1. Law of Large numbers for real-valued random variables.

Let $X_1, X_2, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P. Consider the case $\mathcal{X} = \mathbf{R}$. As $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu = EX, \quad a.s.$$

(Proof sketch: using Chebychev inequality.)

Let the empirical distribution of X be $F_n(t) = \frac{1}{n}\#\{X_i \le t, 1 \le i \le n\}, t \in \mathbf{R}$ and let $F(t) = P(X \le t)$. We also have:

$$F_n(t) \to F(t), \quad a.s. \text{ for all } t$$

Theorem 1.1: Glivenko-Cantelli Uniform Law of Large Numbers. It says the empirical distribution converges to the theoretical distribution uniformly:

$$\sup_t |F_n(t) - F(t)| \to 0, \quad a.s.$$

Applications: *Kolmogorov's goodness-of-fit test.* We want to test whether the theoretical distribution is the same as some specified distribution. $H_0 : F = F_0$. The test statistic is:

$$D_n = \sup_t |F_n(t) - F_0(t)|$$

Reject H_0 for large values of D_n.

Extensions: What if we want to extend the half-intervals $(-\infty, t]$ in $\mathbf{R}$ to sets of higher dimensions? Among what type of sets does Glivenko-Cantelli theorem hold?

The followings are all examples of extensions of half intervals to higher dimensions: quadr-angles in $\mathbf{R}^2$, half spaces in $\mathbf{R}^n$, and monotone sets.

Let us first extend the empirical distribution of one real-valued random variable to the empirical measure based on N random variables. For any measurable set $A \subset \mathcal{X}$, the empirical measure based on $X_1, \ldots, X_n$ is:

$$P_n(A) = \frac{1}{n}\#\{X_i \in A, 1 \le i \le n\}$$

The theoretical distribution of $\{X_1, \ldots, X_n, \ldots\}$ is denoted by

$$\mathbf{P} = P \times \ldots \times P \times \ldots$$

Definition: Glivenko-Cantelli (GC) class of sets. Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$. The collection $\mathcal{D}$ is called a GC class if

$$\sup_{D \in \mathcal{D}} |P_n(D) - P(D)| \to 0, \quad a.s.$$

Example. Let $\mathcal{X} = \mathbf{R}$. The class of half-intervals

$$\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}\}$$

is GC for all distributions. As a counter example, let P be a uniform distribution (or any continuous distribution) on $[0,1]$. The class

$$\mathcal{B} = \{\text{all Borel subsets of } [0,1]\}$$

is not GC, i.e., uniform law of large numbers does not hold on this collection of sets.
(Proof sketch: Take a set to be the collection of all sample points $B = \{X_1, \ldots, X_N\}$. Then

$$\sup_{B \in \mathcal{B}} |P_n(B) - P(B)| = 1)$$

2. What types of sets are of Glivenko-Cantelli class?

Definition: Cardinality of Sets. Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$, and let $\{\xi_1, \ldots, \xi_n\}$ be n points in $\mathcal{X}$ (they do not have to be the sample points). Two sets D_1 and D_2 are equal if the intersection of their symmetric difference with the n points $\{\xi_1, \ldots, \xi_n\}$ are an empty set:

$$D_1 \triangle D_2 \cap \{\xi_1, \ldots, \xi_n\} = \emptyset$$

where $D_1 \triangle D_2 = (D_1 \cap D_2^c) \cup (D_1^c \cap D_2)$ is the symmetric difference between D_1 and D_2. Write the cardinality of sets as

$$\begin{aligned} \triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) &= card(\{D \cap \{\xi_1, \ldots, \xi_n\} : D \in \mathcal{D}\}) \\ &= \text{the number of subsets of } \{\xi_1, \ldots, \xi_n\} \text{ that } \mathcal{D} \text{ can distinguish} \end{aligned}$$

i.e., we count the number of sets up to equivalent classes.

Example. a). Let us first look at half intervals in the real line $\mathbf{R}$. Let $\mathcal{X} = \mathbf{R}$ and $\mathcal{D} = \{1_{(\infty,t]} : \ t \in \mathbf{R}\}$. Then for all $\{\xi_1, \ldots, \xi_n\} \subset \mathbf{R}$,

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) \leq n + 1$$

with "=" when the points $\{\xi_1, \ldots, \xi_n\}$ are distinct.
b) As another example, let $\mathcal{D}$ be the collection of all finite subsets of $\mathcal{X}$. Then if the points $\{\xi_1, \ldots, \xi_n\}$ are distinct,

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = 2^n$$

(Proof sketch: the collection of sets $\mathcal{D}$ contains the 2^n possible combinations of the n points $\{\xi_1\}, \ldots, \{\xi_n\}, \{\xi_1, \xi_2\}, \ldots, \{\xi_1, \ldots, \xi_n\}, \{\emptyset\}$.)

Compared to the previous example, the cardinality of this set is much higher. In fact, it grows exponentially versus polynomially.

Theorem 2.1 (Vapnik and Chervonenkis (1971)). *We have*

$$\sup_{D \in \mathcal{D}} |\mathbf{P}_n(D) - \mathbf{P}(D)| \to 0 \ a.s.$$

if and only if

$$\frac{1}{n} log \triangle^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathrm{P}} 0.$$

Note that $\Delta^{\mathcal{D}}(X_1, \ldots, X_n)$ is random since it depends on the sample. It also depends on the specific probability measure $\mathbf{P}$. What we want is a quantity that is independent of any distribution.

