STAT 583: Advanced Theory of Statistical Inference **Spring 2018**

Lecture 1: Big Picture

Lecturer: Fang Han *March 23*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Lecturer.*

"There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy."

— Hamlet Act 1, scene 5

1.1 Glivenko-Cantalli and Donsker on $\{\mathbb{1}(\cdot \leq t), t \in \mathbb{R}\}$

Classical empirical processes theory deals with the empirical distribution function based on n i.i.d. observations, $X_1, \ldots, X_n$, of $X \in \mathbb{R}$ with distribution function F, corresponding to a measure P on a triplet $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$. Two elementary statistical functionals of everlasting interest are the empirical distribution function

$$\mathbb{P}_n f_t = \mathbb{F}_n(t) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i < t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,t]}(X_i),$$

and the corresponding empirical process, :

$$\mathbb{G}_n f_t := \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n f_t(X_i) - Ef_t\Big) = \sqrt{n}(\mathbb{F}_n(x) - F(x)),$$

where $f_t(x) := \mathbb{1}(x \leq t)$ for $t \in \mathbb{R}$.

We first provide the GC property for $\mathbb{F}_n$.

Theorem 1 (Glivenko-Cantalli Theorem)**.** *We have*

$$\|\mathbb{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \overset{a.s.}{\to} 0.$$

Remark 2. *Obviously, Glivenko-Cantalli is a UNIFORM version of the strong law of large numbers, applied to a particular function that is extremely easy to handle, the indicator function. Of note, in STAT535, we have also rigorously proven that*

$$\|\mathbb{F}_n - F\|_\infty = O_P(\sqrt{\log n/n})$$

using naive VC arguments. Later, we will show $\|\mathbb{F}_n - F\|_\infty = O_P(\sqrt{1/n})$ using the chaining argument introduced in STAT 582. Since this root-n rate is impossible to improve any further (by Donsker Theorem), the story is complete.

The next question is, do we have an analogous uniform central limit theorem? Note that GC Theorem was proved in 1933. This next step takes another 20 years to be resolved (though not quite correctly) by Donsker in 1952, and fully corrected in 1956 by Skorohod and Kolmogorov. This is now called the Donsker Theorem.

Before introducing the Donsker theorem, we have to first rigorously define what is a uniform central limit theorem. For this, it is worthwhile to remind the following equivalent definitions of weak convergence (a.k.a., "convergence in law"):

Lemma 3 (Portmanteau, Lemma 2.2 in V2000). *For any r.v. X_n and X, the following are equivalent:*

(1) $P(X_n \leq x) \to P(X \leq x)$ for all continuity points of $x \to P(X \leq x)$ (written as $X_n \Rightarrow X$);

(2) $Ef(X_n) \to Ef(X)$ for all bounded continuous functions f;

(3) $Ef(X_n) \to Ef(X)$ for all bounded Lipschitz functions f.

Conditions (2) and (3) bear the potential for generalizing the notion of weak convergence w.r.t a single element X_n to a stochastic process $\{\mathbb{G}_n f_t : t \in \mathbb{R}\}$, which takes values in the class of bounded functions from $\mathbb{R}$ to $\mathbb{R}$.

Theorem 4 (Donsker's Theorem, version 1). *Suppose $X \sim \mathrm{Unif}(0,1)$ (so that the law corresponds to the Lebesgue measure on $[0,1]$), then $\{\mathbb{G}_n f_t, t \in [0,1]\} \Rightarrow \mathbb{G}$ as a process in the space $D(0,1)$, where $D(0,1)$ is the space of cadlag functions on $[0,1]$ equipped with the Borel σ-algebra generated by the Skorohod topology, and $\mathbb{G}$ is a tight Brownian bridge process on $[0,1]$, i.e., $\mathbb{G}$ is mean zero Gaussian process indexed by $[0,1]$ with covariance structure $K(s,t) = \mathrm{Cov}(\mathbb{G}(s), \mathbb{G}(t)) = s \wedge t - st$.*

It is obvious that Theorem 4 is difficult to read, and the Skorohod topology is developed to answer some mathematically subtle questions you may never be interested in. Hence, for simplicity, we could instead think about each realization of $\{\mathbb{G}_n f_t, t \in \mathbb{R}\}$ as an element of the space $L^\infty(\mathbb{R})$, the space of bounded functions from $\mathbb{R}$ to $\mathbb{R}$, and impose a topology (for introducing continuity) by using the uniform entropy: for $f, g \in L^\infty(\mathbb{R})$, posing $d(f,g) = \|f-g\|_\infty$. Then, abandoning the measurability issue hereafter, we have the following "nicer" version of the Donsker's Theorem.

