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

Lecture 3: Applications

Lecturer: Fang Han *May 09*

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.*

"Don't get involved in partial problems, but always take flight to where there is a free view over the whole single great problem, even if this view is still not a clear one."

— Ludwig Wittgenstein, on his notebook (November 01, 1914).

3.1 M-estimators

Recalling the theory on M-estimation in Chapter 1, we now finally have the language to state and verify it in its full generality.

Theorem 1 (Linearization of M-estimator, Theorem 3.2.16 in VW1996)**.** *Let Γ_n be a stochastic processes indexed by an open subset Θ of Euclidean space and $\Gamma : \Theta \to \mathbb{R}$ be a deterministic function. Assume $\theta \to \Gamma(\theta)$ is twice continuously differentiable at a point of maximum θ_0 with nonsingular second-derivative matrix V. Suppose that*

$$r_n(\Gamma_n - \Gamma)(\widetilde{\theta}_n) - r_n(\Gamma_n - \Gamma)(\theta_0) = (\widetilde{\theta}_n - \theta_0)^T Z_n + o_P(\|\widetilde{\theta}_n - \theta_0\| + r_n\|\widetilde{\theta}_n - \theta_0\|^2 + r_n^{-1}), \tag{3.1}$$

for every random sequence $\widetilde{\theta}_n = \theta_0 + o_P(1)$ and a uniformly tight sequence of random vectors Z_n. If the sequence $\widehat{\theta}_n \xrightarrow{P} \theta_0$ and satisfies $\Gamma_n(\widehat{\theta}_n) \geq \sup_\theta \Gamma_n(\theta) - o_P(r_n^{-2})$ for every n, then

$$r_n(\widehat{\theta}_n - \theta_0) = -V^{-1} Z_n + o_P(1).$$

Remark 2. *The M-estimation linearization theorem, in its full generality, does not require $\widehat{\theta}_n$ to be a regular root-n type estimator, does not require the data $X_1, \ldots, X_n$ to be either identically or independently distributed. There is, however, one constraint: Equation* (3.1) *does require a certain notion of "stochastic differentiability" for Γ_n, and hence rules out estimators like Manski's score estimator:*

$$\widehat{\beta}_n^{\text{Manski}} := \max_{\beta \in \mathbb{R}^p, \beta_1 = 1} \sum_{i=1}^n \mathbb{1}(Y_i > 0)\mathbb{1}(X_i^T \beta > 0).$$

For this, Pollard's cube root asymptotics kicks in ("Cube Root Asymptotics", Kim and Pollard, AoS 1990).

Proof of Theorem 1. **Step 1.** We first prove $\widehat{\theta}_n$ is r_n-consistent estimator of θ_0. Equation (3.1) and Taylor expanding Γ (in Peano form) yield, for every sequence $\widetilde{h}_n = o_P(1)$,

$$\begin{aligned}\Gamma_n(\theta_0 + \widetilde{h}_n) - \Gamma_n(\theta_0) &= \Gamma(\theta_0 + \widetilde{h}_n) - \Gamma(\theta_0) + r_n^{-1}\widetilde{h}_n^T Z_n + o_P(r_n^{-1}\|\widetilde{h}_n\| + \|\widetilde{h}_n\|^2 + r_n^{-2}) \\ &= \frac{1}{2}\widetilde{h}_n^T V \widetilde{h}_n + r_n^{-1}\widetilde{h}_n^T Z_n + o_P(\|\widetilde{h}_n\|^2 + r_n^{-1}\|\widetilde{h}_n\| + r_n^{-2}).\end{aligned}$$