Definition: Vapnik-Chervonenkis (VC) Class. Let

$$m^{\mathcal{D}}(n) = \sup\{\Delta^{\mathcal{D}}(\xi_1, \ldots, \xi_n) : \xi_1, \ldots, \xi_n \in \mathcal{X}\}$$

We say that $\mathcal{D}$ is a VC class if for certain constants c and r, and for all n,

$$m^{\mathcal{D}}(n) \leq cn^r$$

i.e., if $m^{\mathcal{D}}(n)$ does not grow faster than a polynomial in n.

Conclusion: For any collection of sets, $VC \Rightarrow GC$.

Examples. a) $\mathcal{X} = \mathbf{R}, \mathcal{D} = \{1_{(-\infty,t]} : t \in \mathbf{R}\}$. Since $m^{\mathcal{D}}(n) \leq n + 1$, $\mathcal{D}$ is VC.
b) $\mathcal{X} = \mathbf{R}^d, \mathcal{D} = \{1_{(-\infty,t]} : t \in \mathbf{R}^d\}$. Since $m^{\mathcal{D}}(n) \leq (n+1)^d$, $\mathcal{D}$ is VC.
c) $\mathcal{X} = \mathbf{R}^d, \mathcal{D} = \{\{x : \theta^T x > t\}, \begin{pmatrix}\theta \\ t\end{pmatrix} \in \mathbf{R}^{d+1}\}$. Since $m^{\mathcal{D}}(n) \leq 2^d \binom{n}{d}$, $\mathcal{D}$ is VC.

Lemma 2.2 (Closeness of VC Under Complement, Union and Intersections). *Let $\mathcal{D}$, $\mathcal{D}_1$ and $\mathcal{D}_2$ be VC. Then the following classes are also VC:*

(i) $\mathcal{D}^c = \{D^c : D \in \mathcal{D}\}$,

(ii) $\mathcal{D}_1 \cap \mathcal{D}_2 = \{D_1 \cap D_2 : D_1 \in \mathcal{D}_1, D_2 \in \mathcal{D}_2\}$,

(iii) $\mathcal{D}_1 \cup \mathcal{D}_2 = \{D_1 \cup D_2 : D_1 \in \mathcal{D}_1, D_2 \in \mathcal{D}_2\}$.

Examples of VC Sets: a) the class of intersections of two half-spaces,
b)all ellipsoids,
(Hint: ellipsoids can be expressed as:

$$c(a - x)^2 + d(b - y)^2 \leq r^2, \quad (a, b, c, d, r) \text{ are parameters.}$$

Replace x^2 with x_1, y^2 with y_1, the ellipsoids can also be written as:

$$c \cdot a^2 - 2c \cdot a \cdot x + c \cdot x_1 + d \cdot b^2 - 2 \cdot b \cdot d \cdot y + d \cdot y_1 \leq r^2$$

which is a special half space in $\mathbf{R}^4$.)
c)all half-ellipsoids (intersection of half spaces and ellipsoids),
d)in $\mathbf{R}$, the class $\{\{x : \theta_1 x + \cdots + \theta_r x^r \leq t\} : \begin{pmatrix}\theta \\ t\end{pmatrix} \in \mathbf{R}^{r+1}\}$.

Examples of GC sets which are not VC sets: Let $\mathcal{X} = [0, 1]^2$, and let $\mathcal{D}$ be the collection of all convex subsets of $\mathcal{X}$. Then $\mathcal{D}$ is not VC, but when $\mathbf{P}$ is uniform (or any continuous distribution function defined on $[0, 1]^2$), $\mathcal{D}$ is GC.
(Hint: $\mathcal{D}$ is GC is very hard to prove. See Pollard (1984).)

3. Convergence of Means to Expectations.

Notation. For a function $g : \mathcal{X} \to \mathbf{R}$, we write the expectation of $g(x)$ as

$$\int gdP = Eg(X)$$

and the empirical average as

$$\int gdP_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

3.1 Uniform Law of Large Numbers for Classes of Functions.

Definition: Glivenko-Cantelli (GC) Class for a class of functions. Let $\mathcal{G}$ be a collection of real-valued functions on $\mathcal{X}$. The class $\mathcal{G}$ is GC if

$$\sup_{g\in\mathcal{G}} |\int gdP_n - \int gdP| \to 0, \quad a.s.$$

In empirical estimations, the choice of g typically depends on the random observations, so the uniform convergence on the space of functions is convenient.

Recall that for sets, we uses the number of cardinality of sets to prove GC:

$$\frac{1}{n} log\Delta^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathrm{P}} 0 \Leftrightarrow \sup_{D\in\mathcal{D}} |P_n(D) - P(D)| \to 0 \text{ a.s.}$$

We now introduce a similar notion "entropy" on metric spaces which will help us verify the conditions under which a collection of functions is GC.

