Empirical Processes Reading Group Notes

Manu Navjeevan

June 4, 2021

Contents

1 Math Review

1.1 Vector Spaces and Norms

Definition 1.1 (Vector Space)**.** A vector space X is a set of elements with two operations, addition $(+)$ and scalar multiplication $(\cdot)$, and an additive identity $\mathbf{0} \in X$ satisfying:

1. $x + y = y + x$
2. $(x + y) + z = x + (y + z)$
3. $\mathbf{0} + x = x$, $\forall x \in X$
4. $\alpha(x + y) = \alpha x + \alpha y$
5. $(\alpha + \beta)x = \alpha x + \beta x$
6. $(\alpha\beta)x = \alpha(\beta x)$
7. $0x = \mathbf{0}$ and $1x = x$

Examples include $\mathbb{R}^K$ and $\mathcal{C}[a, b]$, the set of all continuous functions from $[a, b] \to \mathbb{R}$.

Definition 1.2 (Norm)**.** Let X be a vector space. A norm is a functional, $\|\cdot\| : X \to \mathbb{R}$ satisfying

1. $\|x\| \geq 0$, $\forall x \in X$ and $\|x\| = 0$ if and only if $x = \mathbf{0}$.
2. $\|x + y\| \leq \|x\| + \|y\|$ (Triangle Inequality)
3. $\|\alpha x\| = |\alpha|\|x\|$, $\forall \alpha \in \mathbb{R}, x \in X$

Examples of norms include the ℓ^p norms on $\mathbb{R}^K$ or the sup-norm on the space of all bounded, real valued, functions. On $\mathbb{R}^K$ all norms are equivalent, which is to say that for any two norms $\|\cdot\|_1, \|\cdot\|_2$ there are constants C_1, C_2 such that $C_1\|\cdot\|_2 \leq \|\cdot\|_1 \leq C_2\|\cdot\|_2$. However, this is not generally the case for functional vector spaces. For example on $\mathcal{C}[a, b]$ there is no constant c such that, for all f:

$$\sup_{x\in[a,b]} f(x) = \|f\|_\infty \leq c\|f\|_2 = \left(\int_a^b f^2(x)dx\right)^{1/2}.$$

Closely related to a norm is the concept of a metric, which is a way of defining a distance on a space.

Definition 1.3 (Metric)**.** Let X be a vector space. A metric (or distance metric) on X is a functional $d(x, y) : X \times X \to \mathbb{R}$ satisfying:

1. $d(x, y) \geq 0, \forall x, y$ and $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, z) + d(z, y)$, $\forall x, y, z$

It is straightforward to verify that, given a norm on a vector space X, we can generate a valid metric:

$$d_{\|\cdot\|}(x, y) := \|x - y\|.$$

We return to these concepts when discussing a topology.

1.2 Topology and Continuity

A topology is a general structure under which we can discuss concepts such as convergence and continuity. We can start with a general structure and then discuss spaces where the topology is generated by a metric (or norm).

The most recent version of these notes can be found here.

Definition 1.4 (Topology)**.** A topology on a set X is a collection of subsets of X, $\boldsymbol{\tau} \subset 2^X$ satisfying:

1. $\emptyset, X \in \boldsymbol{\tau}$.
2. $\boldsymbol{\tau}$ is closed under finite intersections, if $\{A_k\}_{k=1}^K \in \boldsymbol{\tau}$ then $\bigcap_{k=1}^K A_k \in \boldsymbol{\tau}$.
3. $\boldsymbol{\tau}$ is closed under arbitrary unions. For any index set I, if $\{A_k\}_{k\in I} \in \boldsymbol{\tau}$ then $\bigcup_{k\in I} A_k \in \boldsymbol{\tau}$.

The elements of $A \in \boldsymbol{\tau}$ are called open sets. A set B is closed if it's complement is in $\boldsymbol{\tau}$, $B^c \in \boldsymbol{\tau}$.

Some simple examples include the trivial topology, $\boldsymbol{\tau} = \{X, \emptyset\}$ and the discrete topology $\boldsymbol{\tau} = 2^{\mathcal{X}}$. Given a topology, we can define some familiar terms:

Definition 1.5 (Interior)**.** For a subset $A \subseteq X$, the interior of A, denoted A°, is the largest open set included in A (where largest is defined under the usual subset ordering). We can also express this as the union of all open sets contained by A.

$$A^\circ = \bigcup \{B : B \in \boldsymbol{\tau}, B \subseteq A\}.$$

Note that a set is open if and only if $A = A^\circ$.

Definition 1.6 (Closure)**.** For a subset $A \subseteq X$, the closure of A, denoted $\bar{A}$, is the smallest closed set the covers A. We can express this as the intersection of all closed sets containing A:

$$\bar{A} = \bigcap \{B : B^c \in \boldsymbol{\tau}, A \subseteq B\}.$$

By De-Morgan's law and closure of the topology under arbitrary union we can see that this intersection always gives a closed set. A set is closed if and only if $A = \bar{A}$.

Definition 1.7 (Boundary)**.** The boundary of a set A, denoted δA, is $\bar{A} \setminus A^\circ$.

A useful concept when talking about convergence under a topology is that of a neighborhood of a point $x \in X$.

Definition 1.8 (Neighborhood)**.** For a point $x \in X$ a set V is a neighborhood of X if $x \in V^\circ$.

Lemma 1.1. *Suppose $x \in \bar{A}$, then for every neighborhood of x, V_x, we have that $V_x \cap A \neq \emptyset$.*

Proof. Let $x \in \bar{A}$ and suppose for some neighborhood V_x of x we have that $V_x \cap A = \emptyset$. Then we know that $V_x^\circ \cap A = \emptyset$. Take $\tilde{A} = \bar{A} \cap (V_x^\circ)^c$. We can verify that this is a smaller closed set that also contains A. □

We can now use the topology to define limit points and convergence.

Definition 1.9 (Limit Point)**.** A point $x \in X$ is a limit point of a set $A \subseteq X$ if, for every neighborhood V of x,

$$A \bigcap (V \setminus \{x\}) \neq \emptyset.$$

In other words, every neighborhood of x intersects with A at a point other than x. Let A' be the set of all limit points of $A \subseteq X$.

Lemma 1.2. *If S is a subset of X, then $\bar{S} = S \cup S'$.*

Proof. First show that $\bar{S} \subseteq S \cup S'$. Let $x \in \bar{S}$. If $x \in S$ then we are done. Otherwise, suppose $x \in \bar{S} \setminus S$. This means that for all V_x we have that $S \cap V_x = S \cap (V_x \setminus \{x\})$. By the result of Lemma 1.1, we have that $V_x \cap S \neq \emptyset$. So, $x \in S'$.

Now suppose that $x \in S \cup S'$. Clearly if $x \in S$ then $x \in \bar{S}$. Suppose then that $x \in S' \setminus S$ but $x \notin \bar{S}$. Let $\tilde{S}$ be any closed set containing S, that is $S \subseteq \tilde{S}$. For sake of contradiction, suppose that $x \notin \tilde{S}$ (x is a limit point of S that is not in $\tilde{S}$). Because $\tilde{S}$ is closed we know that $\tilde{S}^c \in \boldsymbol{\tau}$. Further, we know that $x \in \tilde{S}^c$ so that $\tilde{S}^c$ is a neighborhood of x. Since x is a limit point of S, we know that $\tilde{S}^c \cap S = \tilde{S}^c \cap S \setminus \{x\} \neq \emptyset$. However, we also know that $S \subseteq \tilde{S}$ so we have a contradiction. Therefore, it must be that $x \in \bar{S}$ which completes the proof. □

Lemma 1.3 (Characterization of Closed Sets)**.** *A set is closed if and only if it contains all of its limit points.*

Proof. This is a consequence of Lemma 1.2 and the fact that A is closed if and only if $\bar{A} = A$. □

Definition 1.10 (Convergence)**.** We say a sequence $\{x_n\}_{n=1}^{\infty}$ converges to a point $x \in X$ if for every neighborhood V_x of x, there exists a number M such that for all $m \geq M$, $x_m \in V_x$.

Note that under the trivial topology $\boldsymbol{\tau} = \{\emptyset, X\}$ all sequences converge to any point $x \in X$ whereas under the discrete topology on $\mathbb{R}$, $\boldsymbol{\tau} = 2^{\mathbb{R}}$, the only sequences that converges to a point x are the trivially convergent sequences, $x_n = x$ for all $n \geq M$ and some M.

Definition 1.11 (Continuity)**.** Let $(\mathcal{X}, \boldsymbol{\tau}_1)$ and $(\mathcal{Y}, \boldsymbol{\tau}_2)$ be two topological spaces and $f : \mathcal{X} \to \mathcal{Y}$. We say f is continuous if $f^{-1}(A) \in \boldsymbol{\tau}_1$ for all $A \in \boldsymbol{\tau}_2$. That is, a continuous function maps open sets to open sets.

We can now get ready to combine the notions of continuity and convergence coming from a topology with the notions that we are familiar with from metric spaces. First, we need to define the topology generated by a metric.

Definition 1.12 (Generated Topology)**.** Let $\mathcal{A}$ be a collection of subsets of X. The topology generated by $\mathcal{A}$, $\langle \mathcal{A} \rangle$ is the smallest topology that contains $\mathcal{A}$:

$$\langle \mathcal{A} \rangle = \bigcap \{\boldsymbol{\tau} : \mathcal{A} \subseteq \boldsymbol{\tau}\}.$$

We will then define the topology generated by a metric as the topology generated by the collection of open balls $B(x, \epsilon)$.

Definition 1.13 (Open Ball)**.** Let $d(x, y)$ be a metric on a vector space X. For any point $x \in X$ define the open ball of size ϵ around x as:

$$B(x, \epsilon) = \{y : d(x, y) \leq \epsilon\}.$$

In a metric space, we consider the topology generated by all the open balls $\boldsymbol{\tau}_d = \langle \{B(x, \epsilon) : x \in X, \epsilon > 0\} \rangle$. In fact, the set of open balls is a basis for this topology, which means that every open set A in $\boldsymbol{\tau}_d$ and any point $x \in A$, there is an open ball B such that $x \in B \subseteq A$.[1]. Many topological properties such as continuity or convergence can be verified by simply confirming the properties for all members of a basis for the topology. This ties together the "epsilon-delta" notions of continuity and convergence with the more general topological versions given above.

For the rest of this subsection we will talk about separability and compactness, but give examples using normed-metric spaces instead of talking in generality about the topology.

Definition 1.14 (Dense Subset)**.** A topological space $(X, \boldsymbol{\tau})$ has a dense subset $\mathcal{A}$ if $\bar{\mathcal{A}} = X$. Equivalently, by Lemma 1.2, every point of X is either in $\mathcal{A}$ or is a limit point of $\mathcal{A}$.

Informally, all points in X are either in $\mathcal{A}$ or arbitrarily "close" to $\mathcal{A}$. As an example, in the standard topology on $\mathbb{R}$ generated by the metric $d(x, y) = |x - y|$, the rationals $\mathbb{Q}$ are dense. We also have that, for the set of continuous functions under the sup norm, the set of all polynomials is dense, which means that we can approximate a function arbitrarily well with them.

Definition 1.15 (Seperable Space)**.** We say that a topological space $(X, \boldsymbol{\tau})$ is separable if it has a countable dense subset.

As we went over above, the real line with its standard topology is separable. The $L_p[a, b]$ spaces are also generally separable for $1 \leq p \leq \infty$. However L_∞ is not separable, which will cause issues (this is not the example below).

[1] In fact, the set of all open balls with rational ϵ is a basis for the topology

Example 1.1 (Bounded functions with the sup norm is not seperable)**.** Let $\{f_i\}_{i\in\mathbb{N}}$ be a countable set of functions on $B_\infty[a,b]$. Let $\{q_i\}_{i\in\mathbb{N}}$ be some counting of the rational numbers between a and b. Let $\tilde{f}$ be some function that is equal to 0 except on the rational numbers. For each rational number q_i define

$$\tilde{f}(q_i) = \begin{cases} 1 & \text{if } f_i(q_i) \leq 0 \\ -1 & \text{if } f_i(q_i) > 0 \end{cases}.$$

We can see that $\tilde{f}$ is bounded (and integrates to 0), but it is at least distance one from each function in $\{f_i\}_{i\in\mathbb{N}}$.

Initially I thought this example would work for $L_\infty[a,b]$, but this only forces a difference on a set of measure 0 and I believe L_∞ works with an essential supremum norm.

Another important/useful concept is that of compactness. The general notion is given below:

Definition 1.16 (Compact Set)**.** A set A is compact if for every collection of open sets $\{G_i\}$ such that $A \subset \bigcup G_i$, there is a finite subcollection that also covers A.

Example 1.2. The real-line is not compact. Consider the open cover $\big\{(n, n+1)|n \in \mathbb{Z}\big\}$

Example 1.3. The interval $(0, 1]$ is not compact. Consider the open cover $\big\{(1/n, 1+1/n)|n \in \mathbb{N}\big\}$

Theorem 1.1 (Heine-Borel)**.** *For a subset S of the Euclidean Space*[2]*, $\mathbb{R}^n$, the following statements are equivalent:*

- *S is closed and bounded*
- *S is compact*

Compactness is nice because of various extreme value theorems that ensure that a supremum or infimum is attained. Heine-Borel gives a nice way of characterizing compactness for Euclidean Spaces, but there is no equivalent result for general metric spaces. We have to strengthen the boundedness assumption.

Definition 1.17 (Totally Bounded)**.** A set $\mathcal{A}$ is totally bounded if for each $\epsilon > 0$ there exists a finite sequence $\{a_1, \ldots, a_n\}$ such that for $B_i = \big\{a \in \mathcal{A} : \|a - a_i\| \leq \epsilon\big\}$, $\bigcup_{i=1}^N B_i$ covers A.

Intuition: For any precision ϵ, you can find a finite set of points that describe $\mathcal{A}$ arbitrarily well. (much more demanding in infinite dimensions than just bounded).

Theorem 1.2. *In a complete metric space, the following are equivalent:*

- *$\mathcal{A}$ is a compact subset*
- *$\mathcal{A}$ is closed and totally bounded*
- *Every sequence in $\mathcal{A}$ has a convergent subsequence which converges to a point in $\mathcal{A}$.*

For a compact set T, let $C(T)$ be the set of continuous functions from T to $\mathbb{R}$ equipped with the sup norm. We may want to characterize when a subset K of $C(T)$ is compact.

Definition 1.18 (Equicontinuous)**.** A set of functions $K \subseteq C(T)$ is equicontinuous if for every $t_0 \in T$ and $\epsilon > 0$ there is a $\delta > 0$ such that $|f(t) - f(t_0)| < \epsilon$ whenever $\|t - t_0\| < \delta$ **for all** $f \in K$.

This is a bit like to uniformly continuity but adapted a bit to deal with a function space.

Theorem 1.3 (Arzela-Ascoli)**.** *If T is compact, then $K \subseteq C(T)$ is compact (under the sup-norm) if and only if K is bounded and equicontinuous.*

This concludes our discussion of topology and continuity. We now review measurability.

[2]That is the space $\mathbb{R}^n$ equipped by the topology generated by the standard distance metric

1.3 Probability Spaces and Outer Measure

Definition 1.19 (Sigma Algebra)**.** A collection of subsets $\mathcal{F}$ is a sigma-algebra (or sigma-field) if it contains the whole set and is closed under complement and under countable union.

Definition 1.20 (Borel Sigma Algebra)**.** For any collection of sets $\mathcal{A}$, we call the smallest sigma algebra containing $\mathcal{A}$, $\sigma(\mathcal{A})$, the sigma algebra generated by $\mathcal{A}$. The Borel sigma algebra on a topological space is the sigma algebra generated by all the open sets, $\mathcal{B}(X) = \sigma(\boldsymbol{\tau})$.

The Borel sigma algebra is useful as it makes all continuous functions measurable (defined below).

Definition 1.21 (Probability Space)**.** A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a set of elements Ω, a sigma algebra on Ω, $\mathcal{F}$, and a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying:

1. $\mathbb{P}(A) \geq \mathbb{P}(\emptyset) = 0$
2. If $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$
3. $\mathbb{P}(\Omega) = 1$.

A measurable function between two spaces equipped with sigma algebra's is simply one that maps measurable sets to measurable sets, similar to the definition of a continuous function.

Definition 1.22 (Measurable Map)**.** A function $f : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ is measurable if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$

Lemma 1.4 (Lemma 1.3.1 VdV& W)**.** *The Borel σ-field on a metric space $\mathbb{D}$ is the smallest σ-field that makes all elements of $C_b(\mathbb{D})$ measurable (with respect to the Borel sets on $\mathbb{R}$).*[1]*.*

Proof. For any closed set F, F is the null set $\{x : f(x) = 0\}$ of the continuous, bounded function, $x \mapsto d(x, F) \wedge 1$. Since the singleton $\{0\}$ is a closed set in $\mathbb{R}$ (all metric spaces are Hausdorff), F must be in the sigma algebra on $\mathbb{D}$ to make $d(x, F) \wedge 1$ measurable. Since all the closed sets generate the Borel σ-field (because σ-fields are closed under complement), all Borel sets must be included in the sigma-algebra on $\mathbb{D}$. □

Given this, we can abstractly think about a random variable as a measurable map from a probability space into another measurable space (typically the real-line). Measurability ensures that things like expectations and probabilities of random variables are well defined.

However, measurability becomes a problem when we are dealing with random functions. For example, if X is a map from a probability space to $L_\infty[a, b]$, the Borel-sigma algebra on $L_\infty[a, b]$ is quite large (its not separable). This means that measurable sets in $L_\infty[a, b]$ may not map back to measurable sets on the probability space $\Omega, \mathcal{F}, \mathbb{P}$.

This is a problem because L_∞ is typically a useful space to work in for empirical process theory. So we have to find a way to relax measurability. This means that we work with outer expectations and probabilities:

Definition 1.23 (Outer Measure and Inner Measure)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space $T : \Omega \to \mathbb{R}$. Define the outer expectation:

$$\mathbb{E}^\star[T] = \inf\left\{\mathbb{E}[U] : T \leq U, U \text{ is measurable}\right\}.$$

and the inner expectation:

$$\mathbb{E}_\star[T] = \sup\left\{\mathbb{E}[U] : U \leq T, U \text{ is measurable}\right\}.$$

We can use this to define inner and outer probability measures by restricting T to be the indicator function for an arbitrary set B. Inner and outer expectations are generally nicely behaved but they require modified versions of dominated and monotone convergence and Fubini's theorem breaks down.

[1] $C_b(\mathbb{D})$ is the set of all continuous bounded functions from $\mathbb{D} \to \mathbb{R}$, where $\mathbb{R}$ is endowed with the standard topology on the real line

2 Weak Convergence

2.1 Definition and Characterizations

We can now talk about weak convergence of random variables. Let X_n be a real-valued random variable with cdf $F_n(t)$ and let X be a random variable with cdf $F(t)$. The typical definition of weak convergence is that $X_n \xrightarrow{L} X$ if $F_n(t) \to F(t)$ pointwise at all continuity points of F. This is not super general for non-real valued random maps.

Theorem 2.1 (Portmanteau)**.** *For real random variables $X_n \xrightarrow{L} X$ is equivalent to:*

- $\mathbb{E}\left[g(X_n)\right] \to \mathbb{E}[g(X)]$ *for all bounded continuous functions.*
- *For all open sets G,* $\liminf \mathbb{P}(X_n \in G) \geq P(X \in G)$.
- *For all closed sets K,* $\limsup \mathbb{P}(X_n \in K) \leq \mathbb{P}(X \in K)$.

This motivates the theory of weak convergence for general metric spaces. Let $\mathbb{D}$ be a complete metric space with metric d. We can equip $\mathbb{D}$ with it's Borel-sigma algebra as defined in Definition 1.20 and a tight probability measure as defined in Definition 2.1. Let $C_b(\mathbb{D})$ be the set of all continuous and bounded real functions on $\mathbb{D}$. If X is a random variable, $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{D}$ then it's law is given $L = \mathbb{P} \circ X^{-1}$.

Definition 2.1 (Tight Probability Measure)**.** A probability measure is tight if for every $\epsilon > 0$ there is a compact set K_ϵ such that $P(K_\epsilon) \geq 1 - \epsilon$

This is a generalization of bounded in probability I believe.

Definition 2.2 (Borel Law)**.** For a random variable X, we say that X has a Borel Law L if

$$\mathbb{P}(X \in A) = \int_A dL.$$

for all Borel sets A.

Given this setup, we can now define weak convergence:

Definition 2.3 (Weak Convergence)**.** Let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be a sequence of probability spaces and $X_n : \Omega_n \to \mathbb{D}$. Then we say that $X_n \xrightarrow{L} X$ if:

$$\mathbb{E}^\star\left[f(X_n)\right] \to \mathbb{E}[f(X)].$$

for every $f \in C_b(\mathbb{D})$

We can characterize this convergence using another Portmanteau theorem.

Theorem 2.2 (Portmanteau)**.** *The following are equivalent:*

1. $X_n \xrightarrow{L} X$
2. $\liminf \mathbb{P}_\star(X_n \in G) \geq \mathbb{P}(X \in G)$ *for all open sets G.*
3. $\limsup \mathbb{P}^\star(X_n \in F) \leq \mathbb{P}(X \in F)$ *for every closed set F.*
4. $\lim P(X_n \in B) = P(X \in B)$ *for every Borel set B with $P(X \in \delta B) = 0$.*

Proof. This proof is in a few steps.

(4) $\implies$ (3): Suppose that $\lim P(X_n \in B) = P(X \in B)$ for every Borel set B with $\mathbb{P}(X \in \delta B) = 0$. Let F be a closed set and let $F^\epsilon = \{x : d(x, F) < \epsilon\}$. The sets δF^ϵ are disjoint for different values of $\epsilon > 0$ (The boundary of this set is $\delta F^\epsilon = \{x : d(x, F) = \epsilon\}$), so at most countably many of them can have nonzero L-measure (otherwise the measure of the entire space would be infinite). Choose a sequence $\epsilon_m \downarrow 0$ with $L(\delta F^{\epsilon_m}) = 0$ for each m (this is possible because only countably many ϵ have $L(F^\epsilon) \neq 0$). For a fixed m, by (4) we have that:

$$\limsup P^\star(X_\alpha \in F) \leq \limsup P^\star\left(X_\alpha \in \overline{F^{\epsilon_m}}\right) = L\left(\overline{F^{\epsilon_m}}\right).$$

letting $m \to \infty$ gives (3).

(3) $\iff$ (2): Take any closed set F. Its complement F^c is open. If

$$\liminf \mathbb{P}_\star(X_n \in F^c) \geq \mathbb{P}(X \in F^c).$$

Then

$$\begin{aligned}\limsup \mathbb{P}^\star(X_n \in F) &\leq \liminf 1 - \mathbb{P}_\star(X_n \in F^c) \\ &\leq 1 - \mathbb{P}(X \in F^c) \\ &= \mathbb{P}(X \in F)\end{aligned}$$

a symmetric argument shows the backwards direction.

(2)+(3) $\implies$ (4): This is straightforward if we recall that, for any set with $L(\delta B) = 0$ we have that $L(B) = L(\bar{B})$. Then we bound the lim sup by the lim inf:

$$\limsup \mathbb{P}^\star(X \in B) \leq \limsup \mathbb{P}(X \in \bar{B}) \leq \mathbb{P}(X \in \bar{B}) = \mathbb{P}(X \in B) \leq \liminf \mathbb{P}_\star(X_n \in B).$$

which gives (4).

(1) $\implies$ (2): Take any G open and define the sequence of functions:

$$f_m(x) := \min(1, m \cdot d(x, G^c))$$

Notice that $f_m(x) \in C_b(\mathbb{D})$ and $f_m(x) \leq \mathbb{1}\{x \in G\}$. So, for every m we have that

$$\begin{aligned}\liminf \mathbb{P}_\star(X \in G) &= \liminf \mathbb{E}_\star \left[\mathbb{1}\{X \in G\}\right] \\ &\geq \liminf \mathbb{E}_\star \left[f_m(X)\right] \\ &\geq \mathbb{E}[f_m(X)]\end{aligned}$$

since $f_m(x) \uparrow \mathbb{1}\{X \in G\}$ by monotone convergence we get the result in (2).