Choosing $\widehat{h}_n = \widehat{\theta}_n - \theta_0$, using the condition $\Gamma_n(\widehat{\theta}_n) \geq \sup_\theta \Gamma_n(\theta) - o_P(r_n^{-2})$ and $\lambda_{\max}(V) \leq -c$ for some universal constant $c > 0$, we have

$$\begin{aligned} -O_P(r_n^{-2}) &\leq \frac{1}{2}\widehat{h}_n^T V \widehat{h}_n + r_n^{-1}\widehat{h}_n^T Z_n + o_P(\|\widehat{h}_n\|^2 + r_n^{-1}\|\widehat{h}_n\| + r_n^{-2}) \\ &\leq -c\|\widehat{h}_n\|^2 + r_n^{-1}\|\widehat{h}_n\| O_P(1) + o_P(\|\widehat{h}_n\|^2 + r_n^{-2}), \end{aligned}$$

implying

$$\{c + o_P(1)\}\{\|\widehat{h}_n\| - O_P(r_n^{-1})\}^2 \leq O_P(r_n^{-2}),$$

and completes the first part's proof.

Step 2. As soon as we proved $\|\widehat{h}_n\| = O_P(r_n^{-1})$, we have

$$o_P(r_n^{-1}\|\widetilde{h}_n\| + \|\widetilde{h}_n\|^2 + r_n^{-2}) = o_P(r_n^{-2}).$$

We then have

$$\begin{aligned} \Gamma_n(\theta_0 + \widehat{h}_n) - \Gamma_n(\theta_0) &= \frac{1}{2}\widehat{h}_n^T V \widehat{h}_n + r_n^{-1}\widehat{h}_n^T Z_n + o_P(r_n^{-2}), \\ \Gamma_n(\theta_0 - r_n^{-1}V^{-1}Z_n) - \Gamma_n(\theta_0) &= -\frac{1}{2}r_n^{-2} Z_n^T V^{-1} Z_n + o_P(r_n^{-2}). \end{aligned}$$

Subtracting the second from the first and noticing that

$$\Gamma_n(\theta_0 + \widehat{h}_n) - \Gamma_n(\theta_0 - r_n^{-1}V^{-1}Z_n) \geq -o_P(r_n^{-2}),$$

we have

$$0 \geq \frac{1}{2}(\widehat{h}_n + r_n^{-1}V^{-1}Z_n)^T V (\widehat{h}_n + r_n^{-1}V^{-1}Z_n) \geq -o_P(r_n^{-2}).$$

Since V is strictly negative, we conclude

$$r_n(\widehat{\theta}_n - \theta_0) = -V^{-1}Z_n + o_P(1).$$

This completes the proof. □

Under i.i.d. models, usual choices of Γ_n and Γ are

$$\Gamma(\theta) = P m_\theta \quad \text{and} \quad \Gamma_n(\theta) = \mathbb{P}_n m_\theta,$$

where m_θ is a "pseudo-likelihood" function indexed by the parameter $\theta \in \Theta$. For this, picking $r_n = \sqrt{n}$, Equation (3.1) translates to

$$\mathbb{G}_n(m_{\widetilde{\theta}_n} - m_{\theta_0}) = (\widetilde{\theta}_n - \theta_0)^T \mathbb{G}_n \dot{m}_{\theta_0} + o_P(\|\widetilde{\theta}_n - \theta_0\| + \sqrt{n}\|\widetilde{\theta}_n - \theta_0\|^2 + n^{-1/2}), \tag{3.2}$$

which is technically ready to be verified using empirical processes techniques.

Proposition 3 (Lemma 3.2.19 in VW1996)**.** *Suppose there exists a vector-valued function $\dot{m}_{\theta_0}$ such that, for some $\delta > 0$,*

$$\left\{ \frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}}{\|\theta - \theta_0\|} : \ \|\theta - \theta_0\| < \delta \right\} \ \text{is } P\text{-Donsker,}$$

$$\text{and} \quad P\Big[m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}\Big]^2 = o(\|\theta - \theta_0\|^2).$$

Then Equation (3.2) *is satisfied with the remainder as $o_P(\|\widetilde{\theta} - \theta_0\|)$, and hence we have $\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{\theta_0} + o_P(1)$.*

