_{rc}} \mathbf{1}\mathbf{1}^\top.$$

Thus, we get

$$\begin{aligned} W_n &= n(\hat{\theta}_n - \hat{\theta}^0_n)^\top I(\hat{\theta}^0_n)(\hat{\theta}_n - \hat{\theta}^0_n) \\ &= n(\hat{\theta}_n - \hat{\theta}^0_n)^\top \left( \mathrm{diag}((\hat{\theta}^0_i)^{-1}) + \frac{1}{\hat{p}^0_{rc}} \mathbb{1}\mathbb{1}^\top \right) (\hat{\theta}_n - \hat{\theta}^0_n) \end{aligned}$$

$$
\begin{aligned}
&= n\left\{\sum_{(i,j)\neq(r,c)}\frac{(\hat{p}_{ij}-\hat{p}_{ij}^0)^2}{\hat{p}_{ij}^0}+\frac{1}{\hat{p}_{rc}^0}\left(\underbrace{\sum_{(i,j)\neq(r,c)}(\hat{p}_{ij}-\hat{p}_{ij}^0)}_{=\hat{p}_{rc}^0-\hat{p}_{rc}}\right)^2\right\}\\
&= n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{p}_{ij}-\hat{p}_{ij}^0)^2}{\hat{p}_{ij}^0}\\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij}^0)^2}{\hat{E}_{ij}^0},
\end{aligned}
$$

where $\hat{E}_{ij}^0 = n\hat{p}_{ij}^0$. Then

$$\text{dimension of } \Theta =: k = rc-1$$

$$\text{dimension of } \Theta_0 =: k_0 = (r-1)+(c-1)$$

so

$$k-k_0 = (r-1)(c-1),$$

and therefore,

$$W_n = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij}^0)^2}{\hat{E}_{ij}^0} \xrightarrow[n\to\infty]{d} \chi^2((r-1)(c-1))$$

under $H_0$.

**Example 1.3.25** (Homogeneity of contingency table)**.** Now consider the model

$$O_i = (O_{i1},\cdots,O_{ic})^\top \overset{\text{indep}}{\sim} \text{Multi}(n_i,p_i)$$

for $i=1,2,\cdots,r$ in the contingency table. We wish to test

$$H_0 : p_1 = p_2 = \cdots = p_r, \qquad \text{("Homogeneity")}$$

i.e.,

$$H_0 : \begin{pmatrix} p_{11}\\ p_{12}\\ \vdots\\ p_{1c}\end{pmatrix} = \begin{pmatrix} p_{21}\\ p_{22}\\ \vdots\\ p_{2c}\end{pmatrix} = \cdots = \begin{pmatrix} p_{r1}\\ p_{r2}\\ \vdots\\ p_{rc}\end{pmatrix}.$$

Let $\theta = (p_{11},p_{12},\cdots,p_{1,c-1},p_{21},\cdots,p_{2,c-1},\cdots,p_{r1},\cdots,p_{r,c-1})^\top$. Then dimension of the param-

eter space

$$\Theta = \{(p_{ij}) : p_{i\cdot} = 1,\ p_{ij} > 0\}$$

is

$$k := r(c-1),$$

and that of space under null

$$\Theta_0 = \{(p_{ij}) : p_{1j} = p_{2j} = \cdots = p_{rj},\ p_{i\cdot} = 1,\ p_{ij} > 0\}$$

is

$$k_0 := c - 1.$$

Now, note that log likelihood is obtained as

$$l(\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log p_{ij},$$

so

$$\frac{\partial}{\partial\theta} l(\theta) = \begin{bmatrix} \vdots \\ \dfrac{O_{ij}}{p_{ij}} - \dfrac{O_{ic}}{p_{ic}} \\ \vdots \end{bmatrix}.$$

Hence likelihood equation is obtained as

$$\frac{O_{ij}}{p_{ij}} = \frac{O_{ic}}{p_{ic}},$$

and therefore

$$\hat{p}_{ij} = \frac{1}{n_i} O_{ij}.$$

Meanwhile, under the null, likelihood is

$$l(\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log p_{1j} + C = \sum_{j=1}^{c} O_{\cdot j} \log p_{1j} + C,$$

and hence

$$\hat{p}^0_{ij} = \hat{p}^0_{1j} = \frac{1}{n} O_{\cdot j}.$$

Now we obtain the information. Note that $\partial^2 l(\theta)/\partial\theta^2$ has diagonal term

$$-\frac{O_{ij}}{p_{ij}^2}-\frac{O_{ic}}{p_{ic}^2}$$

and other elements of $\ddot{l}$ are

$$-\frac{O_{ic}}{p_{ic}^2},$$

and hence $I(\theta)$ has the form

$$I(\theta)=\begin{bmatrix} I_1 & & & \\ & I_2 & & \\ & & \ddots & \\ & & & I_r \end{bmatrix},$$

where

$$I_i=\begin{bmatrix} \frac{n_i}{p_{i1}} & & \\ & \ddots & \\ & & \frac{n_i}{p_{i,c-1}} \end{bmatrix}+\begin{bmatrix} \frac{n_i}{p_{ic}} & \frac{n_i}{p_{ic}} & \cdots \\ & \vdots & \\ & & \ddots \end{bmatrix}=n_i\left(\mathrm{diag}(\theta_i^{-1})+\frac{1}{p_{ic}}\mathbf{1}_{c-1}\mathbf{1}_{c-1}^\top\right),$$

$$\theta_i=(p_{i1},\cdots,p_{i,c-1})^\top.$$

Note that there are repetition of independent trials. Thus, by addivity, Wald's statistic which is given as

$$\begin{aligned} W_n&=(\hat{\theta}_n-\hat{\theta}_n^0)^\top I(\hat{\theta}_n^0)(\hat{\theta}_n-\hat{\theta}_n^0)\\ &=(\hat{\theta}_n-\hat{\theta}_n^0)^\top n_i\left(\mathrm{diag}(\theta_i^{-1})+\frac{1}{p_{ic}}\mathbf{1}_{c-1}\mathbf{1}_{c-1}^\top\right)(\hat{\theta}_n-\hat{\theta}_n^0)\\ &=\sum_{i=1}^r\sum_{j=1}^{c-1}\frac{n_i(\hat{p}_{ij}-\hat{p}_{ij}^0)^2}{\hat{p}_{ij}^0}+\sum_{i=1}^r\frac{n_i}{\hat{p}_{ic}^0}\left(\underbrace{\sum_{j=1}^{c-1}(\hat{p}_{ij}-\hat{p}_{ij}^0)}_{=\hat{p}_{ic}^0-\hat{p}_{ic}}\right)^2\\ &=\sum_{i=1}^r\sum_{j=1}^{c}\frac{n_i(\hat{p}_{ij}-\hat{p}_{ij}^0)^2}{\hat{p}_{ij}^0} \end{aligned}$$

converges to $\chi^2(k-k_0)$ distribution under $H_0$. Finally, note that

$$O_{ij}=n_i\hat{p}_{ij} \text{ and } \hat{E}_{ij}^0=n_i\hat{p}_{ij}^0,$$

and hence $W_n$ can be represented as

$$W_n = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij}^0)^2}{\hat{E}_{ij}^0}.$$

Therefore, our final result is that

$$W_n = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij}^0)^2}{\hat{E}_{ij}^0} \xrightarrow[n\to\infty]{d} \chi^2((r-1)(c-1))$$

holds under $H_0$.

**Remark 1.3.26.** Actually, it is still not rigorous, because in fact we cannot use the fact

$$W_n(\theta) \xrightarrow[n\to\infty]{d} \chi^2$$

directly in here. Direct representation to a quadratic form may ends the proof. For this, we need the additional assumption

$$\frac{n_i}{n_1 + \cdots + n_r} \to \lambda_i,\ 0 < \lambda_i < 1$$

for any $i = 1, 2, \cdots, r$.

# Chapter 2

# Weak Convergence

## 2.1 Weak Convergence in $\mathbb{R}^k$

**Definition 2.1.1.** *Let $X_n$ and $X$ be random vectors taking values in $\mathbb{R}^k$. Define $F_n(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$, where inequality is defined coordinatewisely, i.e., $P(X \leq x) = P(X_1 \leq x_1, \cdots, X_k \leq x_k)$. Then $X_n$ converges weakly to a random vector $X$, and write $X_n \xrightarrow[n\to\infty]{d} X$, if $P(X_n \in A) \xrightarrow[n\to\infty]{} P(X \in A)$ for any Borel set $A \subseteq \mathbb{R}^k$ with $P(X \in \partial A) = 0$.*

**Remark 2.1.2.** Note that in weak convergence, we only see the "distribution" of random vectors, so it's allowed to be defined on different probability spaces, unlike almost surely convergence. However, just for convenience, we fix probability space.

**Proposition 2.1.3.** *Let $C_f = \{x \in \mathbb{R}^k : f \text{ is continuous at } x\}$. Then followings are equivalent.*

*(i)* $X_n \xrightarrow[n\to\infty]{d} X$.

*(ii)* $F_n(x) \xrightarrow[n\to\infty]{} F(x)$ *for any* $x \in C_F$.

**Theorem 2.1.4** (Skorokhod representation theorem)**.** *Suppose that $X_n \xrightarrow[n\to\infty]{d} X$. Then there exists $X^*$ and $\{X_n^*\}$ defined on "the same probability space" $(\Omega^*, \mathcal{F}^*, P^*)$ s.t.*

$$X_n^* \overset{d}{\equiv} X_n,\ X^* \overset{d}{\equiv} X \text{ and } X_n^* \xrightarrow[n\to\infty]{a.s} X^*.$$

Note that, $X_n$ need not be defined on the same space, but $X_n^*$ should be, because we will say its "almost surely convergence."

**Theorem 2.1.5** (Continuous mapping theorem)**.** *If $X_n \xrightarrow[n\to\infty]{a.s} X$ and $P(X \in C_f) = 1$ for a real valued function $f$, then $f(X_n) \xrightarrow[n\to\infty]{a.s} f(X)$.*

*Proof.* It's clear from

$$P(\lim f(X_n) = f(X)) = P(X \in C_f, \lim f(X_n) = f(X)) \geq P(X \in C_f, \lim X_n = X) = 1.$$

□

**Proposition 2.1.6.** *If* $X_n \xrightarrow[n\to\infty]{a.s} X$, *then* $X_n \xrightarrow[n\to\infty]{d} X$.

*Proof.* (Proof in the lecture note is wrong! It is reported by Shin Jun-ho, and we are searching for alternative proof. Using Portmanteau lemma, it becomes trivial, but I think it breaks our flow.)

However, we cannot say that $X_n \xrightarrow[n\to\infty]{d} X$ implies $X_n \xrightarrow[n\to\infty]{a.s} X$. Following theorem shows some equivalent conditions of weak convergence in $\mathbb{R}^k$.

**Theorem 2.1.7** (Portmanteau lemma)**.** *TFAE.*

*(i)* $F_n(x) \to F(x)$ *for all* $x \in C_F$.

*(ii)* $Ef(X_n) \to Ef(X)$ *for all bounded* $f : \mathbb{R}^k \to \mathbb{R}$ *s.t.* $P(X \in C_f) = 1$.

*(iii)* $Ef(X_n) \to Ef(X)$ *for all bounded continuous* $f : \mathbb{R}^k \to \mathbb{R}$.

*(iv)* $Ef(X_n) \to Ef(X)$ *for all bounded and uniformly continuous* $f : \mathbb{R}^k \to \mathbb{R}$.

*Proof.* Enough to show (i)⇒(ii) and (iv)⇒(i).

(i) ⇒ (ii): Let $f : \mathbb{R}^k \to \mathbb{R}$ be a bounded function with $P(X \in C_f) = 1$, i.e., continuous on the support of $X$. By Skorokhod representation theorem, we can find $X_n^* \overset{d}{\equiv} X_n$ and $X^* \overset{d}{\equiv} X$ satisfying

$$X_n^* \xrightarrow[n\to\infty]{a.s} X^*.$$

Then by continuous mapping theorem,

$$f(X_n^*) \xrightarrow[n\to\infty]{a.s} f(X^*)$$

holds, and thus by bounded convergence theorem,

$$Ef(X_n^*) \xrightarrow[n\to\infty]{} Ef(X^*).$$

Note that since $X_n^*$ and $X^*$ have the same distribution as $X_n$ and $X$, expectation also will be the same, so we get

$$Ef(X_n) \xrightarrow[n\to\infty]{} Ef(X).$$

**Remark 2.1.8.** Note that, if $X_n^*$ and $X^*$ are defined on $(\Omega^*, \mathcal{F}^*, P^*)$, then for expectation $E^*(\cdot)$ under $P^*$, we should write

$$E^* f(X_n^*) \xrightarrow[n\to\infty]{} E^* f(X^*),$$

but just for convenience, we abuse the notation.

(iv) $\Rightarrow$ (i): Fix $x \in C_F$, and define

$$f_{m,x}^+ : \mathbb{R}^k \to \mathbb{R} \text{ and } f_{m,x}^- : \mathbb{R}^k \to \mathbb{R}$$

as

$$f_{m,x}^+(u) = \begin{cases} 1 & \text{if} \quad u \le x \\ \text{linear} & \text{if} \quad x \le u \le x + 1/m \\ 0 & \text{if} \quad u \ge x + 1/m \end{cases}$$

and

$$f_{m,x}^-(u) = \begin{cases} 1 & \text{if} \quad u \le x - 1/m \\ \text{linear} & \text{if} \quad x - 1/m \le u \le x \\ 0 & \text{if} \quad u \ge x. \end{cases}$$

Precisely, they become

$$f_{m,x}^+(u) = \begin{cases} 1 & \text{if} \quad u \le x \\ \dfrac{m}{k}\mathbf{1}^\top \left( x + \dfrac{1}{m} - u \right) & \text{if} \quad x \le u \le x + \dfrac{1}{m} \\ 0 & \text{if} \quad u \ge x + \dfrac{1}{m} \end{cases}$$

and

$$f_{m,x}^-(u) = \begin{cases} 1 & \text{if} \quad u \le x - \dfrac{1}{m} \\ \dfrac{m}{k}\mathbf{1}^\top (x - u) & \text{if} \quad x - \dfrac{1}{m} \le u \le x \\ 0 & \text{if} \quad u \ge x. \end{cases}$$

Both functions are bounded and uniformly continuous, so we get

$$Ef_{m,x}^+(X_n) \to Ef_{m,x}^+(X) \text{ and } Ef_{m,x}^-(X_n) \to Ef_{m,x}^-(X).$$

Further, $0 \le f_{m,x}^- \le f \le f_{m,x}^+ \le 1$, where $f = I_{(-\infty,x]}(\cdot)$ holds, so we get

$$F_n(x) = Ef(X_n) \le Ef_{m,x}^+(X_n),$$

and hence

$$\limsup_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} Ef_{m,x}^+(X_n) = Ef_{m,x}^+(X).$$

Similarly, we get

$$\liminf_{n\to\infty} F_n(x) \ge \liminf_{n\to\infty} Ef_{m,x}^-(X_n) = Ef_{m,x}^-(X).$$

Note that

$$\lim_{m\to\infty} f_{m,x}^+(u) = I_{(-\infty,x]}(u) \text{ and } \lim_{m\to\infty} f_{m,x}^-(u) = I_{(-\infty,x)}(u).$$

Therefore, we get

$$\lim_{m\to\infty} Ef_{m,x}^-(X) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le \lim_{m\to\infty} Ef_{m,x}^+(X),$$

i.e.,

$$F(x-) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x),$$

which yields

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

from the assumption $x \in C_F$. □

**Remark 2.1.9.** Indeed, some more equivalent conditions are known, which is called "*portmanteau lemma.*" Equivalent conditions are given as following:

(i) $F_n(x) \to F(x)$ for all $x \in C_F$.

(ii) $Ef(X_n) \to Ef(X)$ for all bounded and uniformly continuous $f : \mathbb{R}^k \to \mathbb{R}$.

(iii) $\limsup_{n\to\infty} P(X_n \in F) \le P(X \in F)$ for any closed set $F$.

(iv) $\liminf_{n\to\infty} P(X_n \in G) \ge P(X \in G)$ for any open set $G$.

(v) $X_n \xrightarrow[n\to\infty]{d} X$, i.e., $P(X_n \in A) \to P(X \in A)$ for any $X$-continuity sets $A$.

## 2.2 Weak Convergence in Metric Spaces

Let $(\mathbb{S}, \mathcal{S})$ be a metric space, where $\mathcal{S} = \mathcal{B}(\mathbb{S})$ denotes the Borel $\sigma$-field of $\mathbb{S}$. Let $X_n, X$ be *random elements* taking values in $\mathbb{S}$, i.e., $X_n$ and $X$ are "measurable" mappings from $\Omega$ to $\mathcal{S}$. Recall that $X : \Omega \to \mathbb{S}$ is $(\mathcal{F}\backslash\mathcal{S}$-)measurable if

$$X^{-1}(A) \in \mathcal{F} \quad \forall A \in \mathcal{S}.$$

**Definition 2.2.1.** *A sequence $\{X_n\}$ of random elements* ***converges weakly*** *to a random element $X$, which is written as $X_n \xrightarrow[n\to\infty]{d} X$, if*

$$P(X_n \in A) \xrightarrow[n\to\infty]{} P(X \in A) \ \forall A \in \mathcal{S} \text{ (i.e., for any Borel set } A,) \text{ s.t. } P(X \in \partial A) = 0.$$

**Remark 2.2.2.** $X_n \xrightarrow[n\to\infty]{d} X$ if and only if $Ef(X_n) \to Ef(X)$ for any bounded and uniformly continuous real-valued function $f$.

**Example 2.2.3.** In here, we see an example of a metric space $(\mathbb{S}, \mathcal{S})$ which is not Euclidean. Consider $\mathbb{S} = \mathbb{R}^\infty$, i.e., $\mathbb{S}$ is the set of real-valued sequences. Give the metric

$$d(x, y) = \sup_{1\le i} |x_i - y_i|$$

in this space. Then one can observe that $X_n \xrightarrow[n\to\infty]{d} X$ if *every finite-dimensional distribution of $X_n$ converges weakly to corresponding finite-dimensional distribution of $X$*, i.e.,

$$(X_{n,1}, \cdots, X_{n,k})^\top \xrightarrow[n\to\infty]{d} (X_1, \cdots, X_k)^\top \ \forall k \ge 1.$$

It can be also written as

$$P_n \pi_k^{-1} \xrightarrow[n\to\infty]{w} P\pi_k^{-1},$$

where $P_n$ and $P$ are distribution measure of $X_n$ and $X$ respectively, and $\pi_k : \mathbb{R}^\infty \to \mathbb{R}^k$ is a natural projection s.t. $\pi_k(x) = (x_1, \cdots, x_k)^\top$.

Unfortunately, statement given in previous example is not true in general. Especially, we can find a counter-example when $\mathbb{S}$ is a function space, which is introduced in following example.

**Example 2.2.4.** Let $\mathbb{S} = \mathbb{C}[0, 1]$ be the space of continuous functions defined on $[0, 1]$ with

sup-metric $d(x,y) = \sup_{t\in[0,1]} |x(t) - y(t)|$. Define

$$X_n(t) = \begin{cases} nt & \text{if} \quad 0 \le t \le 1/n \\ 2 - nt & \text{if} \quad 1/n \le t \le 2/n \\ 0 & \text{if} \quad 2/n \le t \le 1. \end{cases}$$

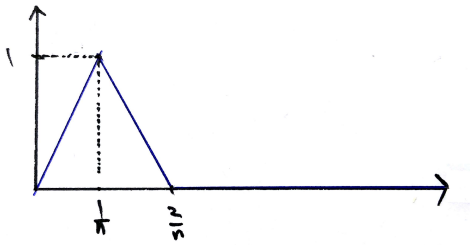

Figure 2.1: ($P$-a.s. unique) sample path of $X_n$.

(Note that $X_n$ is a random element on $\mathbb{C}[0,1]$ which has only one value; it has only one sample path.) Then for $X(t) \equiv 0$, every finite-dimensional distribution of $X_n$ converges weakly to the corresponding finite-dimensional distribution of $X$. It follows from, for any $0 \le t_1 < t_2 < \cdots < t_k \le 1$, (if $0 = t_1$, consider $t_2$ instead of $t_1$, so WLOG $0 < t_1$)

$$(X_n(t_1), X_n(t_2), \cdots, X_n(t_k)) \overset{d}{\equiv} (X(t_1), X(t_2), \cdots, X(t_k))$$

if $t_1 > 2/n$, i.e., for sufficiently large $n$. However, $X_n$ does not converge weakly to $X$. To see this, let $A = B(0, 1/2) = \{y : d(y,0) \le 1/2\}$. Then clearly $A$ is a Borel set, and from

$$d(X_n, X) = 1 \text{ and } d(X, 0) = 0,$$

we get

$$P(X \in \partial A) = P(d(X,0) = 1/2) = 0$$

and

$$P(X_n \in A) = P(d(X_n, 0) \le 1/2) = 0$$

but

$$P(X \in A) = P(d(X,0) \le 1/2) = 1,$$

i.e.,

$$P(X_n \in A) \text{ does not converges to } P(X \in A), \text{ even if } P(X \in \partial A) = 0.$$

⊠

**Remark 2.2.5.** For the natural projection $\pi_{t_1,\cdots,t_k}(x) = (x(t_1), \cdots, x(t_k))^\top$, we get even if

$$P_n \pi^{-1}_{t_1,\cdots,t_k} \xrightarrow[n\to\infty]{w} P\pi^{-1}_{t_1,\cdots,t_k} \ \forall k \geq 1, \ \forall t_i$$

holds, we cannot say that

$$P_n \xrightarrow[n\to\infty]{w} P.$$

Therefore, our main task is to find a plausible set of sufficient conditions that ensures $X_n \xrightarrow[n\to\infty]{d} X$. First, we may introduce following proposition, which is very useful tool.

**Proposition 2.2.6.** *$X_n \xrightarrow[n\to\infty]{d} X$ if and only if*

$$\forall\{n'\} \subseteq \{n\} \ \exists\{n''\} \subseteq \{n'\} \ s.t. \ X_{n''} \xrightarrow[n\to\infty]{d} X.$$

*Proof.* Only $\Leftarrow$ part is non-trivial. Suppose that $X_n$ does not converge to $X$ weakly. Then by Portmanteau lemma, $\exists$bounded, uniformly continuous function $f : \mathbb{S} \to \mathbb{R}$ s.t.

$$Ef(X_n) \not\to Ef(X).$$

(Note that $Ef(X_n)$ is a real sequence) Then $\exists\epsilon > 0$ and $\exists\{n'\} \subseteq \{n\}$ s.t.

$$|Ef(X_{n'}) - Ef(X)| > \epsilon \ \ \forall n'. \tag{2.1}$$

However, by the assumption, $\exists\{n''\} \subseteq \{n'\}$ s.t.

$$X_{n''} \xrightarrow[n''\to\infty]{d} X,$$

i.e.,

$$Ef(X_{n''}) \to Ef(X) \text{ as } n'' \to \infty$$

by Portmanteau lemma, which is contradictory to (2.1). □

Recall that for real-valued random elements,

$$X_n \xrightarrow[n\to\infty]{d} X \Rightarrow X_n = O_P(1)$$

holds. We already know that converse does not hold; if

$$X_n = \begin{cases} Y_n & n = 2k \\ Z_n & n = 2k-1 \end{cases}$$

for $Y_n$ and $Z_n$ s.t. $Y_n \xrightarrow[n\to\infty]{d} Y$, $Z_n \xrightarrow[n\to\infty]{d} Z$, then clearly $X_n = O_P(1)$, but $X_n$ does not converge weakly. Similarly, we will think a general notion of "$O_P(1)$ ," which is called as *relative compactness*, and then we will find a sufficient condition for weakly convergence, *additional to relative compactness.*

**Definition 2.2.7.** *Let $\{X_n\}$ be a sequence of random elements. $\{X_n\}$ is called **relatively compact** if $\forall\{n'\} \subseteq \{n\}$, $\exists\{n''\} \subseteq \{n'\}$ s.t. $X_{n''}$ conveges weakly to some random element.*

Note that in this case, the limit may depend on the choice of subsequence $\{n'\}$! Thus, relative compactness cannot guarantee weak convergence in alone.

**Proposition 2.2.8** (Continuous mapping theorem)**.** *Let $(\mathbb{S}, \mathcal{S})$ and $(\mathbb{S}', \mathcal{S}')$ be metric spaces, and $h : \mathbb{S} \to \mathbb{S}'$ be a measurable function s.t. $P(X \in D_h) = 0$, where $D_h$ denotes the set of discontinuities of $h$. Then $X_n \xrightarrow[n\to\infty]{d} X$ implies $h(X_n) \xrightarrow[n\to\infty]{d} h(X)$.*

*Proof.* Skorokhod representation theorem and continuous mapping theorem introduced in previous section. □

With these, we get the first sufficient condition of weak convergence.

**Theorem 2.2.9.** *Let $\mathcal{H}$ be "a" collection of measurable continuous functions from $(\mathbb{S}, \mathcal{S})$ to $(\mathbb{S}', \mathcal{S}')$. If*

*(i) $\mathcal{H}^{-1}(\mathcal{S}')$ generates $\mathcal{S}$, where $\mathcal{H}^{-1}(\mathcal{S})$ is defined as*

$$\mathcal{H}^{-1}(\mathcal{S}') := \bigcup_{h\in\mathcal{H}} h^{-1}(\mathcal{S}') = \{h^{-1}(A) : A \in \mathcal{S}',\ h \in \mathcal{H}\}.$$

*(ii) $\{X_n\}$ is relatively compact.*

*(iii) $h(X_n)$ converges weakly to $h(X)$ for any $h \in \mathcal{H}$.*

*Then $X_n \xrightarrow[n\to\infty]{d} X$.*

**Remark 2.2.10.** (a) Condition (iii) can be relaxed to:

"$h(X_n)$ converges weakly to some random element $Y_h$ for any $h \in \mathcal{H}$."