Question: How do we know from weak convergence that this sequence converges in inner expectation?

By VdV and Wellner, weak convergence implies (is equivalent to) $\liminf \mathbb{E}_\star \left[f(X_n)\right] \geq \mathbb{E}\left[f(X)\right]$ for every bounded, Lipschitz continuous, non-negative f. I think the argument for why this is the case goes: Let $f \geq 0$ be bounded and continuous. Then by weak convergence

$$\limsup \mathbb{E}^\star[-f(X_n)] = \mathbb{E}[-f(X)].$$

Taking negatives will give:

$$\liminf \mathbb{E}_\star[f(X_n)] \geq -\limsup \mathbb{E}^\star[-f(X_n)] = \mathbb{E}[f(X)].$$

In any case, $f_m(X)$ is Lipschitz continuous which gives the result.

(2) $\implies$ (1): (SKETCH)

- Suppose $f(x) \geq 0$ is continuous and bounded
- Approximate it from above and below by indicator functions of open sets.

□

Weak convergence is nice because it gives the continuous mapping theorem.

Theorem 2.3 (Continuous Mapping Theorem)**.** *Let $g : \mathbb{D} \to \mathbb{E}$ be continuous at every point $\mathbb{D}_0 \subseteq \mathbb{D}$. If $X_n \xrightarrow{L} X$ and $\mathbb{P}(X \in \mathbb{D}_0) = 0$ then $g(X_n) \xrightarrow{L} g(X)$.*

Proof. (Without Discontinuity Points): Let $Z_n = g(X_n)$ and $Z = g(X)$. We want to show that $\mathbb{E}^\star\left[f(Z_n)\right] \to \mathbb{E}\left[f(Z)\right]$ for all $f \in C_b(\mathbb{D};\mathbb{E})$.

$$\lim_{n\to\infty} \mathbb{E}[f(Z_n)] = \lim_{n\to\infty} \mathbb{E}[f(g(X_n))] = \mathbb{E}[f(g(X))] = \mathbb{E}[f(Z)].$$

The main step here is weak convergence of X_n and the stability of $C_b(\mathbb{D};\mathbb{E})$ under composition.

(With Discontinuity Points, from VdV&W): The set D_g of all points at which g is discontinuous can be written

$$D_g = \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} \left\{x : \exists y, z \in B(x, 1/k) \text{ with } d_{\mathbb{E}}(g(y), g(x)) > 1/m\right\}.$$

Intuition: Recall that g is continuous at x if for every $m \in \mathbb{N}$ there exists a $k \in \mathbb{N}$ such that[1]

$$y, z \in B(x, 1/k) \implies d_{\mathbb{E}}(g(y), g(z)) < 1/m$$

If the function is not continuous at x you can find a counterexample for some $k, m \in \mathbb{N}$.

Let $G_k^m = \left\{x : \exists y, z \in B(x, 1/k) \text{ with } d_{\mathbb{E}}(g(y), g(x)) > 1/m\right\}$. Every G_k^m is open (if x is in G_m^k the points just around x will be as well so that we can write G_m^k as a union of open balls) so that D_g is a Borel set. For every closed F we then have that:

$$\overline{g^{-1}(F)} \subseteq g^{-1}(F) \cup D_g.$$

By Portmanteau:

$$\begin{aligned}
\limsup \mathbb{P}^\star\left(g(X_n) \in F\right) \leq \limsup P^\star\left(X_n \in \overline{g^{-1}(F)}\right) &\leq \mathbb{P}\left(X \in \overline{g^{-1}(X)}\right) \\
&= \mathbb{P}\left(X \in g^{-1}(F)\right) \\
&= \mathbb{P}\left(g(X) \in F\right)
\end{aligned}$$

Applying Portmanteau again gives weak convergence. □

Example 2.1. Take $\mathbb{G}_n \in L^\infty(\mathbb{R})$:

$$\mathbb{G}_n(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{1}\{X_i \leq t\} - \mathbb{E}\left[\mathbb{1}\{X \leq t\}\right]\right)$$

and suppose that $\mathbb{G}_n \xrightarrow{L} \mathbb{G}$ where $\mathbb{G}$ is some other element of $L^\infty(\mathbb{R})$. Let $Z : L^\infty(\mathbb{R}) \to \mathbb{R}$ be defined as:

$$Z(f) := \sup_t |f(t)|.$$

this function is continuous. Applying the continuous mapping theorem to Z allows us to build uniform confidence intervals.

Let $\gamma_{1-\alpha}$ be the $1-\alpha$ quantile of $Z := \sup_t |\mathbb{G}(t)|$ and construct a confidence interval (at each t):

$$\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} - \gamma_{1-\alpha}/\sqrt{n}\,,\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} + \gamma_{1-\alpha}/\sqrt{n}\right].$$

Then:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} - \gamma_{1-\alpha}/\sqrt{n} \leq \mathbb{E}\left[\mathbb{1}\{X \leq t\}\right] \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} + \gamma_{1-\alpha}/\sqrt{n} \ : \ \text{for all } t\right) \\
&= \mathbb{P}\left(\left|\mathbb{G}_n(t)\right| \leq \gamma_{1-\alpha}\ \forall t\right) \\
&= \mathbb{P}\left(\sup_t |\mathbb{G}_n(t)| \leq \gamma_{1-\alpha}\right)
\end{aligned}$$

[1] Topologically, this is saying that the inverse map of every open neighborhood of $f(x)$ is an open neighborhood of x

But by continuous mapping theorem and Portmanteau, if $\mathbb{P}(\sup_t |\mathbb{G}| = \gamma_{1-\alpha}) = 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_t \left|\mathbb{G}_n(t)\right| \leq \gamma_{1-\alpha}\right) = \mathbb{P}\left(\sup_t \left|\mathbb{G}(t)\right| \leq \gamma_{1-\alpha}\right) = 1-\alpha.$$

This sort of argument can be applied more generally to functions $\mathbb{G}_n(t) = \hat{m}(t) - m(t)$ to construct uniform confidence intervals.

This shows the usefulness of Portmanteau and Continuous Mapping Theorem. For finite dimension vectors we can use the central limit theorem to establish weak convergence to a normal distribution. However, when X_n is a random element in L^∞ it may be harder to show that $X_n \rightsquigarrow X$ for some other $X \in L^\infty$.

- Don't want to check $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all $f \in C_b(L^\infty)$ [There are at least 20 functions in this class]

Instead we will try to use the structure of L^∞ to show the result.

Definition 2.4 (Asymptotic Tightness)**.** A sequence X_n of random maps is asymptotically tight if for every $\epsilon, \delta > 0$ there is a compact K_ϵ such that

$$\liminf P_\star\left(X_n \in K_\epsilon^\delta\right) \geq 0.$$

where $K_\epsilon^\delta = \{y \in \mathbb{D} : d(y, K_\epsilon) < \delta\}$ is the "δ-enlargement" around K_ϵ.

Definition 2.5 (Asymptotic Measurability)**.** A sequence X_n of random maps is asymptotically measurable if for all $f \in C_b(\mathbb{D})$:

$$\mathbb{E}^\star f(X_n) - \mathbb{E}_\star f(X_n) \to 0.$$

We would like for a sequence X_n that weakly converges to an element X to inherit some properties from X:

Lemma 2.1 (Lemma 1.3.8 VdV& W)**.** *The following are true:*

1. *If $X_n \xrightarrow{L} X$ then X_n is asymptotically measurable*
2. *If $X_n \xrightarrow{L} X$ then X_n is asymptotically tight if and only if X is tight.*

Proof. (1): Take any function $f \in C_b(\mathbb{D})$. By definition of weak convergence we know that

$$\lim \mathbb{E}^\star\left[f(X_n)\right] = \mathbb{E}[f(X)] \quad \text{and} \quad \lim \mathbb{E}^\star\left[-f(X_n)\right] = \mathbb{E}[-f(X_n)].$$

I think we should have that $-\mathbb{E}_\star\left[f(X_n)\right] \geq \mathbb{E}^\star\left[-f(X_n)\right]$ for any f which give the result (I think this holds with equality but I leave it as an inequality since this is all we need for the result).

(2): Fix $\epsilon > 0$. If X is tight then there is a compact K with $\mathbb{P}(X \in K) > 1 - \epsilon$. By Portmanteau:

$$\liminf \mathbb{P}_\star(X_n \in K^\delta) \geq \mathbb{P}(X \in X^\delta).$$

which is larger than $1 - \epsilon$ for every $\delta > 0$.

Conversely, suppose that X_n is asymptotically tight. Then there exists a compact K with $\liminf P_\star(X_n \in K^\delta) \geq 1 - \epsilon$. By Portmanteau,

$$1 - \epsilon \leq \liminf \mathbb{P}_\star(X_n \in K^\delta) \leq \limsup \mathbb{P}^\star(X_n \in \overline{K^\delta}) \leq \mathbb{P}\left(X \in \overline{K^\delta}\right).$$

Let $\delta \to 0$ by monotone convergence to complete the result. [2] □

[2] This proof relies on compact sets being closed in metric spaces. The proof of this is as follows: Let A be compact in a metric space. We wish to show that A is closed. Take a point $x \in X \setminus A$. To show that A is closed, we want to show that there is an open neighborhood of x that is not in A (this will show that A contains all of its limit points). For every $a \in A$, let $U_a = B\left(a, \frac{d(a,x)}{2}\right)$ and $V_a = B\left(x, \frac{d(a,x)}{2}\right)$. By triangle inequality, U_a and V_a are disjoint. The union of all the sets U_a for all points $a \in A$ is an open cover of A. By compactness of A, we can get a finite subcover $U_{a_1}, \ldots, U_{a_n}$. But then $V_{a_1} \cap \cdots \cap V_{a_n}$ is an open neighborhood of x that is disjoint from A. So A is closed. Actually this argument holds in general Hausdorff spaces.

The converse is not generally true. Let $X_n = -1$ if n is odd and $X_n = 1$ if n is even. This sequence is asymptotically measurable and asymptotically tight but clearly does not converge. However, it does converge among a subsequence. This is the idea behind the partial converse to this theorem provided by Pohorov's Theorem.

Theorem 2.4 (Pohorov's Theorem, Theorem 1.3.9 VdV& W)**.** *Let X_n be an asymptotically tight and asymptotically measurable sequence. Then there is a subsequence X_{n_j} that converges weakly to a tight Borel law.*

Example 2.2 (Problem 7; Ch 1.3 VdV& W)**.** Let X_n be a sequence of random elements in $\mathbb{D}$ and $g : \mathbb{D} \to \mathbb{E}$ a continuous function. Want to show that:

1. If X_n is asymptotically tight then $g(X_n)$ is asymptotically tight.
2. If X_n is asymptotically measurable then $g(X_n)$ is asymptotically measurable.

Proof. 1) Suppose that X_n is asymptotically tight. Fix $\epsilon > 0$. We know that there exists a compact set K such that, $\forall \delta_1 > 0$

$$\liminf \mathbb{P}_\star \left(X_n \in K^{\delta_1} \right) \geq 1 - \epsilon.$$

The event $\left\{ X_n \in K^{\delta_1} \right\}$ is a subset of the event that $\left\{ g(X_n) \in g(K^{\delta_1}) \right\}$ so

$$\liminf \mathbb{P}_\star \left(g(X_n) \in g(K^{\delta_1}) \right) \geq \liminf \mathbb{P}_\star \left(X_n \in K_1^{\delta} \right) \geq 1 - \epsilon.$$

To finish recall that $g(K)$ is a compact set and choose δ_1 such that $g(K^{\delta_1}) \subseteq g(K)^{\delta}$ (always possible to do so by continuity of g).

2) Suppose that X_n is asymptotically measurable. This means that, for any $f \in C_b(\mathbb{D})$:

$$\mathbb{E}^\star \left[f(X_n) \right] - \mathbb{E}_\star \left[f(X_n) \right] \to 0.$$

Let $\tilde{f} \in C_B(\mathbb{E})$. For any continuous $g : \mathbb{D} \to \mathbb{E}$, $f \circ g$ is a continuous and bounded function from $\mathbb{D} \to \mathbb{R}$. This completes the proof. □

2.2 Weak Convergence in Space of Bounded Functions

So far, we have defined weak convergence. But, how do we show that $X_n \xrightarrow{L} X$? In $\mathbb{R}^K$ we have the central limit theorem, but no direct analog for random maps into L^∞.

First, some definitions.

Definition 2.6 (Marginal Random Variable)**.** Let X_n be a random map into $L^\infty(T)$ (the space of all bounded functions from $T \to \mathbb{R}$). Then, $X_n(t)$ is the marginal distribution of X_n at t. We can view $X_n(t)$ as the composition of X_n with π_t or directly as a real-valued random variable.

A general strategy will be to deal with the marginals directly. By the central limit theorem, we have conditions for the weak convergence of $X_n(t)$. Want to know what these results imply for the random map X_n.

Lemma 2.2 (Lemma 1.5.1, VdV&W)**.** *Let $X_n : \Omega \to L^\infty(T)$ be asymptotically tight. Then it is asymptotically measurable if and only if $X_n(t)$ is asymptotically measurable for every $t \in T$.*

Lemma 2.3 (Lemma 1.5.3, VdV&W)**.** *Let X and Y be tight Borel measurable maps into $L^\infty(T)$. Then $X \stackrel{L}{=} Y$ if and only if $X(t) \stackrel{L}{=} Y(t)$ for all $t \in T$.*

Theorem 2.5 (Theorem 1.5.4, VdV&W)**.** *Let $X_n : \Omega_n \to L^\infty(T)$ be arbitrary. Then X_n weakly converges to a tight limit if and only if X_n is asymptotically tight and the marginals $\left(X_n(t_1), \ldots, X_n(t_k) \right)$ converge weakly to a limit for every finite subset $t_1, \ldots, t_k$.*

Proof. Forward direction is simple, backwards direction requires more work:

($\implies$) Suppose that $X_n \xrightarrow{L} X$ and X is tight. By Lemma 2.1, this means that X_n is asymptotically tight. Let $T_k : L^\infty(T) \to \mathbb{R}^k$ be the projection onto the coordinates $t_1, \ldots, t_K$. This is a continuous function so by continuous mapping theorem we have convergence of the marginals for any finite collection.

($\impliedby$) Suppose that X_n is asymptotically tight and the marginals converge. Then, by Lemma 2.2, X_n is asymptotically measurable. By Pohorov's theorem, there is a subsequence $X_{n_k} \xrightarrow{L} X$ for some X. Suppose $X_n \not\xrightarrow{L} X$. Then, there is a subsequence $X_{n'_k}$ that stays away from X (in law). However, the marginals converge. This means that the marginals of Y are the same as the marginals of X. By Lemma 2.3, $X \overset{L}{=} Y$. □

Remark (Intuition)**.** Why is tightness and convergence of marginals enough? From tightness we have that $P(X \in K) \geq 1 - \epsilon$ for some compact set K. In a metric space, compact means that for every $\epsilon > 0$ there are a finite set of points that approximate the whole set within an error of ϵ. For a finite set of points we have convergence of marginal distributions by standard central limit theorem.

Showing convergence of marginal distributions is straightforward by CLT. Next, we cover how to show tightness. Then Theorem 2.5 gives convergence of the entire process.To verify tightness we want a better description than the definition of asymptotic tightness. Two approaches

1. Finite Approximation → simpler
2. Arzela-Ascoli Theorem → larger interest (asymptotic equicontinuity)

2.2.1 Finite Approximation

The general idea here is that, for any $\epsilon > 0$, we can partition the index set T (as in $\ell^\infty(T)$) into a finite number of sets T_i so that the variation in each set is $< \epsilon$. Formally, for any $\eta > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_i \sup_{s,t\in T_i} \left|X_n(s) - X_n(t)\right| > \epsilon \right) < \eta.$$

Remark (Intuition)**.** Why should we expect this to work? Tightness means that the probability measure concentrates on a compact set. A compact set in $\ell^\infty(X)$ is well approximated by a finite number of functions.

Theorem 2.6 (Theorem 1.5.6 VdV&W)**.** *A sequence of random maps $X_n \in \ell^\infty(T)$ is asymptotically tight if and only if $X_n(t)$ is asymptotically tight in $\mathbb{R}$ for every t and, for all $\epsilon, \eta > 0$ there is a partition $T = \cup_{i=1}^n T_i$ such that*

$$\limsup_{n\to\infty} \mathbb{P}^\star\left(\max_i \sup_{s,t\in T_i} \left|X_n(s) - X_n(t)\right| > \epsilon \right) < \eta \tag{FA-1}$$

Proof. Cover sufficiency. Necessity follows from Theorem 1.5.7 in Van DerVaart and Wellner. Suppose that (FA-1) holds. Fix $\epsilon > 0$ and let the partition $T = \bigcup_{i=1}^k T_i$ satisfy (FA-1) for some $\eta > 0$. We want to show that $\sup_t \left|X_n(t)\right|$ is asymptotically tight. Then:

$$\begin{aligned}
\limsup \mathbb{P}^\star\left(\sup_{t\in T} \left|X_n(t)\right| > M \right) &\leq \limsup \mathbb{P}^\star\left(\sup_{t\in T} > M, \text{ and (FA-1) holds} \right) \\
&\quad + \limsup \mathbb{P}^\star\left(\text{(FA-1) doesn't hold} \right) \\
&\leq \limsup \mathbb{P}^\star\left(\max_{1\leq i\leq k} \left|x_n(t_i)\right| + \epsilon > M \right) + \eta
\end{aligned}$$

Where in the last line we use the bounded variation within each set T_i and pick some arbitrary elements $t_i \in T_i$. Now note that each $X_n(t_i)$ is asymptotically tight by assumption so that $\max_{1\leq i\leq k_i} \left|X_n(t_i)\right|$ is

asymptotically tight.[1]. This means that we can pick M so that

$$\limsup \mathbb{P}^\star \left(\sup_t \left| X_n(t) \right| > M \right) < \eta.$$

or, to put this another way, for every $\eta > 0$ we can show that there is an M such that:

$$\limsup \mathbb{P}^\star \left(\sup_t \left| X_n(t) \right| > M \right) < \eta.$$

So we have shown that $\sup_t \left| X_n(t) \right|$ is bounded in probability. Since $\sup_t \left| X_n(t) \right|$ is a map onto the real line, bounded in probability coincides with asymptotic tightness (Heine-Borel).

Now we want to construct a candidate compact set K for the process X_n. Fix $\zeta > 0$ and a sequence $\epsilon_n \downarrow 0$. First, pick an M such that

$$\lim \sup_{n \to \infty} \mathbb{P}^\star \left(\sup_t \left| X_n(t) \right| > M \right) < \zeta.$$

we know such an M exists by the above argument. For each ϵ_m partition $T = \bigcup_{i=1}^{K(m)} T_i$ such that

$$\lim \sup_{n \to \infty} \mathbb{P}^\star \left(\sup_{1 \leq i \leq K(m)} \sup_{s,t \in T_i} \left| X_n(s) - X_n(t) \right| > \epsilon_m \right) < \frac{\zeta}{2^m}.$$

For each ϵ_m let $\{z_1, \ldots, z_{p(m)}\}$ be the set of functions in $\ell^\infty(T)$ that are constant on T_i and only take values $0, \pm\epsilon_m, \pm 2\epsilon_m, \ldots, M$. It is only important for now that, for any m, $p(m)$ is finite (though large). Let

$$K_m = \bigcup_{i=1}^{p(m)} \overline{B}(z_i, \epsilon_m).$$

where $\overline{B}(z_i, \epsilon_m)$ is the closed ball of radius ϵ_m around z_i. Note that if $\sup_t \left| X_n(t) \right| \leq M$ and

$$\sup_{1 \leq i \leq k(m)} \sup_{s,t \in T_i} \left| X_n(s) - X_n(t) \right| \leq \epsilon_m$$

then $X_n \in K_m$. Let $K = \bigcap_{m=1}^\infty K_m$. Then K is closed and totally bounded. Closure follows because each K_m is closed (finite union of closed sets) and an arbitrary intersection of closed sets is closed (because the arbitrary union of open sets is open). To see totally bounded fix $\delta > 0$. Then for each $\epsilon_m < \delta$ we have that $K_m = \bigcup_{i=1}^{p(m)} \bar{B}(z_i, \epsilon_m)$. Since $K_m \supset K$ these balls cover K.

We now have a candidate K. We now want to show that, for every $\delta > 0$, $K^\delta \supset \bigcap_{i=1}^m K_i$ for some m. Suppose not. Then there is a sequence $\{z_m\}$ with $z_m \notin K^\delta$ and $z_m \in \bigcap_{i=1}^m K_i$ for every m.[2] This sequence has a subsequence contained in one of the balls making up K_1, this subsequence in one of the balls in K_1 has a further subsequence contained in one of the balls making up K_2, that subsequence contains a subsequence eventually contained in K_3, and so on. [3] Consider the "diagonal" sequence formed by taking the first element of the first subsequence, the second element of the second sequence, and so on. Eventually, this would be contained in a ball of radius ϵ_m for any m.[4]Because $\epsilon_m \downarrow 0$ this means the sequence is Cauchy.

[1]Couple of quick arguments to get this one:

1. If each $X_{i,n}$ in $\{X_{i,n}\}_{i=1}^K$ is asymptotically tight then the vector $\begin{bmatrix} X_1 & \ldots & X_K \end{bmatrix}$ is asymptotically tight. This is because the Cartesian product of a finite number of compact sets is compact (with respect to the product topology).
2. If X_n is asymptotically tight and g is a continuous function then $g(X_n)$ is asymptotically tight. This is shown in Example 2.2 and basically follows from the fact that a continuous function applied to a compact set yields a compact set. The maximum operator is continuous.

[2]Pick $z_m \in \bigcap_{i=1}^m K_i \setminus K^\delta$

[3]Why? Each $\{z_m\}$ is in $\bigcap_{i=1}^m K_i$. Fix some n, then eventually the sequence is contained in $\bigcap_{i=1}^n K_n$ and so is contained in K_n since $K_n \supset \bigcap_{i=1}^n K_n$. This means the sequence $\{z_m\}$ has infinite members in K_n. K_n is the union of a finite number of sets, so one of these sets must contain infinite members

[4]Key here is the boundedness of the functions we are considering.

Since $\ell^\infty(T)$ is a complete (Banach) space this sequence converges and must converge to an element in K. This contradicts the fact that $d(z_m, K) \geq \delta$ for every m.

Finally, combining our previous results, we want to show that $\liminf \mathbb{P}_\star \left(X_n \in K^\delta\right) \geq 1 - 2\zeta$. for every $\delta > 0$. This is equivalent to saying that $\limsup \mathbb{P}^\star \left(X_n \notin K^\delta\right) < 2\delta$. Recall that

$$\sup_t \left|X_n(t)\right| \leq M \quad \text{and} \quad \sup_i \sup_{s,t \in T_i} \left|X_n(s) - X_n(t)\right| \leq \epsilon_m \implies X_n \in K_m.$$

Then, to show asymptotic tightness:

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}^\star \left(X_n \notin \bigcup_{i=1}^n K_i\right) &\leq \limsup \mathbb{P}^\star \left(X_n \notin \bigcup_{i=1}^m K_i; \sup_t \left|X_n(t)\right| \leq M\right) + \underbrace{\limsup \mathbb{P}^\star \left(\sup_t \left|X_n(t)\right| > M\right)}_{<\zeta} \\
&\leq \limsup \mathbb{P}^\star \left(\sup_i \sup_{s,t \in T_i} \left|X_n(s) - X_n(t)\right| > \epsilon_m \text{ for some } m\right) + \zeta \\
&\leq \sum_{j=1}^m \limsup \mathbb{P}^\star \left(\sup_i \sup_{s,t \in T_i} \left|X_n(s) - X_n(t)\right| > \epsilon_j\right) + \zeta \\
&\leq \sum_{j=1}^m \frac{\zeta}{2^j} + \zeta \\
&< 2\zeta
\end{aligned}$$

□

Proof is involved but useful as it shows the equivalence between asymptotic tightness and a finite approximation notion. The proof also builds some intuition for why tightness is important, at each step we are essentially showing that the whole behavior of the set is well describes (up to a tolerance of size ϵ) by a finite set of marginals. Weak convergence of the marginals is much easier to show.