Theorem 5 (Donsker's Theorem, version 2). *Suppose X_i's have a continuous distribution F supported on $\mathbb{R}$. Consider the process $\mathbb{G} \circ F$. Then $\{\mathbb{G}_n f_t, t \in \mathbb{R}\} \Rightarrow \mathbb{G} \circ F$ as a process in $L^\infty(\mathbb{R})$, namely,*

$$EH(\{\mathbb{G}_n f_t, t \in \mathbb{R}\}) \to EH(\mathbb{G} \circ F)$$

for all bounded continuous functions $H : L^\infty(\mathbb{R}) \to \mathbb{R}$.

1.2 Glivenko-Cantalli and Donsker on general function classes

The modern empirical processes theory aims to generalize the classic GC and Donsker results to general function classes besides $\{\mathbb{1}(\cdot \leq t), t \in \mathbb{R}\}$. In other words, the goal is to investigate which function classes $\mathcal{F}$ satisfy the same GC and Donsker properties:

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{a.s.}{\to} 0 \text{ and } \mathbb{G}_n(\mathcal{F}) \Rightarrow \mathbb{G}_P(\mathcal{F}) \text{ in } L^\infty(\mathcal{F}).$$

We will soon realize that verifying the GC and Donsker for a certain function class $\mathcal{F}$ is equivalent to measuring the "complexity" of function classes, usually by the so-called "metric entropy" and VC dimensions.

But first, a concrete probability framework is in line.

Let $(\Omega, \mathcal{A}, P)$ be the studied probability space of the samples, with realizations $X_1, X_2, \ldots$ that are i.i.d. $\mathcal{X}$-valued and could be thought as co-ordinate projections on a product space $\Omega := (\Omega^\infty, \mathcal{A}^\infty, P^\infty)$: thinking about a generic sample point $\omega = (x_1, x_2, x_3, \ldots,)$ we have $X_i(\omega) = x_i$. Consider now a class of functions $\mathcal{F}$ with domain $\mathcal{X}$ and range $\mathbb{R}$ and envelop function F (meaning that $f(x) \leq F(x)$ for any $x \in \mathcal{X}$ and $f \in \mathcal{F}$). Let

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

denote the empirical measure. Hence, $\mathbb{P}_n f - Pf$ could be well understood as

$$\mathbb{P}_n f - Pf = \int f d\mathbb{P}_n - \int f dP = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef.$$

Definition 6. *$\mathcal{F}$ is called P-Glivenko-Cantalli if $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{a.s.}{\to} 0$.*

Similarly, the empirical process $\mathbb{G}_n(\cdot)$ is viewed as a map from Ω to $L^\infty(\mathcal{F})$ and is defined as

$$\mathbb{G}_n^\omega(f) := \sqrt{n}(\mathbb{P}_n^\omega - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i(\omega)) - Pf).$$

Thusly, for each fixed ω, $\mathbb{G}_n^\omega(\cdot)$ is a bounded function from $\mathcal{F}$ to $\mathbb{R}$. We often omit ω when no confusion exists.

We next establish weak convergence regarding $\mathbb{G}_n(\mathcal{F}) := \{\mathbb{G}_n f, f \in \mathcal{F}\}$. By the finite-dimensional central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\{f_1(X_i) - Pf_1(X_i)\}, \ldots, \{f_m(X_i) - Pf_m(X_i)\}\Big] \Rightarrow (\mathbb{G}_P f_1, \ldots, \mathbb{G}_P f_m), \quad f_i \in \mathcal{F}, m \in \mathbb{N},$$

where $\mathbb{G}_P(f), f \in \mathcal{F}$ is a centered Gaussian process with the same covariance as the process $\{f(X) : f \in \mathcal{F}\}$:

$$E\mathbb{G}_P(f)\mathbb{G}_P(g) = P(f - Pf)(g - Pf).$$

Recalling Theorem 4, we may refer to $\mathbb{G}_P(\mathcal{F})$ as the *P-bridge process indexed by $\mathcal{F}$.*

For weak convergence w.r.t. $\mathcal{F}$ to make any sense, we may have to first require the corresponding Gaussian process $\mathbb{G}_P(\mathcal{F})$ to be well-defined and nice in a certain sense.

