Theory of Statistics II (Fall 2016)

J.P.Kim

Dept. of Statistics

Finally modified at September 22, 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.*
- Empirical Processes. Reference: *Empirical Processes in M-estimation, S.A. van de Geer, 2000.*

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*

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 Θ 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_θ 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* $P\left(\lim_{n\to\infty} X_n = X\right) = 1 \Leftrightarrow P(|X_n - X| > \epsilon \text{ i.o.}) = 0 \ \forall \epsilon > 0$
$$\Leftrightarrow \lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty} (|X_n - X| > \epsilon)\right) = 0 \ \forall \epsilon > 0$$

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

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

and hence

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

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 $(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)^T \xrightarrow[n\to\infty]{P} (m_1, m_2, \cdots, m_k)^T$ 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* $(m_1, \cdots, m_k)^T$*.*

We can do similar work in FSE $\nu(F)$. Note that in here, ν 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 ν 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

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

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

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

we obtain FSE for BLP,

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

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

Proposition 1.1.17.

(a) Solution of minimizing problem

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

is

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

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

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

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

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

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

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

It yields that

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

(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 $\mathcal{X}_1 = \boldsymbol{X}_1 - \Pi(\boldsymbol{X}_1|\mathcal{C}(\mathbf{1})) = \boldsymbol{X}_1 - \mathbf{1}\overline{X}^T$,

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

we get

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

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

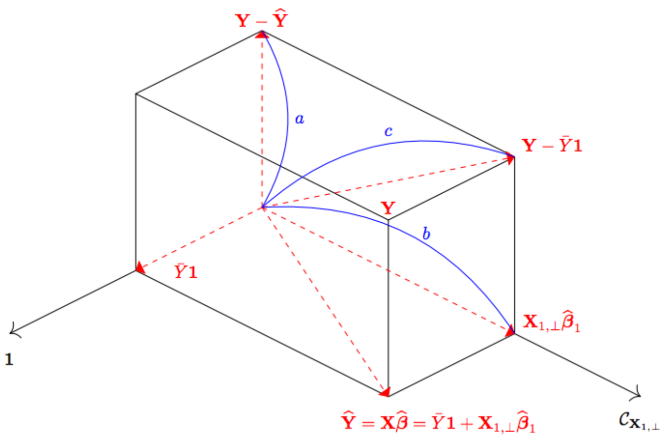

Figure 1.1: Regression with intercept. Image from Lecture Note.

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

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

and sample MCC is

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

Note that,

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

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

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

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 $\mathcal{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y} - \overline{Y}\mathbf{1}|\mathcal{C}(\mathcal{X}_1))$. Thus,

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

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

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

Then,

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

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

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

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

$$\begin{aligned}\sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \epsilon) &\leq \sup_{p\in\Theta} \frac{E|\hat{p}_n - p|^2}{\epsilon^2} \\ &= \sup_{p\in\Theta} \sum_{i=1}^k \frac{p_i(1-p_i)}{n\epsilon^2} \\ &\leq \frac{C}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0\end{aligned}$$

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

(b) Note that q is uniformly continuous on Θ, since Θ 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

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

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*

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

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

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

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* $t_0 \in \partial C$,

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

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

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

(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 $a^T t \geq a^T t_0$ $\forall t \in C$, then

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

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

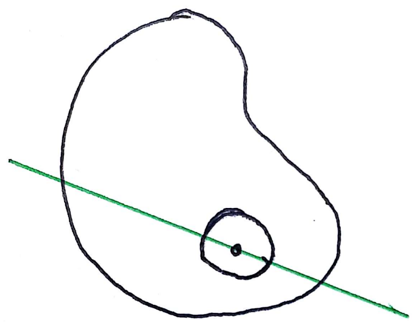

Figure 1.2: Proof of (a)

(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 $a^T t \geq a^T \mu$ for any $t \in C$. It implies that, $\exists a \neq 0$ such that $P(a^T(T-\mu) \geq 0) = 1$, since $\mathcal{T} \subseteq C$. It implies that

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

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

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

<u>Case 1.</u> $|\eta_n| \to \infty$.

Then since

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

if we can conclude

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

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

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

then

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

and from

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

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

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

and the assertion holds. So our claim is:

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

Assume not. If

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

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

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

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

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

which implies that

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

Thus, we get

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

which is equivalent to

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

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

$$u_0^T(t - t_0) \le 0 \ \forall t \in C,$$

however, this yields contradiction to

$$\nexists a \ne 0 \ s.t.\ a^T(t - t_0) \le 0 \ \forall t \in C,$$

from $t_0 \in C^0$.