Definition: Entropy for a General Metric Space. Let (T, τ) be a metric space. Fix $\delta > 0$, let $N(\delta, T, \tau)$ = minimum number of balls with radius δ necessary to cover T. The δ-entropy of set T is defined as:

$$H(\delta, T, \tau) = logN(\delta, T, \tau)$$

Examples. Let us first consider a bounded space in $\mathbf{R}^2$. For example, a square box with length R. With a slight abuse of definitions, let us consider small square boxes with length δ. Then the number of small boxes necessary to cover the big box is:

$$N(\delta, T) = (\frac{R}{\delta})^2$$

The δ-entropy is:

$$H(\delta, T) = logN(\delta, T) = 2log(\frac{R}{\delta})$$

Similarly, in a d-dimensional space, the δ-entropy of d-dimensional cubes with length R is approximately $d\ log(\frac{R}{\delta})$. Since our focus is classes of functions, let us now define entropy for function classes.

Definition: Entropy for Classes of Functions. Suppose $|g| \leq 1 \forall g \in \mathcal{G}$, and g is defined on $\mathcal{X}$. For any $A \subset \mathcal{X}$, let $|g|_{\infty,A} = \sup_{x\in A} |g(x)|$ and $|g|_{\infty,P_n} = \sup_{1\leq i\leq n} |g(X_i)|$. In other words, $|g|_{\infty,A}$ is the sup-norm defined on sets A and

$|g|_{\infty,P_n}$ is the sup-norm for functions of random variables with the empirical distribution P_n. Let $N_\infty(\delta, \mathcal{G}, P_n)$ be the smallest value of N for which a covering of $\mathcal{G}$ by balls with radius δ (using the sup-norm metric) and centers at $g_1, \ldots, g_N$ exists. That is, for each $g \in \mathcal{G}$, there is a $j \in \{1, \ldots, N\}$, such that

$$|g - g_j|_{\infty,P_n} < \delta$$

Then the δ-entropy for $\mathcal{G}$ is denoted as $H_\infty(\delta, \mathcal{G}, P_n)$ and

$$H_\infty(\delta, \mathcal{G}, P_n) = logN_\infty(\delta, \mathcal{G}, P_n)$$

Theorem 3.1: *Suppose* $|g| \leq 1, \forall g \in \mathcal{G}$, *then:*

$$\frac{1}{n} H_\infty(\delta, \mathcal{G}, P_n) \to^P 0, \forall \delta > 0 \Leftrightarrow \sup_{g \in \mathcal{G}} | \int g d(P_n - P)| \to^{a.s.} 0.$$

Example. Let $\mathcal{G}$ be a class of indicator functions of sets: $\mathcal{G} = \{1_D : D \in \mathcal{D}\}$. Then the δ-entropy of this set is:

$$H_\infty(\delta, \{1_D : D \in \mathcal{D}\}, P_n) = log\Delta^{\mathcal{D}}(X_1, \ldots, X_n)$$

From the argument above, we know that if $\mathcal{D}$ is VC, then $\mathcal{G}$ is GC.

In the theorem above, we have assumed that the absolute value of any function g in the class $\mathcal{G}$ is no larger than one. What if we relax the assumption that functions are bounded? For this we need to introduce a new concept: envelope of a class of functions.

Definition: Envelope of a Class of Functions. The envelope of a class of functions $\mathcal{G}$ is:

$$G(x) = \sup_{g \in \mathcal{G}} |g(x)|, x \in \mathcal{X}$$

Theorem 3.2. Let $L_1(P)$ denote the set of functions whose absolute values is P-integrable:

$$L_1(P) = \{g : \int |g| dP \leq c\}$$

Suppose $\mathcal{G}$ *is a subset of* $L_1(P)$. *Then*

$$\left.\begin{array}{l} \int G dP < \infty \\ \frac{1}{n} H_1(\delta, \mathcal{G}, P_n) \to^P 0, \forall \delta > 0 \end{array}\right\} \Leftrightarrow \sup_{g \in \mathcal{G}} | \int g d(P_n - P)| \to^{a.s.} 0$$

where $H_1(\delta, \mathcal{G}, P_n)$ *is the* δ*-entropy with respect to* $|\cdot|_{1,P_n}$, *and*

$$|g|_{1,P_n} = \int |g| dP_n$$

Sometimes the entropy condition is hard to verify because it depends on the empirical distribution P_n and is random. As what we have done to the measurable sets, we will introduce VC classes of functions and explain under what conditions

the class is also GC.

Definition: Subgraph of a function. The subgraph of a function $g : \mathcal{X} \to \mathbf{R}$ is

$$\text{subgraph}(g) = \{(x,t) \in \mathcal{X} \times \mathbf{R} : g(x) \geq t\}.$$

Definition: VC Classes of Functions. A collection of functions $\mathcal{G}$ is called a VC class if the subgraphs $\{\text{subgraph}(g) : g \in \mathcal{G}\}$ form a VC class.

Theorem 3.3: *Suppose $\mathcal{G}$ is VC and that $\int G dP < \infty$. Then $\mathcal{G}$ is GC.*

Time for some exercise to cheer up spirits!

Exercise. Are the following classes of sets (functions) VC? Why or why not?
a) The class of all rectangles in $\mathbf{R}$.
(Proof sketch: Yes. Rectangles are intersections of two quadr-angles.)
b) The classes of all monotone functions on $\mathbf{R}$.
(Proof sketch: No. A VC class can not separate all combinations of the sample points, i.e., there exist(s) some combination(s) of the sample points that can not be distinguished by a VC class. Choose a configuration of the sample points such that they lie on one increasing line. Then it is easy to find increasing functions that separate ANY combination of the sample points.)