Then we can further show that there exists $X$ s.t. $Y_h = h(X)$ for any $h \in \mathcal{H}$. Define $X$ to satisfy

$$P(X \in h^{-1}(A')) = P(Y_h \in A') \ \forall A' \in \mathcal{S}', \ \forall h \in \mathcal{H}.$$

Then we can "define" $X$; because the collection of $h^{-1}(A')$ generates Borel $\sigma$-field $\mathcal{S}$. Next, it is "well-defined," i.e., our definition is "consistent"; for another function $\tilde{h} \in \mathcal{H}$ and corresponding limit $Y_{\tilde{h}}$, if $h^{-1}(A) = \tilde{h}^{-1}(B)$, then

$$\begin{aligned} P(Y_h \in A) &= \lim P(h(X_n) \in A) = \lim P(X_n \in h^{-1}(A)) \\ &= \lim P(X_n \in \tilde{h}^{-1}(B)) = \lim P(\tilde{h}(X_n) \in B) = P(Y_{\tilde{h}} \in B). \end{aligned}$$

Finally, $X$ is uniquely determined, and therefore $Y_h = h(X)$.

(b) Indeed, $\mathcal{H}$ is related to finite-dimensional distribution. For example, on $\mathcal{C}[0,1]$, if $x : [0,1] \to \mathbb{S}$, then for $h(x) = (x(t_1), \cdots, x(t_k))$, $h(X)$ is finite-dimensional distribution of $X$.

*Proof.* By relative compactness, it's enough to show that weak limit of $X_{n''}$ does not depend on the choice of $\{n'\}$. Let $X_{n''} \xrightarrow[n''\to\infty]{d} Y$. Then by continuous mapping theorem, $h(X_{n''}) \xrightarrow[n''\to\infty]{d} h(Y)$. However, by the assumption, $h(X_{n''}) \xrightarrow[n''\to\infty]{d} h(X)$, so we get

$$h(X) \overset{d}{\equiv} h(Y).$$

(Yet $Y$ depends on $\{n'\}$) It implies that

$$P(X \in A) = P(Y \in A) \ \forall A \in \mathcal{H}^{-1}(\mathcal{S}'),$$

and by Dynkin's $\pi$-$\lambda$ theorem (note that $\mathcal{H}^{-1}(\mathcal{S}')$ is $\pi$-system; the collection of such $A$'s is $\lambda$-system), we get

$$P(X \in A) = P(Y \in A) \ \forall A \in \mathcal{S},$$

with the assumption that

$$\sigma\left(\mathcal{H}^{-1}(\mathcal{S}')\right) = \mathcal{S}.$$

Therefore, we get

$$X \overset{d}{\equiv} Y,$$

i.e., weak limit $Y$ does not depend on the choice of subsequence ($\because X$ is predetermined). □

However, relative compactness is hard to show. Thus, we introduce another notion that implies relative compactness; which is *tightness*.

**Definition 2.2.11.** *(a) $A$ is **totally bounded** if $\forall \epsilon > 0$, $A$ is covered by "finitely" many open balls of radius $\epsilon$ in $\mathbb{S}$. (Note that center of each ball need not be contained in $A$)*

*(b) $\mathbb{S}$ is **complete** if every Cauchy sequence has a limit in $\mathbb{S}$. A set $S$ in $\mathbb{S}$ is **complete** if every Cauchy sequence in $S$ converges to a value in $S$.*

*(c) A set $K$ in $(\mathbb{S}, \mathcal{S})$ is **compact** if every open cover of $K$ has finite subcover.*

**Proposition 2.2.12.** *If $A$ is totally bounded and complete, then $A$ is compact.*

**Remark 2.2.13.** Note that, in complete metric space $(\mathbb{S}, \mathcal{S})$, if $A$ is totally bounded but not complete, then think closure of $A$, denoted as $\bar{A}$, and then we get that $\bar{A}$ is compact. ($\because \bar{A}$ is complete, and it is totally bounded; just cover with closed ball instead of open balls!)

**Definition 2.2.14.** *A (single) random element $X$ is **tight** if $\forall \epsilon > 0$, $\exists$compact set $K$ s.t. $P(X \in K) > 1 - \epsilon$.*

**Remark 2.2.15.** Note that tightness implies relative compactness; so one can show tightness to get relative compactness. Also note that, in Euclidean space $\mathbb{R}^k$, or sequence space $\mathbb{R}^\infty$, single random element $\{X\}$ is always tight; in general, it may not be. Hence, we should find a sufficient condition for tightness.

**Definition 2.2.16.** *(a) $D$ is a **dense** subset of $S$ if $\forall x \in S$ $\forall \epsilon > 0$ $\exists y \in D$ s.t. $y \in B(x, \epsilon)$, or equivalently, $d(x, y) < \epsilon$.*

*(b) $\mathbb{S}$ is **separable** space if it contains countable dense subset, i.e.,*

$$\exists \text{countable subset } D \text{ s.t. } \forall \epsilon > 0 \ \forall x \in \mathbb{S} \ \exists y \in D \text{ s.t. } d(x, y) < \epsilon,$$

*or equivalently,*

$$\exists \text{countable subset } D \text{ s.t. } \forall \epsilon > 0 \ \bigcup_{y \in D} B(y, \epsilon) \supseteq \mathbb{S}.$$

*(c) A topological space is called a **Polish space** if it is (i) seperable, and (ii) metrizable in such a way that it becomes complete.*

Following theorem says that, in *separable and complete metric space*, our convention works.

**Theorem 2.2.17.** *If $\mathbb{S}$ is separable and complete, then each random element in $(\mathbb{S}, \mathcal{S})$ is tight.*

*Proof.* Given $\epsilon > 0$. We want to show that:

$$\exists \text{compact set } K \text{ s.t. } P(X \in K) > 1 - \epsilon.$$

Since $\mathbb{S}$ is separable, $\exists$countable number of balls $B_{k,1}, B_{k,2}, \cdots$ with radius $1/k$ that covers $\mathbb{S}(\because \bigcup B(y_k, 1/k) \supseteq \mathbb{S}$ for some $\{y_k : k \geq 1\})$. It implies that

$$P\left(X \in \bigcup_{j=1}^{\infty} B_{k,j}\right) = 1.$$

By continuity of probability measure, we get

$$P\left(X \in \bigcup_{j=1}^{J} B_{k,j}\right) \xrightarrow[J\to\infty]{} 1,$$

and therefore, $\exists J = J(k)$ s.t.

$$P\left(X \in \bigcup_{j=1}^{J(k)} B_{k,j}\right) > 1 - \frac{\epsilon}{2^k}.$$

Now define

$$B = \bigcap_{k=1}^{\infty} \bigcup_{j=1}^{J(k)} B_{k,j},$$

and then $B$ is totally bounded, because $\bigcup_{j=1}^{J(k)} B_{k,j}$ is already totally bounded set, including $B$. Then closure $\bar{B}$ of $B$ is complete, and hence, $\bar{B}$ is compact (cf. remark 2.2.13). Then we get

$$P(X \in \bar{B}^c) \leq P(X \in B^c) \leq \sum_{k=1}^{\infty} P\left(X \in \bigcap_{j=1}^{J(k)} B_{k,j}^c\right) < \sum_{k=1}^{\infty} \frac{\epsilon}{2^k} < \epsilon,$$

and therefore,

$$P(X \in \bar{B}) > 1 - \epsilon$$

holds. Let $K = \bar{B}$. □

Note that, we directly constructed a compact set with probability "near to 1."

## 2.3 Weak Convergence in $\mathbb{C}[0,1]$

Here we consider weak convergence in $\mathbb{C} \equiv \mathbb{C}[0,1]$, *the space of real valued continuous functions defined on the interval* $[0,1]$. (Do not confuse $\mathbb{C}$ with the set of complex numbers!) Let $\mathcal{C}$ be the Borel $\sigma$-field of $\mathbb{C}$, i.e., $\mathcal{C} = \mathcal{B}(\mathbb{C})$. We endow $\mathbb{C}$ with the uniform metric

$$d(x,y) = \sup_{t\in[0,1]} |x(t) - y(t)|.$$

It is well known that $\mathbb{C}$ is separable and complete. Separability and completeness facilitate derivation of a plausible set of sufficient conditions for weak convergence.

**Theorem 2.3.1** (Stone-Weierstrass theorem)**.** *Any continuous function can be approximated by a polynomial function. In other words,* $\mathbb{C}$ *is* ***separable****.*

*Proof.* There is elementary proof introduced by S.N.Bernstein. For this, see K.A.Ross, pp.217-220. □

**Theorem 2.3.2** (Completeness)**.** $\mathbb{C}$ *is* ***complete****.*

*Proof.* Royden, pp.193-194. □

### 2.3.1 Projection from $\mathbb{C}$ to $\mathbb{R}^k$

**Definition 2.3.3.** *For a set of points* $\{t_1, t_2, \cdots, t_k\} \subseteq [0,1]$*, let*

$$\pi_{t_1,\cdots,t_k} : (\mathbb{C}, \mathcal{C}) \to (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$$

*be a* ***projection*** *that carries a point(=function)* $x$ *of* $\mathbb{C}$ *to the points* $(x(t_1), \cdots, x(t_k)) \in \mathbb{R}^k$*.*

The following theorem says that, *the collection of all projections satisfies the conditions of* $\mathcal{H}$ in theorem 2.2.9.

**Theorem 2.3.4.** *The projection* $\pi_{t_1,\cdots,t_k}$ *is measurable and continuous for all* $t_1, \cdots, t_k \in [0,1]$ *and* $k \geq 1$*. Furthermore, all sets of the form*

$$\pi^{-1}_{t_1,\cdots,t_k}(B) \text{ for some } B \in \mathcal{B}(\mathbb{R}^k),\ t_1, \cdots, t_k \in [0,1] \text{ and } k \geq 1$$

*form a field that generates* $\mathcal{C}$*.*

*Proof.* Continuity is obvious: Letting $\delta = \epsilon/k$, we get

$$d(x,y) = \sup_{t\in[0,1]} |x(t)-y(t)| < \delta \Rightarrow |\pi_{t_1,\cdots,t_k}(x)-\pi_{t_1,\cdots,t_k}(y)| = \sqrt{(x(t_1)-y(t_1))^2+\cdots+(x(t_k)-y(t_k))^2} < \epsilon.$$

Now let

$$\mathcal{C}_0 = \{\pi^{-1}_{t_1,\cdots,t_k}(B) : B \in \mathcal{B}(\mathbb{R}^k),\ t_1,\cdots,t_k \in [0,1] \text{ and } k \geq 1\}.$$

Then $\mathcal{C}_0$ is a field from

$$\left[\pi^{-1}_{t_1,\cdots,t_k}(B)\right]^c = \pi^{-1}_{t_1,\cdots,t_k}(B^c),$$

$$\begin{aligned}\pi^{-1}_{t_1,\cdots,t_k}(B_1)\bigcup\pi^{-1}_{s_1,\cdots,s_l}(B_2) &= \pi^{-1}_{t_1,\cdots,t_k,s_1,\cdots,s_l}(B_1\times\mathbb{R}^l)\bigcup\pi_{t_1,\cdots,t_k,s_1,\cdots,s_l}(\mathbb{R}^k\times B_2)\\ &= \pi_{t_1,\cdots,t_k,s_1,\cdots,s_l}\left((B_1\times\mathbb{R}^l)\cup(\mathbb{R}^k\cup B_2)\right).\end{aligned}$$

Since Borel $\sigma$-field $\mathcal{C}$ is generated by closed balls, it's sufficient to show that arbitrary closed ball is contained in $\sigma(\mathcal{C}_0)$. Note that if $d(x,y) \leq \epsilon$, then for any $t \in [0,1]$ we get $|x(t)-y(t)| \leq \epsilon$, and hence, we get

$$\forall n \geq 1,\ 1 \leq i \leq n \Rightarrow \left|y\left(\frac{i}{n}\right) - x\left(\frac{i}{n}\right)\right| \leq \epsilon.$$

Thus we get

$$\bar{B}(x,\epsilon) \subseteq \bigcap_{n=1}^{\infty}\bigcap_{i=1}^{n}\left\{y : \left|y\left(\frac{i}{n}\right) - x\left(\frac{i}{n}\right)\right| \leq \epsilon\right\}.$$

Now our claim is:

**Claim.** $\bar{B}(x,\epsilon) = \bigcap_{n=1}^{\infty}\bigcap_{i=1}^{n}\left\{y : \left|y\left(\frac{i}{n}\right) - x\left(\frac{i}{n}\right)\right| \leq \epsilon\right\}$. To show this, we should prove "$\supseteq$" part. Suppose that

$$y \in \bigcap_{n=1}^{\infty}\bigcap_{i=1}^{n}\left\{y : \left|y\left(\frac{i}{n}\right) - x\left(\frac{i}{n}\right)\right| \leq \epsilon\right\}.$$

By max-min theorem, $\exists t_0 \in [0,1]$ s.t.

$$\sup_{t\in[0,1]} |x(t)-y(t)| = |x(t_0)-y(t_0)|.$$

Since $x$ and $y$ are continuous, $\forall\delta > 0$, $\exists t_\delta \in \{\frac{i}{n} : 1 \leq i \leq n,\ n \geq 1\}$ s.t.

$$|x(t_\delta)-x(t_0)| \leq \frac{\delta}{2},\ |y(t_\delta)-y(t_0)| \leq \frac{\delta}{2}.$$

Now we implement "3-$\epsilon$ argument"; by construction, we get $|x(t_\delta) - y(t_\delta)| \leq \epsilon$, and so

$$\sup_{t\in[0,1]} |x(t) - y(t)| = |x(t_0) - y(t_0)| \leq |x(t_0) - x(t_\delta)| + |x(t_\delta) - y(t_\delta)| + |y(t_\delta) - y(t_0)| \leq \epsilon + \delta$$

holds. Since $\delta > 0$ was arbitrary, we get

$$\sup_{t\in[0,1]} |x(t) - y(t)| \leq \epsilon,$$

i.e., $d(x, y) \leq \epsilon$. Therefore, we get

$$y \in \bar{B}(x, \epsilon),$$

which yields

$$\bar{B}(x,\epsilon) \supseteq \bigcap_{n=1}^{\infty} \bigcap_{i=1}^{n} \left\{ y : \left| y\left(\frac{i}{n}\right) - x\left(\frac{i}{n}\right) \right| \leq \epsilon \right\}.$$

□

Now we get the conclusion.

**Theorem 2.3.5.** *In $\mathcal{C}$, if $\{X_n\}$ is tight, and every finite-dimensional distribution of $X_n$ converges to that of $X$ weakly, then $X_n \xrightarrow[n\to\infty]{d} X$.*

*Proof.* Note that,

$$\mathcal{H} := \{\pi_{t_1,\cdots,t_k} : t_1, \cdots, t_k \in [0,1] \text{ and } k \geq 1\}$$

and

$$\mathcal{H}^{-1}(\mathcal{B}(\mathbb{R}^k)) = \{\pi^{-1}_{t_1,\cdots,t_k}(B) : B \in \mathcal{B}(\mathbb{R}^k),\ t_1, \cdots, t_k \in [0,1] \text{ and } k \geq 1\}$$

satisfies the condition in theorem 2.2.9, by previous theorem. Thus by theorem 2.2.9, we get the conclusion. □

As we have shown, if $\{X_n\}$ is tight, then we can obtain weak convergence of $X_n$ via only showing for that of finite-dimensional distributions. Thus, from now on, our interests are sufficient conditions of tightness of $\{X_n\}$.

### 2.3.2 Conditions for tightness in $\mathbb{C}$

Our first result is a "probabilistic version" of Arzelá-Ascoli argument.

**Theorem 2.3.6** (Arzelá-Ascoli)**.** *A set $A \subseteq \mathbb{C}$ has compact closure (i.e., $\bar{A}$ is compact) if and only if*

*(i)* $\sup_{x\in A}|x(0)|<\infty$

*(ii)* $\lim_{\delta\to 0}\sup_{x\in A} w_x(\delta)=0$, *where*

$$w_x(\delta)=\sup_{(s,t):|s-t|<\delta}|x(s)-x(t)|$$

*is the "modulus of continuity" of* $x$ *in* $\mathbb{C}$.

**Remark 2.3.7.** In fact, these conditions are necessary and sufficient for $A$ to be "totally bounded." Therefore, with completeness of $\mathbb{C}$, we get that $\bar{A}$ is totally bounded and complete, and hence compact.

**Theorem 2.3.8.** *A sequence* $\{X_n\}$ *is tight if and only if*

*(i)* $\{X_n(0)\}$ *is tight in* $\mathbb{R}$.

*(ii) ("asymptotic equicontinuity") For any* $\epsilon>0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P\left(\sup_{(s,t):|s-t|<\delta}|X_n(s)-X_n(t)|\geq\epsilon\right)=0.$$

**Remark 2.3.9.** Recall that:

(i) $f$ is *continuous* at $x$ if

$$\sup_{y:|x-y|<\delta}|f(y)-f(x)|\xrightarrow[\delta\to 0]{}0.$$

(ii) $f$ is *uniformly continuous* if

$$\sup_{(x,y):|x-y|<\delta}|f(y)-f(x)|\xrightarrow[\delta\to 0]{}0.$$

(iii) A sequence $\{f_n\}$ of functions is *equicontinuous* if

$$\sup_{(x,y):|x-y|<\delta}\sup_n|f_n(y)-f_n(x)|\xrightarrow[\delta\to 0]{}0.$$

As we need equicontinuity using Arzelá-Ascoli argument, we need condition (ii), which is stochastic version of equi-continuity condition, to use similar analogy.

*Proof.* $\Rightarrow$) Assume that $\{X_n\}$ is tight. Then for any given $\epsilon>0$, there is a compact set $K$ such that

$$\inf_n P(X\in K)>1-\epsilon.$$

Since $K$ is compact, we apply Arzelá-Ascoli theorem, and then we get

$$\sup_{x\in K}|x(0)|<\infty,\ \lim_{\delta\to 0}\sup_{x\in K}\sup_{(s,t):|s-t|<\delta}|x(s)-x(t)|=0,$$

i.e.,

$$\exists C_0 \text{ and } \delta_0 \text{ s.t. } \sup_{x\in K}|x(0)|<C_0,\ \sup_{x\in K}\sup_{(s,t):|s-t|<\delta_0}|x(s)-x(t)|<\epsilon.$$

(Remark: Always think "sup" as "proposition"!) $\sup_{x\in K}|x(0)|<C_0$ means that

$$x\in K\Rightarrow |x(0)|<C_0,$$

i.e,

$$\{x:x\in K\}\subseteq\{x:|x(0)|<C_0\},$$

so we get

$$\inf_n P(|X_n(0)|<C_0)\geq\inf_n P(X_n\in K)>1-\epsilon.$$

Thus $\{X_n(0)\}$ is tight in $\mathbb{R}$. Similarly, we get

$$x\in K\Rightarrow\sup_{(s,t):|s-t|<\delta_0}|x(s)-x(t)|<\epsilon,$$

and hence

$$\inf_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|<\epsilon\right)\geq\inf_n P(X_n\in K)>1-\epsilon,$$

so that

$$\sup_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|\geq\epsilon\right)<\epsilon \qquad \text{("diagonal probability")}$$

(To get rid of "diagonal problem," we find "very small" $\epsilon$ that is even smaller than $\epsilon_0$) Thus we get

$$\lim_{\delta\to 0}\sup_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|\geq\epsilon\right)\leq\epsilon,$$

and hence for any $\epsilon_0>0$, (we can always find $\epsilon<\epsilon_0$ and so)

$$\lim_{\delta\to 0}\sup_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|\geq\epsilon_0\right)\leq\lim_{\epsilon\to 0}\lim_{\delta\to 0}\sup_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|\geq\epsilon\right)\leq 0,$$

and therefore

$$\lim_{\delta\to 0}\sup_n P\left(\sup_{(s,t):|s-t|<\delta_0}|X_n(s)-X_n(t)|\geq\epsilon_0\right)=0.$$

$\Leftarrow$) Now assume that (i) and (ii) hold. We construct a totally bounded set $K$ such that $\sup_n P(X_n\in K^c)<\epsilon$. By (i), we can find $C_0$ such that

$$\sup_n P(|X_n(0)|>C_0)<\frac{\epsilon}{2}.$$

Also, by (ii), we can choose $\delta_j>0$, $j\geq 1$ such that

$$\limsup_{n\to\infty} P\left(\sup_{(s,t):|s-t|<\delta_j}|X_n(s)-X_n(t)|\geq\frac{1}{j}\right)<\frac{\epsilon}{2^j}.$$

Note that, since each single random element $X_k$ is tight, so by "only if" part that we have shown, we get

$$\lim_{\delta\to 0} P\left(\sup_{(s,t):|s-t|<\delta_j}|X_n(s)-X_n(t)|\geq\frac{1}{j}\right)=0,$$

and consequently, we can choose "even smaller" $\delta_j$¡0 such that

$$\sup_n P\left(\sup_{(s,t):|s-t|<\delta_j}|X_n(s)-X_n(t)|\geq\frac{1}{j}\right)<\frac{\epsilon}{2^j}.$$

Now take

$$K=\{x:|x(0)|\leq C_0\}\bigcap\cap_{j=1}^{\infty}\left\{x:\sup_{(s,t):|s-t|<\delta_j}|x(s)-x(t)|<\frac{1}{j}\right\}.$$

Then by Arzelá-Ascoli theorem, $K$ is totally bounded, and hence $\bar{K}$ is compact. Therefore, we get

$$\sup_n P(X_n\in\bar{K}^c)\leq\underbrace{\sup_n P(|X_n(0)|>C_0)}_{<\epsilon/2}+\sum_{j=1}^{\infty}\underbrace{\sup_n P\left(\sup_{(s,t):|s-t|<\delta_j}|X_n(s)-X_n(t)|<\frac{1}{j}\right)}_{<\epsilon/2^j}<\epsilon$$

from

$$K^c=\{x:|x(0)|>C_0\}\bigcup\cup_{j=1}^{\infty}\left\{x:\sup_{(s,t):|s-t|<\delta_j}|x(s)-x(t)|\geq\frac{1}{j}\right\},$$

which is the desired result. □

Now we get the very important proposition: By theorem 2.3.5, if $\{X_n\}$ is tight, then weak convergence of "every finite dimensional distributions" implies weak convergence "of $X_n$." However,

in previous theorem (theorem 2.3.8), we have a sufficient condition for tightness. Consequently, we have following corollary. Note that tightness of $\{X_n(0)\}$ comes from the weak convergence of "1-dimensional distribution."

**Corollary 2.3.10.** *Let $X_n$ and $X$ be a random element in $\mathbb{C}$. If all finite-dimensional distributions of $X_n$ converge weakly to those of $X$, and if $\forall \epsilon > 0$ $\exists n_0, \delta > 0$ such that*

$$\sup_{n \geq n_0} P\left( \sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \geq \epsilon \right) \leq \epsilon, \tag{2.2}$$

*then $X_n \xrightarrow[n\to\infty]{d} X$.*

*Proof.* It is just re-statement of

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\left( \sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \geq \epsilon \right) = 0.$$

□

However, showing (2.2) is not an easy task. Thus, we rather show alternative condition, which is given in (2.3).

**Theorem 2.3.11.** (2.2) *holds if*

$$\sup_{n \geq n_0} \sup_{0 \leq t \leq 1} P\left( \sup_{s \in [t, t+\delta]} |X_n(s) - X_n(t)| \geq \frac{\epsilon}{3} \right) \leq \delta \epsilon. \tag{2.3}$$

**Remark 2.3.12.** Recall the proof of Glivenko-Cantelli theorem: Give a partition (grid), and find an interval that $x$ belongs to. It is a common technique to handle "sup", and we employee it in here similarly.

*Proof.* WLOG $t < s$ (otherwise, just change the role), and $1/\delta$ is an integer (otherwise, just get smaller $\delta$ to obtain (2.2); we just need to show existence!). Let $t_i = \delta i$ for $i = 0, 1, \cdots, \delta^{-1} =: I_n$. Assume that $|s - t| < \delta$. Then (we supposed $t < s$) $\exists$ grid point $t_i$ such that

(i) $t_{i-1} \leq t \leq t_i \leq s \leq t_{i+1}$ ($s$ and $t$ are located along 2 intervals), or

(ii) $t_i \leq t < s \leq t_{i+1}$ (both $s$ and $t$ are in same interval).

(Check: we used $|s - t| < \delta$.)

Case 1. $t_{i-1} \leq t \leq t_i \leq s \leq t_{i+1}$

In this case, ($t \in [t_{i-1}, t_i]$, $s \in [t_i, t_{i+1}]$) from

$$|X_n(t) - X_n(t_i)| \leq \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)|$$

and

$$|X_n(s) - X_n(t_{i+1}) \leq \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)|,$$

we get

$$\begin{aligned} |X_n(t) - X_n(s)| &\leq |X_n(t) - X_n(t_i)| + |X_n(t_i) - X_n(t_{i+1})| + |X_n(t_{i+1}) - X_n(s)| \\ &\leq 3 \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)|. \end{aligned}$$

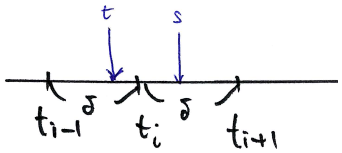


Figure 2.2: Case 1.

Case 2. $t_i \leq t < s \leq t_{i+1}$

In this case, ($t, s \in [t_i, t_{i+1}]$) we get

$$\begin{aligned} |X_n(t) - X_n(s)| &\leq |X_n(t) - X_n(t_i)| + |X_n(t_i) - X_n(s)| \\ &\leq 2 \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)| \\ &< 3 \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)|. \end{aligned}$$

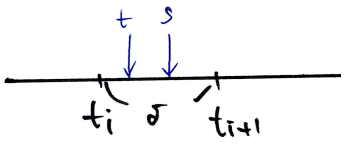


Figure 2.3: Case 2.