<u>CASE 2.</u> $\sup|\eta_n| < \infty$

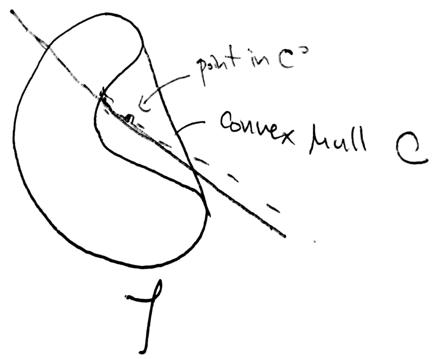

Figure 1.3: Convex hull of $\mathcal{T}$

In this case, we get

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

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*

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

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

(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 η. 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 Ψ_n, Continuity of Ψ_0, Uniqueness of θ_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 θ_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

$$\begin{aligned}
\Delta_{n'} + \inf_{|\theta-\theta_0|>\epsilon} D_{n'}(\theta) &\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} D_{n'}(\theta) \\
&\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} \{|D_{n'}(\theta) - D_0(\theta)| + D_0(\theta)\} \\
&\le 2\Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} D_0(\theta) \\
&= 2\Delta_{n'} + D_0(\theta_0).
\end{aligned}$$

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

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

holds, which implies

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

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 θ_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 θ_0^* ("Bolzano-Weierstrass"), so with the abuse of notation, let $\theta_k \to \theta_0^*$ as $k \to \infty$. Then note that θ_0^* should belong to $\{\theta : |\theta - \theta_0| \ge \epsilon\} \cap C$, 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.

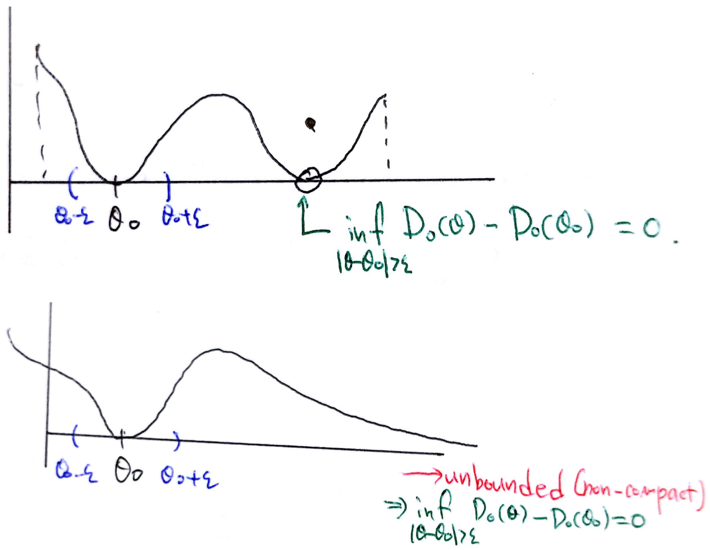

Figure 1.4: Continuity and Compactness are needed.

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 θ_0, then origin sequence should converge to θ_0.) Now for sufficiently large n_k,

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

holds, so

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

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

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

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

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

$\square$

Example 1.1.28 (Consistency of MLE when Θ 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) θ_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 ϵ s.t. $|\theta_n - \theta_0| > \epsilon$ i.o.. Uniqueness of θ_0 implies it clearly, because Θ is finite in here. Note that continuity of D_0 is not considered. $\square$

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

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

and hence $D_0(\theta) \ge 0$. In here $S(\theta_0) = \{x : f_{\theta_0}(x) > 0\}$ and $S(\theta) = \{x : f_\theta(x) > 0\}$. Note that $1 + \log z \le z \Leftrightarrow z = 1$. 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

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

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 θ_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_θ, $\theta \in \Theta \subseteq \mathbb{R}^k$, 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 θ_0 such that

(i) (Unique minimizer) $\theta_0 = \arg\min_{\theta\in K} E_{\theta_0}\rho(X_1,\theta)$, and θ_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 Θ) $P_{\theta_0}(\hat{\theta}_n^{MCE} \in K) \xrightarrow[n\to\infty]{} 1$.

(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 Θ need not be compact. Thus, we may use K instead of Θ. 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

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

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 θ_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 Θ containing θ_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 K.

Then *for any* $\theta_0 \in \Theta$ there exists a compact subset K of Θ containing θ_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* P_{θ_0}*-a.s., 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.