What if we restrict our attention to the class of increasing functions on $[0,1]$: $\mathcal{G} = \{g : \mathbf{R} \to [0,1], g \text{ increasing}\}$? Is $\mathcal{G}$ GC?
(Proof sketch: let us try to find the entropy of $\mathcal{G}$. First fix $\delta > 0$. Approximate the function g with a step function $\tilde{g}$ that is within δ distance of g: $|g(x) - \tilde{g}(x)| \leq \delta$. As g varies, the number of function $\tilde{g}$ needed to cover $\mathcal{G}$ is roughly $\binom{n+\frac{1}{\delta}}{\frac{1}{\delta}}$, which is less than $(\frac{1}{\delta} + n)^{1/\delta}$. Therefore, the entropy is less than a constant times log(n):

$$H_1(\delta, \mathcal{G}, P_n) \leq \frac{1}{\delta} log(n + \frac{1}{\delta})$$

c) The class of all sections in $\mathbf{R}^2$.
(Proof sketch: Yes. Sections are intersections of half spaces and circles.)
d) The class of all star-shaped sets in $\mathbf{R}^2$.
(Proof sketch: No. Similar to the argument of convex sets. You can pick up any collection of the sample points using a star-shaped set.)

Exercise 2. Let $\mathcal{G}$ be the class of all functions g on [0,1] with derivative $\dot{g}$ satisfying $|\dot{g}| \leq 1$. Check that $\mathcal{G}$ is not VC. Show that $\mathcal{G}$ is GC by using partial integration and the Glivenko-Cantelli Theorem.
(Proof sketch:

$$\begin{aligned} \int g d(P_n - P) &= \int_0^1 g(x) d(P_n(x) - P(x)) \\ &= g(1)(P_n(1) - P(1)) - g(0)(P_n(0) - P(0)) - \int_0^1 (P_n(x) - P(x)) \dot{g} dx \end{aligned}$$

Without loss of generality, assume g is bounded. Then all the terms on the right hand side converges to zero uniformly. Therefore the class is GC. To use the

Glivenko-Cantelli theorem, we want to show that the entropy of the class $\mathcal{G}$ divided by n converges to 0. Without loss of generality, let $\mathcal{G} = \{g : [0,1] \to [0,1], |\dot{g}| \leq 1\}$. Fix $\delta > 0$. At the first interval $[0, \frac{1}{\delta}]$, depending on the value of g, the possible number of $\tilde{g}$ is $\frac{1}{\delta}$. At the second interval, the possible number of $\tilde{g}$ reduces to 3, since $|\dot{g}| \leq 1$. Altogether, the number of $\tilde{g}$ to cover $\mathcal{G}$ is less than $(\frac{1}{\delta})3^{\frac{1}{\delta}-1}$.)

As we mentioned earlier, the entropy condition we have introduced is random and can be hard to verify. If every element of $\mathcal{G}$ is also encapsulated by an upper function and a lower function, then there exists a non-random entropy condition that directly implies GC. For this end,we introduce the notion of "entropy with bracketing".

Definition: Entropy with Bracketing. Let $\mathcal{G}$ denote a class of functions. The δ-covering number with bracketing of $\mathcal{G}$ (using $L_1(P)$ norm) is:

$$N_B(\delta, \mathcal{G}, P) = \min N \text{ s.t.}$$
$$\exists\ \{g_j^L, g_j^U\}_{j=1}^N \text{ satisfying } \begin{cases} \int |g_j^U - g_j^L| dP \leq \delta \\ \forall g \in \mathcal{G}, \exists\ j, \text{ s.t. } g_j^L \leq g \leq g_j^U \end{cases}$$

The δ-entropy with bracketing covering $\mathcal{G}$ is defined as:

$$H_B(\delta, \mathcal{G}, P) = log N_B(\delta, \mathcal{G}, P)$$

Theorem 3.4: $H_B(\delta, \mathcal{G}, P) < \infty\ \forall \delta > 0 \Rightarrow \mathcal{G}$ *is GC.*
(Proof sketch:

$$\begin{aligned} \int g d(P_n - P) &= \int g dP_n - \int g dP \\ &\leq \int g_j^U dP_n - \int g_j^L dP \\ &= \int g_j^U d(P_n - P) + \int (g_j^U - g_j^L) dP \\ &\leq \int g_j^U d(P_n - P) + \delta \end{aligned}$$

Similarly, we can show that

$$\int g d(P_n - P) \geq \int g_j^U d(P_n - P) - \delta$$

Since there are only finite pairs of the upper and lower functions $\{g_j^L, g_j^U\}(j = 1, \cdots, N < \infty)$, as $P_n \to P$, we will eventually have:

$$\begin{aligned} \max_{j=1,\ldots,N} |\int g_j^U d(P_n - P)| &\leq \delta, \quad a.s. \\ \max_{j=1,\ldots,N} |\int g_j^L d(P_n - P)| &\leq \delta, \quad a.s. \end{aligned}$$

So eventually,

$$\sup_{g \in \mathcal{G}} |\int g d(P_n - P)| \leq 2\delta, \quad a.s.$$

Next we present a lemma that shows the conditions under which the entropy with bracketing is finite.