Thus, we get

$$|s - t| < \delta \Rightarrow |X_n(t) - X_n(s)| \leq 3 \max_{0 \leq i \leq I_n - 1} \sup_{t_i \leq t \leq t_{i+1}} |X_n(t) - X_n(t_i)|.$$

(Recall: "sup" is always regarded as a proposition!) Hence,

$$\sup_{(s,t):|s-t|<\delta} |X_n(t) - X_n(s)| \leq 3 \max_{0\leq i\leq I_n-1} \sup_{t_i\leq t\leq t_{i+1}} |X_n(t) - X_n(t_i)|$$

is obtained, and it implies

$$\begin{aligned} P\left(\sup_{(s,t):|s-t|<\delta} |X_n(t) - X_n(s)| \geq \epsilon\right) &\leq P\left(\max_{0\leq i\leq I_n-1} \sup_{t_i\leq t\leq t_{i+1}} |X_n(t) - X_n(t_i)| \geq \frac{\epsilon}{3}\right) \\ &\leq \sum_{i=0}^{I_n-1} P\left(\sup_{t_i\leq t\leq t_{i+1}} |X_n(t) - X_n(t_i)| \geq \frac{\epsilon}{3}\right) \\ &\leq \sum_{i=0}^{I_n-1} \delta\epsilon \text{ (Assumption)} \\ &= I_n\delta\epsilon = \epsilon. \end{aligned}$$

□

### 2.3.3 Donsker's Theorem

Consider a random element in $\mathbb{C}$ with the form

$$X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt\rfloor} \xi_i + (nt - \lfloor nt\rfloor)\frac{1}{\sqrt{n}}\xi_{\lfloor nt\rfloor+1},$$

where $\xi_i$ are i.i.d. random variables with $E\xi_1 = 0,\ Var\xi_1 = 1$. Note that it is a "continuous embedding" of elements (See figure 2.4)

$$\tilde{X}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt\rfloor} \xi_i.$$

Note that each sample path of $X_n(t)$ is continuous in $[0,1]$, so it is a random element on $\mathbb{C}[0,1]$ (In the interval $[i/n,(i+1)/n)$, $\tilde{X}_n(t)$ is a constant, and hence

$$X_n(t) = \tilde{X}_n(t) + (nt - i)\sqrt{n}^{-1}\xi_{i+1}$$

becomes a linear function; i.e., $X_n(t)$ is piecewise linear interpolation). Donsker's theorem states that:

**Theorem 2.3.13** (Donsker)**.** *The "partial sum process" $X_n(t)$ converges weakly to the standard*

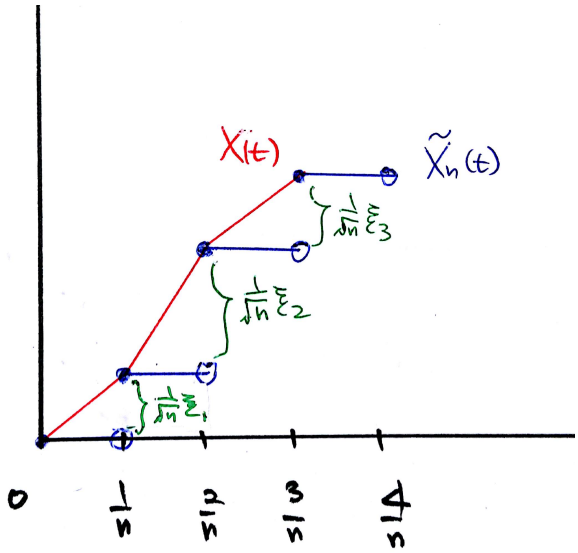


Figure 2.4: $X_n(t)$ and $\tilde{X}_n(t)$

*Brownian motion $W$, i.e.,*

$$X_n \xrightarrow[n\to\infty]{d} W.$$

Before prove Donsker's theorem, we first have to define what the limit process $W$ is.

**Definition 2.3.14.** *A random element $W$ on $\mathbb{C}[0,1]$ is called a (standard)* ***Brownian motion*** *or* ***Wiener process*** *if*

*(i) For each $0 \le t \le 1$, $W(t) \sim N(0,t)$ ("marginally normal").*

*(ii) $W(t_k) - W(t_{k-1}), \cdots, W(t_2) - W(t_1)$ are independent for all $0 \le t_1 \le \cdots \le t_k \le 1$ ("independent increments").*

Note that by definition,

(i) $W(0) = 0$, and

(ii) Each sample path of $W$ is continuous.

*Proof of theorem 2.3.13.* Note that finite-dimensional distributions of $X_n$ converge weakly to those of $W$ by classical CLT. For example, in 1-dimensional case, for given $t \in [0,1]$,

$$(nt - \lfloor nt \rfloor)\frac{1}{\sqrt{n}}\xi_{\lfloor nt \rfloor + 1} \xrightarrow[n\to\infty]{P} 0,$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor}\xi_i = \sqrt{\frac{\lfloor nt \rfloor}{n}}\frac{1}{\sqrt{\lfloor nt \rfloor}}\sum_{i=1}^{\lfloor nt \rfloor}\xi_i \xrightarrow[n\to\infty]{d} \sqrt{t}\cdot N(0,1) = N(0,t)$$

holds by CLT, and

$$W(t) \sim N(0,t).$$

In 2-dimensional case, if $0 \le t_1 < t_2 \le 1$ are given, then we can show that

$$\frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{i=1}^{\lfloor nt_1 \rfloor} \xi_i \\ \sum_{i=1}^{\lfloor nt_2 \rfloor} \xi_i \end{bmatrix} \xrightarrow[n\to\infty]{d} N(0,\Sigma)$$

where

$$\Sigma = \begin{pmatrix} t_1 & t_1 \\ t_1 & t_2 \end{pmatrix},$$

so from

$$\begin{bmatrix} W(t_1) \\ W(t_2) \end{bmatrix} \sim N(0,\Sigma),$$

we get the desired result. For the proof, see Kim (2012), p. 234-235 and exercise 5-15. Thus, we will show that $\forall \epsilon > 0$, $\exists n_0, \delta > 0$ such that (2.3) holds. In here, to show (2.3) we will further assume that $E\xi_1^4 < \infty$ (For the proof of general case, i.e., $E\xi_1^2 < \infty$, see Billingsley (1999), p.89-91). First following lemma is introduced without proof, which is given in Billingsley (1999).

**Lemma 2.3.15.** *Let $\xi_1, \cdots, \xi_m$ be random variables. Let $S_k = \xi_1 + \cdots + \xi_k$ for $k \ge 1$ and put $S_0 = 0$. If*

$$E|S_j - S_i|^\gamma \le (u_{i+1} + \cdots + u_j)^\alpha$$

*for some $\gamma \ge 0,\ \alpha > 1$ and $u_1, \cdots, u_m \ge 0$, then*

$$P\left(\max_{1\le k\le m} |S_k| \ge \lambda\right) \le \frac{C_{\gamma,\alpha}}{\lambda^\gamma}(u_1 + \cdots + u_m)^\alpha,$$

*where $C_{\gamma,\alpha}$ is a constant that depends only on $\gamma$ and $\alpha$.*

Now back to the proof. Fix $t \in [0,1]$. Then for any $n$, $\exists j$ such that

$$t \in \left[\frac{j}{n}, \frac{j+1}{n}\right).$$

(We "partitioned" $[0,1]$ with $n$ subintervals) Let

$$\frac{k}{n} < t + \delta \le \frac{k+1}{n}$$

for some $j \le k \le n-1$, and for $t < s \le t+\delta$, let

$$\frac{i}{n} < s \le \frac{i+1}{n}$$

for $j \le i \le k$. See figure 2.5.

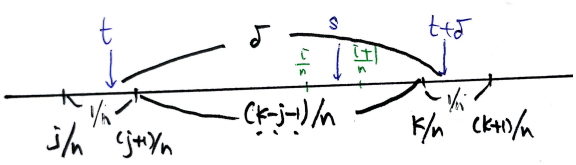


Figure 2.5: Proof of Donsker's theorem

Note that

$$\frac{k-j-1}{n} \le \delta. \qquad (*)$$

Then by triangle inequality and "polygonal character" of $X_n$ (cf. figure 2.4), we get

$$\begin{aligned}
|X_n(s) - X_n(t)| &\le \left|X_n(s) - X_n\left(\frac{i}{n}\right)\right| + \left|X_n\left(\frac{i}{n}\right) - X_n\left(\frac{j}{n}\right)\right| + \left|X_n\left(\frac{j}{n}\right) - X_n(t)\right| \\
&= X_n(s) - X_n\left(\frac{i}{n}\right) + \left|X_n\left(\frac{i}{n}\right) - X_n\left(\frac{j}{n}\right)\right| + X_n(t) - X_n\left(\frac{j}{n}\right) \\
&\le \left|X_n\left(\frac{i+1}{n}\right) - X_n\left(\frac{i}{n}\right)\right| + \left|X_n\left(\frac{i}{n}\right) - X_n\left(\frac{j}{n}\right)\right| + \left|X_n\left(\frac{j+1}{n}\right) - X_n\left(\frac{j}{n}\right)\right|.
\end{aligned}$$

Now note that, since $j \le i \le k \le j+1+n\delta$ from $(*)$, we get $j \le i \le (j+1+n\delta) \wedge (n-1)$. Let

$$I(\delta, j) = (j+1+n\delta) \wedge (n-1).$$

Then we have

$$\begin{aligned}
|X_n(s) - X_n(t)| &\le \max_{j \le i \le I(\delta,j)} \left|X_n\left(\frac{i}{n}\right) - X_n\left(\frac{j}{n}\right)\right| + 2 \max_{j \le i \le I(\delta,j)} \left|X_n\left(\frac{i+1}{n}\right) - X_n\left(\frac{i}{n}\right)\right| \\
&= \max_{j \le i \le I(\delta,j)} \left|\frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}}\right| + 2 \max_{j \le i \le I(\delta,j)} \left|\frac{\xi_i}{\sqrt{n}}\right|.
\end{aligned}$$

(Note that

$$X_n\left(\frac{i}{n}\right) = \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_i)$$

by definition) Thus for $S_k = \xi_1 + \cdots + \xi_k$, we get

$$\begin{aligned} E(S_{i'} - S_i)^4 &= E\left(\xi_{i+1} + \cdots + \xi_{i'}\right)^4 \\ &\le C_1 \left( \sum_{k=i+1}^{i'} \underbrace{E\xi_k^2}_{=1} \right)^2 + \sum_{k=i+1}^{i'} \underbrace{E\xi_k^4}_{\le (E\xi_k^2)^2 = 1} \\ &\le (C_1+1)(i'-i)^2 \ (\because i'-i\text{: integer}) \end{aligned}$$

for $i' > i > j$ ($\because E(\sum_k \xi_k)^4 = \sum_k E\xi_k^4 + \sum_{k \neq k'} (E\xi_k^2)(E\xi_{k'}^2) \le \sum_k E\xi_k^4 + \sum_{k,k'} (E\xi_k^2)(E\xi_{k'}^2)$), and hence by lemma ($\gamma = 4, \alpha = 2, u_k \equiv \sqrt{C_1+1}$), we get

$$\begin{aligned} P\left( \max_{j \le i \le I(\delta,j)} \left| \frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}} \right| \ge \frac{\sqrt{n}\epsilon}{6} \right) &\le \frac{C_{4,2}}{(\epsilon/6)^4} (I(\delta,j) - j)^2 (C_1+1) \\ &\le \frac{C_2 (n\delta+1)^2}{n^2\epsilon^4} \ (\because I(\delta,j) - j \le n\delta + 1) \\ &\le \frac{4C_2\delta^2}{\epsilon^4} \end{aligned}$$

for sufficiently large $n$ (precisely, $n \ge 1/\delta$), and choosing

$$\delta \le \frac{\epsilon^5}{8C_2},$$

we get

$$P\left( \max_{j \le i \le I(\delta,j)} \left| \frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}} \right| \ge \frac{\epsilon}{6} \right) \le \frac{4C_2\delta^2}{\epsilon^4} \le \frac{\delta\epsilon}{2}.$$

Meanwhile, we get

$$\begin{aligned} P\left( \max_{j \le i \le I(\delta,j)} \left| \frac{\xi_i}{\sqrt{n}} \right| \ge \frac{\epsilon}{12} \right) &\le P\left( \bigcup_{i=j}^{I(\delta,j)} \left( \left| \frac{\xi_i}{\sqrt{n}} \right| \ge \frac{\epsilon}{12} \right) \right) \\ &\le \sum_{i=j}^{I(\delta,j)} P\left( \left| \frac{\xi_i}{\sqrt{n}} \right| \ge \frac{\epsilon}{12} \right) \\ &\le \sum_{i=j}^{I(\delta,j)} \frac{C_{4,2}}{(\sqrt{n}\epsilon/12)^4} (C_1+1) \\ &= \underbrace{(I(\delta,j) - j - 1)}_{\le n\delta} \frac{C_{4,2}}{(\sqrt{n}\epsilon/12)^4} (C_1+1) \\ &\le \frac{C_4\delta}{n\epsilon^4} \end{aligned}$$

$$\leq \frac{\delta\epsilon}{2}$$

for sufficiently large $n$, by the lemma again. Therefore, from

$$\forall t \in [0,1], \forall (s,t) : |s-t| < \delta,$$

$$|X_n(s) - X_n(t)| \leq \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}} \right| + 2 \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_i}{\sqrt{n}} \right|,$$

we get

$$\sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \leq \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}} \right| + 2 \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_i}{\sqrt{n}} \right|,$$

and hence

$$\begin{aligned} P\left( \sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \geq \frac{\epsilon}{3} \right) &\leq P\left( \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_{j+1} + \cdots + \xi_i}{\sqrt{n}} \right| \geq \frac{\epsilon}{6} \right) + P\left( 2 \max_{j \leq i \leq I(\delta,j)} \left| \frac{\xi_i}{\sqrt{n}} \right| \geq \frac{\epsilon}{6} \right) \\ &\leq \frac{\delta\epsilon}{2} + \frac{\delta\epsilon}{2} \\ &= \delta\epsilon \end{aligned}$$

for large $n$, i.e.,

$$\sup_{n \geq n_0} \sup_{0 \leq t \leq 1} P\left( \sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \geq \frac{\epsilon}{3} \right) \leq \delta\epsilon.$$

Note that, "sufficiently large" $n_0$ only depends on $\delta$ and $\epsilon$, because $C_1, C_2, C_3$ and $C_4$ are absolutely constant. □

**Remark 2.3.16** (Annotation by compiler)**.** In fact, Donsker's theorem says more general thing: For empirical cdf $F_n(x)$,

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow[n\to\infty]{d} B(F(t)),$$

where weak convergence holds in Skorokhod topology of $\mathbb{D}(-\infty, \infty)$ space. In here, $B$ denotes Brownian bridge. However, empirical cdf is not continuous, so in here, we saw "continuous embedding" of partial sum process. Non-separability of $\mathbb{D}$ space made Donsker's formulation (for empirical cdf) wrong, and implementing Skorokhod metric, Skorokhod and Kolmogorov completed the proof (Wikipedia).

## 2.4 Weak Convergence in $\mathbb{D}[0,1]$

### 2.4.1 $\mathbb{D}$ space and its non-separability

**Definition 2.4.1.** *A function defined on $A \subseteq \mathbb{R}$ is called* ***càdlàg*** *function if it is right continuous and has left limit everywhere in $A$. A function space $\mathbb{D} = \mathbb{D}[0,1]$ is the space of càdlàg functions defined on the interval $[0,1]$.*

**Remark 2.4.2** (Annotation by compiler)**.** A term càdlàg came from French "continue à droite, limite à gauche."

Note that all continuous functions are càdlàg functions, so $\mathbb{C} \subseteq \mathbb{D}$. Furthermore, all distribution functions are also càdlàg functions. However, with respect to sup metric $d_U$, $(\mathbb{D}, d_U)$ becomes *not separable*, which yields difficulty.

**Example 2.4.3** (Non-separability of $(\mathbb{D}, d_U)$)**.** Let $x_\alpha(t) = I(\alpha \le t)$ be a random element on $\mathbb{D}$ (It is a random element with one possible path, i.e., mass on one point). Then if $\alpha \neq \alpha'$, $d_U(x_\alpha, x_{\alpha'}) = 1$. Let $0 < \epsilon \le 1/2$. Then

$$\{x \in \mathbb{D} : d_U(x_\alpha, x) < \epsilon\} \cap \{x \in \mathbb{D} : d_U(x_{\alpha'}, x) < \epsilon\} = \phi,$$

i.e., two open balls are disjoint.

If $\mathbb{D}$ is separable (w.r.t. $d_U$), then there exists a countable dense subset $\mathbb{D}_0$. Then any open ball $\{x \in \mathbb{D} : d_U(x_\alpha, x) < \epsilon\}$ contains a member of $\mathbb{D}_0$ for any $\alpha \in [0,1]$. Let $d_\alpha \in \mathbb{D}_0$ be a member of $\{x \in \mathbb{D} : d_U(x_\alpha, x) < \epsilon\}$. Then

$$\{d_\alpha : \alpha \in (0,1)\}$$

is uncountable ($\because$ such open balls are disjoint), but $\mathbb{D}_0$ is a countable set, which yields contradiction.

**Remark 2.4.4.** Why such separability is important? *If probability space is not separable, then we cannot guarantee measurability of random element.* For example, let

$$X : ([0,1], \mathcal{B}, \mu) \to (\mathbb{D}, \mathcal{D}, d_U)$$

be a random function with

$$X(t, \omega) = I(\omega \le t).$$

Let $H \subseteq [0,1]$. Then

$$\bigcup_{\alpha \in H} B\left(x_\alpha, \frac{1}{2}\right) = \bigcup_{\alpha \in H} \left\{ y \in \mathbb{D} : d_U(y, x_\alpha) < \frac{1}{2} \right\} \in \mathcal{D}$$

is a Borel set, since it is open ($x_\alpha(\cdot) = I(\alpha \leq \cdot)$). Now note that, $X(\cdot, \omega) = x_\omega(\cdot)$, $x_\omega \notin B(x_\alpha, 1/2)$ if $\alpha \neq \omega$, and hence

$$X(\cdot, \omega) \in B\left(x_\alpha, \frac{1}{2}\right) \Leftrightarrow X(\cdot, \omega) = x_\alpha \Leftrightarrow \omega = \alpha.$$

Thus we obtain

$$\begin{aligned} X^{-1}\left(\bigcup_{\alpha \in H} B\left(x_\alpha, \frac{1}{2}\right)\right) &= \left\{ \omega : X(\cdot, \omega) \in \bigcup_{\alpha \in H} B\left(x_\alpha, \frac{1}{2}\right) \right\} \\ &= \{\omega : \omega = \alpha \text{ for some } \alpha \in H\} \\ &= H. \end{aligned}$$

However, if $H \notin \mathcal{D}$, then inverse image of Borel set ($\cup_{\alpha \in H} B(x_\alpha, 1/2)$) is not measurable, i.e., $X$ is not measurable.

Therefore, we need to use the new metric, which makes $\mathbb{D}$ separable, rather than using $d_U$. Skorokhod introduced a metric, so-called *Skorokhod metric*, and showed that $\mathbb{D}$ is separable with such metric.

### 2.4.2 Skorokhod metric

The basic idea is to allow a "deformation' on the time scale (i.e., consider $x_{\alpha'}(\lambda(t))$ instead of $x_{\alpha'}(t)$) to define a distance between two elements in $\mathbb{D}$.

**Definition 2.4.5.** *Let $\Lambda$ be the class of all strictly increasing continuous mapping $\lambda$ such that $\lambda(0) = 0$ and $\lambda(1) = 1$. The **Skorokhod metric**, denoted by $d_S$, is defined by*

$$d_S(x, y) = \inf_{\lambda \in \Lambda} \max\left( \sup_{t \in [0,1]} |\lambda(t) - t|, \sup_{t \in [0,1]} |x(t) - y(\lambda(t))| \right).$$

**Proposition 2.4.6.** *$d_S$ is a metric.*

*Proof.* Billingsley, p.124. □

**Example 2.4.7.** Let $\alpha \neq \alpha'$. Our goal is to compute $d_S(x_\alpha, x_{\alpha'})$. For $\lambda \in \Lambda$, $\lambda^{-1}$ exists and it is also increasing and continuous (indeed, it yields symmetricity of $d_S$), and

$$x_\alpha(t) = \begin{cases} 1 & \text{if} \quad t \geq \alpha \\ 0 & \text{if} \quad t < \alpha \end{cases},$$

$$x_{\alpha'}(t) = \begin{cases} 1 & \text{if} \quad t \geq \lambda^{-1}(\alpha') \\ 0 & \text{if} \quad t < \lambda^{-1}(\alpha') \end{cases}.$$

So we get

$$|x_\alpha(t) - x_{\alpha'}(\lambda(t))| = \begin{cases} 1 & \text{if} \qquad \lambda^{-1}(\alpha') \leq t < \alpha \text{ or } \alpha \leq t < \lambda^{-1}(\alpha') \\ 0 & \text{otherwise} \end{cases}$$

Hence if $\lambda(\alpha) \neq \alpha'$, $\max\left(\sup_{t\in[0,1]} |\lambda(t) - t|, \sup_{t\in[0,1]} |x(t) - y(\lambda(t))|\right) = 1$, so we cannot achieve infimum with such $\lambda$. Thus, it's enough to consider only $\lambda$'s satisfying $\lambda(\alpha) = \alpha'$, i.e.,

$$d_S(x_\alpha, x_{\alpha'}) = \inf_{\lambda:\lambda(\alpha)=\alpha'} \sup_{t\in[0,1]} |\lambda(t) - t|.$$

It's easy to show that such value is $|\alpha' - \alpha|$ (Check!). Consequently, we get

$$d_S(x_\alpha, x_{\alpha'}) = |\alpha' - \alpha|.$$

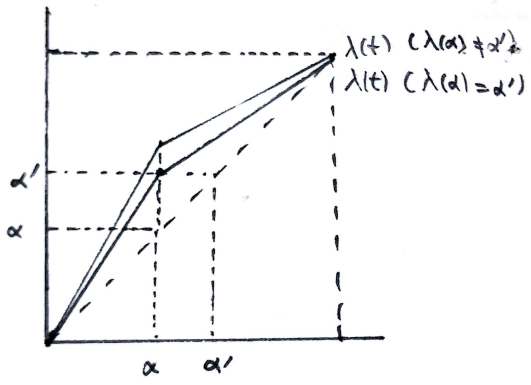


Figure 2.6: Example 2.4.7.

**Theorem 2.4.8.** $(\mathbb{D}, d_S)$ *is separable.*

*Proof.* Billingsley, p.128. □

**Remark 2.4.9.** However, $(\mathbb{D}, d_S)$ is not complete. (Completeness is important to characterize compact sets, for example, theorem 2.3.6.) For example, let

$$x_n(t) = I\left(\frac{1}{2} \le t < \frac{1}{2} + \frac{1}{n}\right) = I\left(t \ge \frac{1}{2}\right) - I\left(t \ge \frac{1}{2} + \frac{1}{n}\right).$$

Then

$$d_S(x_m, x_n) = \left|\frac{1}{m} - \frac{1}{n}\right| \to 0$$

so $\{x_m\}$ is Cauchy (Check! Motivation is: to achieve infimum, we should make $x_m(t) - x_n(\lambda(t)) \overset{t}{\equiv} 0$. When does it occur?) Nevertheless, it does not converge in $\mathbb{D}$. Suppose that $\exists x \in \mathbb{D}$ such that $d_S(x_n, x) \xrightarrow[n\to\infty]{} 0$. Then by definition,

$$\inf_{\lambda\in\Lambda} \max\left(\sup_{t\in[0,1]} |\lambda(t) - t|, \sup_{t\in[0,1]} |x(t) - y(\lambda(t))|\right) \to 0,$$

so $\exists \lambda_n \in \Lambda$ such that

$$\sup_{t\in[0,1]} |\lambda_n(t) - t| \xrightarrow[n\to\infty]{} 0, \quad \sup_{t\in[0,1]} |x_n(\lambda_n(t)) - x(t)| \xrightarrow[n\to\infty]{} 0. \tag{2.4}$$

However ,

$$x_n(\lambda_n(t)) = I\left(\lambda_n^{-1}\left(\frac{1}{2}\right) \le t \le \lambda_n^{-1}\left(\frac{1}{2} + \frac{1}{n}\right)\right),$$

and $x$ is a (uniform) limit of $x_n(\lambda_n(\cdot))$ (from second formula of (2.4)), and thus $x(t)$ is also an indicator function ($\because$ value of $x(t)$ can be only 0 or 1). Note that

$$\left|\lambda_n^{-1}\left(\frac{1}{2}\right) - \frac{1}{2}\right| \le \sup_{t\in[0,1]} |\lambda_n(t) - t| \to 0$$

and

$$\left|\lambda_n^{-1}\left(\frac{1}{2} + \frac{1}{n}\right) - \frac{1}{2}\right| \le \left|\lambda_n^{-1}\left(\frac{1}{2} + \frac{1}{n}\right) - \frac{1}{2} - \frac{1}{n}\right| + \frac{1}{n} \le \sup_{t\in[0,1]} |t - \lambda_n(t)| + \frac{1}{n} \to 0,$$

so we get

$$\lambda_n^{-1}\left(\frac{1}{2}\right) \to \frac{1}{2}, \ \lambda_n^{-1}\left(\frac{1}{2} + \frac{1}{n}\right) \to \frac{1}{2}.$$

Therefore,

$$x(t) = I\left(\frac{1}{2} \le t < \frac{1}{2}\right) \equiv 0.$$

(Note that limit of $I(\lambda_n^{-1}(1/2) \le t < \lambda_n^{-1}(1/2 + 1/n))$ also can be $I(1/2 \le t \le 1/2)$, but it does not belong to $\mathbb{D}$) Then

$$\sup_{t\in[0,1]} |x_n(\lambda_n(t)) - x(t)| = \sup_{t\in[0,1]} |x_n(\lambda_n(t)) - 0| = 1 \ (\because \ x_n(\lambda_n(t)) \text{ is indicator}$$

so $d_S(x_n, x) = 1$, which yields contradiction.