Proof. It is equivalent to proving

$$\lim_{\delta\to 0}\sup_{\theta:\|\theta-\theta_0\|<\delta}\Big|\underbrace{\mathbb{G}_n\Big\{\frac{m_\theta - m_{\theta_0} - (\theta-\theta_0)^T\dot{m}_{\theta_0}}{\|\theta-\theta_0\|}\Big\}}_{Z_n(\theta)}\Big| \xrightarrow{P} 0.$$

Let Z be the Gaussian sequence as the weak convergence limit of Z_n. By the Donsker's property, we remain to verify, defining $\rho^2(\theta_1,\theta_2) = P(Z(\theta_1)-Z(\theta_2))^2$, we have

$$\rho(\theta,\theta_0)\to 0 \text{ if } \theta\to\theta_0.$$

Noticing that

$$P(Z(\theta)-Z(\theta_0))^2 = \frac{P\big[m_\theta - m_{\theta_0} - (\theta-\theta_0)^T\dot{m}_{\theta_0}\big]^2}{\|\theta-\theta_0\|^2},$$

we immediately have the property holds. □

The Donsker property in Proposition 3 is arguably difficult to verify in certain cases. If we have established $\widehat{\theta}_n$ as a $\sqrt{n}$-consistent estimator of θ_0, then a more straightforward criterion could be built. Indeed, if so, verifying (3.2) is equivalent to verifying, for any sequence $\widetilde{\theta}_n = \theta_0 + O_P(1/\sqrt{n})$,

$$\mathbb{G}_n\sqrt{n}(m_{\widetilde{\theta}_n} - m_{\theta_0}) = \sqrt{n}(\widetilde{\theta}_n-\theta_0)^T\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1).$$

The result to verify it is Lemma 3.2.21 in VW1996, which we shall not cover due to the scope limit. However, it renders an important corollary.

Corollary 4 (Lipschitz class, Example 3.2.22 in VW1996)**.** *Let $X_1,\dots,X_n$ be i.i.d. random variables with common law P, and let m_θ be measurable functions indexed by $\theta\in\Theta$. Assume, for every θ_1,θ_2 in a neighborhood of θ_0 (the maximum of Pm_θ),*

$$\begin{aligned} |m_{\theta_1}(x) - m_{\theta_2}(x)| &\le \dot{m}(x)\|\theta_1-\theta_2\|,\\ P[m_\theta - m_{\theta_0} - (\theta-\theta_0)^T\dot{m}_{\theta_0}]^2 &= o(\|\theta-\theta_0\|^2), \end{aligned} \tag{3.3}$$

for functions $\dot{m}$ and m_{θ_0}, with $P\dot{m}^2(x)<\infty$. Further assume $\theta\to Pm_\theta$ is twice continuously differentiable at θ_0 with a nonsingular second-derivative matrix V. Then, if $\widehat{\theta}_n$ maximizes $\theta\to\mathbb{P}_n m_\theta$ (up to an $o_P(1/n)$-term) and is consistent for θ_0, then

$$\sqrt{n}(\widehat{\theta}_n-\theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1).$$

Remark 5. *If we have the Hessian of m_θ to be Lipschitz, then everything is nice and clear. However, it requires a marvelous amount of work to extend the result to the case that the first derivative of m_θ is Lipschitz.*

Example 6 (Example 3.2.23 in VW1996, LAD regression, a proof scratch)**.** *We are now finally ready to study the motivating example: least absolute deviation regression (LAD) estimator for the linear regression model $Y = X^T\beta+\epsilon$:*

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^n |Y_i - X_i^T\theta| = \mathbb{P}_n m_\theta,$$

where $\mathbb{P}_n$ is the empirical measure of the pairs (X_i,Y_i) and $m_\theta(x,y) = |y-x^T\theta|$. We remains to verify

(1) θ_0 is the minimum of $P|Y-\theta^TX| = P|\epsilon-(\theta-\theta_0)^TX|$, which holds if $P|\epsilon|<\infty$ and median$(\epsilon)=0$.