Lemma 3.5: Suppose (Θ, τ) is a compact metric space. Let the class of functions $\mathcal{G}$ be indexed by elements of Θ, i.e. $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$. Suppose the map $\theta \mapsto g_\theta(x)$ is continuous in θ for P-almost all x. Moreover, let us assume that the envelope of $\mathcal{G}$ is P-integrable: $G \in L_1(P)$ (i.e. $\int G dP < \infty$). Then

$$H_{1,B}(\delta, \mathcal{G}, P) < \infty, \ \forall \delta > 0$$

(Proof sketch: Write

$$w(\theta, \rho)(x) = \sup_{\tau(\theta,\tilde{\theta})<\rho} |g_\theta(x) - g_{\tilde{\theta}}(x)|, \theta \in \Theta, \rho > 0$$

Then $w(\theta, \rho)(x) \to 0$ as $\rho \to 0$, for P-almost all x. By dominated convergence,

$$\int w(\theta, \rho) dP \to 0 \quad \text{as } \rho \to 0$$

Fix an arbitrary $\delta > 0$. $\forall \theta, \exists\ \rho_\theta$, such that $\int w(\theta, \rho_\theta) dP < \delta$,. To show that the entropy with bracketing is finite, we want to construct a set of bracketing functions. Let $B_\theta = \{\tilde{\theta} : \tau(\tilde{\theta}, \theta) < \rho_\theta\}$ and let $B_{\theta_1}, \ldots, B_{\theta_N}$ be a finite cover of Θ. Define

$$\begin{aligned} g_j^L &= g_{\theta_j} - w(\theta_j, \rho_{\theta_j}), j = 1, \ldots, N \\ g_j^U &= g_{\theta_j} + w(\theta_j, \rho_{\theta_j}), j = 1, \ldots, N \end{aligned}$$

where N is the number of coverings for set Θ. Then

$$0 \leq \int (g_j^U - g_j^L) dP \leq 2\delta$$

and for $\theta \in B_{\theta_j}$,

$$g_j^L \leq g_\theta \leq g_j^U$$

It follows that

$$H_{1,B}(2\delta, \mathcal{G}, P) \leq log N$$

Theorem 3.6: Suppose $\mathcal{G}$ is VC. Let $\mathcal{D} = \{\text{subgraph of } g : \ g \in \mathcal{G}\}$. By properties of a VC class, $m^{\mathcal{D}}(n) \leq cn^r, \forall\ n$. Let us restrict our attention to bounded functions to simplify proof. The results are essentially the same for unbounded functions. *Assume* $0 \leq g \leq 1, \forall g \in \mathcal{G}$. *Then*

$$N_1(\delta, \mathcal{G}, P) \leq A(\frac{1}{\delta})^{2r}$$

(Proof: exercise.)

4. Uniform Central Limit Theorem

Central Limit Theorem in R. *Suppose* $E(X) = \mu$, *and* $var(X) = \sigma^2$ *exists. Then*

$$Pr(\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \leq z) \to \Phi(z), \ \textit{for all } z,$$

where Φ *is the standard normal distribution function.*

The first step is to extend central limit theorem to higher dimensions. Later on we will extend it to infinite dimensions.

Central Limit Theorem in $\mathbf{R}^d$. Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{R}^d$-valued random variables copies of $X, (X \in \mathcal{X} = \mathbf{R}^d$, with expectation $\mu = EX$, and covariance matrix $\Sigma = EXX^T - \mu\mu^T$.) *We have*

$$\sqrt{n}(\bar{X}_n - \mu) \to^{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

i.e.

$$\sqrt{n}[a^T(\bar{X}_n - \mu)] \to^{\mathcal{L}} \mathcal{N}(0, a^T\Sigma a), \text{ for all } a \in \mathbf{R}^d$$

Central Limit Theorem in Infinite Dimensions. A central limit theorem that holds uniformly in $g \in \mathcal{G}$ is one of the main topics in empirical process theory. Here we briefly discuss the weak convergence of the empirical process. The main concepts are Donsker classes and asymptotic continuity. Let us first give a formal definition of the empirical process.

Definition: Empirical Process. The empirical process indexed by $\mathcal{G}$ is

$$\nu_n(g) = \sqrt{n} \int g d(P_n - P), \ g \in \mathcal{G}$$

For a given function g, central limit theorem for one random variable (here it is $g(X)$) implies that the empirical process converges to a normal distribution:

$$\nu_n(g) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2(g))$$

where $\sigma^2(g)$ is the variance of $g(X)$. The central limit theorem also holds for finitely many g simultaneously. Let $g_k$and g_l be two functions and denote the covariance between $g_k(X)$ and $g_l(X)$ by

$$\sigma(g_k, g_l) = cov(g_k(X), g_l(X)) = Eg_k(X)g_l(X) - Eg_k(X)Eg_l(X)$$

Whenever $\sigma^2(g_k) < \infty$ for $k = 1, \ldots, r$, we will have

$$\begin{pmatrix} \nu_n(g_1) \\ \vdots \\ \nu_n(g_r) \end{pmatrix} \to^{\mathcal{L}} \mathcal{N}(0, \Sigma_{g_1,\ldots,g_r}),$$

where $\Sigma_{g_1,\ldots,g_r}$ is the variance-covariance matrix

$$\Sigma_{g_1,\ldots,g_r} = \begin{pmatrix} \sigma^2(g_1) \ldots \sigma(g_1, g_r) \\ \vdots \ddots \vdots \\ \sigma(g_1, g_r) \ldots \sigma^2(g_r) \end{pmatrix}$$

Before we discuss the uniform central limit theorem, we need to find the limiting process of ν_n for ANY finite collection of g.