Fortunately, there is an equivalent metric $d'_S$ which makes $(\mathbb{D}, d'_S)$ complete. ("Equivalent" means that

$$d_S \le d'_S \le c \cdot d_S$$

holds for some constant $c$. For proof, see Billingsley, p.128.) Thus, we can proceed as if the *Skorokhod space* $(\mathbb{D}, d_S)$ *is separable and complete.*

### 2.4.3 Finite-dimensional distributions

**Theorem 2.4.10.** $\pi_{t_1,t_2,\cdots,t_k} : (\mathbb{D}, \mathcal{D}) \to (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ *is measurable.*

*Proof.* Billingsley, p.134. □

However, unlike $\mathbb{C}$, $\pi_{t_1,\cdots,t_k}$ is not continuous in $\mathbb{D}$.

**Example 2.4.11.** Assume that $\pi_t$ is continuous. Then

$$|\pi_t(x_n) - \pi_t(x)| = |x_n(t) - x(t)| \xrightarrow[n\to\infty]{} 0 \text{ provided that } d_S(x_n, x) \xrightarrow[n\to\infty]{} 0.$$

Now fix $t_0 \in (0, 1)$, and let

$$x(t) = I(t \ge t_0), \ x_n(t) = I\left(t \ge t_0 + \frac{1}{n}\right).$$

Then $d_S(x_n, x) = 1/n \to 0$. However, $x_n(t_0) = 0$ for any $n$, where $x(t_0) = 1$. Thus we get $|x_n(t_0) - x(t_0)| \overset{n}{\equiv} 1$.

Thus we cannot apply theorem 2.2.9 with finite-dimensional distributions. We need following *extended* version of theorem.

**Theorem 2.4.12.** *For a random element $Y$ taking values in a metric space $(\mathbb{S}, \mathcal{S})$, let $\mathcal{H}_Y$ be "a" collection of measurable functions $h$ that map $(\mathbb{S}, \mathcal{S})$ to another metric space $(\mathbb{S}', \mathcal{S}')$ such that $P(Y \in D_h) = 0$, for the set $D_h$ where $h$ is discontinuous. Suppose that $\{X_n\}$ is relatively compact and $h(X_n)$ converges to $h(X)$ for all $h \in \mathcal{H}_X$ for some random element $X$. If*

$$\mathcal{H}_{X,Y}^{-1} := \bigcup_{h \in \mathcal{H}_X \cap \mathcal{H}_Y} h^{-1}(\mathcal{S}')$$

*is a field generating $\mathcal{S}$ "for all" random elements $Y$, then $X_n \xrightarrow[n\to\infty]{d} X$.*

In fact, the requirement that "all finite-dimensional distributions of $X_n$ converge weakly to those of $X$" is too much in $(\mathbb{D}, \mathcal{D}, d_S)$. For example, let $X_n \equiv I([0, 1/2 + 1/n))$ and $X \equiv I([0, 1/2))$. Then $X_n \xrightarrow[n\to\infty]{d} X$ clearly ($\because d_S(X_n, X) = 1/n \to 0$), but $\pi_{1/2}X_n$ does not converge to $\pi_{1/2}X$. Thus we will relax such condition. The first one characterizes the discontinuity sets of $\pi_{t_1,\cdots,t_k}$ in $(\mathbb{D}, d_S)$, which tells that $\pi_{t_1,\cdots,t_k}$ for $0 < t_1 < \cdots < t_k < 1$ is discontinuous at $x$ *if and only if* $x$ is discontinuous at some $t_j$.

**Theorem 2.4.13.** *$\pi_0$ and $\pi_1$ are everywhere continuous. For $0 < t < 1$, $\pi_t$ is continuous at $x$ if and only if $x$ is continuous at $t$.*

*Proof.* First, from $\pi_0(x) = x(0)$, to show first part, we have to show

$$\pi_0(x_n) \to \pi_0(x) \text{ provided that } d_S(x_n, x) \to 0.$$

Since $\lambda(0) = 0$ for any $\lambda \in \Lambda$, $|x_n(0) - x(0)| \le d_S(x_n, x)$ holds, we get the desired result. For second part, suppose that $0 < t < 1$, and $x$ is continuous at $t$. Then for $\{x_n\} \subseteq \mathbb{D}$ s.t. $d_S(x_n, x) \to 0$,

$$\exists \lambda_n \in \Lambda \text{ s.t. } \sup_{t\in[0,1]} |\lambda_n(t) - t| \xrightarrow[n\to\infty]{} 0, \ \sup_{t\in[0,1]} |x_n(\lambda_n(t)) - x(t)| \xrightarrow[n\to\infty]{} 0.$$

Then we get

$$\begin{aligned} |x_n(t) - x(t)| &\le |x_n(t) - x(\lambda_n^{-1}(t))| + |x(\lambda_n^{-1}(t)) - x(t)| \\ &\le \underbrace{\sup_{s\in[0,1]} |x_n(\lambda_n(s)) - x(s)|}_{\to 0} + \underbrace{|x(\lambda_n^{-1}(t)) - x(t)|}_{\to 0 (\star)} \\ &\xrightarrow[n\to\infty]{} 0. \end{aligned}$$

Remark that: We need "continuity (at $t$) of $x$" in $(\star)$, to obtain $x(\lambda_n^{-1}(t)) \to x(t)$ from $\lambda_n^{-1}(t) \to t$. $\lambda_n^{-1}(t) \to t$ comes from $\sup_{t\in[0,1]} |\lambda_n(t) - t| \to 0$. Conversely, suppose that $x$ is discontinuous at $t$. Then define $\lambda_n(t)$ as a linear function on $[0, t]$ and $[t, 1]$, and satisfies $\lambda_n(t) = t - 1/n$. Then letting $x_n(s) = x(\lambda_n(s))$, we get

$$d_S(x_n, x) \le \sup_{t\in[0,1]} |\lambda_n(s) - s| = \frac{1}{n} \to 0.$$

But

$$|\pi_t(x_n) - \pi_t(x)| = |x_n(t) - x(t)| = \left|x\left(t - \frac{1}{n}\right) - x(t)\right| \xrightarrow[n\to\infty]{} |x(t-) - x(t)| \neq 0,$$

since we cannot say that $x$ is left continuous. In other words, $x$ may not be continuous at $t$. □

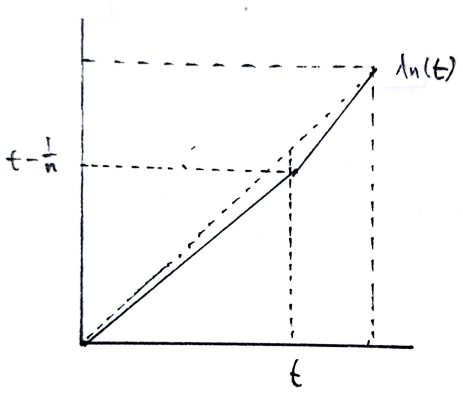


Figure 2.7: $\lambda_n(t)$ in the proof of theorem 2.4.13.

For a point $t \in [0, 1]$, define

$$D(\pi_t) = \{x \in \mathbb{D} : x(t) \neq x(t-)\}.$$

Then $D(\pi_t)$ is the set of all elements of $\mathbb{D}$ which are discontinuous at $t$. By previous theorem, if $0 < t < 1$, $D(\pi_t)$ becomes the set of discontinuity points of $\pi_t$. Also, from the theorem, we can say that

$$\pi_{t_1,\cdots,t_k} \text{ is continuous at } x \Leftrightarrow x \text{ is continuous at } t_1, t_2, \cdots, t_k.$$

**Theorem 2.4.14.** *The "complement of the set* $\{t \in (0, 1) : P(X \in D(\pi_t)) = 0\}$*"(=*$\{t \in (0, 1) : P(X$ *is discontinuous at* $t) > 0\}$*) for random element* $X$ *taking values in* $(\mathbb{D}, \mathcal{D})$ *is at most countable.*

**Remark 2.4.15.** We already know the special case of this theorem: "Distribution function has at most countably many jumps."

*Proof.* Billingsley, p. 138. □

**Definition 2.4.16.** *Define*

$$T_X := \{0,1\} \cup \{t \in (0,1) : P(X \in D(\pi_t)) = 0\}.$$

*Then clearly $T_X$ is dense in $[0,1]$ ($\because$ its complement is countable), and hence*

$$\bigcap_{i=1}^{m} T_{X_i}$$

*is dense in $[0,1]$ for finitely many $m$. (Annotation by compiler: I think countable intersection is also dense.)*

Then we will use following theorem:

**Theorem 2.4.17.** *For a subset $T \subseteq [0,1]$, let $\mathcal{F}_T$ be the class of all sets of the form $\pi^{-1}_{t_1,\cdots,t_k}(A')$ for some $A' \in \mathcal{B}(\mathbb{R}^k)$ , $t_1, \cdots, t_k \in T$ and $k \geq 1$ (i.e., $\mathcal{F}_T$ is the collection of all possible inverse images). If $1 \in T$ and $T$ is dense in $[0,1]$, then $\mathcal{F}_T$ is a field that generates $\mathcal{D}$.*

Using these theorems, we can achieve our goal: To relax sufficient condition of weak convergence.

**Theorem 2.4.18.** *Let $X_n$ and $X$ be random elements in $\mathbb{D}$. If the distribution $(X_n(t_1), \cdots, X_n(t_k))$ converges weakly to the distribution $(X(t_1), \cdots, X(t_k))$ (i.e., $\pi_{t_1,\cdots,t_k} X_n \xrightarrow[n\to\infty]{d} \pi_{t_1,\cdots,t_k} X$) for any $t_1 \cdots, t_k \in T_X$ and for all $k \geq 1$, and if $\{X_n\}$ is tight, then $X_n \xrightarrow[n\to\infty]{d} X$.*

**Remark 2.4.19.** It means that, if one wants to show $X_n \xrightarrow[n\to\infty]{d} X$, then one can show that finite dimensional distribution converges *at specific time points.* not all of them.

**Remark 2.4.20.** This theorem demonstrates that the class of the functions

$$\Pi_X = \{\pi_{t_1,\cdots,t_k} : t_j \in T_X \text{ for all } 1 \leq j \leq k, k \geq 1\}$$

plays the role of $\mathcal{H}_X$ in theorem 2.4.12. Recall that $T_X$ is a dense subset of $[0,1]$ which contains 1. Also, note that

$$D(\pi_{t_1,\cdots,t_k}) = \{x \in (\mathbb{D}, d_S) : \pi_{t_1,\cdots,t_k} \text{ is discontinuous at } x\}$$

$$= \{x \in (\mathbb{D}, d_S) : \pi_{t_j} \text{ is discontinuous at } x \text{ for some } j\}$$
$$= \bigcup_{j=1}^{k} D(\pi_{t_j})$$

if $0 < t_j < 1$. Thus if $t_j \in T_X \ \forall j$, then $P(X \in D(\pi_{t_j})) = 0$, so

$$P(X \in D(\pi_{t_1,\cdots,t_k})) = P\left(X \in \bigcup_{j=1}^{k} D(\pi_{t_j})\right) \leq \sum_{j=1}^{k} P(X \in D(\pi_{t_j})) = 0.$$

It implies that,

$$\forall \pi \in \Pi_X \ P(X \in D(\pi)) = 0,$$

and hence

$$\forall \pi \in \Pi_X \cap \Pi_Y \ P(X \in D(\pi)) = 0 \text{ and } P(Y \in D(\pi)) = 0. \qquad (\star)$$

*Proof.* If $\{X_n\}$ is tight, then it is relatively compact, so $\forall \{n'\} \subseteq \{n\} \ \exists \{n''\} \subseteq \{n'\}$ such that $X_{n''}$ converges weakly. Let

$$X_{n''} \xrightarrow[n\to\infty]{d} Y.$$

($Y$ depends on the choice of subsequence $\{n'\}$) Our goal is to show that $Y$ does not depend on $\{n'\}$. By continuous mapping theorem, we get

$$\forall \pi \in \Pi_Y \ \pi X_{n''} \xrightarrow{d} \pi Y.$$

We already know that

$$\forall \pi \in \Pi_X \ \pi X_{n''} \xrightarrow{d} \pi X$$

by the assumption ($\pi X_n \xrightarrow[n\to\infty]{d} \pi X$). Thus for any $\pi \in \Pi_X \cap \Pi_Y$,

$$\pi Y \overset{d}{\equiv} \pi X$$

holds. It means that

$$P(\pi X \in A) \overset{A}{\equiv} P(\pi Y \in A) \ \forall \pi \in \Pi_X \cap \Pi_Y,$$

i.e.,

$$P(X \in \pi^{-1}(A)) \overset{A}{\equiv} P(Y \in \pi^{-1}(A)) \ \forall \pi \in \Pi_X \cap \Pi_Y,$$

Note that

$$\Pi_X \cap \Pi_Y = \{\pi_{t_1,\cdots,t_k} : t_j \in T_X \cap T_Y\}.$$

Hence we get

$$\begin{aligned} P^X = P^Y \text{ on } \{\pi^{-1}(A) : A \in \mathcal{B}(\mathbb{R}^k), k \geq 1, \pi \in \Pi_X \cap \Pi_Y\} \\ = \{\pi^{-1}_{t_1,\cdots,t_k}(A) : A \in \mathcal{B}(\mathbb{R}^k), k \geq 1, t_1, \cdots, t_k \in T_X \cap T_Y\} \\ =: \mathcal{F}_{T_X \cap T_Y}. \end{aligned}$$

Since $T_X \cap T_Y$ is dense and it contains 1, by theorem 2.4.17, $\mathcal{F}_{T_X \cap T_Y}$ generates $\mathcal{D}$. Therefore, (with Dynkin's $\pi - \lambda$ theorem) we get

$$P^X = P^Y \text{ on } \mathcal{D},$$

i.e.,

$$X \overset{d}{\equiv} Y.$$

□

### 2.4.4 Tightness on $(\mathbb{D}, d_S)$

Let

$$w'_x(\delta) = \inf_{\{t_i\}: t_i - t_{i-1} > \delta} \inf_{1 \leq i \leq r} \sup\{|x(s)xt(t)| : s, t \in [t_{i-1}, t_i)\}$$

be a 'modulus' in $\mathbb{D}$. It plays the role of $w_x(\delta)$ in $\mathbb{C}$. In fact,

**Proposition 2.4.21.** *$w'_x(\delta) \to 0$ as $\delta \to 0$.*

It means that: $\forall \epsilon > 0$ $\exists$partition of $[0,1]$ into finitely many $T_i$ s.t.

$$\max_i \sup_{s,t \in T_i} |x(s) - x(t)| < \epsilon.$$

**Theorem 2.4.22.** *A set $A \subseteq \mathbb{D}$ has a compact closure if and only if*

*(i)* $\sup_{x \in A} \sup_{t \in [0,1]} |x(t)| < \infty$

*(ii)* $\lim_{\delta \to 0} \sup_{x \in A} w'_x(\delta) = 0$

Compare the result with theorem 2.3.8!

Indeed, $w'_x(\delta)$ is hard to deal with, so we often think $w''_x(\delta)$, which is defined as:

$$w''_x(\delta) = \sup\{|x(t) - x(t_1)| \wedge |x(t_2) - x(t)| : 0 \le t_1 \le t \le t_2 \le 1,\ t_2 - t_1 \le \delta\}.$$

**Example 2.4.23.** Let $x_\alpha(t) = I(\alpha \le t)$. Then modulus of continuity is obtained as

$$w_{x_\alpha}(\delta) = \sup_{(s,t):|s-t|\le\delta} |x_\alpha(t) - x_\alpha(s)| = 1\ \forall \delta > 0.$$

Now let's obtain $w'_{x_\alpha}(\delta)$. Let $(t_0 = 0, t_1 = \alpha, t_2 = \alpha + \epsilon, t_3 = 1)$ be a partition. Then if $0 < \delta < \epsilon < 1 - \alpha$,

$$\max_{1\le i\le 3} \sup\{|x_\alpha(s) - x_\alpha(t)| : s, t \in [t_{i-1}, t_i)\} = \sup_{s,t\in[\alpha,\alpha+\epsilon]} |x_\alpha(s) - x_\alpha(t)| = 0.$$

($\because x_\alpha(s) - x_\alpha(t) = 0$ if $s, t \in [0, \alpha]$ and $s, t \in [\alpha + \epsilon, 1]$) Thus, we get

$$w'_{x_\alpha}(\delta) \le \sup_{s,t\in[\alpha,\alpha+\epsilon]} |x_\alpha(s) - x_\alpha(t)| = 0.$$

Consequently, if $\alpha < 1$, $w'_{x_\alpha}(\delta) = 0$ for sufficiently small $\delta > 0$, so $w'_{x_\alpha}(\delta) \xrightarrow[\delta\to 0]{} 0$ (if $\alpha < 1$). Next, from

$$w''_{x_\alpha}(\delta) = \sup_{t\in[\alpha-\delta/2,\alpha+\delta/2]} \left|x_\alpha(t) - x_\alpha\left(\alpha - \frac{\delta}{2}\right)\right| \wedge \left|x_\alpha\left(\alpha + \frac{\delta}{2}\right) - x_\alpha(t)\right|,$$

we also get $w''_{x_\alpha}(\delta) = 0$ for sufficiently small $\delta > 0$ if $0 < \alpha < 1$.

**Proposition 2.4.24.** $\forall x \in \mathbb{D}$, $w''_x(\delta) \le w'_x(\delta) \le w_x(2\delta)$.

The following theorem gives another characterization of compact sets in $(\mathbb{D}, \mathcal{D})$ based on $w''_x(\delta)$. It is sometimes more convenient to work with than the characterization in theorem 2.4.22. We write

$$w_x(T) := \sup_{s,t\in T} |x(s) - x(t)|.$$

**Theorem 2.4.25.** *A set $A \subseteq \mathbb{D}$ has compact closure if and only if*

*(i)* $\sup_{x\in A} \sup_{t\in[0,1]} |x(t)| < \infty$

*(ii)* $\lim_{\delta\to 0} \sup_{x\in A} w''_x(\delta) = 0$

*(iii)* $\lim_{\delta\to 0} \sup_{x\in A} w_x[0, \delta) = 0$

*(iv)* $\lim_{\delta\to 0}\sup_{x\in A} w_x[1-\delta,1)=0$.

Using these theorems, we can find characterizations of a tight sequence in $\mathbb{D}$.

**Theorem 2.4.26.** *A sequence $\{X_n\}$ in $\mathbb{D}$ is tight if and only if*

*(i) the sequence of random variables $\left\{\sup_{t\in[0,1]}|X_n(t)|\right\}$ is tight in $\mathbb{R}$.*

*(ii) for any $\epsilon>0$, $\lim_{\delta\to 0}\limsup_{n\to\infty} P(w'_{X_n}(\delta)\geq\epsilon)=0$.*

*Proof.* $\Rightarrow$) Suppose that $\{X_n\}$ is tight. Given $\epsilon>0$, there exists a compact set $K$ s.t. $\inf_n P(X_n\in K)>1-\epsilon$. Then by theorem 2.4.22, $\exists C_0>0$ s.t.

$$\sup_{x\in K}\sup_{t\in[0,1]}|x(t)|<C_0.$$

Thus

$$(X_n\in K)\subseteq\left(\sup_{t\in[0,1]}|X_n(t)|<C_0\right)$$

holds, which implies

$$P(X_n\in K)\leq P\left(\sup_{t\in[0,1]}|X_n(t)|<C_0\right),$$

i.e.,

$$\sup_n P\left(\sup_{t\in[0,1]}|X_n(t)|\geq C_0\right)\leq\sup_n P(X_n\in K^c)\leq\epsilon.$$

Thus we get (i). By theorem 2.4.22 again,

$$\lim_{\delta\to 0}\sup_{x\in K} w'_x(\delta)=0.$$

It is equivalent to $\exists\delta_0>0$ s.t. $\sup_{x\in K} w'_x(\delta_0)<\epsilon$, i.e.,

$$(X_n\in K)\subseteq(w'_{X_n}(\delta_0)<\epsilon).$$

From this, we get

$$\limsup_{n\to\infty} P(w'_{X_n}(\delta_0)\geq\epsilon)\leq\sup_n P(w'_{X_n}(\delta_0)\geq\epsilon)\leq\sup_n P(X_n\in K^c)<\epsilon,$$

which gives (ii).

$\Leftarrow$) Now assume that (i) and (ii) hold. From (i), given $\epsilon > 0$, $\exists C_0$ s.t.

$$\sup_n P\left(\sup_{t\in[0,1]} |X_n(t)| \geq C_0\right) < \frac{\epsilon}{2},$$

and from (ii), $\forall j \geq 1$ $\exists \delta_j > 0$ s.t.

$$\limsup_{n\to\infty} P\left(w'_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^j}.$$

Actually, it is equivalent to

$$\sup_n P\left(w'_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^j},$$

from tightness of single random element $\{X_k\}$ and $\Rightarrow$ part of this theorem, which is already shown. Define

$$A = \left\{x : \sup_{t\in[0,1]} |x(t)| \leq C_0\right\} \bigcap \bigcap_{j=1}^{\infty} \left\{x : w'_x(\delta_j) < \frac{1}{j}\right\}.$$

Then $A$ has a compact closure by theorem 2.4.22. Letting $K = \bar{A}$, we get

$$\begin{aligned}
\sup_n P(X_n \leq K^c) &\leq \sup_n P(X_n \leq A^c) \\
&= P\left(X_n \in \left\{x : \sup_{t\in[0,1]} |x(t)| > C_0\right\} \bigcup \bigcup_{j=1}^{\infty} \left\{x : w'_x(\delta_j) > \frac{1}{j}\right\}\right) \\
&\leq \sup_n P\left(\sup_{t\in[0,1]} |X_n(t)| > C_0\right) + \sum_{j=1}^{\infty} \sup_n P\left(w'_{X_n}(\delta_j) > \frac{1}{j}\right) \\
&< \epsilon,
\end{aligned}$$

which gives the conclusion. □

**Remark 2.4.27.** Note that (ii) is equivalent to: $\forall \epsilon > 0$ $\forall \eta > 0$ $\exists$ finite partition $\{T_i\}$ s.t.

$$\limsup_{n\to\infty} P\left(\max_i \sup_{s,t\in T_i} |X_n(s) - X_n(t)| \geq \epsilon\right) \leq \eta.$$

This equivalence is clear from that as the partition $\{T_i\}$ becomes finer,

$$\max_i \sup_{s,t\in T_i} |X_n(s) - X_n(t)|$$

becomes smaller, i.e.,

$$P\left(\max_i \sup_{s,t\in T_i} |X_n(s) - X_n(t)| \geq \epsilon\right) \leq \eta$$

becomes smaller. Also note that, such condition holds *if and only if* $\forall \epsilon > 0$ $\exists$ finite partition $\{T_i\}$ s.t.

$$\limsup_{n\to\infty} P\left(\max_i \sup_{s,t\in T_i} |X_n(s) - X_n(t)| \geq \epsilon\right) \leq \eta.$$

*If* part is obtained from applying the condition to $\epsilon^* = \eta \wedge \epsilon$; *only if* part is clear as letting $\eta = \epsilon$.

**Theorem 2.4.28.** *A sequence $\{X_n\}$ in $\mathbb{D}$ is tight if and only if*

*(i) the sequence of random variables $\left\{\sup_{t\in[0,1]} |X_n(t)|\right\}$ is tight in $\mathbb{R}$, and*

*(ii)* $\lim_{\delta\to 0} \limsup_{n\to\infty} P(w''_{X_n}(\delta) \geq \epsilon) = 0$

*(iii)* $\lim_{\delta\to 0} \limsup_{n\to\infty} P\left(\sup_{s,t\in[0,\delta)} |X_n(s) - X_n(t)| \geq \epsilon\right) = 0$

*(iv)* $\lim_{\delta\to 0} \limsup_{n\to\infty} P\left(\sup_{s,t\in[1-\delta,1)} |X_n(s) - X_n(t)| \geq \epsilon\right) = 0$

*hold for any $\epsilon > 0$.*

*Proof.* $\Rightarrow$) Suppose that $\{X_n\}$ is tight in $\mathbb{D}$. For given $\epsilon > 0$, $\exists$a compact set $K$ s.t. $\inf_n P(X_n \in K) > 1 - \epsilon$. Then by theorem 2.4.25, $\exists C_0$ s.t.

$$(X_n \in K) \subseteq \left(\sup_{t\in[0,1]} |X_n(t)| \leq C_0\right),$$

which gives (i). Also, $\exists \delta_0 > 0$ s.t. $\sup_{x\in K} w''_x(\delta_0) < \epsilon$, so we get

$$(X_n \in K) \subseteq (w''_{X_n}(\delta_0) < \epsilon),$$

which gives (ii). For (iii), we use that $\exists \delta_1 > 0$ s.t.

$$\sup_{x\in K} w_x[0,\delta_1) = \sup_{x\in K} \sup_{s,t\in[0,\delta_1)} |x(s) - x(t)| < \epsilon,$$

which gives (iii). Similarly we get (iv).