(2) The first condition in Equation (3.3) *is automatically satisfied.*

(3) The second condition in Equation (3.3) *holds by noticing*

$$P\Big[|Y - X^T\theta| - |Y - X^T\theta_0| - (\theta - \theta_0)^T X \operatorname{sign}(Y - X^T\theta_0)\Big]^2 = o(\|\theta - \theta_0\|^2).$$

We end up proving that $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ *is ASN with mean 0 and covariance* $V^{-1}P(XX^T)V^{-1}$*, with* V *the second derivative matrix of* $\theta \to P|Y - X^T\theta|$*.*

Remark 7. *We only slightly touch the M-estimation theory. As a matter of fact, M-estimation theory is one of the most exciting and fruitful fields in mathematical statistics. People of interest to learn more are highly recommended to attend ECON583, which introduces topics on estimation and testing in linear and nonlinear regression models, with asymptotic theory and bootstrapping.*

3.2 Z-estimation

As has been discussed in Chapter 1, M-estimation usually could be reduced to Z-estimation, which motivates Z-estimation theory. Another reason for considering Z-estimations is the popularity of estimating equation methods, which do not explicitly impose a loss function to solve. As will be seen soon, Z-estimation theory is interestingly much simpler than M-estimation theory provided a simple condition holds, which, however, is arguably strong.

Theorem 8 (Linearization of Z-estimator, Theorem 3.3.1 in VW1996)**.** *Let* Ψ_n *and* Ψ *be random maps and a fixed map, respectively, from* Θ *to a Banach space such that*

$$\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0) = o_P(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_0\|), \tag{3.4}$$

and such that the sequence $\sqrt{n}(\Psi_n - \Psi)(\theta_0)$ *weakly converges to a tight random element* Z*. Let* $\theta \to \Psi(\theta)$ *be Frechet-differentiable at* θ_0*, namely,*

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|),$$

with a continuously invertible derivative $\dot{\Psi}_{\theta_0}$*. If* $\Psi(\theta_0) = 0$ *and* $\widehat{\theta}_n$ *satisfies* $\Psi_n(\widehat{\theta}_n) = o_P(n^{-1/2})$ *and converges in probability to* θ_0*, then*

$$\sqrt{n}\dot{\Psi}_{\theta_0}(\widehat{\theta}_n - \theta_0) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1),$$

and

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\dot{\Psi}_{\theta_0}^{-1}\{\sqrt{n}(\Psi_n - \Psi)(\theta_0)\} + o_P(1).$$

Proof. **Step I.** We first prove that $\widehat{\theta}_n$ is a root-n consistent estimator of θ_0. By the definition of $\widehat{\theta}_n$ and Equation (3.4),

$$\begin{aligned}\sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi_n(\widehat{\theta}_n)) + o_P(1) \\ &= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_0\|). \end{aligned} \tag{3.5}$$

Since the derivative $\dot{\Psi}_{\theta_0}$ is continuously invertible (you could picture $\dot{\Psi}_{\theta_0}$ as the Hessian matrix of a certain loss function), there exists a positive constant c such that

$$\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\| \geq c\|\theta - \theta_0\|$$

for every θ and θ_0. Frechet differentiability condition then yields

$$\|\Psi(\theta) - \Psi(\theta_0)\| \geq c\|\theta - \theta_0\| + o(\|\theta - \theta_0\|).$$

Applying it to (3.5) yields

$$\sqrt{n}\|\widehat{\theta}_n - \theta_0\|(c + o_P(1)) \leq O_P(1) + o_P(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_0\|).$$

This proves that $\widehat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0 in norm.

Step II. We then prove the ASN of $\widehat{\theta}_n$. Applying Frechet differentiability condition again to LHS of (3.5), we obtain

$$\begin{aligned}\sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}\dot{\Psi}_{\theta_0}(\widehat{\theta}_n - \theta_0) + o_P(\sqrt{n}\|\widehat{\theta}_n - \theta_0\|) \\ &= \sqrt{n}\dot{\Psi}_{\theta_0}(\widehat{\theta}_n - \theta_0) + o_P(1),\end{aligned}$$

which is further equal to $-\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1)$. Finally, the continuous mapping theorem closes the proof. □