Definition: P-Brownian bridge. Let ν be a Gaussian process indexed by $\mathcal{G}$. Assume that for each $r \in \mathbf{N}$ and for each finite collection $\{g_1, \ldots, g_r\} \subset \mathcal{G}$, the

r-dimensional vector

$$\begin{pmatrix} \nu(g_1) \\ \vdots \\ \nu(g_r) \end{pmatrix}$$

has a $\mathcal{N}(0, \Sigma_{g_1,\ldots,g_r})$-distribution, with $\Sigma_{g_1,\ldots,g_r}$ defined as above. We then call ν the P-Brownian bridge indexed by $\mathcal{G}$.

Not all classes of functions can have uniform central limit theorem. The "P-Donsker Class" is named for such classes upon which uniform central limit theorem holds.

Definition: P-Donsker Class. Consider ν_n and ν as bounded functions on $\mathcal{G}$. We call $\mathcal{G}$ a P-Donsker class if

$$\nu_n \rightarrow^{\mathcal{L}} \nu,$$

that is, if for all continuous and bounded functions f, we have

$$\mathbf{E}f(\nu_n) \rightarrow \mathbf{E}f(\nu).$$

To check that a class of functions is a P-Donsker class, we can use the "asymptotic continuity" condition. In fact, "asymptotic continuity" is equivalent to "P-Donsker class" for "totally bounded" class of functions.

Definition: Asymptotically Continuous. The process ν_n on $\mathcal{G}$ is called asymptotically continuous if for all $g_0 \in \mathcal{G}$, and all (possible random) sequences $\{g_n\} \subset \mathcal{G}$ with $\sigma(g_n - g_0) \rightarrow^{\mathbf{P}} 0$, we have

$$|\nu_n(g_n) - \nu_n(g_0)| \rightarrow^{\mathbf{P}} 0.$$

Definition: Totally Bounded. The class $\mathcal{G} \subset L_2(P)$ is called totally bounded if for all $\delta > 0$ the number of balls with radius δ necessary to cover $\mathcal{G}$ is finite (using the $L_2(P)$ norm.)

Theorem 4.1. *Suppose that $\mathcal{G}$ is totally bounded. Then $\mathcal{G}$ is a P-Donsker class if and only if ν_n (as a process on $\mathcal{G}$) is asymptotically continuous.*

Theorem 4.1 says that asymptotical continuity implies uniform central limit theorem for totally bounded class. The next theorem shows that for VC-graph class functions with P-square integrable envelopes, uniform central limit theorem also applies.

Theorem 4.2. *Suppose that $\mathcal{G}$ is a VC-graph class with envelope*

$$G = \sup_{g \in \mathcal{G}} |g|$$

satisfying $\int G^2 dP < \infty$. Then $\mathcal{G}$ is P-Donsker.
(Proof: see Van de Geer (2000).)

5. M-estimators.
In this section, we focus on finite-dimension models. We show that consistency and asymptotic normality hold for M-estimators. Later we will extend to models with infinite dimensions.

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P. Let Θ be a parameter space (a subset of some metric space) and let $\gamma_\theta : \mathcal{X} \to \mathbf{R}$ be some loss function. We estimate the unknown parameter

$$\theta_0 = \arg\min_{\theta \in \Theta} \int \gamma_\theta dP$$

by the M-estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \int \gamma_\theta dP_n$$

Here we assume that θ_0 exists and is unique and that $\hat{\theta}_n$ exists.

Examples.
(i) Location Estimators. Let $\mathcal{X} = \mathbf{R}$, $\Theta = \mathbf{R}$.
(i.a) $\gamma_\theta(x) = (x - \theta)^2$ (estimating the mean).
(i.b) $\gamma_\theta(x) = |x - \theta|$ (estimating the median).

(ii) Maximum Likelihood. Let $\{p_\theta : \theta \in \Theta\}$ be a family of densities w.r.t. σ-finite dominating measure μ. The loss function is:

$$\rho_\theta = -log p_\theta$$

If $dP/d\mu = p_{\theta_0}$, $\theta_0 \in \Theta$, then θ_0 is indeed the minimizer of $\int \rho_\theta dP$, $\theta \in \Theta$.
As an exercise, let us find the M-estimator for the logistic distribution:

$$p_\theta(x) = \frac{\exp^{\theta - x}}{(1 + \exp^{\theta - x})^2}, \ \theta \in \mathbf{R}, x \in \mathbf{R}.$$

The loss function is:

$$\rho_\theta = x - \theta + 2log(1 + \exp^{\theta - x})$$

Take derivative of the loss function w.r.t. θ:

$$\frac{d}{d\theta}\rho_\theta = -1 + \frac{2\exp^{\theta - x}}{1 + \exp^{\theta - x}}$$

Then $\hat{\theta}_n$ is the solution to:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{2\exp^{\theta - x}}{1 + \exp^{\theta - x}} - 1\right) = 0$$

Notation. For all $\theta \in \Theta$, denote the theoretical integrated loss function and the empirical average by

$$\begin{aligned} \Gamma(\theta) &= \int \gamma_\theta dP \\ \Gamma_n(\theta) &= \int \gamma_\theta dP_n \end{aligned}$$

We first present an easy proposition on consistency of M-estimators with a very stringent condition.