$\Leftarrow$) Conversely, now suppose that (i)-(iv) hold. Given $\epsilon > 0$, $\exists C_0$ s.t.

$$\sup_n P\left(\sup_{t\in[0,1]} |X_n(t)| > C_0\right) < \frac{\epsilon}{4}.$$

Also, we can find that $\exists \delta_0 > 0$ s.t.

$$\limsup_{n\to\infty} P\left(\sup_{s,t\in[0,\delta_0)} |X_n(s) - X_n(t)| \geq \epsilon\right) < \frac{\epsilon}{4}.$$

Actually, we can find $\delta_0 > 0$ s.t.

$$\sup_n P\left(\sup_{s,t\in[0,\delta_0)} |X_n(s) - X_n(t)| \geq \epsilon\right) < \frac{\epsilon}{4},$$

from tightness of single element and $\Rightarrow$ part of this theorem. In addition, we can further say that $\delta_0$ satisfies

$$\sup_n P\left(\sup_{s,t\in[1-\delta_0,1)} |X_n(s) - X_n(t)| \geq \epsilon\right) < \frac{\epsilon}{4},$$

which comes from (iv). Therefore, if one shows that

$$\forall j \geq 1 \ \exists \delta_j > 0 \text{ s.t. } \sup_n P\left(w''_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^{j+1}},$$

then

$$A = \left\{x : \sup_{t\in[0,1]} |x(t)| \leq C_0\right\} \bigcap \left\{x : \sup_{s,t\in[0,\delta_0)} |x(s) - x(t)| < \epsilon\right\} \bigcap \left\{x : \sup_{s,t\in[1-\delta_0,1)} |x(s) - x(t)| < \epsilon\right\}$$
$$\bigcap \bigcap_{j=1}^{\infty} \left\{x : w''_x(\delta_j) < \frac{1}{j}\right\}$$

has a compact closure, and we get the desired compact set $K := \bar{A}$, which satisfies

$$\begin{aligned}
\sup_n P(X_n \in K^c) &\leq \sup_n P(X_n \in A^c) \\
&\leq \sup_n P\left(\sup_{t\in[0,1]} |X_n(t)| > C_0\right) + \sup_n P\left(\sup_{s,t\in[0,\delta_0)} |X_n(s) - X_n(t)| \geq \epsilon\right) \\
&\quad + \sup_n P\left(\sup_{s,t\in[1-\delta_0,1)} |X_n(s) - X_n(t)| \geq \epsilon\right) + \sum_{j=1}^{\infty} P\left(w''_{X_n}(\delta_j) \geq \frac{1}{j}\right) \\
&< \epsilon.
\end{aligned}$$

Thus our remain part is:

**Claim.** $\forall j \geq 1\ \exists \delta_j > 0$ s.t.

$$\sup_n P\left(w''_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^{j+1}}.$$

By theorem 2.4.22, we can find $\delta_j > 0$ s.t.

$$\limsup_{n\to\infty} P\left(w''_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^{j+1}},$$

i.e., $\exists n_j$ s.t.

$$\sup_{n\geq n_j} P\left(w''_{X_n}(\delta_j) \geq \frac{1}{j}\right) < \frac{\epsilon}{2^{j+1}}.$$

Then tightness of single element and $\Rightarrow$ part of this theorem says that, letting $\delta_j$ smaller, we can guarantee that claim holds. □

### 2.4.5 Weak Convergence in $(\mathbb{D}, \mathcal{D}, d_S)$

Now we reach to our final goal.

**Theorem 2.4.29.** *Let $X_n$ and $X$ be random elements in $\mathbb{D}$. Suppose that*

$$P(X \text{ is discontinuous at } 1) = 0.$$

*(Note that $X$ is clearly continuous at 0, since its sample paths are càdlàg .) If*

$$(X_n(t_1), \cdots, X_n(t_k)) \xrightarrow[n\to\infty]{d} (X(t_1), \cdots, X(t_k))\ \forall t_1, \cdots, t_k \in T_X,$$

*and further if*

$$\lim_{\delta\to 0} \limsup_{n\to\infty} P(w''_{X_n}(\delta) \geq \epsilon) = 0\ \forall \epsilon > 0, \tag{2.5}$$

*then $X_n \xrightarrow[n\to\infty]{d} X$.*

*Proof.* By theorem 2.4.18, we only have to show tightness, so we will show (i), (iii), and (iv) of theorem 2.4.28 ((ii) is given). As you can see below, given condition (2.5) is critical. From now on, let $\epsilon > 0$ be given.

(i) From (2.5), $\exists \delta_0 > 0$ s.t

$$\limsup_{n\to\infty} P(w''_{X_n}(\delta_0) \geq \epsilon) \leq \frac{\epsilon}{2}.$$

Now, let $0 = t_1 < t_2 < \cdots < t_k = 1$ be a partition s.t. $t_i - t_{i-1} \leq \delta_0$, $t_i \in T_X$ for any $i = 1, 2, \cdots, k$ ($T_X$ is dense so we can find such partition). If $t_i$ is the nearest grid point of $t$, then from the definition of $w''$

$$\sup_{t,i} |X_n(t) - X_n(t_i)| \leq \sup_{\substack{0\leq s_1\leq t\leq s_2\leq 1 \\ s_2 - s_1 \leq \delta_0}} |X_n(t) - X_n(s_1)| \wedge |X_n(s_2) - X_n(t)| = w''_{X_n}(\delta_0)$$

holds, and so

$$\sup_{t\in[0,1]} |X_n(t)| \leq w''_{X_n}(\delta_0) + \max_{1\leq i\leq k} |X_n(t_i)|.$$

Since $\{X_n(t_i)\}$ is tight in $\mathbb{R}$ for fixed grid point $t_i$, and we are considering a finite partition, $\{\max_{1\leq i\leq k} |X_n(t_i)|\}$ is also tight in $\mathbb{R}$, and hence $\exists C > 0$ s.t.

$$\limsup_{n\to\infty} P\left(\max_{1\leq i\leq k} |X_n(t_i)| > C\right) \leq \frac{\epsilon}{2}.$$

Therefore, we get

$$\begin{aligned}
\limsup_{n\to\infty} P\left(\sup_{t\in[0,1]} |X_n(t)| > C + \epsilon\right) &\leq \limsup_{n\to\infty} P\left(w''_{X_n}(\delta_0) > \epsilon \text{ or } \max_{1\leq i\leq k} |X_n(t_i)| > C\right) \\
&\leq \limsup_{n\to\infty} P(w''_{X_n} > \epsilon) + \limsup_{n\to\infty} P\left(\max_{1\leq i\leq k} |X_n(t_i)| > C\right) \\
&\leq \epsilon,
\end{aligned}$$

which gives (i).

(iii) To show (iii), we have to show following.

**Claim (iii)-1.** $\forall \epsilon > 0$ $\exists \delta_0$, $n_0$ s.t.

$$\sup_{n\geq n_0} P\left(\sup_{s,t\in[0,\delta_0)} |X_n(s) - X_n(t)| \geq \epsilon\right) \leq \epsilon.$$

However,

$$\begin{aligned}
\sup_{s,t\in[0,\delta)} |X_n(s) - X_n(t)| &\leq \sup_{s\in[0,\delta)} |X_n(s) - X_n(0)| + \sup_{t\in[0,\delta)} |X_n(0) - X_n(t)| \\
&= 2 \sup_{s\in[0,\delta)} |X_n(s) - X_n(0)|
\end{aligned}$$

$$\begin{cases} \leq 2w''_{X_n}(\delta) \leq 2\left(w''_{X_n}(\delta) + |X_n(\delta) - X_n(0)|\right) \text{ if } |X_n(s) - X_n(0)| \leq |X_n(s) - X_n(\delta)| \\ \leq 2\sup_{s\in[0,\delta)} |X_n(s) - X_n(\delta)| + 2|X_n(\delta) - X_n(0)| \leq 2\left(w''_{X_n}(\delta) + |X_n(\delta) - X_n(0)|\right) \\ \text{if } |X_n(s) - X_n(0)| > |X_n(s) - X_n(\delta)| \end{cases}$$
$$\leq 2\left(w''_{X_n}(\delta) + |X_n(\delta) - X_n(0)|\right)$$

holds for any $\delta > 0$ ($|X_n(s) - X_n(0)| \leq w''_{X_n}(\delta)$ if $|X_n(s) - X_n(0)| \leq |X_n(s) - X_n(\delta)|$; $|X_n(s) - X_n(\delta)| \leq w''_{X_n}(\delta)$ if $|X_n(s) - X_n(0)| > |X_n(s) - X_n(\delta)|$). Now let $n_1$ and $\delta_1 \in T_X$ be s.t.

$$\sup_{n\geq n_1} P\left(w''_{X_n}(\delta_1) \geq \frac{\epsilon}{4}\right) \leq \frac{\epsilon}{4}.$$

(Such $n_1$ and $\delta_1$ exist from (2.5)) Thus if one can find $\delta_0 \leq \delta_1$ s.t. $\delta_0 \in T_X$ and

$$P\left(|X(\delta_0) - X(0)| \geq \frac{\epsilon}{12}\right) \leq \frac{\epsilon}{4},$$

then from $\delta_0 \in T_X$ and $0 \in T_X$, by the assumption,

$$X_n(\delta_0) \xrightarrow[n\to\infty]{d} X(\delta_0) \text{ and } X_n(0) \xrightarrow[n\to\infty]{d} X(0)$$

hold. Without loss of generality, we can assume that $X_n(\delta_0)$ and $X(\delta_0)$ are defined on the same probability space with $X_n(\delta_0) \xrightarrow[n\to\infty]{a.s} X(\delta_0)$ by *Skorokhod theorem*, and so on $X_n(0)$. Then we can find $n'_1$ s.t.

$$n > n'_1 \Rightarrow P\left(|X_n(\delta_0) - X(\delta_0)| \geq \frac{\epsilon}{12}\right) \leq \frac{\epsilon}{4},\ P\left(|X_n(0) - X(0)| \geq \frac{\epsilon}{12}\right) \leq \frac{\epsilon}{4},$$

and therefore, we get ($n_0 = n_1 \vee n'_1$)

$$\begin{aligned} \sup_{n\geq n_0} P\left(\sup_{s,t\in[0,\delta_0)} |X_n(s) - X_n(t)| \geq \epsilon\right) &\leq \sup_{n\geq n_0} P\left(2\left(w''_{X_n}(\delta) + |X_n(\delta) - X_n(0)|\right) \geq \epsilon\right) \\ &\leq \underbrace{\sup_{n\geq n_0} P\left(w''_{X_n}(\delta_0) \geq \frac{\epsilon}{4}\right)}_{\leq \epsilon/4} + \sup_{n\geq n_0} P\left(|X_n(\delta_0) - X(\delta_0)| \geq \frac{\epsilon}{12}\right) \\ &+ \sup_{n\geq n_0} P\left(|X(\delta_0) - X(0)| \geq \frac{\epsilon}{12}\right) + \sup_{n\geq n_0} P\left(|X(0) - X_n(0)| \geq \frac{\epsilon}{12}\right) \\ &\leq \epsilon, \end{aligned}$$

i.e., Claim (iii)-1 holds. So the remain part is:

**Claim (iii)-2.** $\exists \delta_0 \leq \delta_1$ s.t. $\delta_0 \in T_X$ and

$$P\left(|X(\delta_0) - X(0)| \geq \frac{\epsilon}{12}\right) \leq \frac{\epsilon}{4}.$$

We obtain this from "right continuity." Since

$$P\left(\lim_{\delta \to 0} \sup_{s \in [0,\delta]} |X(s) - X(0)| = 0\right) = 1,$$

we get

$$P\left(\bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} \left(\sup_{s \in [0,1/l]} |X(s) - X(0)| < \frac{1}{k}\right)\right) = 1,$$

i.e.,

$$\lim_{k \to \infty} \underbrace{P\left(\bigcap_{l=1}^{\infty} \left(\sup_{s \in [0,1/l]} |X(s) - X(0)| < \frac{1}{k}\right)\right)}_{=:(\star)} = 0.$$

However, since $(\star)$ is increasing function of $k$, whose limit is 0, we get $(\star) \equiv 0$, i.e.,

$$P\left(\bigcap_{l=1}^{\infty} \left(\sup_{s \in [0,1/l]} |X(s) - X(0)| < \frac{1}{k}\right)\right) \overset{k \geq 1}{\equiv} 0.$$

It means

$$\lim_{l \to \infty} P\left(\sup_{s \in [0,1/l]} |X(s) - X(0)| < \frac{1}{k}\right) \overset{k \geq 1}{\equiv} 0$$

Thus, $\forall k \geq 1$, $\forall \eta > 0$ $\exists L \geq 1$ s.t.

$$P\left(\sup_{s \in [0,1/L]} |X(s) - X(0)| < \frac{1}{k}\right) \leq \eta.$$

Let $\delta_0$ be s.t. $\delta_0 < 1/L$ and $\delta_0 \in T_X$. Then we get

$$P\left(|X(\delta_0) - X(0)| > \frac{1}{k}\right) \leq P\left(\sup_{s \in [0,1/L]} |X(s) - X(0)| > \frac{1}{k}\right) \leq \eta.$$

Letting $\eta = \epsilon$ and $k > 12/\epsilon$, we get

$$P\left(|X(\delta_0) - X(0)| \geq \frac{\epsilon}{12}\right) \leq P\left(|X(\delta_0) - X(0)| > \frac{1}{k}\right) \leq \epsilon,$$

which proves Claim (iii)-2.

(iv) We can get (iv) in the same way. **BE CAUTIOUS!** In here, we need *"left continuity"*

at 1 to get

$$P\left(|X(1-\delta_0) - X(1)| \geq \frac{\epsilon}{12}\right) \leq \frac{\epsilon}{4},$$

which comes from $P(X \text{ is discontinuous at } 1) = 0$. □

### 2.4.6 Limit process with continuous sample paths

In here, we see a special case of theorem 2.4.29, the case that limit process $X$ has continuous sample paths a.e.. Then $T_X = [0,1]$, so we need weak convergence of *every* finite-dimensional distribution.

**Corollary 2.4.30.** *Let $X_n$ and $X$ be random elements in $\mathbb{D}$, and $P(X \text{ is continuous}) = 1$. If every finite dimensional dimensional distribution of $X_n$ converges weakly to those of $X$, and* (2.5) *holds, then $X_n \xrightarrow[n\to\infty]{d} X$.*

In fact, we needed (2.5) to get tightness. If one can show tightness in other way, then we can replace the condition.

**Theorem 2.4.31.** *Let $\{X_n\}$ be a sequence of random elements in $(\mathbb{D}, \mathcal{D}, d_S)$. Suppose that*

*(i)* $\{X_n(0)\}$ *is tight*

*(ii)* $$\lim_{\delta\to 0} \limsup_{n\to\infty} P\left(\sup_{(s,t):|s-t|<\delta} |X_n(s) - X_n(t)| \geq \epsilon\right) = 0 \ \forall \epsilon > 0$$

*Then $\{X_n\}$ is tight, and hence it is relatively compact. Further, weak limit of its subsequence has continuous sample paths a.e..*

**Remark 2.4.32.** Note that, conditions in previous theorem are condition for weak convergence "in $\mathbb{C}$."

One example of such situation is a partial sum process, defined as in the following theorem.

**Theorem 2.4.33.** *Let $\xi_j \overset{i.i.d.}{\sim} (0,1)$ be random variables, and*

$$X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \xi_i$$

*be a partial sum process in $\mathbb{D}$. Then it converges weakly to the standard BM $W$, i.e.,*

$$X_n \xrightarrow[n\to\infty]{d} W.$$

*Proof.* Weak convergence of finite dimensional distributions comes from classical CLT. Note that $X_n(0) \equiv 0$. Thus it's sufficient to show that $\forall \epsilon > 0$ $\exists \delta_0 > 0$ and $n_0$ s.t.

$$sup_{n \geq n_0} \sup_{t \in [0,1]} P\left( \sup_{s \in [t, t+\delta_0]} |X_n(s) - X_n(t)| \geq \frac{\epsilon}{3} \right) \leq \delta_0 \epsilon.$$

(cf. Theorem 2.3.11) Let $s \in (t, t+\delta_0]$. There exists $i, j$ s.t.

$$\frac{i}{n} \leq s < \frac{i+1}{n}, \ \frac{j}{n} \leq t < \frac{j+1}{n}.$$

Then $k - j - 1 \leq n\delta_0$,

$$X_n(s) = \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_i) \text{ or } \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_{i+1})$$

and

$$X_n(t) = \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_j),$$

so letting $I(\delta, j) = (j + 1 + n\delta_0) \wedge (n-1)$, we get

$$\sup_{s \in [t, t+\delta_0]} |X_n(s) - X_n(t)| \leq \max_{j \leq i \leq I(\delta,j)+1} \frac{1}{\sqrt{n}} |\xi_{j+1} + \cdots + \xi_i|,$$

and we can prove the theorem in the same way as that of Donsker's theorem in $\mathbb{C}$ space, theorem 2.3.11. □

**Remark 2.4.34.** (Open Question) In the lecture note, it also uses theorem 2.4.29 to complete the proof. However, I think this step is unnecessary. Please contact me if I am wrong.

## 2.5 Weak Convergence of Empirical Process

We see an application of the theory we learned. With i.i.d. observations $\xi_1, \cdots, \xi_n$, an *empirical process*

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(\xi_i \leq t)$$

is a random element whose sample paths are cádlág.

### 2.5.1 Uniform Empirical Process

In this section, we see a weak convergence of empirical process based on $U[0,1]$ random sample.

**Definition 2.5.1** (Brownian Bridge)**.** *Let $W$ be a standard BM. Then*

$$B(t) := W(t) - tW(1),\ t \in [0,1]$$

*is called a* ***Brownian Bridge*** *(BB).*

**Remark 2.5.2.** (a) Brownian Bridge $B$ is a random element in $\mathbb{C} = \mathbb{C}[0,1]$.

(b) An equivalent definition is:

$$B(t) := (W(t)|W(1) = 0).$$

**Proposition 2.5.3.** *For $s, t \in [0,1]$,*

$$EB(t) = 0,\ Cov(B(s), B(t)) = s \wedge t - st$$

*hold. Especially,*

$$Var(B(t)) = t(1-t),$$

*and hence*

$$P(B(0) = B(1) = 0) = 1.$$

**Theorem 2.5.4.** *Let $\xi_j \overset{i.i.d.}{\sim} U[0,1]$, and $F_n$ be their empirical distribution function,*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i \le t).$$

*Define*

$$X_n(t) = \sqrt{n}(F_n(t) - t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I(\xi_i \le t) - t).$$

*Then for BB $B$,*

$$X_n \xrightarrow[n\to\infty]{d} B.$$

*Proof.* Note that $P(B \text{ is continuous}) = 1$, and every finite dimensional distribution of $X_n$ converges to that of $B$ by classical CLT. To show tightness, we will use theorem 2.4.31. However, we can prove (ii) of theorem 2.4.31 via showing (2.3) (cf. theorem 2.3.11). So our claim is:

**Claim.** $\forall \epsilon > 0,\ \forall \eta > 0\ \exists \delta > 0$ and $n_0$ s.t.

$$\sup_{n \geq n_0} \sup_{t \in [0,1]} P\left( \sup_{s \in [t, t+\delta]} |X_n(s) - X_n(t)| \geq \eta \right) = 0.$$

Divide $[t, t+\delta]$ into $m$ subintervals $[t+ip, t+(i+1)p]$, $i = 0, 1, \cdots, m-1$, with same width, where $p = \delta/m$. For $s \in [0,1]$, $\exists i$ s.t. $s \in [t+ip, t+(i+1)p]$, so from

$$|X_n(S) - X_n(t)| \leq |X_n(s) - X_n(t+ip)| + |X_n(t+ip) - X_n(t)|,$$

we get

$$\sup_{s \in [t, t+\delta]} |X_n(s) - X_n(t)| \leq \underbrace{\max_{0 \leq i \leq m} |X_n(t+ip) - X_n(t)|}_{=\text{(I)}} + \underbrace{\max_{0 \leq i \leq m-1} \sup_{s \in [t+ip, t+(i+1)p]} |X_n(s) - X_n(t+ip)|}_{=\text{(II)}}.$$

(I) Note that

$$X_n(t+ip) - X_n(t) = \sum_{l=1}^{i} \underbrace{(X_n(t+lp) - X_n(t+(l-1)p))}_{=:\zeta_l} = \zeta_1 + \cdots + \zeta_i.$$

Let

$$S_i = \zeta_1 + \cdots + \zeta_i.$$

Recall the lemma 2.3.15:

$$P\left( \max_{0 \leq i \leq m} |S_i| \geq \lambda \right) \leq \frac{C_{\gamma,\alpha}}{\lambda^\gamma} (u_1 + \cdots + u_m)^\alpha$$

if

$$E|S_j - S_i|^\gamma \leq (u_{i+1} + \cdots + u_j)^\alpha \text{ for some } \alpha > 1, \gamma \geq 0, u_i \geq 0.$$

Now note that

$$\begin{aligned} S_j - S_i &= \zeta_{i+1} + \cdots + \zeta_j \\ &= X_n(t+jp) - X_n(t+ip) \\ &= \sqrt{n}(F_n(t+jp) - (t+jp)) - \sqrt{n}(F_n(t+ip) - (t+ip)) \\ &= \sqrt{n}(F_n(t+jp) - F_n(t+ip) - (j-1)p) \end{aligned}$$

$$= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left( I(\xi_k \le t + jp) - I(\xi_k \le t + ip) - (j-1)p \right),$$

and since $I(\xi_k \le s) - I(\xi_k \le t) = -1$, 0 or 1, we get

$$\begin{aligned} E\left(I(\xi_k \le s) - I(\xi_k \le t)\right)^4 &= E\left(I(\xi_k \le s) - I(\xi_k \le t)\right)^2 \\ &= EI(\xi_k \le s) + EI(\xi_k \le t) - 2EI(\xi_k \le s)I(\xi_k \le t) \\ &= s + t - 2s \wedge t \\ &= |s - t|. \end{aligned} \tag{$\star$}$$

Now consider

$$\begin{aligned} E|S_j - S_i|^4 &= \frac{1}{n^2} E \left[ \sum_{k=1}^{n} \underbrace{\left( \overbrace{I(\xi_k \le t + jp) - I(\xi_k \le t + ip)}^{(A)} - \overbrace{(j-1)p}^{(B)} \right)}_{=:Y_k} \right]^4 \\ &= \frac{1}{n^2} (nEY_1^4 + 3n(n-1)(EY_1^2)^2) \ (\because EY_1 = 0) \end{aligned}$$

Then

$$\begin{aligned} EY_1^4 &= EA^4 - 4EA^3B + 6EA^2B^2 - 4EAB^3 + EB^4 \\ &= EA^4 - 4(j-i)p \cdot EA^3 + 6((j-i)p)^2 \cdot EA^2 - 4((j-i)p)^3 \cdot EA + ((j-i)p)^4 \\ &\underset{(\star)}{=} |(j-i)p| + 6|(j-i)p|^3 + |(j-i)p|^4 \ (\because A^3 = A) \\ &\le C|(j-i)p| \ (\because |(j-i)p| \le \delta < 1) \end{aligned}$$

and

$$\begin{aligned} EY_1^2 &= EA^2 - 2EAB + EB^2 \\ &= |(j-i)p| + |(j-i)p|^2 \\ &\le C'((j-i)p)^2 \end{aligned}$$

so we get

$$E|S_j - S_i|^4 \le \frac{C|(j-i)p|}{n} + \underbrace{\frac{3n(n-1)}{n^2}}_{\le 3} C'^2((j-i)p)^2$$
$$C|(j-i)p|^2 + 3C'^2|(j-i)p|^2 =: C^*((j-i)p)^2,$$

provided that

$$\frac{1}{n} \le |(j-i)p| \ \forall i, j,$$

i.e., $n \ge 1/p$ (♠). Thus, applying lemma 2.3.15 with $\gamma = 4, \alpha = 2$, and $u_i \equiv \sqrt{C^*}p$, we get

$$P\left(\max_{0\le i\le m} |S_i| \ge \lambda\right) \le \frac{C_{4,2}}{\lambda^4}\left(m\sqrt{C^*}p\right)^2 = \frac{C^0}{\lambda^4}m^2p^2.$$

Let $\lambda = \eta/2$. Then using $S_i = X_n(t+ip) - X_n(t)$, we obtain

$$P\left(\max_{0\le i\le m} |X_n(t+ip) - X_n(t)| \ge \frac{\eta}{2}\right) \le 16C^0\eta^{-4}m^2p^2 = C_*\eta^{-4}\delta^2. \tag{2.6}$$

(II) Let $s \in [t+ip, t+(i+1)p]$. Then

$$X_n(s) - X_n(t+ip) = \sqrt{n}(F_n(s) - s) - \sqrt{n}(F_n(t+ip) - (t+ip))$$
$$= \sqrt{n}\left(F_n(s) - F_n(t+ip) - (s - (t+ip))\right)$$

holds. Note that

(i) $t+ip \le s \le t+(i+1)p \Rightarrow F_n(s) \le F_n(t+(i+1)p),\ s-(t+ip) \ge 0$, and hence

$$X_n(s) - X_n(t+ip) = \sqrt{n}\left(F_n(s) - F_n(t+ip) - (s-(t+ip))\right)$$
$$\le \sqrt{n}(F_n(t+(i+1)p) - F_n(t+ip))$$
$$= X_n(t+(i+1)p) - X_n(t+ip) + \sqrt{n}p.$$

(ii) $t+ip \le s \le t+(i+1)p$, again, gives

$$X_n(s) - X_n(t+ip) \ge -\sqrt{n}(s-(t+ip))$$
$$\ge -\sqrt{n}(t+(i+1)p - (t+ip))$$
$$= -\sqrt{n}p \ge -(X_n(t+(i+1)p) - X_n(t+ip)) - \sqrt{n}p.$$

By (i) and (ii), we obtain

$$|X_n(s)-X_n(t+ip)| \le |X_n(t+(i+1)p)-X_n(t+ip)+\sqrt{n}p| \le |X_n(t+(i+1)p)-X_n(t+ip)|+\sqrt{n}p.$$

It implies that,

$$\sup_{s\in[t+ip,t+(i+1)p]} |X_n(s)-X_n(t+ip)| \le |X_n(t+(i+1)p)-X_n(t+ip)|+\sqrt{n}p.$$

Thus,

$$P\left((\mathrm{II}) \ge \frac{\eta}{2}\right) \le P\left(\max_{0\le i\le m-1} |X_n(t+(i+1)p)-X_n(t+ip)| > \frac{\eta}{4}\right)$$

provided that $\sqrt{n}p > \eta/4$ (♣). Using $|X_n(t+(i+1)p)-X_n(t+ip)| \le |X_n(t+(i+1)p)-X_n(t)| + |X_n(t)-X_n(t+ip)|$, we get

$$\begin{aligned} P\left((\mathrm{II}) \ge \frac{\eta}{2}\right) &\le P\left(\max_{0\le i\le m-1} |X_n(t+(i+1)p)-X_n(t+ip)| > \frac{\eta}{4}\right) \\ &\le 2P\left(\max_{0\le i\le m} |X_n(t+ip)-X_n(t)| > \frac{\eta}{8}\right) \\ &\le C_{**}\eta^{-4}\delta^2 \ (\because (2.6)) \end{aligned}$$

provided that (♠) and (♣). Therefore, we get

$$\begin{aligned} \sup_{t\in[0,1]} P\left(\sup_{s\in[t,t+\delta]} |X_n(s)-X_n(t)| \ge \eta\right) &\le \sup_{t\in[0,1]} P\left((\mathrm{I}) \ge \frac{\eta}{2}\right) + \sup_{t\in[0,1]} P\left((\mathrm{II}) \ge \frac{\eta}{2}\right) \\ &\le (C_* + C_{**})\eta^{-4}\delta^2 \\ &=: \tilde{C}\eta^{-4}\delta^2. \end{aligned}$$

We should show existence of $n_0$ and $\delta$ that satisfy Claim, given $\eta > 0$ and $\epsilon > 0$. For any $\delta > 0$, letting $n > (4/\eta)^2$ and $m < n\delta$, both (♠) and (♣) are satisfied. Now, take $\delta$ such that $\tilde{C}\eta^{-4}\delta^2 < \epsilon$, i.e.,

$$\delta \le \sqrt{\frac{\eta^4\epsilon}{\tilde{C}}},$$

we get the conclusion. □

### 2.5.2 Empirical Process in $\mathbb{D}(-\infty, \infty)$

Now we see the general case. Let $\xi_1, \cdots, \xi_n$ be a random sample with cdf $F$. Then

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i \le t)$$

and

$$X_n(t) = \sqrt{n}(F_n(t) - F(t))$$

are random elements in $\mathbb{D}(-\infty, \infty)$, the space of all cádlág functions defined on $(-\infty, \infty)$. Thus we need to extend our theory.