Under i.i.d. models, the usual choices of Ψ_n and Ψ are

$$\Psi_n(\theta) = \mathbb{P}_n\psi_\theta \quad \text{and} \quad \Psi(\theta) = P\psi_\theta,$$

for given measurable functions ψ_θ indexed by Θ. If so, for proving the stochastic differentiable condition (3.4), it is sufficient to prove

$$\|\mathbb{G}_n(\psi_\theta - \psi_{\theta_0})\|_{\Theta_\delta} \to 0 \quad \text{as } \delta \to 0, \tag{3.6}$$

where $\Theta_\delta := \{\theta : \|\theta - \theta_0\| < \delta\}$. The following lemma verifies (3.6).

Lemma 9 (Lemma 3.3.5 in VW1996)**.** *Suppose the class of function*

$$\Big\{\psi_\theta - \psi_{\theta_0} : \|\theta - \theta_0\| < \epsilon\Big\}$$

is P-Donsker for some $\epsilon > 0$ and that

$$P(\psi_\theta - \psi_{\theta_0})^2 \to 0, \quad \text{as } \theta \to \theta_0.$$

Then (3.6) *holds.*

The proof is just rephrasing the definitions.

Corollary 10 (HW problem)**.** *Consider the score function ψ_θ satisfies that, for every θ_1, θ_2 in a neighborhood of θ_0,*

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \leq \dot{\psi}(x)\|\theta_1 - \theta_2\|.$$

Then under some sufficient conditions, we have $\widehat{\theta}_n$ is an ASN estimator of θ_0.

3.3 Bootstrapping theory

We close this chapter with a slight touch on the bootstrap theory. There are essentially three perspectives to understand the advantage of bootstrap inference over the classic ones:

1. Bootstrap has the so-called second-order accuracy for studentized functionals (in Peter Hall's sense, referred to his seminal book "The Bootstrap and Edgeworth Expansion", and three wonderful papers in 1980's);

2. In most applications, bootstrap consistency (meaning that the bootstrapped statistic's distribution is consistent to the statistic's own distribution) holds as soon as ASN holds;

3. There exist cases when bootstrap is consistent while ASN does not hold (Bickel and Freedman, 1983, "Bootstrapping regression models with many parameters").

A good introductory book to bootstrap inference theory is "When does bootstrap work?", written by Professor Enno Mammen.

3.3.1 Peter Hall's view

CLT tells us the limiting behavior of $\overline{X}_n$ as $n \to \infty$. However, it never tells us how fast $\overline{X}_n - \mu$ converges to $N(0, \sigma^2)$. Actually, a result like

$$|P(\sqrt{n}(\overline{X}_n - \mu) \le x) - \Phi(x/\sigma)| = O(1/\log n)$$

would be useless. Gladly, Berry-Esseen Theorem tells us the convergence rate is usually not that disappointing.

Theorem 11 (Berry-Esseen Theorem (Esseen 1956))**.** *Suppose $E_F|X-\mu|^3 < \infty$. We then have*

$$\sup_x |P(\sqrt{n}(\overline{X}_n - \mu)/\sigma \le x) - \Phi(x)| \le \frac{0.4785 \cdot E|X-\mu|^3}{\sigma^3\sqrt{n}}.$$

Let's move on to characterizing the higher-order approximation for CLT. This is known as the Edgeworth expansion, and is celebrated for its application to proving the second-order accuracy of the bootstrap.

Theorem 12 (Edgeworth expansion)**.** *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. Write*

$$\gamma := E_F(X-\mu)^3/\sigma^3 \text{ (skewness)} \quad \text{and} \quad \kappa := E_F(X-\mu)^4/\sigma^4 \text{ (kurtosis)}.$$

We then have

$$\begin{aligned} G_n(x) &:= P_F(\sqrt{n}(\overline{X}_n - \mu)/\sigma \le x) \\ &= \Phi(x) - \phi(x)\left(\frac{\gamma(x^2-1)}{6\sqrt{n}} + \frac{(\kappa-3)(x^3-3x)}{24n} + \frac{\gamma^2(x^5-10x^3+15x)}{72n}\right) + o(1/n). \end{aligned}$$