This being said, the condition in Theorem 2.6 is hard to check. In particular, there is no guidance given on how to select the partition $\{T_i\}_{i=1}^m$. The next way to characterize tightness builds on asymptotic equicontiuity. The idea is the correct way to pick the partition is linked to some form of continuity: pick small T_i so that X_n does not move much on T_i.

Definition 2.7 (Asymptotic ρ-equicontinuity in probability)**.** Suppose ρ is a semimetric on T. Then a sequence of maps $X_n : \Omega_n \to \ell^\infty(T)$ is asymptotically ρ-equicontinuous if for every $\epsilon, \eta > 0$ there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}^\star \left(\sup_{d(s,t)<\delta} \left|X_n(s) - X_n(t) > \epsilon\right|\right) < \eta.$$

Remark. This is basically setting $T_i = \{(s,t) : p(s,t) < \delta\}$

Example. Let $X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\mathbb{1}\{X_i \leq t\} - \mathbb{P}(X \leq t)\right]$. Then $\left|X_n(t) - X_n(t')\right| \approx 0$ for all $|t - t'| < \delta$. Note that here, for every n, $X_n(t)$ is still a discontinuous function of t, it's just that the jumps get closer together or smaller.

Example. Suppose that $\gamma = g(X, \beta_0) + \epsilon$ with $\mathbb{E}[\epsilon|X] = 0$. By the vector LLN, we can say that $\hat{\beta} - \beta_0 \to_p 0$.

In contrast, *asymptotic equicontinuity* will allow to say that:

$$\hat{\beta} \to_p \beta_0 \implies \left|\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{\left(g(x_i, \hat{\beta}) - \mathbb{E}[g(x, \hat{\beta})]\right) - \left(g(x_i, \beta_0) - \mathbb{E}\left[g(x, \beta_0)\right]\right)\right\}\right| = o_p(1).$$

which is a more powerful result.

Theorem 2.7 (Theorem 1.5.7 Vdv&W)**.** *A sequence of random maps, $X_n : \Omega_n \to \ell^\infty(T)$ is asymptotically tight if and only if $X_n(t)$ is asymptotically tight in $\mathbb{R}$ for each t and there exists a semimetric ρ on T such that (T, ρ) is totally bounded and X_n is asymptotically uniformly ρ-equicontinuous.*

Proof. First prove sufficiency then necessity:

($\Longleftarrow$) Fix $\epsilon, \eta > 0$. Then, there is a $\delta > 0$ such that

$$\limsup \mathbb{P}^\star \left(\sup_{\rho(s,t)<\delta} \left| X_n(s) - X_n(t) \right| > \epsilon \right) < \eta.$$

Since T is totally bounded, then there are finitely many balls of radius δ that cover T, $B_1, \ldots, B_{K(\delta)}$. Make these balls disjoint by taking successive "set-minuses" and then we have a partition of T. Then

$$\limsup \mathbb{P}^\star \left(\max_i \sup_{s,t \in T_i} \left| X_n(s) - X_n(t) \right| > \epsilon \right) \leq \limsup \mathbb{P}^\star \left(\sup_{\rho(s,t)<\delta} \left| X_n(s) - X_n(t) \right| > \epsilon \right) < \eta$$

and we can apply the results of Theorem 2.6.

($\Longrightarrow$) If X_n is asymptotically tight, then $g(X_n)$ is asymptotically tight for each continuous function g. Let $K_1 \subset K_2 \subset \ldots$ be compact sets with:

$$\liminf \mathbb{P}_\star \left(X_n \in K_m^\epsilon \right) \geq 1 - 1/m.^{5}$$

For each m define a semimetric ρ_m on T by:

$$\rho_m(s,t) = \sup_{z \in K_m} \left| z(s) - z(t) \right|.$$

Then (T, ρ_m) is totally bounded. How? Cover K_m by finitely many balls of arbitrarily small radius η centered at $z_1, \ldots, z_k$.[6] Partition $\mathbb{R}^k$ into cubes of edge η and for every cube pick at most one $t \in T$ such that $\left(z_1(1), \ldots, z_k(t)\right)$ is in the cube. Since $z_1, \ldots, z_k$ are uniformly bounded,[7] this gives finitely many points $t_1, \ldots, t_p$. Now, the balls $\{t : p_m(t, t_i) < 3\eta\}$ cover T: t is in the ball around t_i for which $\left(z_1(t), \ldots, z_k(t)\right)$ and $\left(z_1(t_i), \ldots, z_k(t_i)\right)$ fall in the same cube. This in turn follows from the fact that $\rho_m(t, t_i)$ can be bounded by $2 \sup_{z \in K_m} \inf_i \|z - z_i\|_T + \sup_j \left| z_j(t_i) - z_j(t) \right|$. [8]

[5] We can choose nested compact sets with this property because the union of a finite number of compact sets is compact and the probability functional is increasing with respect to the subset ordering.

[6] This is possible by compactness. Cover K_m by balls of radius η and then take a finite subcover.

[7] Recall that each z_i is in $\ell^\infty(T)$ which is the space of all bounded functions from $T \to \mathbb{R}$. A finite collection of bounded functions is uniformly bounded

[8] Recall that $\|f\|_T = \sup_{t \in T} |f(t)|$,$\rho_m(t, t_i) = \sup_{z \in K_m} \left| z(t) - z(t_i) \right|$, $z_1, \ldots, z_K$ are the points (bounded functions of T) around which balls of radius η cover K_m, and $t_1, \ldots, t_p$ are points of T such that the vector valued function $\left(z_1(\cdot), \ldots, z_k(\cdot)\right)$ takes values only in cubes of edge length η of which one of $t_1, \ldots, t_p$ is an element. Then, applying the triangle inequality and the above statements:

$$\begin{aligned}
\rho_m(t, t_i) &= \sup_{z \in K_m} \left| z(t) - z(t_i) \right| \\
&\leq \sup_{z \in K_m} \left| z(t) - z_j(t_i) \right| + \left| z_j(t_i) - z(t) \right| \\
&\leq \sup_{z \in K_m} \left| z(t) - z_j(t_i) \right| + \left| z_j(t_i) - z_j(t) \right| + \left| z_j(t) - z(t) \right| \\
&\leq 2 \sup_{z \in K_m} \left\| z - z_j \right\|_T + \left| z_j(t_i) - z_j(t) \right|
\end{aligned}$$

Since this holds for all j, we obtain

$$\rho_m(t, t_i) \leq 2 \sup_{z \in K_m} \inf_j \left\| z - z_j \right\|_T + \sup_j \left| z_j(t) - z_j(t_i) \right|.$$

For any t such that $\left(z_1(t), \ldots, z_k(t)\right)$ falls in the same cube as $\left(z_1(t_i), \ldots, z_k(t_i)\right)$, the first quantity is (strictly) bounded by 2η by the definition of $z_1, \ldots, z_k$ whereas the second quantity is bounded by η because t falls in the same cube as t_i. Now, since, for each $t \in T$, $\left(z_1(t), \ldots, z_k(t)\right) \in T$ must fall in the same cube as $\left(z_1(t_i), \ldots, z_k(t_i)\right)$ for some $i \in \{1, \ldots, p\}$ we have that $t \in \{\tilde{t} : \rho_m(t_i, \tilde{t}) < 3\eta\}$ for some $i \in \{1, \ldots, p\}$. Since η is arbitrary, this shows that (T, ρ_m) is totally bounded.

Now set

$$\rho(s,t) = \sum_{m=1}^{\infty} 2^{-m} \left(\rho_m(s,t) \wedge 1\right).$$

Fix some $\eta > 0$. Take a natural number m with $2^{-m} < \eta$. Cover T with finitely many ρ_m-balls of radius m.[9]. Let $t_1, \ldots, t_p$ be their centers, Since $\rho_1 \leq \rho_2 \leq \ldots$,[10] there is for every t a t_i with

$$\rho(t,t_i) \leq \sum_{k=1}^{m} 2^{-k} \rho_k(t,t_i) + 2^{-m} < 2\eta.^{11}$$

So (T,ρ) is totally bounded as well. It is clear from definitions that $\left|z(s) - z(t)\right| \leq \rho_m(s,t)$ for every $z \in K_m$ and that $\left(\rho_m(s,t) \wedge 1\right) \leq 2^m \rho(s,t)$.[12] Further, if $\|z_0 - z\|_T < \epsilon$ for $z \in K_m$, then $|z_0(s) - z_0(t)| < 2\epsilon + |z(s) - z(t)|$ for any pair s,t. [13] This gives us that

$$K_m^{\epsilon} \subset \left\{ z : \sup_{\rho(s,t)<2^{-m}\epsilon} \left|z(s) - z(t)\right| \leq 3\epsilon \right\}.$$

The system of implications to get this is: if $z \in K_m$ and $\epsilon < 1$ then $\rho(s,t) < 2^{-m}\epsilon \implies \rho_m(s,t) \leq \epsilon \implies \left|z(s) - z(t)\right| \leq \epsilon$. That this holds for all $z \in K_m$ gives that for $z \in K_m^{\epsilon}$, $\rho(s,t) < 2^{-m}\epsilon \implies \left|z(s) - z(t)\right| \leq 3\epsilon$. Taking $\epsilon \leq 1$ is without loss of generality. To finish not that this gives us that, for given ϵ and m and for $\delta < 2^{-m}\epsilon$

$$\liminf \mathbb{P}_\star \left(\sup_{\rho(s,t)<\delta} \left|X_n(s) - X_n(t)\right| < 3\epsilon \right) \geq 1 - \frac{1}{m}.$$

This shows the backwards direction of Theorem 2.6 as well. As a note, this whole argument can be used with nets instead of sequences. □

Remark. Important not to forget the totally bounded part of the theorem. For example, in the example of the empirical CDF case, we need to show that $\mathbb{R}$ is totally bounded. The good news is we have choice of semi-metric.

Remark (Connection to Arzela-Ascoli)**.** Arzela-Ascoli: Let T be a set with metric ρ that is compact. Tet $C(T)$ be the set of all real valued continuous functions on T. Then $A \subset C(T)$ is compact under $|\cdot|_\infty$ if and only if it is equicontinuous and bounded.

We can think of Theorem 2.7 as a stochastic version of this. That is for

$$\liminf \mathbb{P}_\star \left(\sup_{p(s,t)<\delta} \left|X_n(s) - X_n(t)\right| \leq \epsilon \right) \geq 1 - \eta.$$

The set of functions satisfying this condition is equicontinuous. So then, if X_n falls here it is in a compact set by Arzela-Ascoli (Theorem 1.3). Showing this is a focus later.

[9] This is possible because (T,ρ_m) is totally bounded by the above argument
[10] Because $K_1 \subseteq K_2 \subseteq K_3 \ldots$
[11] If t is distance at most η from t_i under ρ_m, it is also distance at most η from t_i under ρ_k for $k \leq m$
[12] In the definition of ρ multiply left side and right side by 2^m. A semimetric is always (weakly) positive.
[13] Same triangle inequality decomposition as above:

$$\begin{aligned} \left|z_0(s) - z_0(t)\right| &\leq \left|z_0(s) - z(s)\right| + \left|z(s) - z_0(t)\right| \\ &\leq \left|z_0(s) - z(z)\right| + \left|z(s) - z(t)\right| + \left|z(t) - z_0(t)\right| \\ &\leq 2\left\|z_0 - z\right\|_T + \left|z(s) - z(t)\right| \end{aligned}$$

3 Empirical Processes

These notes follow Section 2 in VdV&W. So far, we have discussed theory for $X_n \xrightarrow{L} X$ where both X_n and X are random elements in $\ell^\infty(T)$. The classic example that we have kept in mind is convergence of the empirical CDF process, $X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\mathbb{1}\{X_i \leq t\} - \mathbb{P}(X \leq t) \right)$. In this next section we will build on the theory developed to show the convergence of some empirical processes on ℓ^∞.

Definition 3.1 (Empirical Measure)**.** For a random sample $\{X_i\}_{i=1}^n$, the empirical measure $\mathbb{P}_n$ is the measure constructed from the sample (putting mass $1/n$ at each X_i). That is, for any set C:

$$\mathbb{P}_n(C) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \in C\}.$$

We can also write this in terms of the degenerate measures on each X_i:

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

Definition 3.2 (Empirical Process)**.** For a random sample $\{X_i\}_{i=1}^n$ drawn from common distribution P, the empirical process $\mathbb{G}_n$ is the scaled and demeaned measure on X given by:

$$\mathbb{G}_n(C) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\mathbb{1}\{X_i \in C\} - P(X_i \in C) \right).$$

This is often related to the empirical measure in Definition 3.1 by

$$\mathbb{G}_n = \sqrt{n} \left(\mathbb{P}_n - P \right).$$

Or written in terms of the degenerate measures on each X_i:

$$\mathbb{G}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\delta_{X_i} - P \right).$$

Remark (Notation)**.** We will make the following notations to save space later on. For a measure $\mathbb{Q}$ on a space let $\mathbb{Q}f = \mathbb{E}_{\mathbb{Q}}[f(X)]$. E.j: $\mathbb{P}_n f = \mathbb{E}_n[f(X)] = \frac{1}{n} \sum_{i=1}^n f(X_i)$ and $\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(f(X_i) - Pf \right)$.

With this notation:

$$\mathbb{P}_n f \xrightarrow{\text{a.s}} Pf \quad \text{is just saying} \quad \frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{\text{a.s}} \mathbb{E}\left[f(X) \right]$$

$$\mathbb{G}_n f \xrightarrow{L} N(0, \sigma^2) \quad \text{is just saying} \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(f(X_i) - \mathbb{E}\left[f(X) \right] \Big) \xrightarrow{L} N(0, \sigma^2)$$

By LLN and CLT we have that for any function f, $\mathbb{P}_n f \to_{a.s} Pf$ and $\mathbb{G}_n f \xrightarrow{L} N\left(0, P\left(f - Pf\right)^2\right)$

Example 3.1 (Classes of Functions)**.** LLN and CLT establish the behavior of the empirical measure $\mathbb{P}_n f$ and the empirical process $\mathbb{G}_n f$ for a fixed function f (which could even be vector valued). However, we often want to study the behavior of the empirical measure of empirical process over a class of functions $\mathcal{F}$. In this case we can think of $\mathbb{G}_n(\mathcal{F})$ or $\mathbb{P}_n(\mathcal{F})$ as random maps onto $\ell^\infty(\mathcal{F})$. The marginal, $\mathbb{G}_n f$ or $\mathbb{P}_n f$, is then the behavior of the empirical measure/process for a single function $f \in \mathcal{F}$.

Mapping this back to the empirical CDF example of before let $\mathcal{F} = \left\{ f_t : \mathbb{R} \to \mathbb{R} \mid f_t(x) = \mathbb{1}\{x \leq t\}, t \in T \right\}$. Before, we considered convergence of the whole CDF through the map $X_n : \Omega_n \to \ell^\infty(T)$ with the marginals $X_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq t\}$. With these new definitions/notations, we equivalently consider convergence of the entire CDF through the map $\mathbb{P}_n(\mathcal{F}) : \Omega_n \to \ell^\infty(\mathcal{F})$ with marginals $\mathbb{P}_n f_t = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq t\}$.

This sort of notation/generality is useful as we can consider the behavior of the empirical measure or empirical process over a larger class of functions. For example, if we wanted to study an entire semiparametric model we may consider the behavior of $\mathbb{G}_n(\mathcal{F})$ where

$$\mathcal{F} = \left\{ f(x;\theta) \text{ for some } \theta \in \Theta \right\}.$$

Or, if we wanted to consider convergence after imposing some shape restriction, we may take

$$\mathcal{F} = \left\{ f : X \to \mathbb{R} \mid f \text{ is monotonic} \right\}.$$

Remark (Notation)**.** Sometimes we use $\rightsquigarrow$ to denote weak convergence/convergence in law instead of $\xrightarrow{L}$.

Remark (Definition of ℓ^∞ Space)**.** It is useful to review the $\ell^\infty(T)$ space for an arbitrary index space T. Define:

$$\ell^\infty(T) = \left\{ f : T \to \mathbb{R} : \sup_{t \in T} \left| f(t) \right| < \infty \right\} \tag{3.1}$$

and equip this space with the sup-norm, $\|f\|_T = \sup_{t\in T} \left|f(t)\right|$. Note that, for any $\mathcal{F}$, $\mathbb{G}_n(\mathcal{F})$ can be viewed as a random map into $\ell^\infty(\mathcal{F})$ for each n. Boundedness comes from the finiteness of the sample. We will sometimes make the notation $\|\mathbb{Q}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{Q}f|$ for a given measure $\mathbb{Q}$.

Now make some important definitions and then talk about how they relate to what we want to show.

Definition 3.3 (Glivenko-Cantelli Class)**.** A class of functions, $\mathcal{F}$, for which

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \to_p 0 \tag{3.2}$$

is called a Glivenko-Cantelli class, or a P-Glivenko-Cantelli class to emphasize the dependence on the underlying measure P from which the sample is drawn.

Definition 3.4 (Donsker Class)**.** A class of functions, $\mathcal{F}$, for which

$$\mathbb{G}_n(\mathcal{F}) \xrightarrow{L} \mathbb{G}(\mathcal{F}) \tag{3.3}$$

where $\mathbb{G}$ is a tight, Borel measurable element in $\ell^\infty(\mathcal{F})$, is called a Donsker class, or P-Donsker class to emphasize the dependence on the underlying measure P from which the sample is drawn.

A Donsker class is trivially Glivenko-Cantelli.

Example 3.2 (Some Donsker Classes)**.** Some examples of function classes:

1. If $\mathcal{F}$ consists of a single function with finite variance then $\mathcal{F}$ is Donsker by the Central Limit Theorem. That is $\mathbb{G}_n \xrightarrow{L} \mathbb{G}$ where $\mathbb{G}$ is a tight element on $\ell^\infty(\mathcal{F}) = \ell^\infty(\{f\})$
2. The class of functions $\mathcal{F} = \left\{ f(x) = x'\beta : \beta \in \mathcal{B} \right\}$ is Donsker if $\mathcal{B}$ is bounded.
3. The class of monotonic densities on $[0,1]$ is Donsker.
4. The class of square integrable functions is not Donsker (too large).

How do we know if $\mathbb{G}_N \rightsquigarrow \mathbb{G}$ where $\mathbb{G}$ is a tight, Borel measurable element on $\ell^\infty(\mathcal{F})$? By Theorem 2.5 we know that X_n weakly converges if and only if X_n is asymptotically tight and the marginals $\big(X_n(t_1), \ldots, X_n(t_k)\big)$ converge weakly to a limit for every finite subset. Moreover, by Lemma 2.2 asymptotic measurability of the process is equivalent to asymptotic measurability of the marginals. By the Central Limit Theorem, we typically have weak convergence and asymptotic measurability of the marginals, what remains is to show asymptotic tightness.

Theorem 2.7 characterizes asymptotic tightness in terms of ρ-equicontinuity. Much of the work in showing tightness will be to find some semimetric ρ on $\mathcal{F}$ such that for any $\epsilon, \eta > 0$ there is a $\delta > 0$ such that

$$\lim_{n\to\infty} \sup \mathbb{P}^\star \left(\sup_{\rho(f,g)<\delta} \left| \mathbb{G}_n(f) - \mathbb{G}_n(g) \right| > \epsilon \right) < \eta. \tag{3.4}$$

A typical approach will be to let $\mathcal{F}_\delta = \{f, g \in \mathcal{F}, \rho(f,g) < \delta\}$. If we can show that, for some $M(\delta)$ that goes to 0 as $\delta \downarrow 0$

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_\delta}\right] &= \mathbb{E}\left[\sup_{\rho(f,g)<\delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)|\right] \\ &= \mathbb{E}\left[\sup_{\rho(f,g)<\delta} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \{f(X_i) - \mathbb{E}[f(X_i)] - g(X_i) + \mathbb{E}[g(X_i)]\}\right|\right] \\ &\leq M(\delta)\end{aligned}$$

Then, we would get the result in (3.4) by Markov's inequality. This type of result, that $\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_\delta}\right] \leq M(\delta)$ is called a maximal inequality and is immensely useful.

Obtaining such a maximal inequality/establishing asymptotic tightness is dependent on the space not being "too large" (loosely speaking). In the example above, the class $\mathcal{F} = \{f(x) = x'\beta \mid \beta \in \mathcal{B}\}$ is Donsker so long as $\mathcal{B}$ is bounded. To illustrate, see in the single dimensional case that

$$\sup_{b\in\mathcal{B}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i b - \mathbb{E}[xb]\right| = \sup_{b\in\mathcal{B}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i - \mathbb{E}[x]\right| |b|.$$

If we don't impose $|b| \leq M$ then this will blow up to $+\infty$ with probability 1, whereas if we do we have that this is $O_p(1)$. For more involved function classes, we want a way of measuring whether $\mathcal{F}$ is large or not. This motivates the definitions of bracketing and covering numbers below.

Definition 3.5 (Covering Number)**.** The covering number, $\mathcal{N}\left(\epsilon, \mathcal{F}, \|\cdot\|\right)$ of a class of functions $\mathcal{F}$ is the smallest number of balls of radius ϵ under $\|\cdot\|$ needed to cover the set $\mathcal{F}$.

Definition 3.6 (Bracketing Number)**.** Given two functions, ℓ and u, the bracket $[\ell, u]$ is the set of all functions f with $\ell(x) \leq f \leq u(x)$ for all x. An ϵ-bracket is a bracket $[\ell, u]$ with $\|u - \ell\| < \epsilon$. The bracketing number $\mathcal{N}_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|\right)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.

Example 3.3 (Covering Number)**.** Let $A = [0,1]$ and $\|\cdot\|$ be the standard Euclidean norm[1].

1. If $\epsilon \geq 1/2$, then a ball centered at $1/2$ covers the entire interval so $\mathcal{N}\left(\epsilon, A, |\cdot|\right) = 1$.
2. If $\epsilon < 1/2$, then we need $\lceil \frac{1}{2\epsilon} \rceil$ balls to cover A.

Note that (i) in this example the covering number coincides with the bracketing number (ii) in general the balls needed to cover $\mathcal{F}$ need not be centered at points in $\mathcal{F}$ (iii) (in general) as $\epsilon \downarrow 0$ we have that $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \uparrow \infty$.

Example 3.4 (Bracketing Number)**.** Suppose x takes values in $[0,1]$ and let $\mathcal{F} = \{f(x) = x\beta, \text{ for } \beta \in [0,1]\}$. Then, if $\beta_i < \beta_{i+1}$, $[x\beta_i, x\beta_{i+1}]$ forms a bracket containing all functions $f(x) = x\beta$ with $\beta_i \leq \beta \leq \beta_{i+1}$. Further note that

$$\|x\beta_i - x\beta_{i+1}\| = \sup_{x\in[0,1]} |x||\beta_i - \beta_{i+1}| = |\beta_i - \beta_{i+1}|.$$

For any $\epsilon > 0$ break up $[0,1]$ into $[0, \epsilon, 2\epsilon, \dots]$ and take $\beta_i = (i-1)\epsilon$ to get brackets $[x\beta_i, x\beta_{i+1}]$ of size ϵ. We need $\lceil 1/\epsilon \rceil$ of these brackets to cover $\mathcal{F}$ so that $\mathcal{N}_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|_\infty\right) \leq \lceil 1/\epsilon \rceil < 2/\epsilon$.

Remark (Bracketing vs. Covering Numbers)**.** In general we have that $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \mathcal{N}_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|)$, but no opposite relationship. This shows that bracketing numbers are in general stronger than covering numbers and give you better control over the class of functions.

We will see conditions for Glivenko-Cantelli and Donsker properties under both, but in general proving Glivenko-Cantelli involves using bracketing numbers whereas proving Donsker involves using covering numbers.

[1] If we want to view this as a function class we can equivalently say A is the set of constant functions taking values in the interval $[0,1]$ and consider any L_p norm on this class

In general, finding the covering/bracketing number will be difficult but we will learn some tips. Verifying that a set is Donsker will often come down to showing that the covering/bracketing number does not go to infinity "too fast."