Definition 7. *We say that $\mathcal{F}$ is P-pre-Gaussian if the P-bridge process $\mathbb{G}_P(\mathcal{F})$ admits a version (i.e., equivalent in distribution) whose sample paths are all bounded and uniformly continuous for its intrinsic L^2-distance*

$$d_P^2(f, g) := P(f - g)^2 - \{P(f - g)\}^2, \quad f, g \in \mathcal{F},$$

which further produces a pseudo-metric space $(\mathcal{F}, d_P)$.

Definition 8. *We say that the class $\mathcal{F} \subset L^2(\Omega, \mathcal{A}, P)$ satisfying*

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty, \quad \text{for all } x \in \mathcal{X} \tag{1.1}$$

is a P-Donsker class if $\mathcal{F}$ is P-pre-Gaussian and $\mathbb{G}_n(\mathcal{F})$ weakly converges in $L^\infty(\mathcal{F})$ to the Gaussian process $\mathbb{G}_P(\mathcal{F})$ as $n \to \infty$.

We are now finally ready to state the profound Donsker theorem for general function class.

By Theorem 3.7.23 in GN2015 (I also heard that Jon proved it in STAT 522 last year) given below, weak convergence in $L^\infty(T)$ is equivalent to *uniform asymptotic equicontinuity* (w.r.t. the pseudo metric space (T, d)) of $\mathbb{G}_n(\mathcal{F})$, a property naturally characterizing the smoothness of stochastic process.

Theorem 9 (Theorem 3.7.23 in GN2015)**.** *Let $\{X_n(t), t \in T\}_{n\in\mathbb{N}}$ be a sequence of bounded processes. Then the following statements are equivalent:*

(a) *The finite-dimensional distributions of the processes X_n converges in law, and there exists a pseudo-metric d on T such that (T, d) is totally bounded, and*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\Big\{ \sup_{d(s,t) \le \delta} |X_n(t) - X_n(s)| \ge \epsilon \Big\} = 0,$$

for all $\epsilon > 0$.

(b) *There exists a process X whose law is tight and $X_n \Rightarrow X$ in $L^\infty(T)$.*

Theorem 9 immediately gives rise to the following theorem, which states the Donsker theorem for general function class.

Theorem 10 (Theorem 3.7.31 in GN2015)**.** *Assume that $\mathcal{F} \in L^2(\Omega, \mathcal{A}, P)$ satisfying* (1.1). *Then the following three conditions are equivalent:*

(a) *$\mathcal{F}$ is a P-Donsker class.*

(b) *The pseudo-metric space $(\mathcal{F}, d_P)$ is totally bounded and*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\Big\{ \sup_{d_P(f,g) \le \delta} |\mathbb{G}_n f - \mathbb{G}_n g| \ge \epsilon \Big\} = 0,$$

for all $\epsilon > 0$.

(b) *There exists a totally bounded pseudo-metric space $(\mathcal{F}, e)$ such that*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\Big\{ \sup_{e(f,g) \le \delta} |\mathbb{G}_n f - \mathbb{G}_n g| \ge \epsilon \Big\} = 0,$$

for all $\epsilon > 0$. (A typical $e(f, g) = \|f - g\|_{L^2(P)}$.)

Theorem 10 effectively reduces proving Donsker property to proving a maximal inequality, which shall be comparably much easier to handle, and will be the subject of the next chapter using the uniform entropy.

1.3 Examples

You might wonder why we wish to establish such fancy results, especially if you are less exposed to statistics. It turns out that the GC and Donsker properties and the related techniques are the center of modern statistics, and many seemingly simple procedures cannot be well understood without them.

A commonly recurring theme in statistics is to prove (a) consistency; (b) asymptotic normality (ASN) of a given statistic, which is not necessarily the mean of independent random variables.

1.3.1 A toy example

In the future you will see many more statistics without a closed form. But even if the studied statistic enjoys a closed form and looks extremely nice and easy, you might still find yourself powerless in analyzing it unless you master the EP technique.

The following example comes from Pollard.