**Definition 2.5.5.** *Let $\Lambda$ be the class of all strictly increasing and continuous mappings $\lambda$ of $\mathbb{R}$ "onto" $\mathbb{R}$. Then **Skorokhod metric** for $\mathbb{D}(-\infty, \infty)$ is defined by*

$$d_S(x, y) = \inf_{\lambda \in \Lambda} \max\left( \sup_{t \in (-\infty, \infty)} |\lambda(t) - t|, \sup_{t \in (-\infty, \infty)} |x(t) - y(\lambda(t))| \right).$$

With a slight abuse of notation, we continue to denote by $d_S$ the Skorokhod metric for $\mathbb{D}(-\infty, \infty)$.

**Theorem 2.5.6.** *Let $X_n(t) = \sqrt{n}(F_n(t) - F(t))$ be a scaled and centered empirical process. Then for a BB $B$,*

$$X_n \xrightarrow[n\to\infty]{d} B(F(\cdot)).$$

*Proof.* Let $F^{-1}(t) := \inf\{s : t \le F(s)\}$. Then

$$F^{-1}(t) \le s \Leftrightarrow t \le F(s),$$

and

$$F^{-1}(U) \sim F \text{ if } U \sim U[0, 1].$$

Thus we get

$$(\xi_i)_{1 \le i \le n} \overset{d}{\equiv} \left(F^{-1}(u_i)\right)_{1 \le i \le n}, \text{ where } u_i \overset{i.i.d.}{\sim} U[0, 1].$$

Define

$$Y_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(I(u_i \le t) - t\right).$$

Then we get $X_n \overset{d}{\equiv} Y_n(F(\cdot))$, from

$$X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (I(\xi_i \le t) - F(t)) \overset{d}{\equiv} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(I(F^{-1}(u_i) \le t) - F(t)\right)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (I(u_i \le F(t)) - F(t)) = Y_n(F(t)).$$

Define a function

$$\psi : (\mathbb{D}[0,1], d_S) \to (\mathbb{D}(-\infty, \infty) d_S)$$

as

$$x \mapsto x(F(\cdot)),$$

i.e., $(\psi(x))(t) = x(F(t))$. If $\psi$ is continuous on $\mathbb{C}[0,1]$, then from $Y_n \xrightarrow[n\to\infty]{d} B$ and continuous mapping theorem, we get

$$\psi(Y_n) \xrightarrow[n\to\infty]{d} \psi(B),$$

i.e., $X_n \xrightarrow[n\to\infty]{d} B(F(\cdot))$.

**Claim.** $\psi$ is continuous on $\mathbb{C}[0,1]$.

Let $x_n \in \mathbb{D}[0,1]$ be a sequence converging to a point $x$ in $\mathbb{C}[0,1]$, i.e., $d_S(x_n, x) \to 0$. Then $\exists \lambda_n \in \Lambda$ s.t.

$$\sup_{t\in[0,1]} |x_n(\lambda_n(t)) - x(t)| \to 0 \tag{I}$$

$$\sup_{t\in[0,1]} |\lambda_n(t) - t| \to 0 \tag{II}$$

by the definition of $d_S$. So we get

$$|x_n(t) - x(t)| \le |x_n(t) - x(\lambda_n^{-1}(t))| + |x(\lambda_n^{-1}(t)) - x(t)|$$
$$\le \sup_{s\in[0,1]} |x_n(\lambda_n(s)) - x(s)| + |x(\lambda_n^{-1}(t)) - x(t)|.$$

Thus, we get

$$\sup_{t\in[0,1]} |x_n(t) - x(t)| \le \underbrace{\sup_{s\in[0,1]} |x_n(\lambda_n(s)) - x(s)|}_{\to 0 \text{ (I)}} + \underbrace{\sup_{t\in[0,1]} |x(\lambda_n^{-1}(t)) - x(t)|}_{\to 0 \ (\star)}$$
$$\to 0.$$

Note that in $(\star)$, we used (II) and "uniform" continuity of $x$ ($x$ is continuous on the compact

set $[0,1]$, so it is uniformly continuous). Thus we get

$$d_U(x_n, x) \to 0.$$

Since range of $F$ is contained in $[0,1]$, we get

$$d_U(\psi x_n, \psi x) = \sup_{t\in\mathbb{R}} |x_n(F(t)) - x(F(t))| \leq \sup_{t\in[0,1]} |x_n(t) - x(t)| \to 0,$$

and therefore, we get

$$d_S(\psi x_n, \psi x) \leq d_U(\psi x_n, \psi x) \to 0,$$

which yields continuity of $\psi$. □

# Chapter 3

# Nonparametric Estimation

## 3.1 Density Estimation and Histogram

In this section, we use following notations. Let $X_1, X_2, \cdots, X_n$ be an i.i.d. observations with probability density function $f$ (w.r.t. Lebesgue measure). Also, let $B_j$ be the $j$th bin of histogram such that

$$\bigcup_{j\in\mathbb{Z}} B_j = \mathbb{R},\ B_i \cap B_j = \phi \text{ if } i \neq j.$$

And let $b_j = \lambda(B_j)$ be the "binwidth" of $j$th bin. In here, $\lambda(\cdot)$ is a Lebesgue measure.

**Definition 3.1.1.** ***Histogram*** *is defined as*

$$\begin{aligned}\hat{f}(x) &= \sum_{j\in\mathbb{Z}} \left( \frac{1}{nb_j} \sum_{i=1}^{n} I(X_i \in B_j) \right) I(x \in B_j) \\ &= \frac{1}{nb_j} \sum_{i=1}^{n} I(X_i \in B_j) \ \textit{if } x \in B_j.\end{aligned}$$

Note that, by the definition of pdf,

$$P(X_1 \in B_j) \approx b_j f(x) I(x \in B_j)$$

if $b_j \approx 0$. On the other hand, by WLLN,

$$P(X_1 \in B_j) \approx \frac{1}{n}(\#X_i\text{'s in the } j\text{th bin}),$$

when $n \to \infty$. Thus we get $\hat{f}(x) \approx f(x)$. More precisely, following holds.

**Proposition 3.1.2** (Bias and variance of $\hat{f}$)**.** *Let $b_j \equiv h$ and $B_j = [(j-1)h, jh)$ ("equally spaced bin"). Then*

$$bias(\hat{f}(x)) = \left(\frac{2j(x)-1}{2}h - x\right) f'(x) + o(h) \tag{3.1}$$

$$var(\hat{f}(x)) = \frac{1}{nh} f(x) + o\left(\frac{1}{nh}\right) \tag{3.2}$$

*hold, provided that $h \to 0$, $nh \to \infty$ as $n \to \infty$, and $f$ is continuously differentiable on a neighborhood of $x$. In here, $j(x)$ denotes the index of the bin containing $x$.*

*Proof.* (Bias) Let

$$f_j = \sum_{i=1}^{n} I(X_i \in B_j). \qquad \text{("bin frequency")}$$

Then

$$\hat{f}(x) = \sum_{j \in \mathbb{Z}} \frac{1}{nh} f_j I(x \in B_j) = \frac{1}{nh} f_{j(x)} I(x \in B_{j(x)}).$$

Then for

$$p_j = P(X \in B_j) = \int_{B_j} f = \int_{(j-1)h}^{jh} f(t)dt,$$

we get

$$f_j \sim Bin(n, p_j),$$

and hence

$$E\hat{f}(x) = \frac{1}{nh} E f_j = \frac{p_j}{h} = \frac{1}{h} \int_{(j-1)h}^{jh} f(t)dt,$$

with the abuse of notation $j(x) = j$. So we get

$$\begin{aligned}
bias(\hat{f}(x)) &= E\hat{f}(x) - f(x) \\
&= \frac{1}{h} \int_{(j-1)h}^{jh} f(t)dt - f(x) \\
&= \frac{1}{h} \int_{(j-1)h}^{jh} (f(t) - f(x))dt \\
&= \frac{1}{h} \int_{(j-1)h}^{jh} \int_{x}^{t} f'(s)dsdt \\
&= \frac{1}{h} \int_{(j-1)h}^{jh} \int_{x}^{t} (f'(x) + (f'(s) - f'(x))dt \\
&= \frac{1}{h} \int_{(j-1)h}^{jh} \left( f'(x)(t-x) + \int_{x}^{t} f'(s) - f'(x)dx \right) dt
\end{aligned}$$

$$= \underbrace{\frac{1}{h}\int_{(j-1)h}^{jh} f'(x)(t-x)dt}_{=\text{(I)}} + \underbrace{\frac{1}{h}\int_{(j-1)h}^{jh}\int_{x}^{t}(f'(s)-f'(x))dsdt}_{=\text{(II)}}. \tag{3.3}$$

Note that

$$\begin{aligned}
\text{(I)} &= \frac{1}{h}\int_{(j-1)h}^{jh} f'(x)(t-x)dt \\
&= f'(x)\int_{(j-1)h}^{jh}\frac{t-x}{h}dt \ \left(\frac{t-x}{h}=u\right) \\
&= hf'(x)\int_{j-1-x/h}^{j-x/h} u du \\
&= \frac{hf'(x)}{2}\left[\left(j-\frac{x}{h}\right)^2-\left(j-\frac{x}{h}-1\right)^2\right] \\
&= hf'(x)\left[j-\frac{x}{h}-\frac{1}{2}\right]
\end{aligned}$$

holds, which gives the leading term of (3.1). Now the remain part is (II) $= o(h)$. Note that

$$\begin{aligned}
\text{(II)} &= \frac{1}{h}\int_{(j-1)h}^{jh}\int_{x}^{t} f'(s)-f'(x)dsdt \ \left(w=\frac{s-x}{t-x}\right) \\
&= \frac{1}{h}\int_{(j-1)h}^{jh}(t-x)\int_{0}^{1} f'(x+(t-x)w)-f'(x)dwdt \ \left(\frac{t-x}{h}=u\right) \\
&= \int_{j-1-x/h}^{j-x/h} hu\int_{0}^{1} f'(x+whu)-f'(x)dwdu.
\end{aligned}$$

(♠) However, from $x \in [(j-1)h, jh)$, we get

$$\frac{x}{h}-1 < j-1 \le \frac{x}{h} < j \le \frac{x}{h}+1,$$

i.e.,

$$-1 < j-1-\frac{x}{h} \le 0 < j-\frac{x}{h} \le 1,$$

so in the domain of integral $u \in [j-1-h/x, j-h/x)$, $|hu| \le h$ should be held. (♠) So we get

$$\begin{aligned}
|\text{(II)}| &\le \int_{j-1-x/h}^{j-x/h} h|u|\int_{0}^{1}|f'(x+whu)-f'(x)|dwdu \\
&\le \int_{j-1-x/h}^{j-x/h} h\int_{0}^{1}\sup_{|\delta|\le h}|f'(x+\delta)-f'(x)|dwdu \\
&= h\sup_{|\delta|\le h}|f'(x+\delta)-f'(x)|
\end{aligned} \tag{3.4}$$

$$= o(h).$$

We get the last equality because, from continuity of $f'$,

$$\sup_{|\delta|\le h} |f'(x+\delta) - f'(x)| \xrightarrow[h\to 0]{} 0$$

holds. It gives (3.1).

**Note.** We will use similar logic as the part in (♠) repeatedly among all of this chapter.

(Variance) From $f_j \sim Bin(n, p_j)$, we get

$$\begin{aligned}
var(\hat{f}(x)) &= var\left(\frac{1}{nh} f_{j(x)} I(x \in B_{j(x)})\right) \\
&= \frac{np_j(1-p_j)}{n^2h^2} \\
&= \frac{p_j}{nh^2}(1+o(1)) \ (\because p_j \xrightarrow[h\to 0]{} 0) \\
&= \frac{1}{nh} \cdot \frac{1}{h} \int_{(j-1)h}^{jh} f(t)dt \cdot (1+o(1)) \\
&= \frac{1}{nh}\left(f(x) + \frac{1}{h}\int_{(j-1)h}^{jh} f(t) - f(x)dt\right)(1+o(1)) \ \left(\frac{t-x}{h} = u\right) \\
&= \frac{1}{nh}\left(f(x) + \frac{1}{h}\int_{(j-1)-x/h}^{j-x/h} f(x+hu) - f(x)du\right)(1+o(1)).
\end{aligned}$$

Note that,

$$\begin{aligned}
\left|\int_{j-1-x/h}^{j-x/h} (f(x+hu) - f(x))du\right| &\le \int_{j-1-x/h}^{j-x/h} |f(x+hu) - f(x)|du \\
&\le \int_{j-1-x/h}^{j-x/h} \sup_{|\delta|\le h} |f(x+\delta) - f(x)|du \\
&= \sup_{|\delta|\le h} |f(x+\delta) - f(x)| = o(1)
\end{aligned}$$

from continuity of $f$. Thus we get

$$var(\hat{f}(x)) = \left(\frac{1}{nh} f(x) + o\left(\frac{1}{nh}\right)\right)(1+o(1)) = \frac{1}{nh} f(x) + o\left(\frac{1}{nh}\right).$$

□

**Remark 3.1.3.** (a) From this, we can conclude that $\hat{f}(x)$ is a consistent estimator for $f(x)$

when $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(b) With small $h$, bias becomes small, but variance becomes large. With large $h$, variance becomes small, but bias becomes large. In other words, there exists a *variance-bias tradeoff*! Thus, selection of proper $h$ is an important issue. Oftenly, MISE (mean integrated squared error) is used as a measurement of error.

**Definition 3.1.4.** *MISE of $\hat{f}$ is defined as*

$$MISE(\hat{f}) = E\int(\hat{f}(x) - f(x))^2 dx = \int MSE(\hat{f}(x))dx.$$

**Theorem 3.1.5.** *For the histogram estimator $\hat{f}$,*

$$MISE(\hat{f}) = \frac{1}{nh} + \frac{h^2}{12}\int \left(f'(t)\right)^2 dt + o(h^2) + o\left(\frac{1}{nh}\right),$$

*provided that $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $f$ is bounded and continuously differentiable on $\mathbb{R}$, and $\int (f')^2 < \infty$.*

**Remark 3.1.6.** Be careful: We cannot apply proposition 3.1.2 directly! $o$-terms in the proposition depends on $x$, which is integrated in this theorem. Furthermore, we cannot say that $\int g_h(x)$ converges to 0 only from pointwise convergence of $g_h(x)$ (to 0). For this, we need Lebesgue Dominated Convergence Theorem (LDCT).

*Proof.* In this proof, we only handle the leading terms (For the remainder parts, see next remark). We define "asymptotic bias" and "asymptotic variance" as

$$\begin{aligned} Abias^2(\hat{f}(x)) &= h\left(\frac{2j(x)-1}{2} - \frac{x}{h}\right) f'(x) \\ Avar(\hat{f}(x)) &= \frac{1}{nh} f(x). \end{aligned}$$

Bias term becomes

$$\begin{aligned} \int_{-\infty}^{\infty} Abias^2(\hat{f}(x))dx &= \int_{-\infty}^{\infty} h^2 \left(\frac{2j(x)-1}{2} - \frac{x}{h}\right)^2 f'(x)^2 I(x \in B_{j(x)})dx \\ &\underset{(*)}{=} \int_{-\infty}^{\infty} h^2 \left(\frac{2j(x)-1}{2} - \frac{x}{h}\right)^2 f'\left(\frac{2j(x)-1}{2}h\right)^2 I(x \in B_{j(x)})dx + o(h^2) \\ &= \sum_{j\in\mathbb{Z}} \int_{(j-1)h}^{jh} h^2 \left(\frac{2j-1}{2} - \frac{x}{h}\right)^2 f'\left(\frac{2j-1}{2}h\right)^2 dx + o(h^2) \end{aligned}$$

$$= h^2 \sum_{j\in\mathbb{Z}} \frac{h}{12} f'\left(\frac{2j-1}{2}h\right)^2 + o(h^2).$$

Since

$$\sum_{j\in\mathbb{Z}} h f'\left(\frac{2j-1}{2}h\right)^2 = \int (f')^2 + o(1)$$

as $h \to 0$ by *Riemann Integration*, we get

$$int_{-\infty}^{\infty} Abias^2(\hat{f}(x))dx = \frac{h^2}{12}\int (f')^2 + o(h^2).$$

Thus our claim is that $(*)$ holds.

**Claim.** For

$$g_h(x) := \left(\frac{2j(x)-1}{2} - \frac{x}{h}\right)^2 \left(f'(x)^2 - f'\left(\frac{2j-1}{2}h\right)^2\right) I(x \in B_{j(x)}),$$

we get

$$\int g_h(x)dx \xrightarrow[h\to 0]{} 0.$$

First, from continuity of $f'$, and the definition of $j(x)$,

$$\left|\frac{2j(x)-1}{2} - \frac{x}{h}\right| \le 1$$

and hence

$$g_h(x) \xrightarrow[h\to 0]{} 0$$

holds. Next, note that $(2j(x)-1)h/2$ is "the median of $j$th bin $B_j$." Thus, if we define

$$g(x) = 2 \max_{x\in[(j-1)h,jh)} f'(x)^2 \text{ if } x \in [(j-1)h, jh),$$

then we get $|g_h| \le g$ and $\int g < \infty$ ($\because \int g/2$ is a Riemann supremum sum of $(f')^2$).

Thus by LDCT, we get the claim. For variance,

$$\int_{-\infty}^{\infty} Avar(\hat{f}(x))dx = \frac{1}{nh}\int f(x)dx = \frac{1}{nh},$$

so we achieve our goal. □

**Remark 3.1.7.** In fact, we should deal with remainder term, too.

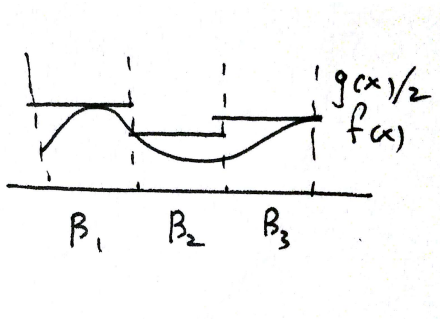


Figure 3.1: Characterization of $g$

(Bias) From (3.3), we get

$$bias^2(\hat{f}(x)) = ((\mathrm{I}) + (\mathrm{II}))^2 = (\mathrm{I})^2 + 2(\mathrm{I})(\mathrm{II}) + (\mathrm{II})^2.$$

$(\mathrm{I})^2$ is the leading term, so we should show that

$$2(\mathrm{I})(\mathrm{II}) + (\mathrm{II})^2 = o(h^2).$$

However note that

$$|(\mathrm{I})(\mathrm{II})| \leq \sqrt{\int (\mathrm{I})^2}\sqrt{\int (\mathrm{II})^2} = O(h^{1/2}) \cdot \sqrt{\int (\mathrm{II})^2};$$

so if one can show that

$$\int (\mathrm{II})^2 = o(h^2),$$

then we obtain

$$\int bias^2(\hat{f}(x))dx = \frac{h^2}{12}\int (f')^2 + o(h^2).$$

Also note that

$$(\mathrm{II})^2 \leq h^2 \sup_{|\delta|\leq h} |f'(x+\delta) - f'(x)|^2$$

from (3.4). Thus we should show that

$$\int \sup_{|\delta|\leq h} |f'(x+\delta) - f'(x)|^2 dx = o(1)$$

as $h \to 0$. It's clear that $\sup_{|\delta|\leq h} |f'(x+\delta) - f'(x)|^2 = o(1)$, so our claim is:

**Claim.** $\sup_{|\delta|\leq h} |f'(x+\delta) - f'(x)|^2$ has a integrable majorant.

If one proves this claim, then by LDCT, we get the conclusion. Since $\sup_{|\delta|\le h}|f'(x+\delta)-f'(x)|^2$ is a decreasing function of $h$, it's enough to show that

$$\int \sup_{|\delta|\le h_0}|f'(x+\delta)-f'(x)|^2dx<\infty$$

for some $h_0>0$. Now, for sufficiently small $h_0>0$,

$$\begin{aligned}
\int \sup_{|\delta|\le h_0}|f'(x+\delta)-f'(x)|^2dx &= \sum_{j\in\mathbb{Z}}\int_{(j-1)h_0}^{jh_0}\sup_{|\delta|\le h_0}|f'(x+\delta)-f'(x)|^2dx\\
&\le 4\sum_{j\in\mathbb{Z}}\int_{(j-1)h_0}^{jh_0}\sup_{x\in[(j-2)h_0,(j+1)h_0)}|f'(x)|^2dx\\
&\le 4\sum_{j\in\mathbb{Z}}\int_{(j-2)h_0}^{(j+1)h_0}\sup_{x\in[(j-2)h_0,(j+1)h_0)}|f'(x)|^2dx\\
&= 4\left(\sum_{j\equiv 0 \pmod 3}\int_{(j-2)h_0}^{(j+1)h_0}\sup_{x\in[(j-2)h_0,(j+1)h_0)}|f'(x)|^2dx\right.\\
&+\sum_{j\equiv 1 \pmod 3}\int_{(j-2)h_0}^{(j+1)h_0}\sup_{x\in[(j-2)h_0,(j+1)h_0)}|f'(x)|^2dx\\
&\left.+\sum_{j\equiv 2 \pmod 3}\int_{(j-2)h_0}^{(j+1)h_0}\sup_{x\in[(j-2)h_0,(j+1)h_0)}|f'(x)|^2dx\right)\\
&<\infty\ (\because \text{Riemann supremum sum with binwidth } 3h_0)
\end{aligned}$$

holds, which ends the proof.

(Variance) Recall that

$$\begin{aligned}
var(\hat f(x)) &= \frac{p_j(1-p_j)}{nh^2}\\
&= \frac{1}{nh}\cdot\frac{1}{h}\int_{(j-1)h}^{jh}f(t)dt-\frac{1}{n}\left(\frac{1}{h}\int_{(j-1)h}^{jh}f(t)dt\right)^2\\
&= \frac{1}{nh}\left(f(x)+\int_{j-1-x/h}^{j-x/h}(f(x+hu)-f(x))du\right)-\frac{1}{n}\left(f(x)+\int_{j-1-x/h}^{j-x/h}(f(x+hu)-f(x))du\right)^2.
\end{aligned}$$

Thus we have to show that

$$\int\left(\int_{j-1-x/h}^{j-x/h}(f(x+hu)-f(x))du+h\left(f(x)+\int_{j-1-x/h}^{j-x/h}(f(x+hu)-f(x))du\right)^2\right)dx=o(1).$$

Again, we use LDCT. First, note that

$$\begin{aligned}
&\left| \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du + h\left( f(x) + \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du \right)^2 \right| \\
&\leq \int_{j-1-x/h}^{j-x/h} |f(x+hu)-f(x)|du + 2h\left( f(x)^2 + \left( \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du \right)^2 \right) \\
&\leq \sup_{|\delta|\leq h} |f(x+\delta)-f(x)| + 2h\left( f(x)^2 + \sup_{|\delta|\leq h} |f(x+\delta)-f(x)|^2 \right) \\
&\xrightarrow[h\to 0]{} 0.
\end{aligned}$$

Next, we will show that

$$\int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du + h\left( f(x) + \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du \right)^2$$

has an integrable majorant. Note that

$$\begin{aligned}
&\left| \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du + h\left( f(x) + \int_{j-1-x/h}^{j-x/h} (f(x+hu)-f(x))du \right)^2 \right| \\
&\leq \sup_{|\delta|\leq h} |f(x+\delta)-f(x)| + 2h\left( f(x)^2 + \sup_{|\delta|\leq h} |f(x+\delta)-f(x)|^2 \right),
\end{aligned}$$

and since $f^2$ is integrable ($\because$from boundedness of $f$, $\int f^2 \leq \int \|f\|_\infty \cdot |f| < \infty$), $\sup_{|\delta|\leq h} |f(x+\delta) - f(x)|^2$ is also integrable (think Riemann supremum sum again!), and from integrability of $f$, $\sup_{|\delta|\leq h} |f(x+\delta)-f(x)|^2$ is integrable. Therefore,

$$\sup_{|\delta|\leq h} |f(x+\delta)-f(x)| + 2h\left( f(x)^2 + \sup_{|\delta|\leq h} |f(x+\delta)-f(x)|^2 \right)$$

is integrable, and since it is a decreasing function of $h$, by similar logic as handling bias term, we get the desired result.