Remark 13. *When F is symmetric, we have $\gamma = 0$, so that $\Phi(x)$ approximates $G_n(x)$ in the rate $O(1/n)$ (This justifies the intuition that "30 is good enough for CLT to work").*

Remark 14. *When F is asymmetric, generally $\gamma \neq 0$ and the CLT can only attain $O(1/\sqrt{n})$ rate of convergence. However, say, if we are interested in calculating the confidence interval of $\sqrt{n}(\overline{X}_n - \mu)/\sigma$, a balanced interval gives us*

$$G_n(x) - G_n(-x) = \Phi(x) - \Phi(-x) + O(1/n),$$

with the first term cancelled out. This is the intuition why balanced confidence interval is more preferred.

Remark 15. *The first two-order approximation is involved with κ only through $\kappa - 3$. This is the intuition why the excess kurtosis is defined as $\kappa - 3$.*

We then move on to prove bootstrap consistency and its second-order accuracy. Suppose $T(X_1, \ldots, X_n; F)$ is a functional (e.g., $T(X_1, \ldots, X_n; F) = \sqrt{n}(\overline{X}_n - \mu)$). Each time, the (nonparametric, multinomial, or Efron) bootstrapped sample $X_1^*, \ldots, X_n^*$ is sampled from $X_1, \ldots, X_n$ with replacement. In other words, the

bootstrap sample is drawn from the ECDF F_n with point mass on $X_1, \ldots, X_n$. The corresponding statistic is $T(X_1^*, \ldots, X_n^*; F_n)$. It is set up to approximate the true distribution of $T(X_1, \ldots, X_n; F)$.

Let's consider the simplest case, where $T(X_1, \ldots, X_n; F) = \sqrt{n}(\overline{X}_n - \mu)$. The bootstrap consistency is established as follows. Its proof is due to Professor Anirban DasGupta.

Theorem 16. *Provided $E_F X^2 < \infty$ and $T(X_1, \ldots, X_n; F) := \sqrt{n}(\overline{X}_n - \mu)$, we have*

$$\sup_x |P_F(T_n \le x) - P_*(T_n^* \le x)| \overset{a.s.}{\to} 0,$$

where P_* corresponds to the uniform distribution over all the n^n possible replacement resamples from $(X_1, \ldots, X_n)$, and $T_n^* := \sqrt{n}(\sum X_i^*/n - \overline{X}_n)$.

Proof. By triangle inequality, we have

$$\begin{aligned}\sup_x |P_F(T_n \le x) - P_*(T_n^* \le x)| &\le \sup_x |P_F(T_n/\sigma \le x/\sigma) - \Phi(x/\sigma)| + \sup_x |\Phi(x/\sigma) - \Phi(x/s)| \\ &\quad + \sup_x |\Phi(x/s) - P_*(T_n^*/s \le x/s)| \\ &= A_n + B_n + C_n,\end{aligned}$$

where s is the sample standard deviation, and is the standard deviation of $(X_1, \ldots, X_n)$ under P_*. Here $A_n \to 0$ by CLT. $B_n \to 0$ by the fact $s \overset{a.s.}{\to} \sigma$ and the continuous mapping theorem. Finally, applying the Berry-Esseen theorem to P_*, we have

$$C_n \le \frac{C}{\sqrt{n}} \cdot \frac{E_{F_n}(X_1^* - \overline{X}_n)^3}{[\mathrm{Var}_{F_n}(X_1^*)]^{3/2}} = \frac{C}{\sqrt{n}} \cdot \frac{\sum |X_i - \overline{X}_n|^3}{ns^3} \le \frac{8C}{n^{3/2}s^3} \cdot (\sum |X_i - \mu|^3 + n|\overline{X}_n - \mu|^3),$$

where in the last inequality we use the fact $(a+b)^3 \le 8(a^3 + b^3)$ for any $a, b > 0$. We then continue to have

$$\frac{8C}{n^{3/2}s^3} \cdot (\sum |X_i - \mu|^3 + n|\overline{X}_n - \mu|^3) \le \frac{C'}{s^3}\left(\frac{1}{n^{3/2}} \sum |X_i - \mu|^3 + \frac{|\overline{X}_n - \mu|^3}{\sqrt{n}}\right).$$