Proposition 5.1 *Suppose that $\theta \mapsto \Gamma(\theta)$ is continuous. Assume moreover that*

$$\sup_{\theta\in\Theta} |\Gamma_n(\theta) - \Gamma(\theta)| \to 0, \quad a.s.$$

i.e., that $\{\gamma_\theta : \theta \in \Theta\}$ is a GC class. Then $\hat{\theta}_n \to \theta_0$ a.s.
(Proof: Since θ_0 is the minimizer of $\Gamma(\theta)$ and θ_n is the minimizer of $\Gamma_n(\theta)$, we have:

$$\begin{aligned} \Gamma_n(\hat{\theta}_n) &\leq \Gamma_n(\theta_0) \\ \Gamma(\theta_0) &\leq \Gamma(\hat{\theta}_n) \end{aligned}$$

It follows that

$$\begin{aligned} 0 &\leq \Gamma(\hat{\theta}_n) - \Gamma(\theta_0) \\ &\leq -\{[\Gamma_n(\hat{\theta}_n) - \Gamma(\hat{\theta}_n)] - [\Gamma_n(\theta_0) - \Gamma(\theta_0)]\} \\ &\to 0 \ a.s. \end{aligned}$$

by the sup condition. The continuity assumption implies that $\hat{\theta}_n \to \theta_0 a.s.$

The following example illustrates why this proposition is not very useful by implicitly assuming something close to the compactness of the parameter space. In empirical estimations, we do not want to impose compactness a priori. Following the example, we will give a lemma that relax the compactness assumption.

Example. Let the loss function be $\gamma_\theta(x) = (x - \theta)^2$. The theoretical integration and the empirical average of the loss function are:

$$\begin{aligned} \Gamma(\theta) &= E(X - \theta_0)^2 - 2(\theta - \theta_0)E(X - \theta_0) + (\theta - \theta_0)^2 \\ \Gamma_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta_0)^2 - 2(\theta - \theta_0)\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta_0) + (\theta - \theta_0)^2 \end{aligned}$$

Since Θ is unbounded, the term $2(\theta - \theta_0)$ is also unbounded. By assuming $\sup_{\theta\in\Theta} |\Gamma_n(\theta) - \Gamma(\theta)| \to 0$, *a.s.*, it is close to requiring the compactness of Θ. The next lemma replaces the sup condition with a convexity assumption, which works well when Θ is of finite dimension.

Lemma 5.2 Consistency of M-Estimators. *Suppose that Θ is a subset of an open convex set of $\mathbf{R}^r$, and that $\theta \mapsto \gamma_\theta$, $\theta \in \Theta$ is convex. Then $\hat{\theta}_n \to \theta_0$, a.s.*
(Proof sketch: By convexity, for any $\alpha > 0$,

$$\begin{aligned} \Gamma_n(\alpha\hat{\theta}_n + (1 - \alpha)\theta_0) &\leq \alpha\Gamma_n(\hat{\theta}_n) + (1 - \alpha)\Gamma_n(\theta_0) \\ &\leq \Gamma_n(\theta_0) \end{aligned}$$

Take $\hat{\alpha}_n = \frac{1}{1+\|\hat{\theta}_n - \theta_0\|}$ (where $\|\cdot\|$ denotes the norm of a vector), and $\tilde{\theta}_n = \hat{\alpha}_n\hat{\theta}_n + (1 - \hat{\alpha}_n)\theta_0$. Then

$$\|\tilde{\theta}_n - \theta_n\| = \frac{\|\hat{\theta}_n - \theta_0\|}{1 + \|\hat{\theta}_n - \theta_0\|} \leq 1$$

which leads to:

$$0 \leq \Gamma(\tilde{\theta}_n) - \Gamma(\theta_0) \leq -\{[\Gamma_n(\tilde{\theta}_n) - \Gamma(\tilde{\theta}_n)] - [\Gamma_n(\theta_0) - \Gamma(\theta_0)]\} \to 0 \ a.s.$$

using the argument that point-wise convergence implies uniform convergence for convex functions. It follows then

$$\|\tilde{\theta}_n - \theta_0\| \to 0 \ a.s.$$

and simple calculation shows that

$$\|\hat{\theta}_n - \theta_0\| \to 0 \ a.s.$$

Having discussed about consistency of the M-estimators, we next show the conditions under which the M-estimators are asymptotically normal. We prove normality via asymptotic linearity condition. Throughout the discussion of asymptotical normality, we assume that $\hat{\theta}$ is consistent, and that θ_0 is an interior point of $\theta \subset \mathbf{R}^r$.