**Remark 3.1.8.** How to select $h$? One method is to choose one that minimizes "asymptotic MISE"

$$AMISE(\hat{f}, h) = \frac{h^2}{12}\int (f')^2 + \frac{1}{nh}. \tag{3.5}$$

It's easy to show that minimizer $h_{opt}$ of (3.5) is

$$h_{opt} = \left( \frac{6}{n \int (f')^2} \right)^{1/3},$$

and then

$$AMISE(\hat{f}, h_{opt}) \sim n^{-2/3}.$$

Note that MISE of $N(\overline{X}, 1)$ is proportional to $n^{-1}$ when $f \in \{N(\theta, 1) : \theta \in \mathbb{R}\}$. Since histogram is nonparametric approach, its convergence rate ($n^{-2/3}$) is inferior then that of parametric one ($n^{-1}$).

## 3.2 Univariate Kernel Density Estimation

In this section, again assume that

$$X_1, \cdots, X_n \overset{i.i.d.}{\sim} f,$$

where $f$ is a pdf w.r.t. Lebesgue measure. In addition, let $K$ be a *kernel* satisfying

$$\int K = 1.$$

Usually, a symmetric density function is used for $K$. We consider a *Kernel Density Estimator (KDE)*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - X_i}{h} \right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i).$$


Figure 3.2: Histogram vs KDE

**Definition 3.2.1.** *(a) For any function $g$, we define the notation $g_h$ as*

$$g_h(x) = \frac{1}{h} g\left( \frac{x}{h} \right).$$

*(b) Let*

$$f * g(x) = \int f(x-t)g(t)dt = \int f(t)g(x-t)dt$$

*be a convolution.*

Our goal is to derive bias and variance of KDE, as we did in previous section. Note that expectation of KDE is represented as a convolution,

$$E\hat{f}(x) = E\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\right] = EK_h(x-X_1) = \int K_h(x-t)f(t)dt = K_h * f(x).$$

Therefore, we may expand $K_h * f$ first.

**Proposition 3.2.2** (Expansion of $K_h * f$)**.** *Let $f^{(r)}$ be the $r$th derivative of $f$. Assume that*

*(i) $f^{(r)}$ is continuous at $x$, and $f^{(r)} \in \mathcal{L}^1$;*

*(ii) $|t|^{r+1}K(t) \to 0$ as $|t| \to \infty$;*

*(iii) $\int |t|^r K(t)dt < \infty$.*

*Then*

$$K_h * f(x) = \sum_{j=0}^{r} \frac{(-1)^j \mu_j(K)}{j!} f^{(j)}(x)h^j + o(h^r)$$

*as $h \to 0$, where*

$$\mu_j(K) = \int t^j K(t)dt.$$

**Remark 3.2.3.** For the proof, we need Taylor theorem with integral formed remainder:

$$f(x+h) = f(x) + \sum_{j=1}^{r} \frac{h^j}{j!} f^{(j)}(x) + R_r^0(x,h)$$

where

$$R_r^0(x,h) = \frac{h^r}{(r-1)!}\int_0^1 (1-w)^{r-1}(f^{(r)}(x+wh) - f^{(r)}(x))dw.$$

It can be derived as following.

$$\begin{aligned} f(x+h) - f(x) &= \int_0^h f'(x+u_1)du_1 \\ &= hf'(x) + \int_0^h (f'(x+u_1) - f')du_1 \\ &= hf'(x) + \int_0^h \int_0^{u_1} f''(x+u_2)du_2du_1 \end{aligned}$$

$$
\begin{aligned}
&= hf'(x) + \frac{h^2}{2}f''(x) + \int_0^h \int_0^{u_1} f''(x+u_2) - f''(x) du_2 du_1 \\
&= hf'(x) + \frac{h^2}{2}f''(x) + \int_0^h \int_0^{u_1} \int_0^{u_2} f^{(3)}(x+u_3) du_3 du_2 du_1 \\
&= \cdots \\
\Rightarrow R_r^0(x,h) &= \int_0^h \int_0^{u_1} \int_0^{u_2} \cdots \int_0^{u_{r-1}} f^{(r)}(x+u_r) - f^{(r)}(x) du_r \cdots du_2 du_1 \\
&= \int_0^h \int_{u_r}^h \int_{u_{r-1}}^h \cdots \int_{u_2}^h f^{(r)}(x+u_r) - f^{(r)}(x) du_1 du_2 \cdots du_r \\
&= \int_0^h \int_{u_r}^h \int_{u_{r-1}}^h \cdots \int_{u_3}^h (h-u_2) du_2 \cdots du_{r-1} (f^{(r)}(x+u_r) - f^{(r)}(x)) du_r \\
&= \int_0^h \int_{u_r}^h \int_{u_{r-1}}^h \cdots \int_{u_4}^h \frac{(h-u_3)^2}{2} du_3 \cdots du_{r-1} (f^{(r)}(x+u_r) - f^{(r)}(x)) du_r \\
&= \cdots \\
&= \int_0^h \frac{(h-u_r)^{r-1}}{(r-1)!} (f^{(r)}(x+u_r) - f^{(r)}(x)) du_r \\
&= \frac{h^r}{(r-1)!} \int_0^1 (f^{(r)}(x+wh) - f^{(r)}(x)) dw.
\end{aligned}
$$

*Proof.* By Taylor's theorem, we get

$$
\begin{aligned}
K_h * f(x) &= \int_{-\infty}^{\infty} K_h(x-u) f(u) du \\
&= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u) du \\
&= \int_{-\infty}^{\infty} f(x-ht) K(t) dt \\
&= \int_{-\infty}^{\infty} \left( \sum_{j=0}^{r} \frac{(-1)^j}{j!} f^{(j)}(x) (ht)^j + (-1)^r R_r^0(x, ht) \right) K(t) dt \\
&= \sum_{j=0}^{r} \int_{-\infty}^{\infty} \frac{(-1)^j}{j!} f^{(j)}(x) h^j t^j K(t) dt + \int_{-\infty}^{\infty} (-1)^r R_r^0(x, ht) K(t) dt \\
&= \sum_{j=0}^{r} \frac{(-1)^j}{j!} f^{(j)}(x) \mu_j(K) h^j + (-1)^r \int_{-\infty}^{\infty} R_r^0(x, ht) K(t) dt.
\end{aligned}
$$

Representing the remainder term as an integral form, we get

$$
\begin{aligned}
R_r(x) &:= \left| (-1)^r \int_{-\infty}^{\infty} R_r^0(x, ht) K(t) dt \right| \\
&= \left| \int_{-\infty}^{\infty} \frac{h^r t^r}{(r-1)!} \int_0^1 (1-w)^{r-1} (f^{(r)}(x - wht) - f^{(r)}(x)) dw K(t) dt \right|.
\end{aligned}
$$

To show $R_r(x) = o(h^r)$, we should show that

$$(\bigstar) := \left| \int_{-\infty}^{\infty} t^r K(t) \int_0^1 (1-w)^{r-1} (f^{(r)}(x - wht) - f^{(r)}(x)) dw dt \right| \xrightarrow[h \to 0]{} 0.$$

It can be proven as following. First note that

$$\begin{aligned}
&\left| \int_{-\infty}^{\infty} t^r K(t) \int_0^1 (1-w)^{r-1} (f^{(r)}(x - wht) - f^{(r)}(x)) dw dt \right| \\
&\leq \int_{-\infty}^{\infty} |t|^r K(t) \int_0^1 (1-w)^{r-1} |f^{(r)}(x - wht) - f^{(r)}(x)| dw dt \\
&= \int_0^1 \int_{|wht| \leq \epsilon} |t|^r K(t) (1-w)^{r-1} |f^{(r)}(x - wht) - f^{(r)}(x)| dt dw \qquad (:=\text{(A)}) \\
&+ \int_0^1 \int_{|wht| > \epsilon} |t|^r K(t) (1-w)^{r-1} |f^{(r)}(x - wht) - f^{(r)}(x)| dt dw \qquad (:=\text{(B)})
\end{aligned}$$

for any $\epsilon > 0$. Terms (A) and (B) can be dominated as

$$\begin{aligned}
\text{(A)} &= \int_0^1 \int_{|wht| \leq \epsilon} |t|^r K(t) (1-w)^{r-1} |f^{(r)}(x - wht) - f^{(r)}(x)| dt dw \\
&\leq \sup_{|\delta| \leq \epsilon} |f^{(r)}(x + \delta) - f^{(r)}(x)| \int_{-\infty}^{\infty} |t|^r K(t) dt
\end{aligned}$$

and

$$\begin{aligned}
\text{(B)} &= \int_0^1 \int_{|wht| > \epsilon} |t|^r K(t) (1-w)^{r-1} |f^{(r)}(x - wht) - f^{(r)}(x)| dt dw \\
&\leq \int_0^1 \int_{|t| > \frac{\epsilon}{wh}} \frac{wh}{\epsilon} |t|^{r+1} K(t) |f^{(r)}(x - wht)| dt dw + |f^{(r)}(x)| \int_0^1 \int_{|t| > \frac{\epsilon}{wh}} |t|^r K(t) dt dw \\
&\leq \sup_{|t| > \epsilon/h} |t|^{r+1} K(t) \cdot \frac{h}{\epsilon} \int_0^1 \int_{|t| > \frac{\epsilon}{wh}} w |f^{(r)}(x - wht)| dt dw + \int_{|t| > \frac{\epsilon}{h}} |t|^r K(t) dt \ (wht = s) \\
&= \sup_{|t| > \epsilon/h} |t|^{r+1} K(t) \cdot \frac{1}{\epsilon} \int_0^1 \int_{|s| > \epsilon} |f^{(r)}(x - s)| ds dw + \int_{|t| > \frac{\epsilon}{h}} |t|^r K(t) dt \\
&\leq \sup_{|t| > \epsilon/h} |t|^{r+1} K(t) \cdot \frac{1}{\epsilon} \int_{-\infty}^{\infty} |f^{(r)}(x - s)| ds + \int_{|t| > \frac{\epsilon}{h}} |t|^r K(t) dt \\
&= \sup_{|t| > \epsilon/h} |t|^{r+1} K(t) \cdot \frac{1}{\epsilon} \int_{-\infty}^{\infty} |f^{(r)}(s)| ds + \int_{|t| > \frac{\epsilon}{h}} |t|^r K(t) dt.
\end{aligned}$$

Since $|t|^{r+1} K(t)$ decays to 0 and $|t|^r K(t)$ is integrable, we get

$$\begin{aligned}
(\bigstar) &\leq \text{(A)} + \text{(B)} \\
&\leq \sup_{|\delta| \leq \epsilon} |f^{(r)}(x + \delta) - f^{(r)}(x)| \int_{-\infty}^{\infty} |t|^r K(t) dt
\end{aligned}$$

$$+ \sup_{|t|>\epsilon/h} |t|^{r+1}K(t) \cdot \frac{1}{\epsilon}\int_{-\infty}^{\infty} |f^{(r)}(s)|ds + \int_{|t|>\frac{\epsilon}{h}} |t|^r K(t)dt$$
$$\xrightarrow[h\to 0]{} \sup_{|\delta|\le\epsilon} |f^{(r)}(x+\delta) - f^{(r)}(x)| \int_{-\infty}^{\infty} |t|^r K(t)dt.$$

As $\epsilon > 0$ was arbitrary, letting $\epsilon \searrow 0$, we get

$$\limsup_{h\to 0}(\bigstar) \le 0,$$

i.e.,

$$\left|\int_{-\infty}^{\infty} t^r K(t) \int_0^1 (1-w)^{r-1}(f^{(r)}(x-wht) - f^{(r)}(x))dwdt\right| \xrightarrow[h\to 0]{} 0.$$

□

**Remark 3.2.4.** There are *very many* typos or errors in the note of T.A.'s lesson! Readers are recommended to verify all of details line-by-line. If there is any wrong modification in this note, please feel free to contact me.

We can obtain bias and variance term directly from the expansion of convolution $K_h * f$.

**Proposition 3.2.5.** *(a) If*

*(i)* $f''$ *is continuous at* $x$*, and* $f'' \in \mathcal{L}^1$*;*

*(ii)* $|t|^3 K(t) \to 0$ *as* $|t| \to \infty$*;*

*(iii)* $K$ *is symmetric, and satisfies* $\int |t|^2 K(t) < \infty$*,*

*then*

$$bias(\hat{f}(x;h)) = \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2).$$

*(b) If*

*(i)* $f$ *is continuous at* $x$*;*

*(ii)* $|t|K(t) \to 0$ *as* $|t| \to \infty$*, and* $\int K^2 < \infty$*,*

*then*

$$var(\hat{f}(x;h)) = \frac{1}{nh}\mu_0(K^2)f(x) + o\left(\frac{1}{nh}\right).$$

*Proof.* (Bias) From the assumption $\mu_0(K) = 1$, and from symmetricity, $\mu_1(K) = 0$. Thus by previous proposition ($r = 2$),

$$\begin{aligned} E\hat{f}(x) &= K_h * f(x) \\ &= f(x) - \mu_1(K)f'(x)h + \frac{\mu_2(K)}{2}f''(x)h^2 + o(h^2) \\ &= f(x) + \frac{1}{2}f''(x)h^2 + o(h^2), \end{aligned}$$

so we get the conclusion.

(Variance) Using previous proposition ($r = 0$) again, we obtain

$$\begin{aligned} var(\hat{f}(x;h)) &= \frac{1}{nh^2}var\left(K\left(\frac{x - X_1}{h}\right)\right) \\ &= \frac{1}{n}\left[\int K_h^2(x-t)f(t)dt\right] - \frac{1}{n}\left[\int K_h(x-t)f(t)dt\right]^2 \\ &= \frac{1}{n}K_h^2 * f(x) - \frac{1}{n}\left(K_h * f(x)\right)^2 \\ &= \frac{1}{nh}\left(\mu_0(K^2)f(x) + o(1)\right) - \frac{1}{n}\left(\mu_0(K)f(x) + o(1)\right)^2 \\ &= \frac{1}{nh}\left(\mu_0(K^2)f(x) + o(1)\right) - \underbrace{\frac{1}{n}\left(f(x)^2 + o(1)\right)}_{=O(n^{-1})} \\ &= \frac{1}{nh}\mu_0(K^2)f(x) + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n}\right). \end{aligned}$$

Since

$$nh \cdot O\left(\frac{1}{n}\right) = h \cdot O(1) = o(1) \ (h \to 0),$$

we get

$$O\left(\frac{1}{n}\right) = o\left(\frac{1}{nh}\right),$$

thus the assertion holds. □

**Remark 3.2.6.** Verify that conditions in proposition 3.2.2 are satisfied to apply it in the proof of proposition 3.2.5!

Again, selection of bandwidth $h$ is a rising problem. For this, we will find the leading term of MISE.

**Theorem 3.2.7** (Asymptotic MISE of KDE)**.**

*(a) If*

*(i)* $f''$ *is continuous on* $\mathbb{R}$ *and* $f'' \in \mathcal{L}^2$*;*

*(ii)* $|t|^3 K(t) \to 0$ *as* $|t| \to \infty$*;*

*(iii)* $K$ *is a symmetric density, and it satisfies* $\int |t|^2 K(t)dt < \infty$*,*

*then*

$$\int bias^2(\hat{f}(x;h))dx = \frac{h^4}{4}\mu_2(K)^2 \int f''^2 + o(h^4).$$

*(b) If*

*(i)* $f$ *is continuous and bounded on* $\mathbb{R}$*;*

*(ii)* $|t|K(t) \to 0$ *as* $|t| \to \infty$*;*

*(iii)* $\int K^2(x)dx < \infty$*,*

*then*

$$\int var(\hat{f}(x;h))dx = \frac{1}{nh}\mu_0(K^2) + o\left(\frac{1}{nh}\right).$$

*Proof.* Note that

$$bias(\hat{f}(x;h)) = K_h * f(x) - f(x) = \frac{f''(x)}{2}\mu_2(K)h^2 + R_2(x), \ (\because \mu_1(K) = 0)$$

where

$$R_2(x) = h^2 \int_{-\infty}^{\infty} t^2 K(t) \int_0^1 (1-w)\left(f''(x-wht) - f''(x)\right) dwdt.$$

Thus we get

$$\int bias^2(\hat{f}(x;h))dx = \int \frac{h^4}{4} f''(x)^2 \mu_2(K)^2 dx + \int f''(x)\mu_2(K)h^2 \cdot R_2(x)dx + \int R_2^2(x)dx.$$

If one shows that

$$\int R_2^2(x)dx = o(h^4),$$

then from

$$\left|\int f''(x)\mu_2(K)h^2 \cdot R_2(x)dx\right| \leq \int |f''(x)\mu_2(K)h^2 \cdot R_2(x)|dx \leq \underbrace{\sqrt{\int h^4 f''(x)^2 \mu_2(K)^2 dx}}_{=O(h^2)} \sqrt{\int R_2^2(x)dx},$$

we get

$$\int f''(x)\mu_2(K)h^2 \cdot R_2(x)dx + \int R_2^2(x)dx = o(h^4)$$

and hence the assertion holds.

**Claim A.** $\int R_2^2(x)dx = o(h^4)$.

By the proof of proposition 3.2.2, $R_2(x) = o(h^2)$, and so we get $R_2^2(x) = o(h^4)$, i.e.,

$$\lim_{h\to 0} h^{-4}R_2^2(x) = 0. \tag{3.6}$$

Next, by considering $t^2K(t)$ as a (positive) finite measure on $\mathbb{R}$,

$$\begin{aligned}
|h^{-4}R_2^2(x)| &\le \left(\int_{-\infty}^{\infty} t^2K(t)\int_0^1 (1-w)\left(f''(x-wht)-f''(x)\right)dwdt\right)^2 \\
&\le \int_{-\infty}^{\infty}\left(\int_0^1 (1-w)\left(f''(x-wht)-f''(x)\right)dw\right)^2 t^2K(t)dt\cdot\int_{-\infty}^{\infty} t^2K(t)dt \\
&\text{(Cauchy-Schwarz)} \\
&\le \int_{-\infty}^{\infty}\int_0^1 \left((1-w)\left(f''(x-wht)-f''(x)\right)\right)^2 dwt^2K(t)dt\cdot\int_{-\infty}^{\infty} t^2K(t)dt \\
&\text{(Cauchy-Schwarz)} \\
&\le \int_{-\infty}^{\infty}\int_0^1 |f''(x-wht)-f''(x)|^2dw\ t^2K(t)dt\cdot\int_{-\infty}^{\infty} t^2K(t)dt \\
&\le \int_{-\infty}^{\infty}\int_0^1 2(f''(x-wht)^2+f''(x)^2)dw\ t^2K(t)dt\cdot\underbrace{\int_{-\infty}^{\infty} t^2K(t)dt}_{=\mu_2(K)}\ (\because (a-b)^2\le 2(a^2+b^2)) \\
&= 2\mu_2(K)\int_{-\infty}^{\infty}\int_0^1 f''(x-wht)^2dw\ t^2K(t)dt + 2\mu_2(K)f''(x)^2\int_{-\infty}^{\infty} t^2K(t)dt \\
&= 2\mu_2(K)\underbrace{\int_{-\infty}^{\infty}\int_0^1 f''(x-wht)^2dw\ t^2K(t)dt}_{=:g_h(x)} + 2\mu_2(K)^2f''(x)^2
\end{aligned}$$

holds. Note that,

$$\begin{aligned}
g_h(x) &= \int_0^1\int_{-\infty}^{\infty} f''(x-wht)^2t^2K(t)dtdw \\
&= \int_0^1\int_{|wht|<\epsilon} f''(x-wht)^2t^2K(t)dtdw + \int_0^1\int_{|wht|\ge\epsilon} f''(x-wht)^2t^2K(t)dtdw \\
&\le \sup_{|\delta|<\epsilon} f''(x+\delta)^2\int t^2K(t)dt + \int_0^1\int_{|wht|\ge\epsilon}\frac{wh|t|}{\epsilon}f''(x-wht)^2t^2K(t)dtdw \\
&\le \sup_{|\delta|<\epsilon} f''(x+\delta)^2\int t^2K(t)dt + \sup_{|t|\ge\epsilon/wh}|t|^3K(t)\int_0^1\int_{|wht|\ge\epsilon}\frac{wh}{\epsilon}f''(x-wht)^2dtdw\ (wht=s) \\
&= \sup_{|\delta|<\epsilon} f''(x+\delta)^2\int t^2K(t)dt + \sup_{|t|\ge\epsilon/wh}|t|^3K(t)\int_0^1\int_{|s|\ge\epsilon}\frac{1}{\epsilon}f''(x-s)^2dsdw
\end{aligned}$$

$$\leq \sup_{|\delta|<\epsilon} f''(x+\delta)^2 \int t^2 K(t)dt + \sup_{|t|\geq \epsilon/wh} |t|^3 K(t) \int \frac{1}{\epsilon} f''(s)^2 ds$$

$$\xrightarrow[h\to 0]{} \sup_{|\delta|<\epsilon} f''(x+\delta)^2 \mu_2(K)$$

for any $\epsilon > 0$, from the assumption that $|t|^3 K(t) \xrightarrow[|t|\to\infty]{} 0$. Since $\epsilon > 0$ was arbitrary, letting $\epsilon \searrow 0$, from continuity of $f''$, we get

$$g_h(x) \xrightarrow[h\to 0]{} f''(x)^2 \mu_2(K).$$

Now, by MVT, $\exists w(t) \in [0,1]$ s.t.

$$g_h(x) = \int_{-\infty}^{\infty} f''(x - w(t)\cdot ht)^2 t^2 K(t)dt,$$

and by Fubini's theorem,

$$\begin{aligned}
\int_{-\infty}^{\infty} g_h(x)dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f''(x - w(t)\cdot ht)^2 t^2 K(t)dtdx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f''(x - w(t)\cdot ht)^2 dx\ t^2 K(t)dt \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f''(x)^2 dx\ t^2 K(t)dt \\
&= \mu_2(K)\int f''(x)^2 dx
\end{aligned}$$

holds. Therefore, with this and (3.6), by (generalized) LDCT (see next remark), we get

$$\int h^{-4} R_2^2(x)dx \xrightarrow[h\to 0]{} 0,$$

which yields claim A.

(Variance) First note that

$$var(\hat{f}(x;h)) = \frac{1}{n}\left(\int K_h^2(x-t)f(t)dt\right) - \frac{1}{n}\left(\int K_h(x-t)f(t)dt\right)^2.$$

For the first term,

$$\begin{aligned}
\int \frac{1}{n}\left(\int K_h^2(x-t)f(t)dt\right)dx &= \frac{1}{n}\int\int K_h^2(x-t)dx\ f(t)dt \\
&= \frac{1}{n}\int K_h^2(x)dx \int f(t)dt
\end{aligned}$$

$$= \frac{1}{nh}\int K^2(x)dx \ (\because f \text{ is density})$$

holds. For the second term, by Cauchy-Schwarz, we get

$$\begin{aligned}\left(\int K_h(x-t)f(t)dt\right)^2 &\leq \left(\int K_h(x-t)f^2(t)dt\right)\underbrace{(K_h(x-t)dt)}_{=1}\\ &\leq \max_y f(y)\cdot\int K_h(x-t)f(t)dt,\end{aligned}$$

and hence

$$\begin{aligned}\int \frac{1}{n}\left(\int K_h(x-t)f(t)dt\right)^2 dx &\leq \frac{1}{n}\max_y f(y)\int\int K_h(x-t)f(t)dtdx\\ &= \frac{1}{n}\max_y f(y)\int\int K_h(x-t)dx\ f(t)dt\\ &= \frac{1}{n}\max_y f(y)\\ &= O(n^{-1}) = o(n^{-1}h^{-1}),\end{aligned}$$

by boundedness of $f$. Therefore, we get

$$\int var(\hat{f}(x;h))dx = \frac{1}{nh}\int K^2(x)dx + o\left(\frac{1}{nh}\right).$$

□

**Remark 3.2.8.** Actually, in the proof of previous theorem, following *generalized* LDCT is used: *Let $\{f_n\}$ be a sequence of measurable functions that converges pointwisely a.e. to $f$. Suppose that there is a sequence of nonnegative integrable functions $\{g_n\}$ that converges pointwisely a.e. to an integrable function $g$, and dominates $\{f_n\}$ on $\mathbb{R}$ in the sense that*

$$|f_n| \leq g_n \ \text{on } \mathbb{R}\ \forall n.$$

*If*

$$\lim_{n\to\infty}\int g_n = \int g < \infty,$$

*then*

$$\lim_{n\to\infty}\int f_n = \int f.$$

**Remark 3.2.9.** How to select the bandwidth $h$? Note that the leading term of MISE becomes

$$AMISE(\hat{f}, h) := \frac{h^4}{4}\mu_2(K)^2 \int f''^2 + \frac{1}{nh}\int K^2.$$

Minimizer $h_{opt}$ of AMISE is

$$h_{opt} = \left(\frac{\int K^2}{n\mu_2(K)^2 \int f''^2}\right)^{1/5},$$

and for such bandwidth, we get

$$MISE(\hat{f}, h_{opt} = \frac{5}{4}\left(\mu_2(K)^2\mu_0(K^2)^4\|f''\|_2^2\right)^{1/5} n^{-4/5} + o(n^{-4/5}).$$

However, in practice, it's impossible to obtain $h_{opt}$, because it contains the term of unknown $f''$. In this sense, we use "cross-validation" measure with one-leave-out techinque.

**Remark 3.2.10.** How to select an optimizing kernel $K$? To minimize

$$AMISE(\hat{f}, h_{opt}) = \frac{5}{4}\left(\mu_2(K)^2\mu_0(K^2)^4\|f''\|_2^2\right)^{1/5} n^{-4/5},$$

we should control both $\mu_2(K)$ and $\mu_0(K^2)$, but it is very hard. Instead, we only consider "standardized kernels," i.e., kernels that satisfy $\mu_2(K) = 1$. Note that

(i) If $K$ minimizes $\mu_0(G^2)$ among $G$'s under the constraint $\mu_2(G) = 1$, then $K$ is the minimizer of $\mu_2(G)\mu_0(G^2)^2$ among all $G$'s without the constraint;

(ii) If $K$ minimizes $\mu_0(G^2)$ among $G$'s under the constraint $\mu_2(G) = 1$, then for any $a > 0$, $K_a$ is the minimizer of $\mu_2(G)\mu_0(G^2)^2$ among all $G$'s without the constraint.