Viewing the SLLN:

Theorem 17 (SLLN)**.** *(i) If $E_F|X| < \infty$, then $\overline{X}_n \overset{a.s.}{\to} E_F X$. In other words, for arbitrary $\epsilon > 0$,*

$$P_F(\lim_{n\to\infty} X_n = E_F X) = 1.$$

(ii) (Zygmund-Marcinkiewicz SLLN.) If for some $0 < \delta < 1$, $E_F|X|^\delta < \infty$, then we have

$$n^{-1/\delta} \sum X_i \overset{a.s.}{\to} 0.$$

Clearly, these two terms will vanish by Zygmund-Marcinkiewicz SLLN. □

We then move to study the so-called second-order accuracy of the bootstrap. In short, under some assumptions, the bootstrap convergence rate is $O(1/n)$ compared to $O(1/\sqrt{n})$ for CLT. The following argument is due to Eric Lehmann.

Consider $T = \sqrt{n}(\overline{X}_n - \mu)/\sigma$. By Edgeworth expansion, we have

$$\begin{aligned}P_F(T \le x) &= \Phi(x) + \phi(x)(p_1(x|F)/\sqrt{n} + p_2(x|F)/n) + o(1/n) \\ P_{F^*}(T^* \le x) &= \Phi(x) + \phi(x)(p_1(x|F_n)/\sqrt{n} + p_2(x|F_n)/n) + o(1/n) \\ P_F(T \le x) - P_{F^*}(T^* \le x) &= \phi(x)\left(\frac{p_1(x|F) - p_1(x|F_n)}{\sqrt{n}} + \frac{p_2(x|F) - p_2(x|F_n)}{n}\right) + o(1/n),\end{aligned}$$

with

$$p_1(x|F) = \frac{\gamma}{6}(1-x^2), \quad p_2(x|F) = \frac{\kappa-3}{24}(3x-x^3) - \frac{\gamma^2}{72}(x^5 - 10x^3 + 15x).$$

Hence, since $\gamma_{F_n} - \gamma_F = O_P(1/\sqrt{n})$, we obtain $O(1/n)$ rate of convergence given the finiteness of the moments, which is called the second-order accuracy, in comparison to the first-order accuracy ($O(1/\sqrt{n})$) in CLT.

However, when we do not standardize the data, the second-order accuracy is lost, since additional effort is required to bound $\Phi(x/\sigma) - \Phi(x/s)$. Therefore, a rule of thumb is as follows:

Proposition 18 (DasGupta)**.** *If $T(X_1, \ldots, X_n; F) \xrightarrow{d} N(0, \tau^2)$ with τ independent of F and an Edgeworth expansion is available to T, then the second order accuracy is likely.*

3.3.2 Empirical processes' view

Bootstrap usually works as long as ASN holds. This could be rigorously stated in the following grand theorem. But before that, let's first introduce the multiplier (wild) bootstrap: thinking about the example in the last section, we could rewrite the nonparametric bootstrapped sample mean as:

$$\frac{1}{n}\sum_{i=1}^n X_i^* = \frac{1}{n}\sum_{i=1}^n W_{ni} X_i,$$

where $W_n = (W_{n1}, \ldots, W_{nn})^T$ is a multinomial vector with probability $(1/n, \ldots, 1/n)$ and number of trials n, and W_n is independent of $X_1, \ldots, X_n$. This will then give rise to the nonparametric bootstrap empirical measure:

$$\widehat{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n W_{ni} f(X_i).$$

An alternative to nonparametric bootstrap is the multiplier (wild) bootstrap: Let $\xi_1, \xi_2, \ldots,$ be an infinite sequence of nonnegative i.i.d. random variables, independent of $X_1, \ldots, X_n$, having mean μ and variance τ^2, and satisfying $\|\xi\|_{2,1} := \int_0^\infty \sqrt{P(|\xi| > x)}dx < \infty$. This will give rise to the multiplier bootstrap empirical measure:

$$\widetilde{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n (\xi_i/\overline{\xi}_n) f(X_i),$$

where $\overline{\xi}_n := n^{-1}\sum_{i=1}^n \xi_i$. Note that the weights add up to n for both bootstraps.