3.1 Maximal Inequalities for Finite Classes

For an arbitrary set of functions, $\mathcal{F}$, want to develop an inequality that looks something like:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f(x_i)-\mathbb{E}[f(x)]\right)\right|\right]\leq \text{size}\left(\mathcal{F}\right).$$

Or, rewriting in the notation of above:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{G}_n f\right|\right]\leq \text{size}\left(\mathcal{F}\right).$$

This sort of inequality is useful as it can be used to show the uniform law of large numbers:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\left(\mathbb{P}_n-P\right)f\right|\right]\leq \frac{1}{\sqrt{n}}\text{size}\left(\mathcal{F}\right) \quad + \quad \text{Markov's Inequality.}$$

Or show asymptotic tightness through stochastic equicontinuity:

$$\mathbb{E}\left[\sup_{\rho(f,g)<\delta}\left|\mathbb{G}_n\left(f-g\right)\right|\right]\leq \text{size}\left(\mathcal{F}_\delta\right) \quad + \quad \text{Theorem 2.7.}$$

However, often we may need to change the exact application of these maximal inequalities. We will work out where these come from as we go along. The inequality will be presented for general stochastic processes (for our purposes, a stochastic process is a random map into $\ell^\infty(T)$). To build the maximal inequality, we will need to define a new norm which generalizes the L_p norms. We do so quickly below.

3.1.1 Orlicz Norm

Definition 3.7 (Orlicz Norm)**.** Let ψ be a non-decreasing, convex function with $\psi(0)=0$ and X a random variable. Then, the Orlicz norm $\|X\|_\psi$ is defined as

$$\|X\|_\psi=\inf\left\{C>0:\mathbb{E}\psi\left(\frac{|X|}{C}\right)\leq 1\right\} \tag{3.5}$$

Where here the infimum over the empty set is taken to be $+\infty$.

Remark (Orlicz norms generalize L_p)**.** Note that for any $p\geq 1$ the function $f(x)=x^p$ is convex and non-decreasing. With this in mind we can view the Orlicz norms as a generalization of the L_p norms to general convex and non-decreasing functions functions.

Remark (Orlicz p-norms)**.** Of particular interest will be the Orlicz norms generated by the functions

$$\psi_p=e^{x^p}-1.$$

for $p\geq 1$. The Orlicz norm in this case is often denoted $\|\cdot\|_{\psi_p}$. These norms give more weight to the tails of X than the standard L_p norms. It is not the case that these norms are uniformly larger than all L_p norms, however, we do have the inequalities

$$\|X\|_{\psi_p}\leq \|X\|_{\psi_q}(\log 2)^{p/q}$$
$$\|X\|_p\leq p!\,\|X\|_{\psi_1}$$

Remark (Orlicz Norms and Markov's Inequality)**.** Any Orlicz norm can be used to bound tail probabilities. Using Markov's inequality:

$$\mathbb{P}\left(|X| > x\right) \leq \mathbb{P}\left(\psi\left(|X|/\left\|X\right\|_{\psi}\right) \geq \psi\left(x/\left\|X\right\|_{\psi}\right)\right) \leq \frac{1}{\psi\left(x/\|X\|_{\psi}\right)}.$$

For $\psi_p(x) = e^{x^p} - 1$ this leads to tail estimates like $\exp\left(-Cx^p\right)$ for any random variable with a finite ψ_p-norm. Conversely, an exponential tail bound of this type shows that $\left\|X\right\|_{\psi_p}$ is finite.

Lemma 3.1 (Lemma 2.2.1 VdV&W)**.** *Let X be a random variable with $\mathbb{P}\left(|X| > x\right) \leq Ke^{-Dx^p}$ for every x and some (fixed) constants K and D and for some $p \geq 1$. Then, the Orlicz norm of X satisfies*

$$\left\|X\right\|_{\psi_p} \leq \left((1+K)/D\right)^{1/p}$$

In particular, this will mean that for $C = \left((1+K)/D\right)^{1/p}$

$$\mathbb{E}\left[\psi\left(\frac{|X|}{C}\right)\right] \leq 1.$$

Proof. By Fundamental Theorem of Calculus and Tonelli's Theorem, for any constant B:

$$\mathbb{E}\left[e^{B|X|^p} - 1\right] = \mathbb{E}\int_0^{|X|^p} Be^{Bs}\, ds = \int_0^{\infty} \mathbb{P}\left(|X > s^{1/p}\right) Be^{Bs}\, ds$$

Now use the inequality on the tails of $|X|$, plug in $B = C^{-p} = D/(1+K)$, and see that the final equality is bounded by 1. □

Using the fact that $\max |X_i|^P \leq \sum |X_i|^p$ we obtain for the L_p norms, the result that

$$\left\|\max_{1\leq i\leq m} X_j\right\|_p = \left(\mathbb{E}\max_{1\leq i\leq m}|X_i|^p\right)^{1/p} \leq m^{1/p}\max_{1\leq i\leq m}\left\|X_i\right\|_p.$$

We can generalize this for the Orlicz norm.

Lemma 3.2 (Lemma 2.2.2 VdV&W)**.** *Let ψ be a convex, non-decreasing, nonzero function with $\psi(0) = 0$ and $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ for some constant c. Then, for any random variables $X_1, \ldots, X_m$,*

$$\left\|\max_{1\leq i\leq m} X_i\right\|_{\psi} \leq K\psi^{-1}(m)\max_{1\leq i\leq m}\left\|X_i\right\|_{\psi} \tag{3.6}$$

For a constant K depending only on ψ.

Proof. Without loss of generality, assume that $\psi(x)\psi(y) \leq \psi(cxy)$ for all $x, y \geq 1$ and that $\psi(1) \leq 1/2$.[1] In this case, $\psi(x/y) \leq \psi(cx)/\psi(y)$ for all $x \geq y \geq 1$.[2] Thus, for $y \geq 1$ and any D;

$$\begin{aligned}\max_{1\leq i\leq m}\psi\left(\frac{|X_i|}{Dy}\right) &\leq \max_{1\leq i\leq m}\left[\frac{\psi(c|X_i|/D)}{\psi(y)} + \psi\left(\frac{|X_i|}{Dy}\right)\mathbb{1}\left\{\frac{|X_i|}{Dy} < 1\right\}\right] \\ &\leq \sum_{i=1}^{m}\frac{\psi\left(c|X_i|D\right)}{\psi(y)} + \psi(1)\end{aligned}$$

[1] If this is not the case there are constants $\sigma \leq 1$ and $\tau > 0$ such that $\phi(x) = \sigma\psi(\tau x)$ satisfies these conditions. Apply the inequality to ϕ and note that

$$\left\|X\right\|_{\psi} \leq \left\|X\right\|_{\phi}/(\sigma\tau) \leq \left\|X\right\|_{\psi}/\sigma.$$

[2] $x/y \geq 1$ so $\psi(x/y)\psi(y) \leq \psi\left(c(x/y)y\right)$

Let $D = c \max_{1\le i\le m} \|X_i\|_\psi$, and take expectations to get:

$$\mathbb{E}\psi\left(\frac{\max|X_i|}{Dy}\right) \le \frac{m}{\psi(y)} + \psi(1).$$

When $\psi(1) \le 1/2$ take $y = \psi^{-1}(2m)$. Then:

$$\left\|\max_{1\le i\le m}|X_i|\right\|_\psi \le \psi^{-1}(2m) c \max_{1\le i\le m}\|X_i\|_\psi .$$

By the convexity of ψ and the fact that $\psi(0) = 0$, it follows that $\psi^{-1}(2m) \le 2\psi^{-1}(m)$. This gives the result. □

To review, we have established the following inequalities above:

1. For maximums of a finite number of random variables

$$\mathbb{E}\left[\max_{1\le i\le m}|X_i|\right] \le m \max_{1\le i\le m}\mathbb{E}\left[|X_i|\right].$$

2. Then, generalized this to the L_p norms

$$\left\|\max_{1\le i\le m}|X_i|\right\|_{L_p} \le m^{1/p} \max_{1\le i\le m}\|X_i\|_{L_p} .$$

3. Then, generalized this using the Orlicz norm (Definition 3.7)

$$\left\|\max_{1\le i\le m}|X_i|\right\|_\psi \le K\psi^{-1}(m) \max_{1\le i\le m}\|X_i\|_\psi .$$

   In particular, taking $\psi(a) = e^{a^2} - 1$, we have that $\mathbb{E}\left[\max_{1\le i\le m}|X_i|\right] \le C\sqrt{\log(m+1)}$ for any C such that $\max_{1\le i\le m}\mathbb{E}\left[\psi\left(\frac{|X_i|}{C}\right)\right] \le 1$. Lemma 3.1 gives a condition for the existence of such a C.

3.2 Chaining and Inequalities for Infinite Classes

So far, we have developed inequalities that deal with finite number of random variables. These inequalities are useful for showing Donsker/Glivenko-Cantelli property for finite classes of functions, $|\mathcal{F}| < \infty$, just set $X_i = \mathbb{G}_n f_i$. However, we often want to show uniform convergence for (uncountably) infinite classes of sets, $|\mathcal{F}| = |\mathbb{Q}|$ or $|\mathcal{F}| = |\mathbb{R}|$. To do this, we will use a technique called *chaining*.

Roughly speaking, this will work whenever our class of functions $\mathcal{F}$ is "separable", with respect to the empirical process $\mathbb{G}_n$ (or empirical measure $\mathbb{P}_n$). This means there is a countable subset $\tilde{\mathcal{F}}$ of $\mathcal{F}$ such that $\sup_{\mathcal{F}}|\mathbb{G}_n(f)| = \sup_{\tilde{\mathcal{F}}}|\mathbb{G}_n(f)|$. What does this buy us? If $\tilde{\mathcal{F}}_0 \subset \tilde{\mathcal{F}}_1 \subset \tilde{\mathcal{F}}_2 \cdots \subset \tilde{\mathcal{F}}$ is an infinite sequence of sets whose union is $\tilde{\mathcal{F}}$ and where each $\tilde{\mathcal{F}}_i$ is finite, then:

$$\lim_{k\to\infty}\sup_{\tilde{\mathcal{F}}_k}|\mathbb{G}_n(f)| \overset{a.s}{=} \sup_{\tilde{\mathcal{F}}}|\mathbb{G}_n(f)| \overset{\text{monotone convergence}}{\Longrightarrow} \lim_{k\to\infty}\mathbb{E}\left[\sup_{\tilde{\mathcal{F}}_k}|\mathbb{G}_n(f)|\right] = \mathbb{E}\left[\sup_{\tilde{\mathcal{F}}}|\mathbb{G}_n(f)|\right].$$

and by separability, the last expectation is equal to the expectation of the supremum over the whole class $\mathcal{F}$. To make this work, we want to make sure that we can apply the inequalities that we developed in the past section. Specifically, we want to make sure that the conditions of Lemma 3.1 hold. To do so, make a definition.

Definition 3.8 (Subgaussian Process)**.** Let $\mathbb{G}$ be a stochastic process on a space $\mathcal{F}$ equipped with a metric $d(\cdot,\cdot)$. Then $\mathbb{G}$ is subgaussian if

$$\mathbb{P}\left(\left|\mathbb{G}(f)-\mathbb{G}(g)\right|>x\right)\leq 2e^{-1/2x^2/d^2(f,g)} \tag{3.7}$$

for all $f,g\in F$ and any $x\geq 0$.

Also define a separable function as an analytic concept and then extend this to the case of stochastic processes.

Definition 3.9 (Separable Function)**.** A function $f:A\to B$ from a topological space A into a topological space B is separable if there is a countable, dense, subset $S\subset A$ such that for any closed $F\subset B$ and any open $I\subset A$, if $f(t)\in F$ for all $t\in F\cap S$ then $f(t)\in F$ for all $t\in I$. This is often denoted as an S-separable function to emphasize the dependence on the countable, dense subset S.

Lemma 3.3 (Continuity and Separability)**.** *A continuous function $f:\mathcal{X}\to\mathcal{Y}$ from a separable space $\mathcal{X}$ onto $\mathcal{Y}$ is separable.*

Definition 3.10 (Separable Process; Shalizi 2007)**.** A stochastic process on a topological space $\mathcal{F}$, $\mathbb{G}(\cdot,\omega):\Omega\to\ell^\infty(\mathcal{F})$, is separable if there is a countable, dense, subset of $\mathcal{F}$, $\tilde{\mathcal{F}}$, and a measure zero set N such that for all $\omega\notin N$, $\mathbb{G}(\cdot,\omega)$ is $\tilde{\mathcal{F}}$-separable.[1]

Separability can be roughly interpreted as ensuring that the behavior of the function (and therefore the stochastic process) can be well described by its behavior on countable subset. This ensures some of the properties that we've seen above, namely that $\sup_{f\in\tilde{\mathcal{F}}}|\mathbb{G}(f)|\overset{a.s}{=}\sup_{f\in\mathcal{F}}|\mathbb{G}(f)|$. We are now ready for the main theorem of this subsection, the proof of which will rely on the chaining argument roughly discussed above.

Theorem 3.1 (Theorem 2.2.4 VdV&W)**.** *Let $\mathbb{G}$ be a separable subgaussian process on a space $\mathcal{F}$ equipped with a metric $d(\cdot,\cdot)$ and let* $\operatorname{diam}(\mathcal{F})=\sup_{f,g\in\mathcal{F}}d(f,g)$*. Then*

$$\mathbb{E}\sup_{f,g\in\mathcal{F}}\left|\mathbb{G}(f)-\mathbb{G}(g)\right|\leq K\int_0^{\operatorname{diam}(\mathcal{F})}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d)}\,d\epsilon \tag{3.8}$$

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\mathbb{G}(f)\right|\leq\mathbb{E}\left|\mathbb{G}(f_0)\right|+K\int_0^{\operatorname{diam}(\mathcal{F})}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d)}\,d\epsilon,\ \forall f_0\in\mathcal{F} \tag{3.9}$$

Proof. Proof proceeds in steps. Let $M=\operatorname{diam}(\mathcal{F})=\sup_{f,g\in\mathcal{F}}d(f,g)$. For any $f_0,f\in\mathcal{F}$ we have that $d(f_0,g)\leq M$. First step will be to build a "chain" to almost any point in $\mathcal{F}$. Further, let $\tilde{\mathcal{F}}$ be the dense subset as described in Definitions 3.9 and 3.10.

Step 1: Building a Chain. Pick any $f_0\in\tilde{\mathcal{F}}$ and let $\tilde{\mathcal{F}}_0=\{f_0\}$. Build nesting sets, $\mathcal{F}_0\subset\tilde{\mathcal{F}}_1\subset\tilde{\mathcal{F}}_2\subset\cdots\subset\tilde{\mathcal{F}}$, such that for each $k\in\mathbb{N}$ $\tilde{\mathcal{F}}_k=\{f_1,\ldots,f_{m(k)}\}$ is a maximal collection of points such that $d(f_k,g_k)>\frac{M}{2^K}$ for any $f_k,g_k\in\mathcal{F}_k$. By definition of the packing numbers we know that $\mathcal{N}\left(\frac{M}{2^{k+1}},\mathcal{F},d\right)$ balls cover $\mathcal{F}$. Putting a point at the center of each of these balls creates points that are at least distance $\frac{M}{2^k}$ from each other. Similarly, if we could fit more points at least distance $\frac{M}{2^k}$ distance away from each other than we could pack more balls of radius $\frac{M}{2^{k+1}}$ into $\mathcal{F}$ by centering a ball at each point. So, $|\tilde{\mathcal{F}}_K|\leq\mathcal{N}\left(\frac{M}{2^{k+1}},\mathcal{F},d\right)$ (Inequality comes because each $\tilde{\mathcal{F}}_k$ has to contain all previous sets).

Finally, link each point $f_k\in\tilde{\mathcal{F}}_k$ to a unique point $f_{k-1}\in\tilde{\mathcal{F}}_{k-1}$ such that $d(f_k,f_{k-1})\leq\frac{M}{2^{k-1}}$.[2]

[1] Note that this requires a topology on $\mathcal{F}$. In the applications we will be talking about $\mathcal{F}$ will be equipped with a metric d. This will generate a topology.

[2] I found it helpful to remember here that $\tilde{\mathcal{F}}_{k-1}\subset\tilde{\mathcal{F}}_k$. If no such f_{k-1} exists we could add f_k to $\tilde{\mathcal{F}}_{k-1}$, a contradiction. If $f_k\in\tilde{\mathcal{F}}_{k-1}$ we can link it to itself.

Step 2: Use the chain to build a bound. Using these links, for any $f_k, g_k \in \tilde{\mathcal{F}}_k$ we can build a chain back to f_0:

$$\begin{aligned}\left|\mathbb{G}(f_k) - \mathbb{G}(g_k)\right| &= \left|\left(\mathbb{G}(f_k) - \mathbb{G}(f_0)\right) - \left(\mathbb{G}(g_k) - \mathbb{G}(f_0)\right)\right| \\ &= \left|\sum_{j=0}^{k}\left(\mathbb{G}(f_i) - \mathbb{G}(f_{i-1})\right) - \sum_{j=0}^{k}\left(\mathbb{G}(g_i) - \mathbb{G}(g_{i-1})\right)\right|\end{aligned}$$

By the triangle inequality:

$$\mathbb{E}\left[\max_{g_k, f_k \in \tilde{\mathcal{F}}_k}\left|\mathbb{G}(f_k) - \mathbb{G}(g_k)\right|\right] \leq 2\sum_{j=0}^{K}\mathbb{E}\left[\max_{s_i \in \tilde{\mathcal{F}}_i}\left|\mathbb{G}(s_i) - \mathbb{G}(s_{i-1})\right|\right] \tag{P-1}$$

With this setup, we can use the maximal inequalities developed above, applying them to the finite sets $\tilde{\mathcal{F}}_k$.

Step 3: Try to control the jumps. Recall that there are at most $\mathcal{N}\left(\frac{M}{2^{k+1}}, \mathcal{F}, d\right)$ points in $\mathcal{F}_k$ and that $d(s_k, s_{k-1}) \leq \frac{M}{2^{k-1}}$. By our maximal inequality in Lemma 3.2, taking $\psi(a) = e^{a^2} - 1$ we have that

$$\mathbb{E}\left[\max_{s_j \in \tilde{\mathcal{F}}_j}\left|\mathbb{G}(s_j) - \mathbb{G}(s_{j-1})\right|\right] \leq C_j\sqrt{\log\left(\mathcal{N}\left(\frac{M}{2^{j+1}}, \mathcal{F}, d\right) + 1\right)}.$$

For any constant C_j such that

$$\mathbb{E}\left[\exp\left(\frac{\left(\mathbb{G}(s_j) - \mathbb{G}(s_{j-1})\right)^2}{c_j^2}\right) - 1\right] \leq 1, \quad \forall s_j \in \tilde{\mathcal{F}}_j.$$

Since $\mathbb{G}$ is subgaussian we know that $\mathbb{P}\left(\left|\mathbb{G}(f) - \mathbb{G}(g)\right| > x\right) \leq 2e^{-\frac{1}{2}x^2/d^2(f,g)}$. By construction, we know that $d(s_j, s_{j-1}) \leq \frac{M}{2^{j-1}}, \forall s_j \in \tilde{\mathcal{F}}_j$. So

$$\mathbb{P}\left(\left|\mathbb{G}(s_j) - \mathbb{G}(s_{j-1})\right| > x\right) \leq 2e^{-\frac{1}{2}\frac{x^2}{\lceil M/2^{j-1}\rceil^2}}.$$

By Lemma 3.1 we can take $C_j = \frac{\sqrt{3}M}{2^{j-1}}$ and combine with the other results in this section to get

$$\mathbb{E}\left[\max_{s_j \in \tilde{\mathcal{F}}_j}\left|\mathbb{G}(s_j) - \mathbb{G}(s_{j-1})\right|\right] \leq \frac{\sqrt{3}M}{2^{j-1}}\sqrt{\log\left(\mathcal{N}\left(\frac{M}{2^{j+1}}, \mathcal{F}, d\right) + 1\right)} \tag{P-2}$$

Step 4: Combine Results of Previous Steps. Combine the inequalities from (P-1) and (P-2) to get

$$\mathbb{E}\left[\max_{g_k, f_k \in \tilde{\mathcal{F}}_k}\left|\mathbb{G}(f_g) - \mathbb{G}(g_k)\right|\right] \leq \sqrt{12}M\sum_{j=0}^{k}\frac{1}{2^{j-1}}\sqrt{\log\left(\mathcal{N}\left(\frac{M}{2^{j+1}}, \mathcal{F}, d\right) + 1\right)}$$

With some complex rearranging of squares, we can bound the sum in the display by it's integral up to a constant scale, dropping the added 1 in the log in the process.[3] That is, we ultimately obtain for some constant K:

$$\mathbb{E}\left[\max_{g_k, f_k \in \tilde{\mathcal{F}}_k}\left|\mathbb{G}(f_g) - \mathbb{G}(g_k)\right|\right] \leq K\int_0^M \sqrt{\log\left(\mathcal{N}\left(\epsilon, \mathcal{F}, d\right)\right)}\, d\epsilon \tag{P-3}$$

[3] Here we use the fact that $\log(1+m) \leq 2\log(m)$ for $m \geq 2$

Step 5: Conclude by Separability. $\left\{\tilde{\mathcal{F}}_k\right\}_{k=1}^{\infty}$ is an increasing sequence of sets that approaches $\tilde{\mathcal{F}}$ and note that the bound in (P-3) does not depend on (little) k. So, invoking monotone convergence and separability of $\mathbb{G}$:

$$\begin{aligned}\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\mathbb{G}(f)-\mathbb{G}(g)\right|\right] &= \mathbb{E}\left[\sup_{f,g\in\tilde{\mathcal{F}}}\left|\mathbb{G}(f)-\mathbb{G}(g)\right|\right] \\ &= \lim_{k\to\infty}\mathbb{E}\left[\max_{f,g\in\tilde{\mathcal{F}}}\left|\mathbb{G}(f)-\mathbb{G}(g)\right|\right] \\ &\leq K\int_0^M \sqrt{\log\mathcal{N}\left(\epsilon,\mathcal{F},d\right)}\,d\epsilon\end{aligned}$$

This is the inequality in equation (3.8). To get equation (3.9) fix any f_0 and apply triangle inequality. □

Remark (Comments on Theorem 3.1)**.** Theorem 3.1 is an involved result. Some remarks below.

1. We have shown that if $\mathbb{G}_n$ is a separable subgaussian process then

$$\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\mathbb{G}_n(f)-\mathbb{G}_n(g)\right|\right] \leq K\int_0^{\operatorname{diam}(\mathcal{F})}\sqrt{\log\mathcal{N}\left(\epsilon,\mathcal{F},d\right)}\,d\epsilon.$$

   Note that the right hand side does not depend on $\mathbb{G}_n$ at all! Only on the "size" of $\mathcal{F}$.

2. So, suppose we want to show that $\mathbb{G}_n$ is an asymptotically tight process on $\mathcal{F}$. By Theorem 2.7 it is sufficient (and necessary) to show that for every $\epsilon,\eta>0$ there is a $\delta>0$ such that:

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

   Let $\mathcal{F}_\delta=\left\{s=f-g:f,g\in\mathcal{F}\text{ and }\rho(f,g)\leq\delta\right\}$. The above can be restated as showing that $\exists\delta>0$ such that:

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{s\in\mathcal{F}_\delta}\left|\mathbb{G}_n(s)\right|>\epsilon\right)<\eta.$$

   By Markov's inequality we can bound the probability in the above display by

$$\frac{1}{\epsilon}\mathbb{E}\left[\sup_{s\in\mathcal{F}_\delta}\left|\mathbb{G}_n(s)\right|\right] \leq \frac{K}{\epsilon}\int_0^{\delta}\sqrt{\log\mathcal{N}\left(\epsilon,\mathcal{F},d\right)}\,d\epsilon$$

   And then we can sent the RHS to 0 by sending $\delta\downarrow 0$ as long as the integral on the RHS is finite. Asymptotic tightness plus convergence of marginals will give convergence to a tight element in $\ell^\infty(\mathcal{F})$ by Theorem 2.5. What remains is to show the conditions on $\mathbb{G}$, separability and subgaussian.

3.3 Symmetrization

Symmetrization is a technique that will allow us to get/show(?) a subgaussian process. This follows the discussion in Chapter 2.2.1 and 2.3 in VanDerVaart and Wellner.

What sort of variables are subgaussian? A classic example below.

Definition 3.11 (Rademachar Random Variable)**.** Random variable $\epsilon_i:\Omega_i\to\mathbb{R}$ is a Rademachar random variable if $\mathbb{P}(\epsilon_i=1)=\mathbb{P}(\epsilon_i=-1)=1/2$.