It is clear from following "scale invariance"

$$\mu_2(K)\mu_0(K^2)^2 \overset{a>0}{\equiv} \mu_2(K_a)\mu_0(K_a^2)^2.$$

Then our optimization problem is

$$\text{minimize } \mu_0(K^2) \text{ subject to } \mu_2(K) = 1.$$

With the method of undetermined multiplier (cf. *Variational Methods in Statistics*, Rustagi.),

we can find the solution of such optimization problem is given as following *Epanechnikov kernel*

$$K^*(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right) I_{(-\sqrt{5},\sqrt{5})}(u),$$

or its scaled version

$$K^*_{1/\sqrt{5}}(u) = \frac{3}{4}(1-u^2) I_{(-1,1)}(u).$$

**Remark 3.2.11** (Annotation by compiler)**.** Epanechnikov kernel gives optimal fit with least (asymptotic) MISE, but since it has a compact support, sometimes it gives wiggly fit.

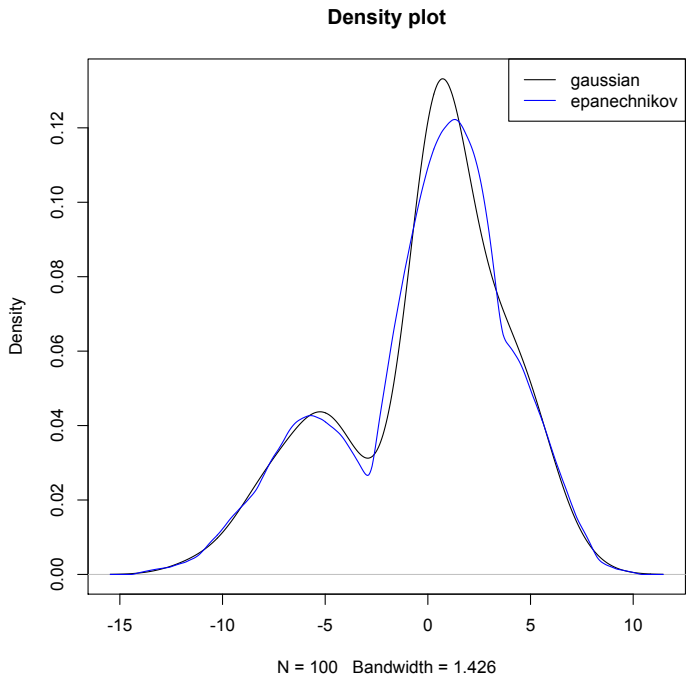


Figure 3.3: KDE with Gaussian kernel vs Epanechnikov kernel

Our last issue is asymptotic behavior of KDE: Under some mild conditions, it has an asymptotic normality.

**Theorem 3.2.12** (Asymptotic Normality of KDE, I)**.** *Suppose that*

$$\begin{cases} f \text{ is continuous at } x \\ |t|K^2(t) \to 0 \text{ as } |t| \to \infty, \ \int K^2 < \infty \end{cases} \tag{3.7}$$

*Then*

$$\frac{\hat{f}(x;h) - E\hat{f}(x;h)}{\sqrt{var(\hat{f}(x;h))}} \xrightarrow{d} N(0,1).$$

*Proof.* We will show Lindeberg condition (cf. Lecture Note of Probability Theory II). Define

$$Y_{ni} = K_h(x - X_i) - EK_h(x - X_i),\ i = 1, 2, \cdots, n.$$

Then $EY_{ni} = 0$ and $Var(Y_{ni}) = VarK_h(x - X_i)$. Thus we get

$$\begin{aligned} s_n^2 &= \sum_{i=1}^{n} Var(Y_{ni}) \\ &= nVar(K_h(x - X_1)) \end{aligned}$$

By the proof of proposition 3.2.5,

$$Var(K_h(x - X_1)) = \frac{1}{h}\mu_0(K^2)f(x) + o\left(\frac{1}{h}\right),$$

so we get

$$s_n^2 = \frac{n}{h}\mu_0(K^2)f(x) + o\left(\frac{n}{h}\right).$$

Now we should show that

$$\frac{1}{s_n^2}\sum_{i=1}^{n} EY_{ni}^2 I(|Y_{ni}| \geq \epsilon s_n) = \frac{n}{s_n^2} EY_{n1}^2 I(|Y_{n1}| \geq \epsilon s_n) \xrightarrow[n\to\infty]{} 0\ \forall \epsilon > 0.$$

Note that for any $x, y \geq 0$,

$$\begin{aligned} (x + y)I(|x + y| \geq h) &\leq (x + y)\left(I\left(|x| \geq \frac{h}{2}\right) + I\left(|y| \geq \frac{h}{2}\right)\right) \\ &\leq 2\max(x, y) \cdot 2I\left(\max(x, y) \geq \frac{h}{2}\right) \\ &= 4\max(x, y)I\left(\max(x, y) \geq \frac{h}{2}\right) \\ &\leq 4\left(xI\left(x \geq \frac{h}{2}\right) + yI\left(y \geq \frac{h}{2}\right)\right). \end{aligned}$$

From this and

$$Y_{n1}^2 = (K_h(x - X_1) - EK_h(x - X_1))^2 \leq 2|K_h(x - X_1)|^2 + 2|EK_h(x - X_1)|^2,$$

we get

$$Y_{n1}^2 I(|Y_{n1}| \geq \epsilon s_n) \leq 4\left(2K_h(x - X_1)^2 I\left(2K_h(x - X_1)^2 \geq \frac{\epsilon^2 s_n^2}{2}\right)\right.$$

$$+2(EK_h(x-X_1))^2 I\left(2(EK_h(x-X_1))^2 \geq \frac{\epsilon^2 s_n^2}{2}\right)\Big)$$
$$= 8K_h(x-X_1)^2 I\left(K_h(x-X_1) \geq \frac{\epsilon s_n}{2}\right) + 8\left(EK_h(x-X_1)\right)^2 I\left(EK_h(x-X_1) \geq \frac{\epsilon s_n}{2}\right)$$

and hence

$$\frac{n}{s_n^2} EY_{n1}^2 I(|Y_{n1}| \geq \epsilon s_n) \leq \underbrace{\frac{8n}{s_n^2} EK_h(x-X_1)^2 I\left(K_h(x-X_1) \geq \frac{\epsilon s_n}{2}\right)}_{=(A)}$$
$$+ \underbrace{\frac{8n}{s_n^2}\left(EK_h(x-X_1)\right)^2 I\left(EK_h(x-X_1) \geq \frac{\epsilon s_n}{2}\right)}_{=(B)}.$$

First note that $K_h * f(x) = f(x) + o(1)$, so $\exists C_1$ s.t. $K_h * f(x) \leq C_1$. Since $EK_h(x - X_1) = K_h * f(x)$, we get

$$(B) \leq \frac{8n}{s_n^2} C_1^2 = \frac{8nC_1^2}{\frac{n}{h}\mu_0(K^2) f(x) + o\left(\frac{n}{h}\right)} = \frac{8C_1^2 h}{\mu_0(K^2) f(x) + o(1)} = O(h) = o(1) \ (h \to 0).$$

Next, we get

$$(A) = \frac{8n}{s_n^2} E\left[K_h(x-X_1)^2 I\left(K_h(x-X_1) \geq \frac{\epsilon s_n}{2}\right)\right]$$
$$= \frac{8n}{s_n^2} \int_{-\infty}^{\infty} \left(K_h(x-\xi)\right)^2 I\left(K_h(x-\xi) \geq \frac{\epsilon}{2} h s_n\right) f(\xi) d\xi \ \left(t = \frac{x-\xi}{h}\right)$$
$$= \frac{8n}{s_n^2} \frac{1}{h} \int_{-\infty}^{\infty} K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2} h s_n\right) f(x-ht) dt.$$

Since

$$\frac{8n}{s_n^2} \frac{1}{h} = \frac{8}{\mu_0(K^2) f(x) + o(1)} = O(1),$$

we have to show

$$\int_{-\infty}^{\infty} K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2} h s_n\right) f(x-ht) dt = o(1) \ (h \to 0)$$

to get Lindeberg's condition. From

$$h^2 s_n^2 = h^2\left(\frac{n}{h}\mu_0(K^2) f(x) + o\left(\frac{n}{h}\right)\right) = nh(\mu_0(K^2) f(x) + o(1)),$$

we can say that there exists $C_2$ s.t.

$$\mu_0(K^2)f(x) + o(1) \geq C_2,$$

i.e.,

$$h^2 s_n^2 \geq C_2 nh,$$

provided that $f(x) \neq 0$. It implies that

$$K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2} h s_n\right) f(x-ht) \leq K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2}\sqrt{C_2 nh}\right) f(x-ht) \to 0$$

as $n \to \infty$, since we assume that $nh \to \infty$. Also, note that

$$K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2} h s_n\right) f(x-ht) \leq K(t)^2 f(x-ht),$$

whose integral is

$$\int K^2(t) f(x-ht)dt = (K^2)_h * f(x) = f(x)\mu_0(K^2) + o(1)$$

by proposition 3.2.2. Thus by generalized LDCT, we get

$$\int K(t)^2 I\left(K(t) \geq \frac{\epsilon}{2} h s_n\right) f(x-ht)dt \to 0.$$

□

**Corollary 3.2.13** (Asymptotic Normality of KDE, II)**.** *Suppose that* (3.7) *and additionally*

$$\begin{cases} f'' \text{ is continuous at } x \text{ and } f'' \in \mathcal{L}^1. \\ |t|^3 K(t) \to 0 \text{ as } |t| \to \infty \\ K \text{ is symmetric, and satisfies } \int |t|^2 K(t) < \infty. \end{cases} \tag{3.8}$$

*Then*

$$\hat{f}(x;h) - f(x) = \frac{1}{2}\mu_2(K) f''(x) h^2 + o(h^2) + (nh)^{-1/2}(\mu_0(K^2)f(x) + o(1))^{1/2} Z_n,$$

*where $Z_n$ obeys asymptotically the standard normal distribution.*

*Proof.* By proposition 3.2.5, we get

$$E\hat{f}(x;h) - f(x) = \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2)$$

and

$$var(\hat{f}(x;h)) = \frac{1}{nh}\mu_0(K^2)f(x) + o(n^{-1}h^{-1}).$$

Letting

$$Z_n := \frac{\hat{f}(x;h) - E\hat{f}(x;h)}{\sqrt{var(\hat{f}(x;h))}},$$

we get

$$\begin{aligned}\hat{f}(x;h) - f(x) &= \hat{f}(x;h) - E\hat{f}(x;h) + E\hat{f}(x;h) - f(x) \\ &= \sqrt{var(\hat{f}(x;h))}Z_n + \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2) \\ &= \sqrt{(nh)^{-1}}(\mu_0(K^2)f(x) + o(1))^{1/2}Z_n + \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2).\end{aligned}$$

□

**Corollary 3.2.14** (Asymptotic Normality of KDE, III)**.** *Suppose that* (3.7) *and* (3.8) *hold, and additionally,* $h = Cn^{-1/5}$ *(e.g..* $h = h_{opt}$*). Then*

$$n^{2/5}(\hat{f}(x;h) - f(x)) \xrightarrow[n\to\infty]{d} N(\beta(x), \sigma^2(x)),$$

*where*

$$\beta(x) = \frac{C^2}{2}\mu_2(K)f''(x) \text{ and } \sigma^2(x) = C^{-1}\mu_0(K^2)f(x).$$

*Proof.* Note that

$$\begin{aligned}\hat{f}(x;h) - f(x) &= \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2) + (nh)^{-1/2}(\mu_0(K^2)f(x) + o(1))^{1/2}Z_n \\ &= \frac{1}{2}\mu_2(K)f''(x)C^2n^{-2/5} + o(n^{-2/5}) + C^{-1/2}n^{-2/5}(\mu_0(K^2)f(x) + o(1))^{1/2}Z_n,\end{aligned}$$

so

$$\begin{aligned}n^{2/5}(\hat{f}(x;h) - f(x)) &= \frac{C^2}{2}\mu_2(K)f''(x) + C^{-1/2}(\mu_0(K^2)f(x) + o(1))^{1/2}Z_n + o(1) \\ &\xrightarrow[n\to\infty]{d} N\left(\frac{C^2}{2}\mu_2(K)f''(x), C^{-1}\mu_0(K^2)f(x)\right).\end{aligned}$$

□

**Remark 3.2.15** (Open Question)**.** In T.A.'s note, statement of previous corollary is:

*Suppose that* (3.7) *and* (3.8) *hold, and additionally,* $h \sim n^{-1/5}$*. Then*

$$n^{2/5}(\hat{f}(x;h) - f(x)) \xrightarrow[n\to\infty]{d} N(\beta(x), \sigma^2(x)),$$

*where*

$$\beta(x) = \frac{1}{2}\mu_2(K)f''(x) \text{ and } \sigma^2(x) = \mu_0(K^2)f(x).$$

However, I think that it holds only when $h = n^{-1/5}$, not $h \sim n^{-1/5}$. Further, I cannot understand the statement in the lecture note (which considers the situation that $h = o(n^{-1/5})$, i.e., $h \lesssim n^{-1/5}$). I welcome the discussion.

## 3.3 Nonparametric Regression

### 3.3.1 Introduction

Recall that regression setting is to estimate

$$m(x) = E(Y|X = x) \tag{3.9}$$

based on observations $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, which are i.i.d. copies of $(X, Y)$. In this section, we assume that $X, Y \in \mathbb{R}$.

**Example 3.3.1.** *Parametric regression* is the model that $m(x) = g(\theta, x)$ with a fully specified $g$. In this model, $\theta$ is unknown, so estimation of $m$ is reduced to the problem of estimating $\theta$. For example,

$$g(\theta, x) = \theta_0 + \theta_1 x + \cdots + \theta_p x^p$$

gives polynomial regression model.

In nonparametric model, there is no functional form on $m$, only smoothness conditions on $m$ exist. In this case, we assume that $X_i$ has compact support, WLOG supported on $[0, 1]$. Note that it is impossible to estimate $m(x)$ which is defined on whole $\mathbb{R}$.

**Example 3.3.2.** We often use least square method in the regression.

(a) Consider the case of parametric polynomial regression. Then we minimize

$$\sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 X_i - \cdots - \theta_p X_i^p)^2$$

with respect to $\theta_0, \theta_1, \cdots, \theta_p$. For minimizing solutions $\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_p$, we get

$$\hat{m}(x) = \hat{\theta}_0 + \hat{\theta}_1 x + \cdots + \hat{\theta}_p x^p.$$

(b) Now for fixed $x$, consider the minimizing problem

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1(X_i - x) - \cdots - \beta_p(X_i - x)^p)^2.$$

It means that we consider the model

$$m(u) = \beta_0 + \beta_1(u - x) + \cdots + \beta_p(u - x)^p,$$

and hence, for minimizing solution $\hat{\beta}_0(x), \hat{\beta}_1(x), \cdots, \hat{\beta}_p(x)$, we get

$$\hat{m}(x) = \beta_0(x).$$

It will be called *local polynomial regression*, and we will handle it in this section.

(c) In general, consider the problem that minimizing

$$\sum_{i=1}^{n}(Y_i - m(X_i))^2$$

over a function class

$$\mathcal{M} := \{m : m \text{ is "smooth"}\}.$$

Note that $\mathcal{M}$ is infinite dimensional, whereas that of our data is finite. Thus it is ill-posed problem; we always have an "interpolater" that makes the empirical loss exactly zero. As $n$ becomes large, it yields *overfitting* problem. Thus we use following "sieve approach"; For finite dimensional set $\mathcal{M}_n$ (with dimension less than $n$), we approximate $\mathcal{M}$ as $\mathcal{M}_n$, and assume that "$\mathcal{M}_n \to \mathcal{M}$" (There is no rigorous definition to such convergence), and find the solution on $\mathcal{M}_n$.

**Example 3.3.3.** For example, let $\mathcal{M} = \mathcal{L}^2$. Then for $m \in \mathcal{M}$, we can expand as

$$m(x) = \sum_{j=1}^{\infty} \theta_j b_j(x)$$

for some basis $b_j(x)$. Let

$$\mathcal{M}_n := \left\{ \sum_{j=1}^{d_n} \theta_j b_j(x) : \theta_j \in \mathbb{R} \right\}, \ \dim(\mathcal{M}_n) = d_n$$

s.t.

$$\frac{d_n}{n} \xrightarrow[n\to\infty]{} 0 \text{ and } d_n \to \infty.$$

(In this sense, we can argue that $\mathcal{M}_n \to \mathcal{M}$) If $b_j(x)$ is a spline basis, then it yields spline regression (Takezawa (2005); Ruppert *et al.* (2003)); if $b_j(x)$ is a wavelet basis, then it yields multiresolution analysis (Nason (2010)).

In this section, we only consider the local approximation problem, just at (b) in example 3.3.2.

### 3.3.2 Nadaraya-Watson Estimator

From now on, assume that kernel function $K$ is a symmetric density with support $[0, 1]$.

Note that $m(x)$ in (3.9) is the minimizer of $E(Y - f(X))^2$. Thus we find $\hat{m}(x)$ as minimizer of "empirical loss"

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

In here, assume that for each fixed $x$,

$$m(u) \approx m(x) \text{ for } u \approx x. \qquad \text{("local constant")}$$

Then minimizing criterion is

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x))^2 I_{[x-h,x+h]}(X_i);$$

assume that $m(X_i) \approx m(x)$ if $X_i \approx x$, in precise, $|X_i - x| \leq h$. Generalizing indicator window $I_{[x-h,x+h]}(\cdot)$ to a compact supported kernel $\frac{1}{h} K\left(\frac{\cdot - x}{h}\right)$ yields a criterion

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x))^2 K_h(X_i - x).$$

**Definition 3.3.4.** *For each fixed $x$,*

$$\begin{aligned}\hat{m}(x) &= \underset{f(x)\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x))^2 K_h(X_i - x) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)}\end{aligned}$$

*is called* ***Nadaraya-Watson estimator****, or* ***local constant estimator****.*

Note that denominator of $\hat{m}(x)$ is a KDE.

**Remark 3.3.5** (Annotation by compiler)**.** There exists another formulation of Nadaraya-Watson estimator. Recall that

$$m(x) = E(Y|X = x) = \int y\frac{\pi(x,y)}{\pi(x)}dy.$$

Plug-in KDE of each density,

$$\hat{\pi}(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)$$

$$\hat{\pi}(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)K_y\left(\frac{y - Y_i}{h_y}\right),$$

and we get

$$\frac{\hat{\pi}(x,y)}{\hat{\pi}(x)} = \frac{1}{h_y}\frac{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)}{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)K_y\left(\frac{y - Y_i}{h_y}\right)}.$$

Using symmetricity of $K$, finally we get

$$\begin{aligned}\hat{m}(x) &= \frac{\sum K_x\left(\frac{x - X_i}{h_x}\right)\int \frac{y}{h_y}K_y\left(\frac{y - Y_i}{h_y}\right)dy}{\sum K_x\left(\frac{x - X_i}{h_x}\right)} \\ &= \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)},\end{aligned}$$

letting $K_x = K$ and $h_x = h$. For details, see Takezawa (2005).

Now our interest is asymptotic behavior. Especially, we wonder if $\hat{m}(x)$ has following properties:

(1) **Pointwise consistency**, i.e., $\hat{m}(x) \xrightarrow[n\to\infty]{P} m(x)$ "for each" $x \in [0,1]$.

(2) **Uniform consistency**, i.e., $\sup_{x\in[0,1]} |\hat{m}(x) - m(x)| \xrightarrow[n\to\infty]{P} 0$.

(3) **$\mathcal{L}^2$-consistency**, i.e., $E\left[(\hat{m}(x) - m(x))^2\right] \xrightarrow[n\to\infty]{} 0$.

(4) **Pointwise asymptotic normality**.

(It's impossible to consider asymptotic normality as a process, just as we handled in chapter 2.)

**Theorem 3.3.6** (Pointwise consistency of N-W estimator)**.** *Assume that $X_i$'s are supported on $[0,1]$, and $K$ is a symmetric density supported on $[-1,1]$. Also assume that*

*(i) density of $X$, denote as $p$, is continuously differentiable at $x$, and $p(x) > 0$;*

*(ii) $v(x) := Var(Y|X = x)$ is continuous at $x$;*

*(iii) $m$ is twice continuously differentiable at $x$.*

*Then*

$$bias(\hat{m}(x;h)|X_1,\cdots,X_n) = \frac{h^2}{2}\left(m''(x) + 2m'(x)\frac{p'(x)}{p(x)}\right)\mu_2(K) + o_P(h^2) + O_P(n^{-1/2}h^{1/2})$$

*and*

$$var(\hat{m}(x;h)|X_1,\cdots,X_n) = \frac{1}{nh}\frac{v(x)}{p(x)}\mu_0(K^2) + o_P(n^{-1}h^{-1}).$$

**Remark 3.3.7.** Note that we cannot define $E(\hat{m}(x))$; denominator of $\hat{m}(x)$ is zero with positive probability. It is clear from

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) = 0\right) &= P\left(\text{There exists no } X_i \text{ in } [x-h, x+h]\right) \\ &= \left(1 - \int_{x-h}^{x+h} p(u)du\right)^n \\ &\geq \left(1 - 2h \max_{u\in[x-h,x+h]} p(u)\right)^n > 0. \end{aligned}$$

To define $\hat{m}(x)$ with probability tending to 1, we assume $nh \to \infty$, so that

$$\left(1 - \int_{x-h}^{x+h} p(u)du\right)^n \approx (1 - 2hp(x))^n \leq c_1 e^{-c_2 nh} \xrightarrow[n\to\infty]{} 0,$$

for some constants $c_1$ and $c_2$. Also, since we cannot define $E(\hat{m}(x))$, we see "conditional moment," $E(\hat{m}(x)|X_1, \cdots, X_n)$.

*Proof.* (Bias) Note that

$$bias(\hat{m}(x)|X_1, \cdots, X_n) = E(\hat{m}(x)|X_1, \cdots, X_n) - m(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)(m(X_i) - m(x))}{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)},$$

because our model is

$$Y_i = m(X_i) + \epsilon_i, \ E(\epsilon_i|X_i) = 0.$$

If

$$\frac{Z_n - EZ_n}{\sqrt{var(Z_n)}} \xrightarrow[n\to\infty]{d} N(0,1),$$

then we can rewrite such formula as

$$Z_n = EZ_n + O_P(\sqrt{var(Z_n)}).$$

Now, using this and CLT, we can write the denominator of bias term as

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) = E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\right) + O_P\left(\sqrt{var\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\right)}\right).$$

Note that

$$E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\right) = K_h * p(x)$$

and

$$var\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\right) \leq \frac{1}{n} E K_h^2(X_1 - x) = o_P(1).$$

Next, see decimator. Since

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)(m(X_i) - m(x))\right) &= \left(K_h * (m(\cdot) - m(x))p(\cdot)\right)(x) \\ &= \frac{h^2}{2}\mu_2(K)\left((m(\cdot) - m(x))p(\cdot)\right)''(x) + o(h^2), \end{aligned}$$

($\because (m(u)-m(x))p(u)|_{u=x}=0$, $\mu_1(K)=0$, proposition 3.2.2) we can find

$$E\left(\frac{1}{n}\sum_{i=1}^{n}K_h(X_i-x)(m(X_i)-m(x))\right)=\frac{h^2}{2}\mu_2(K)(m''(x)p(x)+2m'(x)p'(x))+o(h^2)$$

with easy calculation. Further, note that $K_h(u-x)=0$ only when $|u-x|\le h$, and so we can obtain

$$\begin{aligned}
var\left(\frac{1}{n}\sum_{i=1}^{n}K_h(X_i-x)(m(X_i)-m(x))\right)&=\frac{1}{n}var\left(K_h(X_1-x)(m(X_i)-m(x))\right)\\
&\le\frac{1}{n}E\left(K_h(X_1-x)(m(X_i)-m(x))\right)^2\\
&\le\frac{1}{n}E\left(K_h(X_1-x)^2(m'(x)h+o(h))^2\right)\\
&=\frac{1}{n}E\left(K_h(X_1-x)^2\right)\cdot O(h^2).
\end{aligned}$$

Now, note that

$$\begin{aligned}
EK_h^2(X_1-x)&=\frac{1}{h^2}\int K^2\left(\frac{u-x}{h}\right)f(u)du\ \ \left(\frac{u-x}{h}=\xi\right)\\
&=\frac{1}{h}\int_{-1}^{1}\underbrace{K^2(\xi)f(x+\xi h)}_{\text{conti on }[-1,1]\Rightarrow\text{ bdd}}\ d\xi\\
&=O(h^{-1}).
\end{aligned}$$

Thus we get

$$(\text{decimator})=\frac{h^2}{2}\mu_2(K)(m''(x)p(x)+2m'(x)p'(x))+o(h^2)+O_P(n^{-1/2}h^{1/2}),$$

and therefore,

$$\begin{aligned}
bias(\hat{m}(x)|X_1,\cdots,X_n)&=\frac{\frac{h^2}{2}\mu_2(K)(m''(x)p(x)+2m'(x)p'(x))+o(h^2)+O_P(n^{-1/2}h^{1/2})}{K_h*p(x)+o_P(1)}\\
&=\frac{\frac{h^2}{2}\mu_2(K)(m''(x)p(x)+2m'(x)p'(x))+o(h^2)+O_P(n^{-1/2}h^{1/2})}{p(x)+o_P(1)}\\
&\quad\text{(Proposition 3.2.2)}\\
&=\frac{h^2}{2}\mu_2(K)\left(m''(x)+2m'(x)\frac{p'(x)}{p(x)}\right)+o(h^2)+O_P(n^{-1/2}h^{1/2}).
\end{aligned}$$

(Check the condition for proposition 3.2.2! In here, $K$ has a compact support.)