Let $X_1, \ldots, X_n$ be i.i.d. P on the real line. Set $\mu = EX_1$. Consider the mean absolute sample deviation,

$$M_n = \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n|.$$

A near-trivial argument shows that $M_n \overset{P}{\to} M := E|X - \mu|$. The next question is, how to prove $\sqrt{n}(M_n - M)$ is asymptotically normal (which is a natural conjecture)? Surprisingly, it is very difficult to prove without resorting to EP techniques. Let's outline a proof below.

Consider $\mathcal{F} := \{|x - t| =: f_t(x) : t \in [\mu - \delta_0, \mu + \delta_0]\}$ for some $\delta_0 > 0$. We then have

$$\begin{aligned}
\sqrt{n}(M_n - M) &= \sqrt{n}(\mathbb{P}_n f_{\overline{X}_n} - P f_\mu) \\
&= \sqrt{n}(\mathbb{P}_n - P) f_\mu + \sqrt{n}[\mathbb{P}_n f_{\overline{X}_n} - \mathbb{P}_n f_\mu] \\
&= \sqrt{n}(\mathbb{P}_n - P) f_\mu + \sqrt{n}(\mathbb{P}_n - P)(f_{\overline{X}_n} - f_\mu) + \sqrt{n}(\psi(\overline{X}_n) - \psi(\mu)) \\
&= A_n + B_n + C_n,
\end{aligned}$$

where $\psi(t) = P f_t = E_P|X - t|$. Assume that P has a density. Then

$$\psi(t) = \mu - 2\int_{-\infty}^{t} x f(x) dx - t + 2tF_P(t)$$

with derivative $2F_P(t) - 1$. Hence, the delta method implies

$$C_n = \sqrt{n}(\overline{X}_n - \mu)\psi'(\mu) + o_P(1),$$

yielding

$$A_n + C_n = \mathbb{G}_n(f_\mu(x) + \psi'(\mu)x).$$

Lastly, we claim $B_n = o_P(1)$ to finish the proof. This is where the empirical processes techniques kick in. To this end, we need the following nice proposition of the Donsker class.

Proposition 11. *Let $\mathcal{F}$ be a P-Donsker class. Let f_0 be a fixed function and $\widehat{f}_n$ be a random function depending on $X_1, \ldots, X_n$ such that $d_P(\widehat{f}_n, f_0) \overset{P}{\to} 0$. Then*

$$|\mathbb{G}_n \widehat{f}_n - \mathbb{G}_n f_0| \overset{P}{\to} 0.$$

Proof. Let $\eta, \epsilon > 0$ be given. Since $\mathcal{F}$ is Donsker, we have

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(\sup_{d_P(f,g) < \delta} |\mathbb{G}_n(f - g)| > \eta) = 0.$$

Thusly, we can find $\eta_0 > 0$ such that

$$\limsup_{n \to \infty} P(\sup_{d_P(f,g) < \delta} |\mathbb{G}_n(f - g)| > \eta) < \epsilon,$$

implying that, for all sufficiently large n,

$$P(\sup_{d_P(f,g) < \delta} |\mathbb{G}_n(f - g)| > \eta) < 2\epsilon.$$

Let $\Omega_n := \{d_P(\widehat{f}_n, f_0) < \delta_0\}$. By hypothesis, $P(\Omega_n) \geq 1 - \epsilon$ eventually. Let $\widetilde{\Omega}_n := \{\sup_{d_P(f,g) < \delta_0} |\mathbb{G}_n(f - g)| \leq \eta\}$. Then $P(\widetilde{\Omega}_n) \geq 1 - 2\epsilon$. Thus,

$$P(\Omega_n \cap \widetilde{\Omega}_n) \geq 1 - 3\epsilon$$

eventually. However, we know $\Omega_n \cap \widetilde{\Omega}_n \subset \{|\mathbb{G}_n(\widehat{f}_n - f_0)| \leq \eta\}$. This implies that $P(|\mathbb{G}_n(\widehat{f}_n - f_0)| > \eta)$ is eventually less than 3ϵ. □

We are now technically ready to prove $B_n = o_P(1)$. To this end, we take $\Theta = [\mu - \delta_0, \mu + \delta_0]$. Then $\{f_\theta(x) : \theta \in \Theta\}$ is a P-Donsker class of functions provided $E_P X^2 < \infty$ (as it is a VC-class of square-integrable envelop). Let $\widehat{\theta}_n = \overline{X}_n \mathbb{1}(\overline{X}_n \in [\mu - \delta_0, \mu + \delta_0]) + \mu \mathbb{1}(\overline{X}_n \notin [\mu - \delta_0, \mu + \delta_0])$ be the truncated version of $\overline{X}_n$ (you will repeatedly see this simple technique everywhere), and $\theta_0 = \mu$. We then have

$$B_n = \mathbb{G}_n(f_{\widehat{\theta}_n} - f_{\theta_0}) + \mathbb{G}_n(f_{\overline{X}_n} - f_{\widehat{\theta}_n}).$$