The following lemma shows that a particular process consisting of Rademachar random variables is subgaussian.

Lemma 3.4 (Hoeffding's Inequality)**.** *Let $a_1, \ldots, a_n$ be constants and $\epsilon_1, \ldots, \epsilon_n$ be independent Rademachar random variables. Then*

$$\mathbb{P}\left(\left| \sum_{i=1}^n a_i \epsilon_i \right| > x \right) \leq 2e^{\frac{1}{2}\frac{x^2}{\|a\|^2}}.$$

where $\|a\|$ denotes the Euclidean norm of a.

Proof. (From VdV&W, Lemma 2.2.7) For any λ and any Rademachar random variable ϵ one has that $\mathbb{E}e^{\lambda\epsilon} = \left(e^{\lambda} + e^{-\lambda}\right)/2$. By power series expansion:

$$\begin{aligned}
e^{\lambda} &= 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \\
e^{-\lambda} &= 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \ldots \\
\implies \left(e^{\lambda} + e^{-\lambda}\right)/2 &= 1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} \ldots \\
&\leq 1 + \frac{\lambda^2}{2} + \frac{\lambda^4}{2^2 \cdot 2!} + \frac{\lambda^6}{2^3 \cdot 3!} \\
&= e^{\lambda^2/2}
\end{aligned}$$

where in the last inequality we use that $2^k \cdot k! \leq (2k)!$ so that in total we have that $\mathbb{E}e^{\lambda\epsilon} = \left(e^{\lambda} + e^{-\lambda}\right)/2 \leq e^{\lambda^2/2}$. Take $\lambda = x/\|a\|^2$ and apply Markov's inequality to get the result. □

Example. For any functions f, g we have that

$$\mathbb{P}\left(\left| \sum_{i=1}^n \frac{\epsilon_i}{\sqrt{n}} \left(f(x_i) - g(x_i)\right) \right| > x \mid \{X_i\} \right) \leq 2e^{-\frac{1}{2}\frac{x^2}{d_n^2(f,g)}}.$$

where $d_n^2(f,g) := \frac{1}{N}\sum_{i=1}^n \left(f(x_i) - g(x_i)\right)^2$ is the square of the prediction norm.

We would like to use the maximal inequality in Theorem 3.1 to control $\mathbb{E}[\sup_{f,g} |\mathbb{G}_n(f) - \mathbb{G}_n(g)|]$, but the problem is that $\mathbb{G}_n$ is not (in general), subgaussian. However, from Lemma 3.4 we know that, at least conditional on our data, $\mathbb{G}_n^\circ := \frac{1}{\sqrt{n}}\sum \epsilon_i (f(x_i) - g(x_i))$ is. Strategy will be to relate the two processes, $\mathbb{G}_n$ and $\mathbb{G}_n^\circ$.

Before starting, it is useful to formally define the probability space that we are working with. Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d Rademachar random variables that are generated independent of $(X_1, \ldots, X_n)$, our observed data. Define the symmetrized process:

$$\mathbb{P}_n^\circ f = \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i).$$

Because $\mathbb{P}_n^\circ$ is subgaussian, conditional on $X_1, \ldots, X_n$, it can be easier to study. We want to bound supremum of the process $\mathbb{P}_n - P$ by that of the symmetrized process. To formalize these bounds, we have to be careful about the non-measurability of supremum like $\|\mathbb{P}_n - P\|_{\mathcal{F}}$.[1]

In the following discussion, outer expectations of functions of $X_1, \ldots, X_n$ are assumed to be taken with respect to the coordinate projection of the infinite product space $(\mathcal{X}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, P^{\mathbb{N}})$ onto its first n coordinates, $(\mathcal{X}^n, \mathcal{A}^n, P^n)$.[2]. When auxiliary variables, independent of the X's are involved, as in the next lemma, we can use a similar convention. The underlying probability space is assumed to be of the form $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times$

[1] Even if $\mathcal{F}$ is a class of measurable functions, the supremum may not be measurable.

[2] That is the outer expectation is taken relative to P^n where P^n is defined from the projection of the infinite product space onto its first n coordinates

$(\mathcal{Z}, \mathcal{C}, Q)$. Independence is understood in terms of a product probability space.[3]. To manage all this, we take advantage of a modified Fubini's theorem for outer expectations, stated here without proof.

Lemma 3.5 (Fubini's Theorem, Lemma 1.2.6 VdV&W)**.** *Let T be defined on a product probability space. Then*

$$\mathbb{E}_\star T \leq \mathbb{E}_{1\star}\mathbb{E}_{2\star}T \leq \mathbb{E}_1^\star\mathbb{E}_2^\star T \leq \mathbb{E}^\star T.$$

Proof. For the last inequality, we can assume that $\mathbb{E}^\star T < \infty$ so that $\mathbb{E}^\star T = \mathbb{E}T^\star$. Since $T^\star$ is jointly measurable with respect to the product σ-field, the map $\omega_2 \mapsto T^\star(\omega_1, \omega_2)$ is a measurable majorant of $\omega_2 \mapsto T(\omega_1, \omega_2)$ for P_1 almost all ω_1. Hence $\int T^\star(\omega_1, \omega_2) dP_2(\omega_2) \geq (\mathbb{E}_2^\star T)(\omega_1)$ for P_1 almost all ω_1. Further, by Fubini's theorem for standard integrals, this is a measurable function of ω_1. Thus the integral of this with respect to P_1 is an upper bound for $\mathbb{E}_1^\star\mathbb{E}_2^\star T$. Since $T^\star$ is jointly measurable, by another application Fubini's theorem for standard integrals:

$$\mathbb{E}^\star T = \mathbb{E}T^\star = \int \left(\int T^\star(\omega_1, \omega_2) dP_2(\omega_2) \right) dP_1(\omega_1) \geq \mathbb{E}_1^\star\mathbb{E}_2^\star T.$$

The inequalities for inner expectations hold by considering $-T$. □

Lemma 3.6 (Symmetrization)**.** *For every non-decreasing, convex, $\Phi : \mathbb{R} \to \mathbb{R}$ and class of measurable functions $\mathcal{F}$:*

$$\mathbb{E}^\star \Phi\left(\|\mathbb{P}_n - P\|_\mathcal{F} \right) \leq \mathbb{E}^\star \Phi\left(2 \|\mathbb{P}_n^\circ\|_\mathcal{F} \right).$$

Where outer expectations are calculated as described above.

Proof. Let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$ (independently drawn from the same joint distribution as $X_1, \ldots, X_n$, defined formally as the coordinate projections on the last n coordinates in the product space $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{C}, Q) \times (\mathcal{X}^n, \mathcal{A}^n, P^n)$).

For fixed values $X_1, \ldots, X_n$ applying Jensen's inequality to the absolute value gives:

$$\|\mathbb{P}_n - P\|_\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \left[f(X_i) - \mathbb{E}f(Y_i) \right] \right| \leq \mathbb{E}_Y^\star \sup_{f\in\mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \left[f(X_i) - f(Y_i) \right] \right|.$$

where $\mathbb{E}_Y^\star$ is the outer expectation with respect to $Y_1, \ldots, Y_n$ computed for P^n. Again applying Jensen's inequality gives:

$$\Phi\left(\|\mathbb{P}_n - P\|_\mathcal{F} \right) \leq \mathbb{E}_Y \Phi\left(\left\| \frac{1}{n} \sum_{i=1}^n \left[f(X_i) - f(Y_i) \right] \right\|_\mathcal{F}^{\star Y} \right).$$

where $f^{\star Y}$ is the minimal measurable majorant of f with respect to the distribution of Y. Because Φ is non-decreasing and continuous, the $\star Y$ inside Φ can be moved to $\mathbb{E}_Y^\star$. In total then:

$$\Phi\left(\|\mathbb{P}_n - P\|_\mathcal{F} \right) \leq \mathbb{E}_Y^\star \Phi\left(\left\| \frac{1}{n} \sum_{i=1}^n \left[f(X_i) - f(Y_i) \right] \right\|_\mathcal{F} \right).$$

Next, take the expectation with respect to $X_1, \ldots, X_n$ of the above quantity to get:

$$\mathbb{E}^\star \Phi\left(\|\mathbb{P}_n - P\|_\mathcal{F} \right) \leq \mathbb{E}_X^\star \mathbb{E}_Y^\star \Phi\left(\frac{1}{n} \left\| \sum_{i=1}^n \left[f(X_i) - f(Y_i) \right] \right\| \right).$$

[3]Two sub-sigma algebras, $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{A}$ are considered independent if $\mathbb{P}(A_1 A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ for any $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$. The sigma algebra generated by a random map $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{X}, \mathcal{B})$ is the smallest sigma algebra on Ω that makes X measurable,

$$\sigma(X) := \{X^{-1}(B) : B \in \mathcal{B}\}.$$

Two random variables, X, Y, defined on the same probability space are independent if their generated sigma algebras, $\sigma(X), \sigma(Y)$, are independent. In the context of having independent draws $X_1, \ldots, X_n$ we can think of this as the projection mappings $\pi_i(\mathcal{X}^n)$ being independent.

Adding a minus sign in front of the term $\big[f(X_i) - f(Y_i)\big]$ has the effect of exchanging X_i and Y_i. By construction of the underlying probability space this does not change the expectation. Hence, the expression

$$\mathbb{E}^\star \Phi\left(\frac{1}{n}\left\|\sum_{i=1}^n e_i\big[f(X_i) - f(Y_i)\big]\right\|\right).$$

is the same for any n-tuple $(e_1, \ldots, e_n) \in \{-1, 1\}^n$. So:

$$\mathbb{E}^\star \Phi\left(\left\|\mathbb{P}_n - P\right\|_{\mathcal{F}}\right) \leq \mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \Phi\left(\left\|\sum_{i=1}^n \epsilon_i\big[f(X_i) - f(Y_i)\big]\right\|_{\mathcal{F}}\right).$$

where each ϵ_i is an independent Rademachar random variable and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. By triangle inequality and convexity of the Φ:

$$\begin{aligned}
&\mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \Phi\left(\left\|\sum_{i=1}^n \epsilon_i\big[f(X_i) - f(Y_i)\big]\right\|_{\mathcal{F}}\right) \\
&\quad\leq \mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \Phi\left(\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}} + \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right\|_{\mathcal{F}}\right) \\
&\quad\leq \frac{1}{2}\mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \Phi\left(2\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|\right) + \frac{1}{2}\mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \Phi\left(2\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right\|\right) \\
&\quad\leq \mathbb{E}^\star \Phi\left(2\left\|\mathbb{P}_n^\circ\right\|_{\mathcal{F}}\right)
\end{aligned}$$

where we use the face that a repeated outer expectation can be bounded above by a joint outer expectation, $\mathbb{E}_\epsilon \mathbb{E}^\star_{X,Y} \leq \mathbb{E}^\star_{\epsilon,X,Y} (= \mathbb{E}^\star)$ using Lemma 3.5. □

Corollary 3.1 (Symmetrization of Empirical Process, Andres' Notes)**.** *For real valued processes as described above:*

$$\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i) - Pf(X_i)\right|\right] \leq 2\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right].$$

Proof. Take $\Phi(x) = |x|$. All norms on $\mathbb{R}$ are equivalent to $|\cdot|$. Lemma 3.6 then gives us that $\mathbb{E}^\star \left\|\mathbb{P}_n - P\right\|_{\mathcal{F}} \leq 2\mathbb{E}^\star \left\|\mathbb{P}_n^\circ\right\|_{\mathcal{F}}$. Expand this out and scale by $\sqrt{n}$ to get the result. For non real-valued processes (vector valued, function valued, etc.) we can replace $|\cdot|$ with $\|\cdot\|$ above. □

Remark. The proof of Corollary 3.1 uses the fact that Lemma 3.6 is not an asymptotic bound, it holds in every finite sample.

We know have the pieces to show a class of functions $\mathcal{F}$ is either

- Glivenko-Cantelli, i.e that $\left\|\mathbb{P}_n - P\right\|_{\mathcal{F}} = o_p(1)$. We will do this by placing conditions on the bracketing/covering numbers.
- Donsker, i.e that $\mathbb{G}_n(\mathcal{F}) \xrightarrow{L} \mathbb{G}(\mathcal{F})$ for some tight $\mathbb{G}$. To do so, we will use covering numbers. The system of arguments needed to show this is usually as follows:
  - By Theorem 2.5 weak convergence to a tight limit is equivalent to asymptotic tightness and weak convergence of the marginals.
  - Weak convergence of the marginals is generally provided by CLT.Theorem 2.7 shows that asymptotic tightness is equivalent to uniform ρ-equicontinuity (Definition 1.18)

- Asymptotic equicontinuity holds if $\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}|\mathbb{G}_n(f)|\right]$ goes to 0 as $\delta \downarrow 0$. Theorem 3.1 gives conditions where this is possible for separable, subgaussian processes.
- Lemma 3.3 suggests separability if $\mathcal{F}$ is separable. Lemma 3.4 gives us that the Rademachar process is subgaussian conditional on $X_1, \ldots, X_n$. Combining with Theorem 3.1 gives

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \leq \int_0^{\text{diam}(\mathcal{F}_\delta)} \sqrt{\log \mathcal{N}\left(s, \mathcal{F}_\delta, L_2(\mathbb{P}_n)\right)}\, ds$$

  where $L_2(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$ is the L_2 norm with respect to the empirical measure. Since X is random this norm will also end up random. This seems like it will make dealing with

$$\mathbb{E}\left[\int_0^{\text{diam}(\mathcal{F}_\delta)} \sqrt{\log \mathcal{N}\left(s, \mathcal{F}_\delta, L_2(\mathbb{P}_n)\right)}\, ds\right]$$

  painful, but we end up having good bounds for this.
- Lemma 3.6, and in particular Corollary 3.1, relates the empirical process to the Rademachar process. Take expectations with respect to X in the above bound to bound the and apply the symmetrization lemma to get bounds on the empirical process of interest.

We next move to verifying the various conditions and applying them to show that some specific processes are Glivenko-Cantelli or Donsker.

3.4 Glivenko-Cantelli

This subsection follows Section 2.4 in Van DerVaart and Wellner. Goal is to establish conditions for a uniform law of large numbers using bracketing and covering numbers.

Theorem 3.2 (Bracketing Glivenko-Cantelli Theorem)**.** *Let $\mathcal{F}$ be a class of measurable functions such that*

$$\mathcal{N}_{[]}\left(\epsilon, \mathcal{F}, L_1(P)\right) < \infty$$

for every $\epsilon > 0$. Then $\mathcal{F}$ is Glivenko-Cantelli.

Proof. Fix $\epsilon > 0$. Choose finitely many ϵ-brackets $[l_i, u_i]$ whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) < \epsilon$ for every i. Then, for every $f \in \mathcal{F}$ there is a bracket, $l_i \leq f \leq u_i$, such that:

$$\begin{aligned}(\mathbb{P}_n - P) f \leq \mathbb{P}_n u_i - P f \leq (\mathbb{P}_n - P) u_i + P(u_i - f) \leq (\mathbb{P}_n - P) u_i + \epsilon \\ (\mathbb{P}_n - P) f \geq \mathbb{P}_n l_i - P f \geq (\mathbb{P}_n - P) l_i + P(l_i - f) \geq (\mathbb{P}_n - P) l_i - \epsilon\end{aligned}$$

Consequently,

$$\begin{aligned}\sup_{f\in\mathcal{F}} (\mathbb{P}_n - P) f \leq \max_i (\mathbb{P}_n - P) u_i + \epsilon \\ \inf_{f\in\mathcal{F}} (\mathbb{P}_n - P) f \geq \min_i (\mathbb{P}_n - P) l_i - \epsilon\end{aligned}$$

By the strong law of large numbers, both the maximums and the minimums on the right hand side of the inequalities above converge almost surely to 0. Combination these yields that $\limsup \|\mathbb{P}_n - P\|_{\mathcal{F}}^\star \leq \epsilon$ almost surely for every $\epsilon > 0$. Take $\epsilon \downarrow 0$ to see that the $\limsup$ must be 0 almost surely. □

Remark. Some comments on Theorem 3.2:

1. Proof is really quite straightforward. Bracketing gives pointwise control so just use the upper and lower bounds.
2. No measurability condition is needed and no requirements on the rate of growth of $\mathcal{N}_{[]}(\epsilon, \cdot, \cdot)$ as $\epsilon \downarrow 0$.

Example (Empirical CDF is Glivenko-Cantelli)**.** Let X be a scalar random variable.[1]. We want to show that

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{X_i\le t\}-P(X_i\le t)\right|=o_p(1).$$

Let $\mathcal{F}=\left\{f(x)=\mathbb{1}\{X_i\le t\}:t\in\mathbb{R}\right\}$. Partition $\mathbb{R}$ into grids $-\infty=t_0<t_1<\cdots<t_m=\infty$ such that $\mathbb{P}\left(t_i\le X\le t_{i+1}\right)<\epsilon$ for each i. Then the finitely many brackets $\left[\mathbb{1}\{X_i\le t_i\},\mathbb{1}\{X_i\le t_{i+1}\}\right]$ cover $\mathcal{F}$ and are "size" ϵ under P. So, $\mathcal{N}_{[]}\left(\epsilon,\mathcal{F},L_1(P)\right)<\infty$ for every $\epsilon>0$. So $\mathcal{F}$ is Glivenko-Cantelli (i.e, we have a uniform law of large numbers).

The requirement on the bracketing numbers can in general be hard. Would like a result for the covering numbers as well. This will make showing that some classes are Glivenko-Cantelli easier later on. Before doing so, we need to make a couple definitions:

Definition 3.12 (Envelope)**.** A class $\mathcal{F}$ has envelope F if $|f(x)|\le F(x)$ for all x and all $f\in\mathcal{F}$.

Definition 3.13 (Truncated Class)**.** Let $\mathcal{F}$ be a class of functions. Then the truncated class $\mathcal{F}_M$ is given

$$\mathcal{F}_M=\left\{f(x)\mathbb{1}\{f\le M\}:f\in\mathcal{F}\right\}.$$

Definition 3.14 (P-Measurable Class)**.** A class $\mathcal{F}$ is P-measurable if $\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(x_i)\epsilon_i\right|$ is measurable with respect to the product measure on $(\mathcal{X}^m,\mathcal{A}^m,P^n)\times(\mathcal{Z},\mathcal{C},Q)$, where $(\mathcal{Z},\mathcal{C},Q)$ denotes the probability space that the Rademachar random variables are defined on.

Definition 3.15 ($L_p(\mathbb{P}_n)$-norm)**.** We have that $\|f-g\|_{L_1(P)}=\mathbb{E}_P\left[\left|f(x)-g(x)\right|^p\right]^{1/p}$, similarly we can define

$$\|f-g\|_{L_p(\mathbb{P}_n)}=\mathbb{E}_{\mathbb{P}_n}\left[\left|f(x)-g(x)\right|^p\right]^{1/p}.$$

and through this define $\mathcal{N}_{[]}\left(\epsilon,\mathcal{F},L_p(\mathbb{P}_n)\right)$.

Theorem 3.3 (Covering Glivenko-Cantelli Theorem)**.** *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope F such that $\mathbb{P}^\star F<\infty$. If $\log\mathcal{N}\left(\epsilon,\mathcal{F}_M,L_1\left(\mathbb{P}_n\right)\right)=o_{P^\star}(n)$ for every ϵ and $M>0$, then $\|\mathbb{P}_n-P\|_{\mathcal{F}}^{\star}\to 0$ almost surely and in mean.*

Proof. Idea will be to apply the maximal inequality in Theorem 3.1.

Step 1: Symmterization. First, we will apply symmetrization (Corollary 3.1)

$$\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-Pf(X_i)\right|\right]\le 2\cdot\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right]$$

And then truncate the functions, $f=f\mathbb{1}\{f\le M\}+f\mathbb{1}\{f>M\}$, apply triangle inequality, and bound the functions not in $\mathcal{F}_M$ with the envelope F.

$$\begin{aligned}&\le 2\mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right]+2\mathbb{E}^\star\left[\epsilon_i F(X_i)\mathbb{1}\{F\ge M\}\right]\\&=2\mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right]+2P^\star F(X_i)\mathbb{1}\{F\ge M\}\end{aligned}$$

Note the argument that allows us to replace the first outer expectation with iterated expectations over X and ϵ: each of the functions in $\mathcal{F}$ are measurable and $\mathcal{F}_M$ is uniformly bounded, which means that the supremum will be measurable and bounded with probability 1 in any finite sample (with respect to the empirical measure/conditional on the X data).

Since $P^\star F<\infty$ we can choose M so that the term on the right is arbitrarily small. [2] That is, for any $\delta>0$

[1] This generalizes easily for a vector valued random variable

[2] I am sort of using $P^\star$ and $\mathbb{E}_X^\star$ interchangeably here, which I apologize for

we can pick M such that

$$\mathbb{E}^\star\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - Pf(X_i)\right|\right] \leq \mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] + \delta.$$

Step 2: Deal with the term that is conditional on $\{X_i\}$. Let $\mathcal{G}_\delta = \{g_1, \ldots, g_{K(\delta)}\}$ be such that for every $f \in \mathcal{F}_M$ there is a $g \in \mathcal{G}_\delta$ such that $\|f - g\|_{L_1(\mathbb{P}_n)} < \delta$. Since $\log \mathcal{N}\left(\delta, \mathcal{F}_m, L_1(\mathbb{P}_n)\right) = o_p(n)$, we know that it is possible to pick a $\mathcal{G}_\delta$ in this fashion with probability approaching 1. Note that:

- Cardinality of $\mathcal{G}_\delta$: $|\mathcal{G}_\delta| = \mathcal{N}\left(\delta, \mathcal{F}_M, \|\cdot\|_{L_1(\mathbb{P}_n)}\right)$.
- Envelope of $\mathcal{G}_\delta$: by construction $\mathcal{F}_M \leq M$ so we can assume that $\mathcal{G}_\delta \leq M$.

Then, for all $f \in \mathcal{F}_M$ we have that, for some $g \in \mathcal{G}_\delta$:

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i) - g(X_i)\right)\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right| + \delta\end{aligned}$$

This gives us that

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \leq \mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right|\right] + \delta.$$

Step 3: Apply the Maximal Inequality. We bound the first term in the last display using the maximal inequality in Lemma 3.2, for the particular case of the ψ_2 Orlicz norm [3]: if $D = \{f_1, \ldots, f_M\}$ then

$$\mathbb{E}\left[\sup_{f\in D}\left|f(X_i)\right|\right] \leq C\sqrt{1 + \log m}$$

for any C with $\mathbb{E}\left[\exp\left(\frac{f(X_i)}{C^2}\right) - 1\right] \leq 1$ for all $f \in D$. In our setting we will apply this to the functions $\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)$ for $g \in \mathcal{G}_\delta$, with the $g(X_i)$ treated as fixed so that these are considered random variables in ϵ_i. In our setting we can bound:

$$\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right|\right] \leq C_\delta\sqrt{1 + \log \mathcal{N}\left(\delta, \mathcal{F}_M, \|\cdot\|_{L_1(\mathbb{P}_n)}\right)}.$$

for such a C_δ such that, for all $g \in \mathcal{G}_\delta$

$$\mathbb{E}_\epsilon\left[\exp\left(\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right)^2 / C_\delta^2\right) - 1\right] \leq 1.$$

Hoeffding's inequality (Lemma 3.4) gives us that, for a general Rademachar process:

$$\mathbb{P}_\epsilon\left(\left|\sum_{i=1}^n \epsilon_i a_i\right| > x\right) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{\|a\|^2}\right)$$

[3] We know that, in any finite sample, this Orlicz norm exists because our empirical expectation is bounded.