Definition: Asymptotically linear. The (sequence) of estimator(s) $\hat{\theta}$ of θ_0 is asymptotically linear if we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n} \int l dP_n + o_{\mathbf{P}}(1),$$

where

$$l = \begin{pmatrix} l_1 \\ \vdots \\ l_r \end{pmatrix} : \mathcal{X} \to \mathbf{R}^r$$

satisfies $\int l dP = 0$ and $\int l_k^2 dP < \infty, \ k = 1, \ldots, R$. The function l is called the **influence function**. For the case $r = 1$, we call $\sigma^2 = \int l^2 dP$ the asymptotic variance.

Conditions a, b, c for Asymptotic Normality. We start with conditions a, b, and c, which are easier to check but more stringent. Later we relax them to conditions A, B, and C.

Condition a. There exists an $\epsilon > 0$ such that $\theta \mapsto \gamma_\theta$ is differentiable for all $|\theta - \theta_0| < \epsilon$ and all x, with derivative

$$\psi_\theta(x) = \frac{\partial}{\partial \theta} \gamma_\theta(x), \ x \in \mathcal{X}.$$

Condition b. As $\theta \to \theta_0$, we have

$$\int (\psi_\theta - \psi_{\theta_0}) dP = V(\theta - \theta_0) + o(1)|\theta - \theta_0|$$

where V is a positive definite matrix.

Condition c. There exists an $\epsilon > 0$ such that the class

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

has envelope $\Psi \in L_2(P)$ and is Donsker. Moreover,

$$\lim_{\theta - \theta_0} \|\psi_\theta - \psi_{\theta_0}\} = 0$$

Lemma 5.3 *Suppose conditions a, b, and c hold. Then $\hat{\theta}_n$ is asumptotically linear with influence function*

$$l = -V^{-1}\psi_{\theta_0}$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow^{\mathcal{L}} \mathcal{N}(0, V^{-1}JV^{-1}),$$

where $J = \int \psi_{\theta_0}\psi_{\theta_0}^T dP$.

(Proof sketch:

$$\text{condition (a)} \Rightarrow \int \psi_{\hat{\theta}_n} dP_n = 0$$

$$\text{condition (a) and (c)} \Rightarrow \int \psi_{\theta_0} dP = 0$$

We now have:

$$\begin{aligned}
0 &= \int \psi_{\hat{\theta}_n} dP_n - \int \psi_{\theta_0} dP \\
&= \int \psi_{\hat{\theta}_n} d(P_n - P) + \int (\psi_{\hat{\theta}_n} - \psi_{\theta_0}) dP \\
&= \int \psi_{\hat{\theta}_n} d(P_n - P) + V(\hat{\theta}_n - \theta_0) + o_P(1)|\hat{\theta}_n - \theta_0| \text{(by condition b)} \\
&= \int \psi_{\theta_0} d(P_n - P) + o_P(\frac{1}{\sqrt{n}}) + V(\hat{\theta}_n - \theta_0) + o_P(1)|\hat{\theta}_n - \theta_0| \text{(by condition c)}
\end{aligned}$$

which leads to:

$$\begin{aligned}
(V + o_P(1))(\hat{\theta}_n - \theta_0) &= -\int \psi_{\theta_0} d(P_n - P) + o_P(\frac{1}{\sqrt{n}}) \\
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \sqrt{n} \int l d(P_n - P) + o_P(1)
\end{aligned}$$

where $l = -V^{-1}\psi_{\theta_0}$.

In the following discussion, we relax the differentiability assumption of the loss functions.

Conditions A, B, C for Asymptotic Normality.

Condition A: Differentiability in Quadratic Mean. There exists a function $\psi_0 : \mathcal{X} \rightarrow \mathbf{R}^r$, with components in $L_2(P)$, such that

$$\lim_{\theta \rightarrow \theta_0} \frac{\|\gamma_\theta - \gamma_{\theta_0} - (\theta - \theta_0)^T \psi_0\|}{|\theta - \theta_0|} = 0$$

Condition B. As $\theta \rightarrow \theta_0$, we have

$$\Gamma(\theta) - \Gamma(\theta_0) = \frac{1}{2}(\theta - \theta_0)^T V(\theta - \theta_0) + o(1)|\theta - \theta_0|^2,$$

with V a positive definite matrix.

Condition C. For $\theta \neq \theta_0$, define

$$g_\theta = \frac{\gamma_\theta - \gamma_{\theta_0}}{|\theta - \theta_0|}$$

Suppose that for some $\epsilon > 0$, the class $\{g_\theta : 0 < |\theta - \theta_0| < \epsilon\}$ has envelope $G \in L_2(P)$ and that it is a Donsker class.

Lemma 5.4 *Suppose conditions A, B, and C hold. Then $\hat{\theta}_n$ has influence function*

$$l = -V^{-1}\psi_0$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{\mathcal{L}} \mathcal{N}(0, V^{-1}JV^{-1}),$$

where $J = \int \psi_0 \psi_0^T dP$.
(Proof: see Van de Geer (2000).)

References

[1] Pollard, David. (1984) *Convergence of Stochastic Processes.* Springer Verlag, New York.

[2] Pollard, David. (1990), "Empirical Processes: Theory and Applications". *NSFCBMS Regional Conference Series in Probability and Statistics 2.* Institute of Mathematical Statistics and American Statistical Association.

[3] Van de Geer, Sara. (2000) *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge, UK.