The second term is clearly $o_P(1)$ since $P(\widehat{\theta}_n \neq \overline{X}_n) = o(1)$. For the first term, Proposition 11 directly applies to solve the problem by noticing

$$d_P^2(f_{\theta_1}, f_{\theta_2}) = \mathrm{Var}_P(|X - \theta_1| - |X - \theta_2|) \leq P(|X - \theta_1| - |X - \theta_2|)^2 \leq (\theta_1 - \theta_2)^2.$$

This finishes the proof.

1.3.2 M-estimators

Later, following VW1996, we will give the analysis of M-estimators in its full power. Some popular regression procedures, such as the least absolute deviation regression,

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n |Y_i - X_i^T \beta|, \tag{1.2}$$

have to be analyzed using these powerful EP techniques.

However, a more elementary version does exist and is very revealing. The following framework is explicitly stated in Sherman (1993), credited to Pollard, and originally comes from Huber.

Suppose θ_0 maximizes a function $\Gamma(\theta)$ defined on Θ, and $\widehat{\theta}_n$ maximizes $\Gamma_n(\theta)$, which is a sample analogue of Γ. The following three theorems separately establish consistency, rate of convergence, and ASN.

Theorem 12 (Theorem 5.7 in V2000). *Suppose for every $\epsilon > 0$,*

$$\sup_{\theta \in \Theta} |\Gamma_n(\theta) - \Gamma(\theta)| \xrightarrow{P} 0 \quad \text{and} \quad \sup_{\theta : d(\theta, \theta_0) \geq \epsilon} \Gamma(\theta) < \Gamma(\theta_0). \tag{1.3}$$

Then $d(\widehat{\theta}_n, \theta_0) \xrightarrow{P} 0$.

Remark 13. *The first equation in (1.3) is the GC property, and calls for EP techniques to verify.*

Theorem 14. *Suppose that $\widehat{\theta}_n$ converges in probability to θ_0, and also*

(1) **(Strong convexity on $\Gamma(\cdot)$)** *There exists a neighborhood $\mathcal{N}$ of θ_0 and an absolute constant $\kappa > 0$ such that*

$$\Gamma(\theta) - \Gamma(\theta_0) \leq -\kappa \|\theta - \theta_0\|^2$$

for all θ in $\mathcal{N}$.

(2) **(Uniform smoothness)** *Uniformly over $o_P(1)$ neighborhood of θ_0,*

$$\Gamma_n(\theta) - \Gamma_n(\theta_0) = \Gamma(\theta) - \Gamma(\theta_0) + O_P(\|\theta - \theta_0\|/\sqrt{n}) + o_P(\|\theta - \theta_0\|^2) + O_P(1/n).$$

Then

$$\|\widehat{\theta}_n - \theta_0\| = O_P(1/\sqrt{n}).$$

Proof. In-class presentation. □

Theorem 15. *Suppose $\widehat{\theta}_n$ is $\sqrt{n}$-consistent for θ_0, an interior point of Θ. Suppose also that uniformly over $O_P(1/\sqrt{n})$ neighborhood of θ_0,*

$$\Gamma_n(\theta) - \Gamma_n(\theta_0) = \frac{1}{2}(\theta - \theta_0)^T V(\theta - \theta_0) + \frac{1}{\sqrt{n}}(\theta - \theta_0)^T W_n + o_P(1/n),$$

where V is a negative definite matrix, and W_n weakly converges to a $N(0, \Delta)$ random vector. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow N(0, V^{-1}\Delta V^{-1}).$$