Where the norm above is the standard Euclidean norm. In our setting:

$$\mathbb{P}_\epsilon\left(\left|\sum_{i=1}^n \epsilon_i \frac{g(X_i)}{n}\right| > x\right) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{n^{-2}\sum g(X_i)^2}\right)$$

As discussed above, we can uniformly bound $\mathcal{G}_\delta$ by M. The exponential is negative so it is decreasing in the numerator and increasing in the denominator. This allows us to get:

$$\mathbb{P}_\epsilon\left(\left|\sum_{i=1}^n \epsilon_i \frac{g(X_i)}{n}\right| > x\right) \leq 2\exp\left(-\frac{nx^2}{2M^2}\right)$$

Now apply Lemma 3.1. If $\mathbb{P}\left(|X| > x\right) \leq Ke^{-Dx^2}$ then the ψ_2-Orlicz norm of X is less than $\sqrt{(1+K)/D}$. Using the above take $K = 2$ and $D = n/(2M^2)$ to get $C_\delta = \sqrt{6M^2/n}$. Putting this together with the original application of the maximal inequality towards the top of Step 3, we get:

$$\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right|\right] \leq \sqrt{6M^2/n + \log\mathcal{N}\left(\delta, \mathcal{F}_M, \|\cdot\|_{L_1(\mathbb{P}_n)}\right)/n}$$

By assumption $\log\mathcal{N}(\cdot)/n = o_P(1)$ and the first term under the square root is $o(1)$ so that this whole thing is $o_p(1)$. All together, combining this with the result of Step 2, we have shown that, for any $\delta > 0$:

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \leq \delta + o_p(1).$$

So that the whole thing ($\mathbb{E}_\epsilon[\sup_{f\in\mathcal{F}_M}\dots]$) is $o_p(1)$.

Step 4: Put all the parts together.

By the symmetrization at the top of step 1, we have that, for every $\delta > 0$

$$\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \leq 2\mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] + \delta.$$

In Step 3, we showed that the inner expectation is $o_p(1)$. Combining this with the fact that $\mathcal{F}_M \leq M$ gives us that[4]:

$$\mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] = o(1).$$

which we can combine with Markov's inequality to get that

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left\{f(X_i) - Pf(X_i)\right\}\right| = o_p(1).$$

This shows that $\|\mathbb{P}_n - P\|_{\mathcal{F}}^\star \to 0$ in mean. From VdV&W: That it also converges almost surely follows from the fact that the sequence $\|\mathbb{P}_n - P\|_{\mathcal{F}}^\star$ is a reverse martingale with respect to a suitable filtration.[5] □

Remark. A couple comments on Theorem 3.3:

- Proof is harder than that using bracketing numbers (Theorem 3.2). However, the technique is much closer to what will be used for the Donsker Theorems.
- Note how the measurability is obtained using the ϵ_i and conditioning on $\{X_i\}$.
- The conditions look cryptic, but we will find ways of verifying them.

[4] Convergence in probability to 0 implies convergence in distribution to 0 implies convergence in bounded moments

[5] This part may depend on i.i.d. I am not familiar with the martingale convergence theorems.

3.5 Donsker Theorems

This subsection follows section 2.5 in Van DerVaart and Wellner. In this subsection we will establish conditions for $\mathcal{F}$ to be Donsker (Definition 3.4). We will present two main results, one that relies on the covering numbers (through a Uniform Entropy condition) and another using the bracketing numbers.

Definition 3.16 (Uniform Entropy Condition)**.** Let $\mathcal{F}$ be a class of functions with envelope F and let $\mathcal{Q}$ be the set of all finitely discrete probability measures on $(\mathcal{X}, \mathcal{A})$. We say that $\mathcal{F}$ satisfies a uniform entropy bound if:

$$\int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log\left(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)}\, d\epsilon < \infty \tag{UEB}$$

Similarly to before define

$$\mathcal{F}_\delta = \left\{f - g : f, g \in \mathcal{F} \text{ and } \|f - g\|_{P,2} < \delta\right\} \tag{3.10}$$

$$\mathcal{F}_\infty^2 = \left\{(f - g)^2 : f, g \in \mathcal{F}\right\} \tag{3.11}$$

Theorem 3.4 (Covering Donsker Theorem)**.** *Let $\mathcal{F}$ be a class of measurable functions with envelope F that satisfies the uniform entropy bound,* (UEB)*. Let the classes $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ also be P-measurable for every $\delta > 0$. If $P^\star F^2 < \infty$ then $\mathcal{F}$ is P-Donsker.*

Proof. Because the envelope F has a bounded second moment we can apply CLT to get convergence of the marginals for any finite collection $f_1, \ldots, f_k$. That is

$$(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_n) \xrightarrow{L} (\mathbb{G} f_1, \ldots, \mathbb{G} f_k)\,.$$

for some tight limit $\mathbb{G}$. By Theorem 2.5 it is now sufficient (and necessary) to show that $\mathbb{G}_n(\mathcal{F})$ is asymptotically tight. We do this by way of Theorem 2.7, showing that $\mathbb{G}_n$ is asymptotically ρ-equicontinuous and $\mathcal{F}$ it totally bounded for some ρ.[1] That is, we want to show that there is some ρ semimetric with

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

Goal will be to show this for $\rho(f, g) = \left(P(f - g)^2\right)^{1/2}$.

Step 1: Use Symmetrization to apply Maximal Inequality. Apply Markov's inequality and then Lemma 3.6 (or Corollary 3.1) to the class $\sqrt{n}\mathcal{F}_\delta$ to get:

$$\begin{aligned}\mathbb{P}\left(\sup_{\rho(f,g)<\delta} |\mathbb{G}_n f - \mathbb{G}_n g| > \epsilon\right) &\leq \frac{1}{\epsilon} \cdot \mathbb{E}\left[\sup_{\rho(f,g)<\delta} |\mathbb{G}_n f - \mathbb{G}_n g|\right] \\ &\leq \frac{2}{\epsilon} \cdot \mathbb{E}_X \mathbb{E}_\epsilon \left[\sup_{\rho(f,g)<\delta} \left|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \left(f(x_i) - g(x_i)\right)\right|\right] \end{aligned} \tag{D-0}$$

Note that the inside is measurable, so we can use iterated expectations. Recall by the maximal inequality (Theorem 3.1) if D is a set equipped with metric d and $\mathbb{G}$ is a subgaussian process, then

$$\mathbb{E}\left[\sup_{f,g \in D} |\mathbb{G} f - \mathbb{G} g|\right] \leq C \int_0^{\operatorname{diam}(D)} \sqrt{\log \mathcal{N}(\epsilon, D, d)}\, d\epsilon.$$

[1] We know that the marginals of $\mathbb{G}_n$ are asymptotically tight because they converge to a tight limit (Lemma 2.1)

where subgaussian is defined as $\mathbb{P}\left(|\mathbb{G}f - \mathbb{G}f| > x\right) \leq 2\exp\left(-\frac{1}{2}x^2/d^2(f,g)\right)$. Lemma 3.4 (Hoeffding's) gives us that the Rademachar process is subgaussian conditional on $\{X_i\}$ for any class of functions $\mathcal{F}$ equipped with the $L_2(\mathbb{P}_n)$ norm.[2] This gives us that

$$\mathbb{E}_\epsilon\left[\sup_{\rho(f,g)<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\left(f(x_i)-g(x_i)\right)\right|\right] \lesssim \int_0^{\text{diam}(\mathcal{F}_\delta)} \sqrt{\log \mathcal{N}\left(\epsilon, \mathcal{F}_\delta, \|\cdot\|_{L_2(\mathbb{P}_n)}\right)}\, d\epsilon \tag{D-1}$$

where note that the diameter on the RHS is calculated with respect to $L_2(P_n)$ not the $\rho(f,g) = |f-g|_{L_2(P)}$ on the right hand side.[3]

Step 2: Make Sense of the Upper Bound. Let $\theta_n^2 := \text{diam}^2(\mathcal{F}_\delta) = \sup_{\rho(f,g)<\delta} \frac{1}{n}\sum_{i=1}^n \left(f(x_i)-g(x_i)\right)^2$. Let $u := \epsilon/\|F\|_{L^2(\mathbb{P}_n)}$ and rewrite the above[4]

$$\begin{aligned}\int_0^{\text{diam}(\mathcal{F}_\delta)} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_\delta, L_2(\mathbb{P}_n))}\, d\epsilon &= \int_0^{\theta_n} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_\delta, L_2(\mathbb{P}_n))}\, d\epsilon \\ &= \|F\|_{\mathbb{P}_n,2}\int_0^{\theta_n/\|F\|_{\mathbb{P}_n,2}} \sqrt{\log \mathcal{N}\left(u\|F\|_{\mathbb{P}_n}, \mathcal{F}_\delta, L_2(\mathbb{P}_n)\right)}\, du\end{aligned}$$

Since $\mathbb{P}_n$ is a discrete probability measure

$$\leq \|F\|_{\mathbb{P}_n,2}\int_0^{\theta/\|F\|_{\mathbb{P}_n,2}} \sup_{\mathcal{Q}} \sqrt{\log \mathcal{N}\left(u\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{L^2(Q)}\right)}\, du$$

But, since $\mathcal{F}_\delta \subseteq \mathcal{F}_\infty$ we get that $\mathcal{N}(\epsilon, \mathcal{F}_\delta, L_2(Q)) \leq \mathcal{N}(\epsilon, \mathcal{F}_\infty, L_2(Q))$. Also, see that $\mathcal{N}(\epsilon, \mathcal{F}_\infty, L_2(Q)) \leq \mathcal{N}^2(\epsilon/2, \mathcal{F}, L_2(Q))$. (Why?) In total:

$$\int_0^{\text{diam}(\mathcal{F}_\delta)} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_\delta, L_2(\mathbb{P}_n))}\, d\epsilon \leq \|F\|_{\mathbb{P}_n,2}\int_0^{\theta_n/\|F\|.} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\left(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\right)}\, du \tag{D-2}$$

Step 3: Go Back to the Full Expectation. Combining the inequality directly above and (D-1) and applying to the symmetrization inequality in (D-0), we get that

$$\mathbb{E}\left[\sup_{\rho(f,g)<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\left(f(x_i)-g(x_i)\right)\right|\right] \leq \mathbb{E}\left[\|F\|_{\mathbb{P}_n,2}\int_0^{\theta_n/\|F\|.} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\left(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\right)}\, du\right]$$

Where the expectations above are with respect to X. Apply Cauchy-Schwarz to upper bound the above by

$$\mathbb{E}_X\left[\left(\int_0^{\theta_n/\|F\|.} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\left(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\right)}\, du\right)^2\right]^{1/2} \mathbb{E}_X\left[\|F\|^2_{\mathbb{P}_n,2}\right]^{1/2} \tag{D-3}$$

Note that $\mathbb{E}_X[\|F\|^2_{\mathbb{P}_n,2}] = \mathbb{E}_X[n^{-1}\sum F^2(x_i)] \leq P^\star F^2 < \infty$. What is left is to show that the expectation of the integral in (D-3) converges to zero provided that $\theta_n/\|F\|_{\mathbb{P}_n,2} \to_{P^\star} 0$.

Step 4: Figure out what is happening with $\theta_n/\|F\|_{\mathbb{P}_n,2}$. Note that $\|F\|_{\mathbb{P}_n,2}$ is bounded below by $\|F_\star\|_{\mathbb{P}_n,2}$ which converges almost surely to its expectation. Recall that

$$\theta_n^2 = \sup_{\rho(f,g)<\delta} \frac{1}{n}\sum_{i=1}^n \left(f(x_i)-g(x_i)\right)^2 \leq \underbrace{\sup_{\rho(f,g)<\delta} \|(\mathbb{P}_n - P)(f-g)^2\|}_{=\|\mathbb{P}_n - P\|_{\mathcal{F}_\delta^2}} + \underbrace{\sup_{\rho(f,g)<\delta} P(f-g)^2}_{<\delta^2}.$$

[2] Note that

$$\left\|\left[\frac{1}{\sqrt{n}}f(x_i) - \frac{1}{\sqrt{n}}g(x_i)\right]_{i=1}^n\right\|^2 = \frac{1}{n}\sum_{i=1}^n \left(f(x_i)-g(x_i)\right)^2 = \|f-g\|_{L_2(\mathbb{P}_n)}.$$

[3] It also may be helpful to recall that $\mathcal{F}_\delta = \{f-g : \rho(,g) < \delta\}$

[4] Going to use $L_2(\mathbb{P}_n)$ or just $\|\cdot\|_{\mathbb{P}_n,2}$ instead of $\|\cdot\|_{L_2(\mathbb{P}_n)}$ to save some space. Also the empirical measure is bounded and F has bounded second outer moment so we know that $\|F\|_{L_2(\mathbb{P}_2)} < \infty$ almost surely.

Since $\mathcal{F}_\delta^2 \subseteq \mathcal{F}_\infty^2$ we want to show to show that $\mathcal{F}_\infty^2$ is Glivenko-Cantelli. Theorem 3.3 gives that is enough to show that $\log \mathcal{N}(\epsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) = o_{P^\star}(n)$.[5] For any pair of functions $f, g \in \mathcal{F}_\infty$

$$\mathbb{P}_n|f^2 - g^2| = \mathbb{P}_n|(f-g)(f+g)| \leq \mathbb{P}_n|f-g|4F \leq 2\|F\|_{\mathbb{P}_n,2}\|f-g\|_{\mathbb{P}_n,2}.^{6}$$

where in the above we use that $2F$ is an envelope for $\mathcal{F}_\infty$. This gives us that $\mathcal{N}(\epsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) \leq \mathcal{N}(\epsilon/\|2F\|_{\mathbb{P}_n,2}), \mathcal{F}_\infty, L_2(\mathbb{P}_n))$. Why? Suppose we cover $\mathcal{F}_\infty^2$ with N balls of size ϵ. By the inequalities above, we can find equivalent (centered at say, $(f-g)/2$ instead of $(f^2-g^2)/2$) balls that will cover $\mathcal{F}_\infty^2$ with radius $\epsilon/\|F\|_{\mathbb{P}_n,2}$ in $L_2(\mathbb{P}_n)$. As argued above, we know that this is less than $\mathcal{N}^2(\epsilon/4\|F\|_{L_2(\mathbb{P}_n)}, \mathcal{F}, L_2(\mathbb{P}_n))$ which we know has to be finite for any ϵ in order for (UEB) to be satisfied. Since $\mathcal{N}(\epsilon, \mathcal{F}_\infty, L_1(\mathbb{P}_n))$ is bounded by a constant, it's logarithm is surely $o_{p^\star}(n)$.

So, combining with the display at the beginning of this set we have that

$$\theta_n^2 \leq \sup_{\rho(f,g)<\delta} \left|(\mathbb{P}_n - P)(f-g)^2\right| + \delta^2 \to_{a.s} \delta.$$

Together this gives us that

$$\frac{\theta_n}{\|F\|_{\mathbb{P}_n,2}} \leq \frac{\theta_n}{\|F_\star\|_{\mathbb{P}_n,2}} \to_{a.s} \frac{\delta}{\|F_\star\|_{P,2}}. \tag{D-4}$$

Step 5: Put Together to get Asymptotic Equicontinuity. Take $C = (P^\star F^2)^{1/2}$ and combine the inequalities in (D-1), (D-2), (D-3) with the convergence result in (D-4).

$$\begin{aligned}
&\limsup \mathbb{P}\left(\sup_{\rho(f,g)<\delta} |\mathbb{G}_n f - \mathbb{G}_n g| > \epsilon\right) \\
&\leq \frac{1}{\epsilon} \limsup \mathbb{E}\left[\|F\|_{\mathbb{P}_n,2} \int_0^{\theta_n/\|F\|_{\mathbb{P}_n,2}} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\big(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\big)}\, du\right] \\
&\leq \frac{C}{\epsilon} \limsup \mathbb{E}\left[\left(\int_0^{\theta_n/\|F\|_{\mathbb{P}_n,2}} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\big(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\big)}\, du\right)^2\right]^{1/2} \\
&\to_{a.s} \frac{C}{\epsilon} \int_0^{\delta/\|F_\star\|_{P,2}} \sup_{\mathcal{Q}} \sqrt{2\log \mathcal{N}\big(u\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)\big)}\, du
\end{aligned}$$

However, by the uniform entropy bound in (UEB) the integral up to infinity in the above display is finite so the display above converges to 0 as $\delta \downarrow 0$. So we have verified that $\mathbb{G}_n$ is asymptotically equicontinuous.

Step 6: Verify that $\mathcal{F}$ is totally bounded. Finally what remains is to show that $\mathcal{F}$ is totally bounded in $L_2(P)$ (Definition 1.17). Take a sequence of discrete measures $\mathbb{P}_n$ such that $\|\mathbb{P}_n - P\|_{\mathcal{F}_\infty}$ converges to 0. Since we know that $\mathcal{F}_\infty$ is Glivenko-Cantelli, this is always possible. Pick n large enough such that $\|\mathbb{P}_n - P\|_{\mathcal{F}_\infty} < \delta^2$. By triangle inequality

$$P(f-g)^2 = \mathbb{P}_n(f-g)^2 - (\mathbb{P}_n - P)(f-g)^2 \leq \frac{1}{n}\sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2 + \|\mathbb{P}_n - P\|_{\mathcal{F}_\infty}.$$

This implies that $\mathcal{F}$ is totally bounded under $L_2(P)$ since we are totally bounded under $L_2(\mathbb{P}_n)$ (Finite Envelope + Finite Measure). □

Example 3.5 (Cells in $\mathbb{R}$ are Donsker)**.** Let X be a scalar and suppose that we want to show a functional CLT for

$$\mathbb{G}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{\mathbb{1}[X_i \leq t] - P(X \leq t)\right\}.$$

[5] Inherits bounded envelope by triangle inequality. If F bounds $\mathcal{F}$, then $4F^2$ bounds $\mathcal{F}_\infty^2$.

[6] By Cauchy-Schwarz: $\frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i) \leq \left[\frac{1}{n}\sum_{i=1}^n f^2(x_i)\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^n g^2(x_i)\right]^{1/2}$

Main challenge is showing the uniform entropy bound (UEB). Recall that for any $\|\cdot\|$, $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \mathcal{N}_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|)$. Apply the above to $\|\cdot\|_{Q,2}$. Partition $\mathbb{R}$ into $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $\mathbb{P}(t_i \leq X \leq t_{i+1}) \leq 4\epsilon^2$. These cover $\mathcal{F} = \{f(x) = \mathbb{1}\{x \leq t\}, t \in \mathbb{R}\}$. Further

$$\mathbb{E}\left[\left(\mathbb{1}[X \leq t_i] - \mathbb{1}[X \leq t_{i+1}]\right)^2\right]^{1/2} = \mathbb{P}\left(t_i \leq X \leq T_{i+1}\right)^{1/2} = 2\epsilon.$$

In order to cover $\mathcal{F}$ we need $\lceil 1/4\epsilon^2 \rceil \leq 2\epsilon^2$. This process can be done for any probability measure Q so that if $\epsilon < 1$:

$$\mathcal{N}(\epsilon, \mathcal{F}, L_2(Q)) \leq \mathcal{N}_{[]}(2\epsilon, \mathcal{F}, L_2(Q)) \leq 2/\epsilon^2 \wedge 1.$$

If $\epsilon \geq 1$ only one ball is needed. Then, since $F(x) = 1$ and if $\epsilon > 1$ one ball is enough (and $\log 1 = 0$):

$$\begin{aligned}\int_0^\infty \sup_Q \sqrt{\log \mathcal{N}(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon &= \int_0^1 \sup_Q \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(Q))}\, d\epsilon \\ &\leq \int_0^1 \sqrt{\log(2/\epsilon)}\, d\epsilon\end{aligned}$$

This is easily Donsker by Theorem 3.4 and in fact the argument hold for cells in $\mathbb{R}^K$. However, note that this seems like cheating a bit, since we are using the bracketing numbers to get the covering numbers.

We now want to show a Donsker Theorem using bracketing numbers. The proof is more involved than for the Glivenko-Cantelli Theorem (Theorem 3.2) using bracketing numbers, so we may not get too far into it here.

To show this theorem we make minor use of the following related statements from VdV&W.

Lemma 3.7 (Problem 2.5.5 VdV&W)**.** *If X is a positive random variable, $\|X\|_2^2 \leq \sup_{t>0} t\mathbb{E}X\mathbb{1}\{X > t\} \leq 2\|X\|_2^2$.*

Lemma 3.8 (Problem 2.5.6 VdV&W)**.** *Any random variable X with a finite second moment satisfies $E|X|\{X > t\} = o(t^{-1})$ as $t \to \infty$.*

Lemma 3.9 (Equation 2.5.5 VdV&W)**.** *For a finite set $\mathcal{F}$ of cardinality $\mathcal{F} \geq 2$,*

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{n}} \log|\mathcal{F}| + \max_f \|f\|_{P,2}\sqrt{\log|\mathcal{F}|}.$$

Theorem 3.5 (Bracketing Donsker Theorem)**.** *Let $\mathcal{F}$ be a class of measurable functions with an envelope F such that $P^\star F^2 < \infty$ and*

$$\int_0^\infty \sqrt{\log \mathcal{N}_{[]}\left(\epsilon, \mathcal{F}, L_2(P)\right)}\, d\epsilon < \infty \tag{3.12}$$

then $\mathcal{F}$ is Donsker.

Proof. The proof of this is roughly based on the steps in Theorem 2.5.6 in VanDerVaart and Wellner. I attempt to replicate the argument below:

For each $q \in \mathbb{N}$ there is a partition $\mathcal{F} = \bigcup_{i=1}^{N_q} \mathcal{F}_{qi}$ of $\mathcal{F}$ into N_q disjoint subsets such that

$$\begin{aligned}\sum_{i=1}^\infty 2^{-q}\sqrt{\log N_q} &< \infty \\ \|\big(\sup_{f,g \in \mathcal{F}_{qi}} |f-g|\big)^\star\|_{P,2} &< 2^{-q} \\ \sup_{f,g \in \mathcal{F}_{qi}} \|f-g\|_{P,2} &< 2^{-q}\end{aligned}$$

To see this, cover $\mathcal{F}$ with a minimal numbers of $L_2(P)$ balls and $L_2(P)$ brackets of size 2^{-q}, disjointify and take the intersection of the two partitions. By definition of the 2^{-q} balls and 2^{-q} brackets, the last two

conditions hold. To see that the first condition holds note that the integral of the bracketing number being finite implies that the integral of the covering number is finite, and that $\mathcal{N}_{[]}$ and $\mathcal{N}$ are decreasing in ϵ. For any decreasing function f it is clear that

$$\sum_{i=1}^{n} 2^{-q} f(q) \leq \int_0^\infty f(q)\, dq.$$

which gives us the first condition. This sequence of partitions can, without loss of generality, be chosen as successive refinements. [7]

For each q choose a fixed element f_{qi} from each partitioning set $\mathcal{F}_{qi}$ and define $\pi_q f = f_{qi}$ if $f \in \mathcal{F}_{qi}$. Further define $\Delta_q f = \sup_{f,g \in \mathcal{F}_{q,i}} |f-g|^\star$ if $f \in \mathcal{F}_{qi}$.

Note that $\pi_q f$ and $\Delta_q f$ take on one of N_q values as f ranges through $\mathcal{F}$. In view of Theorem 2.6 it suffices to show that the sequence $\|\mathbb{G}_n(f - \pi_{q_0} f)\|_{\mathcal{F}}$ converges in probability to zero as $n \to \infty$ for an arbitrary q_0 and then take $q_0 \to \infty$.

Define for each fixed n and $q \geq q_0$ the following numbers and indicator-type functions:

$$\begin{aligned} a_q &= 2^{-q}/\sqrt{\log N_{q+1}} \\ A_{q-1} f &= \mathbb{1}\{\Delta_{q0} f \leq \sqrt{n} a_{q_0} \wedge \cdots \wedge \Delta_{q-1} f \leq \sqrt{n} a_{q-1}\} \\ B_q f &= \mathbb{1}\{\Delta_{q_0} f \leq \sqrt{n} a_{q_0} \wedge \cdots \wedge \Delta_{q-1} f \leq \sqrt{n} a_{q-1} \wedge \Delta_q f > \sqrt{n} a_q\} \\ B_{q_0} &= \mathbb{1}\{\delta_{q_0} f > \sqrt{n} a_{q_0}\} \end{aligned}$$

Note that $A_q f$ and $B_q f$ are constant in f on each of the partitioning sets $\mathcal{F}_{qi}$ at level q because the partitions are nested.[8]

Now, pointwise in x, decompose

$$f - \pi_{q_0} f = (f - \pi_{q_0} f) B_{q_0} f + \sum_{q=q_0+1}^{\infty} (f - \pi_q f) B_q f + \sum_{q=q+1}^{\infty} (\pi_q f - \pi_{q-1} f) A_{q-1} f \tag{B-1}$$

Here note that we are essentially decomposing the event space into B_{q_0} and $B_{q_0}^c = A_{q_0}$. We can think of B_q as f being "between" A_{q-1} and A_q. That is, if all the conditions for B_q hold except for the last one, then A_{q-1} is equal to 1 and B_q is equal to zero. Conversely if B_q is equal to one, then all the conditions for A_q hold except for the last one ($\Delta_q g \leq \sqrt{n} a_q$) so A_q is equal to zero and B_q is equal to one. Equivalently $A_q + B_q = A_{q-1}$ or $A_{q-1} - A_q = B_q$. Combine this with the fact that sets indicated A_q are nested and telescope to get the decomposition above.