Let $\widehat{\mathbb{G}}_n = \sqrt{n}(\widehat{\mathbb{P}}_n - \mathbb{P}_n)$, $\widetilde{\mathbb{G}}_n = \sqrt{n}(\mu/\tau)(\widehat{\mathbb{P}}_n - \mathbb{P}_n)$, and $\mathbb{G}$ be the standard P-bridge in $L^\infty(\mathcal{F})$.

Theorem 19 (Theorem 2.6 in K2008)**.** *The following are equivalent:*

(i) $\mathcal{F}$ is P-Donsker;

(ii) $\widehat{\mathbb{G}}_n$ weakly converges to $\mathbb{G}$ conditionally on the data, and the sequence $\widehat{\mathbb{G}}_n$ is asymptotically measruable;

(iii) $\widetilde{\mathbb{G}}_n$ weakly converges to $\mathbb{G}$ conditionally on the data, and the sequence $\widetilde{\mathbb{G}}_n$ is asymptotically measruable.

Let's consider $\theta \in \Theta \subset \mathbb{R}^p$, $\psi_\theta : \mathcal{X} \to \mathbb{R}^p$, $\Psi(\theta) = P\psi_\theta$, $\Psi_n(\theta) = \mathbb{P}_n\psi_\theta$, and $\Psi_n^b(\theta) = \mathbb{P}_n^b\psi_\theta$, where $\mathbb{P}_n^b$ could be either $\widehat{\mathbb{P}}_n$ or $\widetilde{\mathbb{P}}_n$. A simple Z-estimator bootstrap consistency theorem to close the whole chapter is as follows.

Theorem 20 (Z-estimator master theorem, Theorem 10.16 in K2008)**.** *Let $\Theta \subset \mathbb{R}^p$ be open, and assume $\theta_0 \in \Theta$ satisfies $\Psi(\theta_0) = 0$. Also assume the following:*

(i) (Identifiability condition) For any sequence $\{\theta_n\} \in \Theta$, $\Psi(\theta_n) \to 0$ implies $\|\theta_n - \theta_0\| \to 0$;

(ii) The class $\{\psi_\theta : \theta \in \Theta\}$ is P-GC;

(iii) For some $\delta > 0$, the class $\mathcal{F}_\delta := \{\psi_\theta - \psi_{\theta_0} : \theta \in \Theta, \|\theta - \theta_0\| \leq \delta\}$ is P-Donsker and $P\|\psi_\theta - \psi_{\theta_0}\|^2 \to 0$ as $\|\theta - \theta_0\| \to 0$;

(iv) $P\|\psi_{\theta_0}\|^2 < \infty$, and $\Psi(\theta)$ is Frechet differentiable (in finite-dimensional real space, reduces to classic differentiability) at θ_0 with nonsingular derivative matrix V_{θ_0};

(v) $\Psi_n(\widehat{\theta}_n) = o_P(n^{-1/2})$ and $\Psi_n^b(\widehat{\theta}_n^b) = o_P(n^{-1/2})$.

Then, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N(0, V_{\theta_0}^{-1} P[\psi_{\theta_0}\psi_{\theta_0}^T][V_{\theta_0}^{-1}]^T)$$

and conditionally on the data,

$$\sqrt{n}(\widehat{\theta}_n^b - \widehat{\theta}_n) \xrightarrow{P} k_0 Z,$$

with $k_0 = 1$ for nonparametirc bootstrap, and $k_0 = \tau/\mu$ for the multiplier bootstrap.

The whole proof is just a combination of the proof of Theorem 8 and the result in Theorem 19.