Proof. Let's construct $\widetilde{\theta}_n := \theta_0 - V^{-1}W_n/\sqrt{n}$ so that

$$-V(\widetilde{\theta}_n - \theta_0) = W_n/\sqrt{n}.$$

Our aim is to prove

$$\|\widehat{\theta}_n - \widetilde{\theta}_n\| = o_P(1/\sqrt{n}).$$

By the given condition, with high probability we have

$$\begin{aligned}\Gamma_n(\widehat{\theta}_n) - \Gamma_n(\theta_0) &= \frac{1}{2}(\widehat{\theta}_n - \theta_0)^T V(\widehat{\theta}_n - \theta_0) + \frac{1}{\sqrt{n}}(\widehat{\theta}_n - \theta_0)^T W_n + o_P(1/n) \\ &= \frac{1}{2}(\widehat{\theta}_n - \theta_0)^T V(\widehat{\theta}_n - \theta_0) - (\widehat{\theta}_n - \theta_0)^T V(\widetilde{\theta}_n - \theta_0) + o_P(1/n)\end{aligned}$$

and also

$$\Gamma_n(\widetilde{\theta}_n) - \Gamma_n(\theta_0) = -\frac{1}{2}(\widetilde{\theta}_n - \theta_0)^T V(\widetilde{\theta}_n - \theta_0) + o_P(1/n).$$

By definition, $\Gamma_n(\widehat{\theta}_n) \geq \Gamma_n(\widetilde{\theta}_n)$ so that

$$\frac{1}{2}(\widehat{\theta}_n - \theta_0)^T V(\widehat{\theta}_n - \theta_0) - (\widehat{\theta}_n - \theta_0)^T V(\widetilde{\theta}_n - \theta_0) + \frac{1}{2}(\widetilde{\theta}_n - \theta_0)^T V(\widetilde{\theta}_n - \theta_0) + o_P(1/n) \geq 0,$$

or in other words,

$$o_P(1/n) \leq \frac{1}{2}(\widehat{\theta}_n - \widetilde{\theta}_n)^T V(\widehat{\theta}_n - \widetilde{\theta}_n) \leq 0,$$

where the last inequality is due to negative-definiteness of V. This implies the desired result. □

The above framework can work perfectly when the hessian of Γ_n enjoys a Lipschitz property, i.e., $\|\nabla_2\Gamma_n(\theta_1) - \nabla_2\Gamma_n(\theta_2)\|_{\mathsf{F}} \leq M(X_1, \ldots, X_n)\|\theta_1 - \theta_2\|$, which applies to simple linear regression. However, it cannot handle the quantile regression estimators given in (1.2). For this, we have to stick to the strongest version, which we will introduce in the third chapter. People of interest could read Chapter 3.2.4 in VW1996.

1.3.3 Z-estimators

Think about the generalized estimating equations (GEE). It belongs to a large family of estimators, called the Z-estimators. Z-estimators are naturally connected to M-estimators, though not every M-estimator could be written as a Z-estimator (e.g., Manski's rank estimator).

Let's nail down the framework. Consider $\Psi : \mathbb{R} \times \Theta \to \mathbb{R}$. $\widehat{\theta}_n$ solves $\mathbb{P}_n\Psi(\cdot, \theta) = 0$, while θ_0 satisfies $P\Psi(\cdot, \theta_0) = 0$. The purpose is to establish ASN of $\widehat{\theta}_n$ to θ_0 provided that we know $d(\widehat{\theta}_n, \theta_0) \xrightarrow{P} 0$.

To this end, expanding the identity $\sqrt{n}\mathbb{P}_n\Psi(\cdot,\widehat{\theta}_n)=0$ yields

$$\mathbb{G}_n\Psi(\cdot,\theta_0)+\mathbb{G}_n(\Psi(\cdot,\widehat{\theta}_n)-\Psi(\cdot,\theta_0))+\sqrt{n}(P_{\theta_0}\Psi(\cdot,\widehat{\theta}_n)-P_{\theta_0}\Psi(\cdot,\theta_0))=0.$$

Assuming $\{\Psi(,\theta)\}$ is P_{θ_0}-Donsker in a neighborhood of Θ, similar to the argument in Section 1.3.1, it is straightforward to establish $\mathbb{G}_n(\Psi(\cdot,\widehat{\theta}_n)-\Psi(\cdot,\theta_0))=o_P(1)$, which then yields

$$\mathbb{G}_n\Psi(\cdot,\theta_0)=-\sqrt{n}\{P_{\theta_0}\Psi(\cdot,\widehat{\theta}_n)-P_{\theta_0}\Psi(\cdot,\theta_0)\}+o_P(1).$$