Now we will apply the empirical process $\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - P)$ to each of the terms in (B-1) and take the suprema over $f \in \mathcal{F}$. We will show that each of the resulting 3 variables converge to zero in probability and then take $q_0 \to \infty$.

[7] To see this, construct a sequence of partitions $\mathcal{F} = \bigcup_{i=1}^{\bar{N}_q} \bar{\mathcal{F}}_{qi}$ without this property. Next, take the partition at stage q, $\mathcal{F} = \bigcup_{i=1}^{N_q} \mathcal{F}_{qi}$ to consists of all intersections of the form $\bigcap_{p=1}^{q} \bar{\mathcal{F}}_{p,i_p}$ (basically take the intersection of all partitions up till q) so that $|N_q| = \prod_{p=1}^{q} \bar{N}_p$. Using the inequality $\left(\log \prod \bar{N}_p\right)^{1/2} \leq \sum \left(\log \bar{N}_p\right)^{1/2}$ we get that:

$$\begin{aligned} \sum_{q=1}^{\infty} 2^{-q} \sqrt{\log N_q} &\leq \sum_{q=1}^{\infty} 2^{-q} \sum_{p=1}^{q} \log \sqrt{\bar{N}_p} \\ &= \sum_{q=1}^{\infty} 2^{-q} \sum_{p=1}^{\infty} 2^{-p} \sqrt{\log \bar{N}_p} \\ &< \infty \end{aligned}$$

The equality is a bit difficult to see but follows from the preceding line after rewriting the double summation as a triangular array. To simplify, consider an infinite sequence $\{a_i\}_{i=1}^n$. Then $\sum_{i=1}^{\infty} 2^{-i} \sum_{j=1}^{i} a_i = a_1 + \frac{1}{2}(a_1 + a_2) + \frac{1}{2^2}(a_1 + a_2 + a_3) + \ldots$. Rearrange to get the result. This gives us that the first condition still holds for the new nesting sequence of partitions. The second and third conditions trivially still hold as the new partitions are finer than the previous ones.

[8] Recall that $\Delta_q f$ is the same for all elements in $\mathcal{F}_{q,i}$ and that $\mathcal{F}_{qi} \subset \mathcal{F}_{(q-1)i}$

First, since $|f - \pi_{q_0} f| B_{q_0} f \leq 2F \mathbb{1}\{2F > \sqrt{n} a_{q_0}\}$ one has that[9]

$$\mathbb{E}^\star \|\mathbb{G}_n (f - \pi_{q_0} f) B_{q_0} f\|_{\mathcal{F}} \leq 4\sqrt{n} P^\star F \mathbb{1}\{2F > \sqrt{n} a_{q_0}\}.$$

The right hand size converges to zero as $n \to \infty$ by Lemma 3.8 because F has a finite second outer moment.

By applying Lemma 3.7 and noting that $B_q f \leq \mathbb{1}\{\Delta_q f > \sqrt{n} a_q\}$:

$$\sqrt{n} a_q P \Delta_q f B_q f \leq \sqrt{n} a_q P \Delta_q f \mathbb{1}\{\Delta q f > \sqrt{n} a_q\} \leq 2 \|\Delta_q f\|_2^2 \leq 2\, 2^{-2q}.$$

Applying once that $\Delta_{q-1} f B_q f$ is bounded by $\sqrt{n} a_{q-1}$ for $q > q_0$, multiplying and dividing by a_q, and applying the inequality from above, we obtain the that inequality that we will need below:

$$P\left(\Delta_q f B_q f\right)^2 \leq \sqrt{n} a_{q-1} P \Delta_q f \mathbb{1}\{\Delta_q f > \sqrt{n} a_q\} \leq 2 \frac{a_{q-1}}{a_q} 2^{-2q}.$$

And now applying the triangle inequality, using that $|ab| = \big||a|b\big|$ for $a, b \in \mathbb{R}$:

$$\mathbb{E}^\star \left\| \sum_{q=q_0+1}^{\infty} \mathbb{G}_n (f - \pi_q) B_q f \right\|_{\mathcal{F}} \leq \sum_{q=q_0+1}^{\infty} \mathbb{E}^\star \left\| \mathbb{G}_n \Delta_q f B_q f \right\|_{\mathcal{F}}$$

Now applying Lemma 3.9. Note that 1) the implicit constant in the inequality can be taken to be universal by the nesting property of the subsets, 2) that $\Delta_q f B_q f \leq \sqrt{n} a_{q-1}$, the bounds derived above, and 3) that the supremum is taken over N_q functions at each level q:

$$\lesssim \sum_{q=q_0+1}^{\infty} a_{q-1} \log N_q + 2^{-q} \sqrt{2 \frac{a_{q-1}}{a_q}} \sqrt{\log N_q}$$

Note that a_q is decreasing in q (as q increases the top of the fraction is getting smaller and the bottom of the fraction is getting larger) so that the quotient can be replaced by it's square. Now use the definition of a_q to bound this sum:

$$\lesssim \sum_{q=q_0+1}^{\infty} 2^{-q} \sqrt{\log N_q}$$

This bound is independent of n and converges to zero as $q_0 \to \infty$.

To bound the third term note that there are at most N_q functions $(\pi_q - \pi_{q-1}) f$[10] and at most N_{q-1} functions $A_{q-1} f$. Since the partitions are nested ($\mathcal{F}_{qi} \subset \mathcal{F}_{q-1,i}$), the function $|\pi_q f - \pi_{q-1} f| A_{q-1} f$ is bounded by $\Delta_{q-1} f A_{q-1} f \leq \sqrt{n} a_{q-1}$. Also by nesting, the $L_2(P)$ norm of $\pi_q f - \pi_{q-1} f$ is bounded by $2^{-(q-1)}$. Apply the inequality in Lemma 3.9 to get

$$\mathbb{E}^\star \left\| \sum_{q=q_0+1}^{\infty} \mathbb{G}_n (\pi_q - \pi_{q-1} f) A_{q-1} f \right\|_{\mathcal{F}} \lesssim \sum_{q=q_0+1}^{\infty} a_{q-1} \log N_q + 2^{-q} \sqrt{\log N_q}.$$

As before, this upper bound is independent of n and converges to zero as $q_0 \to \infty$.

Putting this all together we get that as $n, q_0 \to \infty$

$$\mathbb{E}^\star \sup_{f,g \in \mathcal{F}} |f - g| \leq 2 E^\star \|f - \pi_{q_0} f\|_{\mathcal{F}} \longrightarrow 0.$$

which allows us to apply Theorem 2.6 and establish asymptotic equicontinuity. Because the envelope has finite second outer moment we can apply CLT to the marginals to get convergence to a tight distribution and apply Theorem 2.5 to get weak convergence of the whole process $\mathbb{G}_n$. □

[9] Recall the definition of the "$\mathcal{F}$" norm, pull out the $\sqrt{n}$, apply triangle inequality and note that $\mathbb{E}^\star \mathbb{P}_n F \leq \mathbb{E}^\star F$ because $(S+T)^\star \leq S^\star + T^\star$ for any functions S, T.

[10] each π_q has a specific π_{q-1} attached to it

Remark (Comments on Donsker Theorems)**.** Note that the integral in the bracketing Donsker Theorem above does not require taking the supremum over all finitely discrete probability measures Q, instead we only deal with the underlying measure P. The idea here is the bracketing gives us more control on the arguments.

Also, note that there are no measurability concerns for the bracketing Donsker theorem. This is because we did not have to use symmetrization, instead relying on Lemma 3.9.

Remark (Summary: Glivenko-Cantelli vs. Donsker)**.** Conditions for:

Glivenko-Cantelli:

- If $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every ϵ then $\mathcal{F}$ is (P)-Glivenko-Cantelli.
- If $\mathcal{F}$ is P-measurable with envelope F satisfying $P^\star F < \infty$ and for every $M, \epsilon > 0$ we have that $\log \mathcal{N}(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_{p^\star}(n)$ then $\mathcal{F}$ is (P)-Glivenko-Cantelli.

Donsker:

- If $\mathcal{F}$ has envelope F with $P^\star F^2 < \infty$ and
$$\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon < \infty.$$
then $\mathcal{F}$ is Donsker.
- If $\mathcal{F}$ has envelop F with $P^\star F^2 < \infty$, $\mathcal{F}_\infty^2$ and $\mathcal{F}_\delta$ are P-measurable for every δ and the (UEB) is satisfied, that is
$$\int_0^\infty \sup_{\mathcal{Q}} \sqrt{\log \mathcal{N}(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon < \infty.$$
then $\mathcal{F}$ is Donsker.

The bracketing numbers give you pointwise control of functions between the brackets, this gives us cleaner conditions for Glivenko-Cantelli and Donsker that depend only on the true measure. The covering number proofs require symmetrization and then applying Fubini's theorem, so we need measurability assumptions. The tradeoff is that bracketing numbers are larger than covering numbers, so the conditions may be harder to satisfy.

We have now reduced the conditions that we need for uniform convergence to conditions in terms of bracketing/covering numbers. Next we will turn to verifying these conditions and applying the uniform convergence results.

3.6 Covering Numbers

This section roughly covers section 2.6 in Van Der Vaart and Wellner.

Recall that for the covering Donsker Theorem, Theorem 3.4, we require the uniform entropy bound:

$$\int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log \mathcal{N}\left(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)}\, d\epsilon < \infty. \tag{UEB}$$

where the supremum under the integral is taken over the set of all finitely discrete probability measures $\mathcal{Q}$. If we can show that $\mathcal{F}$ is such that $\sup_{\mathcal{Q}} \log \mathcal{N}\left(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right) \lesssim \left(\frac{1}{\epsilon}\right)^{2-\delta}$ for some $\delta > 0$ then we are fine as the square root of this will integrate to a finite number. In fact, we will often be able to verify a much stronger condition, that:

$$\sup_{\mathcal{Q}} \mathcal{N}\left(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right) \leq K \left(\frac{1}{\epsilon}\right)^V, \quad 0 < \epsilon < 1.$$

3.6.1 VC Classes of Sets

A typical way to show that a class of sets $\mathcal{F}$ satisfies the (UEB) will be to show that it has a limited "VC-Dimension", where "VC" stands for Vapnick and Cervonenkis. What does this mean?

Let $\mathcal{C}$ be a collection of subsets of some set $\mathcal{X}$, that is $\mathcal{C} \subseteq 2^{\mathcal{X}}$. An arbitrary set of points $\{x_1, \ldots, x_n\}$ possesses 2^n subsets.

Definition 3.17 (Picking Out)**.** Let $A \subset \{x_1, \ldots, x_n\}$. The collection $\mathcal{C}$ picks out A if $A = C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$. We define the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by $\mathcal{C}$ as

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \#\left\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\right\} \tag{3.13}$$

Definition 3.18 (Shattering)**.** A collection $\mathcal{C}$ shatters $\{x_1 \ldots, x_n\}$ if it can pick out all of its 2^n subsets.

Definition 3.19 (VC Index)**.** The VC-Index, $V(\mathcal{C})$ is the smallest $n \in \mathbb{N}$ such that *no* set of size n is shattered by $\mathcal{C}$. Equivalently

$$V(\mathcal{C}) = \inf\left\{n : \max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n\right\} \tag{3.14}$$

Definition 3.20 (VC Class)**.** A collection $\mathcal{C}$ of measurable sets is called VC Class if $V(\mathcal{C}) < \infty$.

Remark. Notice that in the Definition of the VC Index, we require that *no* set of size n is shattered by $\mathcal{C}$ rather than requiring that n be the smallest number such that there exits *any* set of size n that is not shattered by $\mathcal{C}$.

Example. Suppose $\mathcal{C} = \left\{(-\infty, c] : \text{for some } c \in \mathbb{R}\right\}$. Then any set $\{x_1\}$ has subsets $\emptyset$ and $\{x_1\}$. For $\emptyset$ we have that $\emptyset \subset (-\infty, c] \cap \{x_1\}$ for any $c < x_1$ while for $\{x_1\}$ we have that $\{x_1\} \subseteq (-\infty, c] \cap \{x_1\}$ for any $c \geq x_1$. Thus we have that $\mathcal{C}$ shatters any single element subset of $\mathbb{R}$.

However, take a two element subset $\{x_1, x_2\} \subset \mathbb{R}$. Without loss of generality take $x_1 < x_2$. There are four subsets to consider, $\emptyset, \{x_1\}, \{x_2\}, \{x_1, x_2\}$. We can see that there is no set $C \in \mathcal{C}$ such that $C \cap \{x_1, x_2\} = \{x_2\}$, if we take $C = (\infty, c]$ for some $c < x_2$ we get that $C \cap \{x_1, x_2\} = \{x_1\}$ whereas if we take $C = (-\infty, c]$ for some $c \geq x_2$ we get that $C \cap \{x_1, x_2\} = \{x_1, x_2\}$. This exhausts all sets in $\mathcal{C}$.

Since $\{x_1, x_2\}$ is arbitrary, we can conclude that the VC Index of $\mathcal{C}$ is two, $V(\mathcal{C}) = 2$.

Example. Let $\mathcal{C}_1 = \left\{(a, b] : a < b \text{ for } a, b \in \mathbb{R}\right\}$. Note that this collection is larger than the collection from the prior example. Now however, given a set $\{x_1, x_2\}$ with $x_1 < x_2$ we can pick out $\{x_2\}$ with the set $(x_1, x_2] \in \mathcal{C}_1$.

However, now consider a three element subset $\{x_1, x_2, x_3\}$. Without loss of generality suppose $x_1 < x_2 < x_3$. We will try to pick out the set $\{x_1, x_3\}$. Consider any set $C \in \mathcal{C}_1$ such that $\{x_1, x_3\} \subset C$, a necessary condition for $\{x_1, x_2\} = C \cap \{x_1, x_2, x_3\}$. The set C is of the form $(a, b]$ for some $a < x_1$ and $b \geq x_3$. However, this means that $x_2 \in C$. So, we cannot pick out $\{x_1, x_3\}$ with $\mathcal{C}_1$.

Since we can pick out one and two element subsets of $\mathbb{R}$ with $\mathcal{C}_1$, but not arbitrary three element subsets, we get that $V(\mathcal{C}_1) = 3$.

Lemma 3.10 (Lemma 2.6.2 VdV&W)**.** *Let $\{x_1, \ldots, x_n\}$ be arbitrary points in $\mathcal{X}$ and $\mathcal{C}$ some collection of subsets of $\mathcal{X}$. Then the total number of subsets picked out by $\mathcal{C}$, $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ picked out by $\mathcal{C}$ is bounded above by the total number of subsets of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{C}$.*

Proof. Without loss of generality assume that every $C \in \mathcal{C}$ is a subset of the given set of points so that $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ is the cardinality of $\mathcal{C}$.

Call the class $\mathcal{C}$ *hereditary* if it is closed under subsetting. That is $C \in \mathcal{C}$ and $B \subset C \implies B \in \mathcal{C}$. Each of the sets in a hereditary collection of sets is shattered[1] so that a hereditary collection shatters at least

[1] For any $B \subset C$, $B \in \mathcal{C}$ and $B \cap C = B$.

$|\mathcal{C}|$ sets and the assertion of the lemma is certainly true for hereditary collections.[2] The goal now will be to show that an arbitrary collection $\mathcal{C}$ can be transformed into a hereditary collection without changing its cardinality or increasing the number of shattered sets.

Given $1 \leq i \leq n$ and $C \in \mathcal{C}$ define the set

$$T_i(C) = \begin{cases} C \setminus \{x_i\} & \text{if } C \setminus \{x_i\} \notin \mathcal{C} \\ C & \text{otherwise} \end{cases}$$

The map $T_i(x)$ is injective (one-to-one) so the collections $\mathcal{C}$ and $T_i(\mathcal{C}) = \{T_i(C), C \in \mathcal{C}\}$ have the same cardinalities.[3] Furthermore, every subset $A \subset \{x_1, \ldots, x_n\}$ that is shattered by $T_i(\mathcal{C})$ is shattered by $\mathcal{C}$. To see this note that if $x_i \notin A$ then $\{\mathcal{C} \cap A\} = \{T_i(\mathcal{C}) \cap A\}$. Conversely if $x_i \in A$ and A is shattered by $T_i(\mathcal{C})$ then for every $B \subset A$ there is a $C \in \mathcal{C}$ with $B \cup \{x_i\} = T_i(C) \cap A$.[4] This implies that $x_i \in T_i(C)$ so that $T_i(C) = C$. This in turns gives that $C \setminus \{x_i\} \in \mathcal{C}$ since otherwise T_i would not have a fixed point at C. Thus both $B \cup \{x_i\}$ and $B \setminus \{x_i\} = (C \setminus \{x_i\}) \cap A$ are picked out by $\mathcal{C}$. One of these sets equals B.

So the assertion of the lemma is true for $\mathcal{C}$ if it is true for $T_i(\mathcal{C})$. Furthermore the assertion of the lemma is true for $\mathcal{C}$ if it is true for $T(\mathcal{C})$ where $T = T_1 \circ T_2 \circ \cdots \circ T_n$; by repeatedly applying the argument above we have that if a set is shattered by $T(\mathcal{C})$ it is shattered by $\mathcal{C}$. Apply T repeatedly until the collection of sets does not change anymore. This happens after at most $\sum_{C \in \mathcal{C}} |C|$ steps (finite) since $\sum_{C \in \mathcal{C}} |T_i(C)| < \sum_{C \in \mathcal{C}} |C|$ whenever the collections $T_i(\mathcal{C})$ and $\mathcal{C}$ are different[5]. The collection $\mathcal{D}$ obtained in this manner has the property that $D \setminus \{x_i\} \in \mathcal{D}$ for every $D \in \mathcal{D}$ and every x_i. So $\mathcal{D}$ is hereditary. □

Corollary 3.2 (Corollary 2.6.3 VdV&W)**.** *For a VC-class of sets of index $V(\mathcal{C})$ one has*

$$\max_{x_1, \ldots, x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j}.$$

Consequently, the numbers on the left hand side grow polynomially of order at most $O\left(n^{V(\mathcal{C})-1}\right)$ as $n \to \infty$.

Proof. The RHS of the corrolary above is the number of subsets of size at most $V(\mathcal{C})-1$. A VC-class shatters no set of $V(\mathcal{C})$ points. All shattered sets are of size at most $V(\mathcal{C}) - 1$. The number of shattered sets gives an upper bound on Δ_n by Lemma 3.10. □

Theorem 3.6 (Theorem 2.6.4 VdV&W)**.** *There exists a universal constant K such that, for any VC-class $\mathcal{C}$ of sets, any probability measure Q, any $r \geq 1$ and $0 < \epsilon < 1$,*

$$\mathcal{N}(\epsilon, \mathcal{C}, L_r(Q)) \leq KV(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{r(V(C)-1)} \tag{3.15}$$

Proof. The proof of this Theorem takes some 3 pages in VanDerVaart so I am leaving it for now. It can be found on pages 137-139. □

Remark. For practical purposes, if $V(\mathcal{G}) < \infty$, then $\sup_{Q \in \mathcal{Q}} \log \mathcal{N}(\epsilon, \mathcal{G}, L_2(Q)) \lesssim \log\left(\frac{1}{\epsilon}\right)$ which will mean that the (UEB) is satisfied.

[2] Taking a look at (3.13) we see that all sets picked out by $\mathcal{C}$ are all subsets of elements of $\mathcal{C}$. Since all subsets of elements of $\mathcal{C}$ are also elements of $\mathcal{C}$, the number of sets picked out by $\mathcal{C}$ is bounded by the number of sets in $\mathcal{C}$. Every element of $\mathcal{C}$ is clearly shattered by $\mathcal{C}$ which gives the statement.

[3] Recall that $\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = |\mathcal{C}|$ since by assumption (without loss of generality) every $C \in \mathcal{C}$ is a subset of the given set of points. This holds as well for $T_i(\mathcal{C})$

[4] This is just because $B \cup \{X_i\} \subset A$ and A is shattered by $T_i(\mathcal{C})$

[5] We can think of repeatedly applying T as "pruning" the collection $\mathcal{C}$, removing elements from sets whose subsets are not contained in $\mathcal{C}$.

Example. Suppose $\mathcal{G}$ is a VC-class, that is the functions $\mathbb{1}_G$ are measurable for $G \in \mathcal{G}$ and $V(\mathcal{G}) < \infty$. Since this set of indicator functions is bounded by 1, this class is clearly Donsker. In the example above, we showed that the set $\mathcal{G} = \{(a, b] : a < b, a, d \in \mathbb{R}\}$ has VC-index, $V(\mathcal{G}) = 3 < \infty$. By Theorem 3.6 (and subsequently Theorem 3.4), this is a Donsker class. That is uniformly over $a, b \in \mathbb{R}$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \mathbb{1}[a < X_i \leq b_i] - P(a < X_i < b_i) \right\} \rightsquigarrow \mathbb{G}.$$

for some tight element $\mathbb{G}$ on $\ell^\infty(\mathcal{G})$.

All together, this is interesting for showing that collections of indicator functions are Donsker, but what about arbitrary classes of functions?

3.6.2 VC Classes of Functions

Definition 3.21 (Subgraph)**.** The subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by

$$\left\{ (x, t) : t < f(x) \right\}.$$

Remark. Note that the subgraph does not include the points $\{(x, y) : y = f(x)\}$

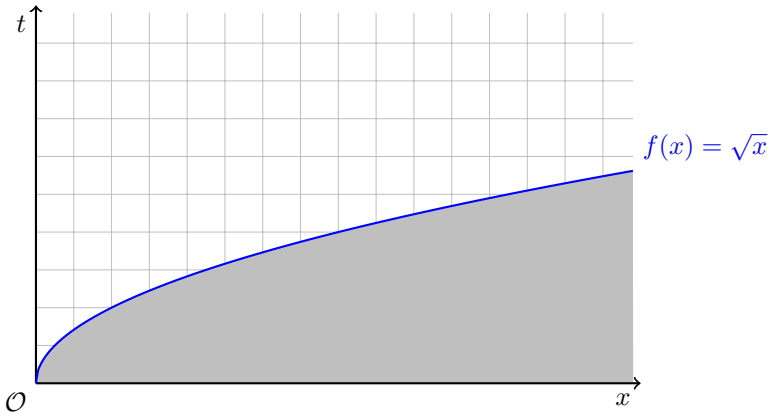

Figure 3.1: The subgraph of $f : \mathbb{R}^+ \to \mathbb{R}^+$, $x \mapsto \sqrt{x}$ is shaded in gray.

Definition 3.22 (VC Class for Functions)**.** A collection of functions $\mathcal{F}$ is called a VC-subgraph class, or just a VC-class if the collections of all subgraphs of the functions in $\mathcal{F}$ forms a VC-class of sets in $\mathcal{X} \times \mathbb{R}$. Let $V(\mathcal{F})$ be the VC-index of the set of subgraphs of functions in $\mathcal{F}$.

Just as for sets, the covering numbers of VC-classes of functions grow at a polynomial rate.

Theorem 3.7 (Theorem 2.6.7 VdV&W)**.** *Let $\mathcal{F}$ be a VC-class of functions with measurable envelope F. Then for $r \geq 1$ and any probability measure Q with $\|F\|_{Q,r} > 0$,*

$$\mathcal{N}\left(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right) \leq KV(\mathcal{F}) \left(16e\right)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)} \tag{3.16}$$

for a universal constant K and $0 < \epsilon < 1$.