Let's write $m(\theta)=P_{\theta_0}\Psi(\cdot,\theta)$ and $\sigma^2(\theta)=\mathrm{Var}_{\theta_0}(\Psi(\cdot,\theta))$. We then have

$$\mathbb{G}_n\Psi(\cdot,\theta_0)=-\sqrt{n}(\widehat{\theta}_n-\theta_0)\cdot\frac{m(\widehat{\theta}_n)-m(\theta_0)}{\widehat{\theta}_n-\theta_0}+o_P(1). \tag{1.4}$$

It is very easy (by Taylor expansion) to check that

$$\frac{m(\widehat{\theta}_n)-m(\theta_0)}{\widehat{\theta}_n-\theta_0}\xrightarrow{P}-I(\theta_0):=-m'(\theta_0),$$

where $I(\theta_0)$ is known as the Fisher information matrix when Ψ is set to be the score function (derivative of the likelihood). We then have the desired result that

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)\Rightarrow N(0,I(\theta_0)^{-1}),$$

which also proves the ASN of MLE.

Remark 16. *For simplicity, we only prove the case when θ is one-dimensional. This is crucial in establishing (1.4). For multidimensional θ, in order to establish a similar result, we have to "Taylor expand" $P_{\theta_0}\Psi(\cdot,\widehat{\theta}_n)-P_{\theta_0}\Psi(\cdot,\theta_0)$ in a vector space, which requires a certain version of differentiability. Details will be seen in the near future. Readers of interest could read Chapter 3.3 in VW1996.*

1.3.4 Von-mises calculus and functional delta method

We only give a very brief introduction to this topic. It is understood that $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P)$ and hence $\mathbb{P}_n=P+n^{-1/2}\mathbb{G}_n$. Then, for any function θ of measures as input, we have

$$\sqrt{n}(\theta(\mathbb{P}_n)-\theta(P))=\sqrt{n}(\theta(P+n^{-1/2}\mathbb{G}_n)-\theta(P))=\frac{\theta(P+n^{-1/2}\mathbb{G}_n)-\theta(P)}{n^{-1/2}}.$$

For any fixed measure Q, it is understood that

$$\lim_{n\to\infty}\frac{\theta(P+n^{-1/2}Q)-\theta(P)}{n^{-1/2}}=\dot{\theta}(Q)$$

when θ is smooth enough (a subtle concept since it involves arguments of infinite dimensions). By standard functional analysis argument (or just think about the Taylor expansion), $\dot{\theta}(\cdot)$ is a linear operator, which means $\dot{\theta}(Q)$ has to take the form $Q\Phi$ for some unknown function Φ that depends on P. Φ is then called the "influence function" of the operator θ (slightly different from the original definition though: Φ versus $\Phi-P\Phi$).

To continue, it has to be shown that Q could be taken to be $\mathbb{G}_n$. This is through the idea of P-Donsker and Hadamard differentiability. In detail, assuming the function $\theta:\mathcal{Q}\to\mathbb{R}$ to be Hadamard differentiable at P with respect to a certain metric $d(\cdot,\cdot)$, i.e., there exists $\dot{\theta}(Q)$ continuous and linear such that

$$\Big|\frac{\theta(P+tQ)-\theta(P)}{t}-\dot{\theta}(Q)\Big|=o(1)$$

for all $\{Q\}$ such that $d(Q, \Delta) \to 0$ for some measure Δ. Under some conditions, we then have, with probability tending to 1, that $d(\mathbb{G}_n, \mathbb{G} \circ F) \to 0$ by taking $t = n^{-1/2}$ and using the Donsker property of $\mathcal{F} = \{f_t, t \in \mathbb{R}\}$. We thus have

$$\sqrt{n}(\theta(\mathbb{P}_n) - \theta(P)) = \dot{\theta}(\mathbb{G}_n) + o_P(1) = \mathbb{G}_n\Phi + o_P(1).$$

Let's give an example. Consider the sample mean $\overline{X}_n := \int xd\mathbb{P}_n = \theta(\mathbb{P}_n)$ with $\theta(Q) := \int xdQ$. We then have

$$\lim_{\epsilon \to 0} \frac{\theta(P + \epsilon H) - \theta(P)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\int xd(\epsilon H)}{\epsilon} = \int xdH = Hx.$$

This implies that the influence function for the sample mean is just $\Phi(x) = x$, and

$$\sqrt{n}(\overline{X}_n - \mu) = \mathbb{G}_n x + o_P(1),$$

which is a trivial argument.