Proof. Let $\mathcal{C}$ be the set of all subgraphs C_f of functions $F \in \mathcal{F}$. That is

$$C_f = \left\{ (x, t) : t < f(x) \right\}$$
$$\mathcal{C} = \left\{ C_f : f \in \mathcal{F} \right\}$$

By Fubini's Theorem, $Q|f-g| = (Q \times \lambda)\left(C_f \triangle C_g\right)$ where $A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$ is the symmetric difference of two sets.[6] Renormalize $Q \times \lambda$ to a probability measure on the set $\{(x,t) : |t| \leq F(x)\}$

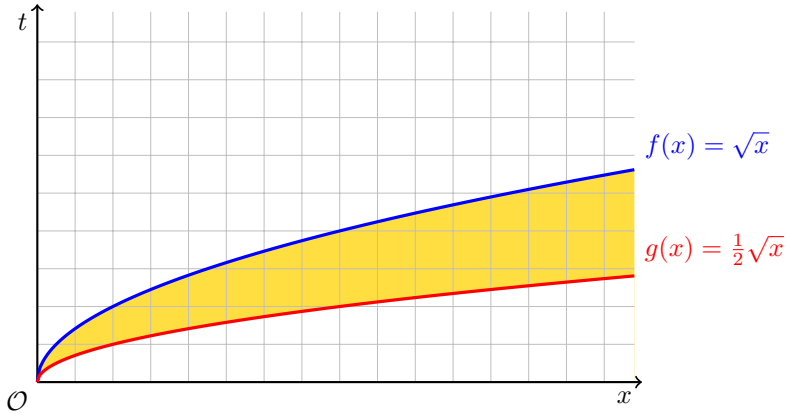

Figure 3.2: The symmetric difference $C_f \triangle C_g$ is shaded in yellow

by defining $P = \frac{(Q \times \lambda)}{2QF}$. Then, by the result in Theorem 3.6 we have that

$$\mathcal{N}\left(\epsilon(2QF), \mathcal{F}, L_1(Q)\right) = \mathcal{N}\left(\epsilon, \mathcal{C}, L_1(P)\right) \leq KV(\mathcal{F})\left(\frac{4e}{\epsilon}\right)^{V(\mathcal{F})-1}.$$

By adjusting the constant K to convert $(4e)^{V(\mathcal{F})-1}$ to a $(16e)^{V(\mathcal{F})}$ this concludes the proof for $r = 1$.

For $r > 1$ note that $2F$ is an envelope for $\mathcal{F}_\infty$, a property that we have used before. Define R to be the probability measure with density F^{r-1}/QF^{r-1} with respect to Q. That is $R\bar{f} = Q\left(\bar{f}\frac{F^{r-1}}{QF^{r-1}}\right)$ for any function $\bar{f} : \mathcal{X} \to \mathbb{R}$. Then:

$$Q|f-g|^r \leq Q|f-g|(2F)^{r-1} = 2^{r-1}R|f-g|QF^{r-1}.$$

Thus the $L_r(Q)$ distance is bounded by the distance $2(QF^{r-1})^{1/r}\|f-g\|_{R,1}^{1/r}$. This gives that

$$\mathcal{N}\left(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right) \leq \mathcal{N}\left(\epsilon^r \|F\|_{R,1}, \mathcal{F}, L_1(R)\right).$$

which can be uniformly bounded by the result for $r = 1$. This gives the result after nothing that $\frac{1}{\epsilon^r} = \left(\frac{1}{\epsilon}\right)^r$. □

If the conditions of Theorem 3.7 are satisfied with $V(\mathcal{F}) < \infty$ then we can bound the (UEB) with

$$\begin{aligned}\int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log \mathcal{N}\left(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)}\, d\epsilon &\lesssim \int_0^1 \sqrt{\log \mathcal{N}\left(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_1(Q)\right)}\, d\epsilon \\ &\lesssim \int_0^1 \sqrt{\log \tilde{K}\left(\frac{1}{\epsilon}\right)^{2(V(\mathcal{F})-1)}}\, d\epsilon < \infty\end{aligned}$$

where the first step is using the fact that if $\epsilon > 1$ we only need one ball of radius $\epsilon\|F\|_{Q,2}$ to cover $\mathcal{F}$ under Q and then noting that the covering numbers under L_1 differ from the covering numbers under L_2 by a bounded constant.

[6] It is useful to recall that

$$C_f \cup C_g = \{(x,t) : t < f(x) \text{ or } t < g(x)\}$$
$$C_f \cap C_g = \{(x,t) : t < f(x) \text{ and } t < g(x)\}$$

and also that Q is a probability measure over $\mathcal{X}$ whereas λ is a measure over $\mathbb{R}$. Both C_f and C_g are subsets of $\mathcal{X} \times \mathbb{R}$.

Example. Suppose $\mathcal{F} = \left\{\sum_{j=1}^{K} \beta_j \phi_j(x), (\beta_1, \ldots, \beta_K) \in \mathbb{R}^K\right\}$. Then $V(\mathcal{F}) \leq K + 2$. How? Recall that we are considering the VC-index of the collection

$$\mathcal{C} = \left\{\left\{(x,t) : t \leq \beta'\phi(x)\right\} : \beta \in \mathbb{R}^K\right\} \subseteq 2^{\mathcal{X}\times\mathbb{R}}.$$

Pick $(x_1, t_1), \ldots, (x_{k+2}, t_{k+2})$ distinct points in $\mathcal{X} \times \mathbb{R}$. Then for any $f \in \mathcal{F}$:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_{n+2}) \end{bmatrix} = \beta_1 \begin{bmatrix} \phi_1(x_1) \\ \vdots \\ \phi_1(x_{n+2}) \end{bmatrix} + \cdots + \beta_K \begin{bmatrix} \phi_K(x_1) \\ \vdots \\ \phi_K(x_{n+2}) \end{bmatrix}.$$

so $\big(f(x_1), \ldots, f(x_K)\big)$ is in a subspace of $\mathbb{R}^{K+2}$ of dimension K for any $f \in \mathcal{F}$. Similarly, we can show that $f(x) - t$ is in a subspace of dimension $K + 1$ for any fixed $f \in \mathcal{F}$ and $t_1, \ldots, t_{K+2}$.

To show that $V(\mathcal{F}) \leq K + 2$ we want to show that $\left\{(x_1, t_1), \ldots, (x_{K+2}, t_{K+2})\right\}$ cannot be shattered by subgraphs of $\mathcal{F}$. That is we want a subset $A \subseteq \left\{(x_1, t_1), \ldots, (x_{K+2}, t_{K+2}\right\}$ such that $\nexists f \in \mathcal{F}$ with

$$A = \left\{(x_1, t_1), \ldots, (x_{K+2}, t_{K+2})\right\} \cap \left\{(x,t) : t < f(x)\right\}.$$

Since for any fixed $f \in \mathcal{F}, t_1, \ldots, t_{K+2}$, $f(x_1) - t_1, \ldots, f(x_{K+2}) - t_{K+2})$ is in a subspace of dimension $K + 1$ of $\mathbb{R}^{K+2}$ for any $x_1, \ldots, x_{K+2}$, there exists a vector $a \neq 0$ orthogonal to this subspace such that

$$\sum_{j=1}^{K+2} a_j(f(x_j) - t_j) = 0 \implies \sum_{a_j>0} a_j\big(f(x_j) - t_j\big) = \sum_{a_j\leq 0} (-a_j)(f(x_j) - t_j).$$

Pick the indices corresponding to the positive $a_j > 0$ values (the indices on the LHS after the implication) and consider $A = \{(x_j, t_j) : a_j > 0\}$. Since $a \neq 0$ we know this set is non-empty. If $A = \{(x_1, t_1), \ldots, (x_{K+2}, t_{K+@})\} \cap \{(x,t) : t < f(x)\}$ for some $f \in \mathcal{F}$ then

$$\sum_{a_j>0} a_i(f(x_j) - t_j) > 0 \geq \sum_{a_j\leq 0} (-a_j)(f(x_j) - t_j).$$

which is a contradiction as these two are equal as demonstrated above. Thus $V(\mathcal{F}) \leq K + 2$. However, notice that unless the functions $\phi_j(x)$ are all uniformly equal to 0, the class $\mathcal{F}$ does not have an envelope with $P^\star F < \infty$, so it is neither Glivenko-Cantelli nor Donsker unless we restrict the parameter space (the range of $\beta_1, \ldots, \beta_K$). It is useful to consider the following sanity check:

$$\sup_{\beta_1,\ldots,\beta_K} \left| \frac{1}{n} \sum_{i=1}^{n} \big(\beta_1\phi_1(x_i) + \cdots + \beta_k\phi_K(x_i)\big) - \mathbb{E}[\ldots] \right| = \infty.$$

This example can be generalized and outlines the proof for the next lemma

Lemma 3.11 (VC Dimension of Finite Dimensional Vector Spaces)**.** *Any finite dimensional vector space $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is a VC function class with $V(\mathcal{F}) \leq dim(\mathcal{F}) + 2$.*

Proof. Follows the example above. Can also be found as Lemma 2.6.16 in VanDerVaart and Wellner. □

Another examples of a VC function class is given below:

Lemma 3.12 (VC Dimension of Translates)**.** *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a monotonic function and let $\mathcal{F}$ be the set of all translates of ψ, $\mathcal{F} = \big\{\psi(x - h) : h \in \mathbb{R}\big\}$. Then $V(\mathcal{F}) = 2$.*

Proof. This follows similarly from the example above where we considered the VC-dimension of $\mathcal{C} = \big\{(-\infty, b) : a < b\big\}$. Without loss of generality suppose that ψ is decreasing and denote $\psi_h = \psi(x - h)$. The subgraphs of ψ_h are nested in that if $h > h'$, $C_{h'} \subseteq C_h$ since if $t < \psi_{h'}(x)$ then $t < \psi_h(x)$. Take any two element subset of $\mathbb{R} \times \mathbb{R}$: $\{(x_1, t_1), (x_2, t_2)\}$. By considering the single element subsets of this set we see that we cannot shatter this set without breaking the nesting of the subgraphs so $V(\mathcal{F}) = 2$. □

3.7 Bracketing Numbers

This discussion roughly follows Van Der Vaart and Wellner Chapter 2.7. Results on bracketing numbers rely on approximation theory.

Definition 3.23 (Differential Operator)**.** For a vector $K = (K_1, \ldots, K_d) \in \mathbb{N}^d$ let $|K| = \sum_{j=1}^d K_j$. For any $|K|$ times differentiable function $f : \mathcal{X} \to \mathbb{R}$ define

$$D^K f(x) = \frac{\partial^{|K|}}{\partial x_1^{K_1} \partial x_2^{K_2} \cdots \partial x_d^{K_d}} f(x).$$

Definition 3.24 (Differential Norm)**.** For any $\alpha > 0$ let $\underline{\alpha} = 1 \vee \lfloor \alpha \rfloor$, the smallest positive integer less than α. Then, for a function $f : \mathcal{X} \to \mathbb{R}$ let

$$\|f\|_\alpha = \max_{|k| \leq \underline{\alpha}} \sup_x \left| D^{|K|} f(x) \right| + \max_{|K| = \underline{\alpha}} \sup_{x,y \in \mathcal{X}^\circ} \frac{\left| D^K f(x) - D^K f(y) \right|}{\|x - y\|^{\alpha - \underline{\alpha}}}.$$

Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous function $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_\alpha \leq M$.

Example (Differential Norm)**.** Let $\mathcal{X} = \mathbb{R}$. Then, $\|f\|_2 \leq M$ means that

- (Bounded Function): $\sup_{\mathcal{X}} \left| f(x) \right| \leq M$
- (Bounded Derivative): $\sup_{\mathcal{X}} \left| f'(x) \right| \leq M$
- (Lipschitz Condition): $|f'(x) - f'(y)| \leq |x - y| M$

Example (Differential Norm)**.** Let $\mathcal{X} = \mathbb{R}$. Then, $\|f\|_{0.5} \leq M$ means that

- (Bounded Function): $\sup_{\mathcal{X}} \left| f(x) \right| \leq M$
- (Hölder Condition): $\left| f(x) - f(y) \right| \leq \sqrt{|x - y|}$

Theorem 3.8 (Theorem 2.7.1 VdV&W)**.** *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Then, there exists a constant K depending only on α and d such that*

$$\log \mathcal{N}\left(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty\right) \leq K \lambda(\mathcal{X}^1) \left(\frac{1}{\epsilon}\right)^{d/\alpha},$$

where λ denotes Lebesgue measure and $\mathcal{X}^1 = \{x : d(x, \mathcal{X}) < 1\}$.

Corollary 3.3 (Bracketing Numbers for α-smooth Functions)**.** *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. There exists a constant K depending only on $\alpha, diam(\mathcal{X})$ and d such that*

$$\log \mathcal{N}_{[]}\left(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)\right) \leq K \left(\frac{1}{\epsilon}\right)^{d/\alpha}.$$

Proof. Let $f_1, \ldots, f_p$ be the centers of $\|\cdot\|_\infty$ balls of radius ϵ that cover $C_1^\alpha(\mathcal{X})$. The brackets $[f_i - \epsilon, f_i + \epsilon]$ cover C_1^α. Each bracket has $L_r(Q)$ size at most 2ϵ, for any r. Apply Theorem 3.8. □

Remark (Relaxing the Bound in C_1^α)**.** Suppose we want to apply the results of Corollary 3.3 but to the slightly larger set $C_M^\alpha(\mathcal{X})$. Pick an $\epsilon, \|\cdot\|_\infty$ ball cover of $C_1^\alpha(\mathcal{X})$ with centers at $g_1, \ldots, g_k$ and consider $Mg_1, \ldots, Mg_k \in C_M^\alpha(\mathcal{X})$. For every $f \in C_M^\alpha(\mathcal{X})$, $\|f - Mg_i\|_\infty = M\|f/M - g_i\| < \epsilon M$ for some $1 \leq i \leq k$ so $Mg_1, \ldots, Mg_k$ is an ϵM cover of $C_M^\alpha(\mathcal{X})$. Applying Corollary 3.3 gives

$$\begin{aligned} \log \mathcal{N}\left(\epsilon, C_M^\alpha(\mathcal{X}), \|\cdot\|_\infty\right) &\leq \log \mathcal{N}\left(\epsilon/M, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty\right) \\ &\lesssim \left(\frac{1}{\epsilon/M}\right)^{d/\alpha} \lesssim \left(\frac{1}{\epsilon}\right)^{d/\alpha} \end{aligned}$$

Example (Glivenko-Cantelli)**.** To apply the bracketing Glivenko-Cantelli Theorem, Theorem 3.2, we require that $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for all $\epsilon > 0$. Hence for $\mathcal{F} = C_M^\alpha(\mathcal{X}) = \left\{f : \mathcal{X} \to \mathbb{R} : \|f\|_\alpha \leq M, \alpha > 0\right\}$ and $\mathcal{X}$ bounded and convex, by Corollary 3.3 that this is finite for any $\epsilon > 0$.

Example (Donsker)**.** For the bracketing Donsker Theorem, Theorem 3.5, we required that

$$\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon < \infty.$$

Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ and consider $C_m^\alpha(\mathcal{X})$. Then by Corollary 3.3:

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\epsilon, C_m^\alpha(\mathcal{X}), L_2(P))}\, d\epsilon &= \int_0^{2M} \sqrt{\log \mathcal{N}_{[]}(\epsilon, C_M^\alpha(\mathcal{X}), L_2(P))}\, d\epsilon \\ &\lesssim \int_0^{2M} \left(\frac{1}{\epsilon}\right)^{\frac{d}{2\alpha}}\end{aligned}$$

So long as $\frac{d}{2\alpha} < 1$ or (equivalently) $d < 2\alpha$, this will be finite and the class will be Donsker. If $d = 1$ then this holds for $\alpha > 1/2$ so the set of all bounded Lipschitz function is Donsker. If $d = 2$ then this holds for $\alpha > 1$ so the set of all level functions with bounded derivatives and Lipschitz first order derivatives is bounded. In general in higher dimensions we need to add smoothness.

Theorem 3.9 (Monotone Donsker Class)**.** *The class $\mathcal{F}^m$ of monotone functions $f : \mathbb{R} \to [0, 1]$ satisfies for every $r \geq 1$:*

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right).$$

For a constant K depending only on r and every Q.

Remark. The above gives us that the class of monotone functions into $[0, 1]$ is Glivenko-Cantelli and Donsker. This gives us another way of showing that the empirical CDF is Donsker as we can take

$$\mathcal{F}^{\text{ind}} = \left\{\mathbb{1}[x \leq b] : b \in \mathbb{R}\right\}.$$

Since $\mathcal{F}^{\text{ind}} \subset \mathcal{F}^m$, $\mathcal{F}^{\text{ind}}$ is Donsker by Theorem 3.9.

We now turn to some examples to demonstrate the usefulness of the above results and demonstrate some other useful relationships.

Example (Classes of Differences)**.** Let $\mathcal{G} = \{f - g : f, g \in \mathcal{F}\}$. How do we find $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|)$?[1] Let $f_1, \ldots, f_k$ be a minimal set of $\epsilon/2$ balls in $\|\cdot\|$ that cover $\mathcal{F}$.[2]. Form the functions

$$\left.\begin{array}{cccc} f_1 - f_1 & f_2 - f_1 & \cdots & f_k - f_1 \\ f_1 - f_2 & f_2 - f_2 & \cdots & f_k - f_2 \\ \vdots & \ddots & \ddots & \vdots \\ f_1 - f_K & f_2 - f_k & \cdots & f_k - f_k \end{array}\right\} \begin{array}{l}\text{Label these} \\ g_1, \ldots, g_{k^2}\end{array}$$

Then $g_1, \ldots, g_{k^2}$ is an ϵ-cover for $\mathcal{G}$. If $\phi = f - g \in \mathcal{G}$ then there is a g_ℓ function in $\{g_1, .., g_{k^2}\}$ such that

$$\|\phi - g_\ell\| = \|f - g - (f_i - f_j)\| \leq \|f - f_i\| + \|g - f_j\| < \epsilon$$

So that $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|) \leq \mathcal{N}^2(\epsilon/2, \mathcal{F}, \|\cdot\|)$.

Example (Power Rules)**.** Recall the definition of $\|f\|_\alpha$ (Definition 3.24). Suppose we are interested in $\mathcal{F} = \{f^2 : \|f\|_\alpha \leq M\}$. If $g_1, g_2 \in \mathcal{F}$ then $g_1 = f_1^2$ and $g_2 = f_2^2$ for $\|f_1\|_\alpha \leq M$ and $\|f_2\|_\alpha \leq M$. Then

$$\|g_1 - g_2\|_\infty = \|f_1^2 - f_2^2\|_\infty = \|(f_1 - f_2)(f_1 + f_2)\|_\infty \leq \|f_1 - f_2\|_\infty \|f_1 + f_2\|_\infty \leq 2M\|f_1 - f_2\|_\infty$$

Therefore, if $g_1, \ldots, g_K^2$ provide an $\epsilon/2M$ cover of C_M^α under $\|\cdot\|_\infty$, then $g_1, \ldots, g_K^2$ provide an ϵ cover of $\mathcal{F}$. This argument was also used in the proof of Theorem 3.4.

[1] Recall that we used an argument of this nature in the proof of Theorem 3.4.
[2] That is take $k = \mathcal{N}(\epsilon/2, \mathcal{F}, \|\cdot\|)$

Remark. In the above example it is important that the functions are bounded. Otherwise, we may run into trouble when showing Donsker properties as the fourth moment of the envelope may not be finite, even if the second moment is.

An example of when we are all ok is when the functions are Lipschitz.

Theorem 3.10 (Lipschitz Combination of Donsker Classes)**.** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be Donsker classes with envelopes $F_1, \ldots, F_k$ and $\varphi : \mathbb{R}^K \to \mathbb{R}$ a Lipschitz function, i.e φ is such that*

$$\left|\varphi(x_1, \ldots, x_k) - \varphi(y_1, \ldots, y_k)\right| \lesssim \|x - y\|^2.$$

Then if $\mathbb{E}\left[\phi^2(F_1(x), \ldots, F_k(x))\right] < \infty$, the class

$$\mathcal{G} = \left\{\phi(f_1, \ldots, f_k) : f_i \in \mathcal{F}_i\right\}.$$

is Donsker.

Example (Specification Testing)**.** Suppose we estimate the model

$$Y = f(X, \theta) + \epsilon \quad \text{with} \quad \mathbb{E}[\epsilon|X] = 0.$$

But, after we estimate the model, we want to test whether the model is correctly specified. That is we want to test

$$H_0 : \Pr\left(\mathbb{E}[Y|X = x] = f(x, \theta)\right) = 1, \text{ for some } \theta \in \Theta$$
$$H_1 : \Pr\left(\mathbb{E}[Y|X = x] = f(x, \theta)\right) < 1, \text{ for all } \theta \in \Theta$$

First Approach: If $f(x, \theta_0) = \mathbb{E}[Y|X = x]$, then $\mathbb{E}\left[(Y - f(X, \theta_0)\varphi(X)\right] = 0$ for all (integrable) $\varphi(X)$. A first guess at a test would be to pick a set of functions $\varphi_1, \ldots, \varphi_k$ and test $\mathbb{E}[(Y_i - f(X, \theta_0))\varphi_j(X)] = 0$ for all j. A feasible way of doing so would be to stack $\varphi = (\varphi_1, \ldots, \varphi_k)'$ and use the test statistic

$$F = \left(\mathbb{G}_n(Y_i - f(X_i, \hat{\theta})\varphi(X_i)\right)' W \mathbb{G}_n(Y_i - f(X_i, \hat{\theta})\varphi(X_i).$$

For an appropriate choice of W, F will be distributed χ^2_k under the null hypothesis. The problem is that this test does not exhaust all the moment restrictions and so will not in general be consistent.

Second Approach: Insight from Bierens: How do we use an infinite number of moments?

Lemma 3.13 (Bierens 1990)**.** *Let V be a random scalar with $\mathbb{E}[|V|] < \infty$ and X be a bounded random vector in $\mathbb{R}^K$ such that $\Pr(\mathbb{E}[V|X] = 0) < 1$. Then*

$$\mathcal{S} = \left\{t \in \mathbb{R}^K : \mathbb{E}\left[ve^{t'X}\right] = 0\right\}$$

is a set of Lebesgue measure zero.[3]

What does this mean? If $\mathbb{E}[Y|X = x] = f(x, \theta_0)$ then $\mathbb{E}[(Y - f(X, \theta_0)e^{t'X}] = 0$ for (almost) all $t \in T$, which is some appropriately chosen compact set with positive Lebesgue measure. On the other hand, if $\mathbb{E}[Y|X] \neq f(x, \theta_0)$ then $\mathbb{E}[(Y - f(X, \theta_0)e^{t'X}] \neq 0$ for "most" $t \in T$.

The goal will be to build a test statistic based on this observation. The requirement that X is bounded is not restrictive as $\mathbb{E}[V|X] = \mathbb{E}[V|\tanh(X)]$.

Test Statistic: Is $\mathbb{E}[(Y - f(X, \theta_0)\exp(t'X)] = 0$ for all $t \in T$? Take

$$T_n = \max_T \left|\mathbb{G}_n(Y - f(X_i, \hat{\theta}))e^{t'X_i}\right|.$$

Under the null hypothesis T_n should be well behaved whereas under the alternative it may diverge. Analysis is in a few steps.

[3] This results is generalized in Stinchcombe and White (1998)

Step 1: Deal with the difference between $\hat{\theta}$ and θ_0.

Suppose we estimate $\hat{\theta}$ via nonlinear least squares. That is

$$\hat{\theta} = \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i, \theta)\right)^2.$$

Using the first order condition and a (first order) Taylor expansion we see that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \left[\mathbb{P}_n \frac{\partial f(X_i, \hat{\theta})}{\partial \theta} \frac{\partial f(X_i, \bar{\theta})}{\partial \theta'}\right]^{-1} \mathbb{G}_n \frac{\partial f(X_i, \hat{\theta})}{\partial \theta} (Y_i - f(X_i, \theta_0)).$$

Taking $A = \mathbb{E}\left[\frac{\partial f(X_i, \hat{\theta})}{\partial \theta} \frac{\partial f(X_i, \bar{\theta})}{\partial \theta'}\right]$ allows us to rewrite the above as

$$\sqrt{n}(\hat{\theta} - \theta_0) = A^{-1} \mathbb{G}_n \frac{\partial f(X_i, \theta_0)}{\partial \theta} (Y_i - f(X_i, \theta_0)) + o_p(1) \tag{S1}$$

where we note that there is a uniform consideration being hidden here in dealing with $\bar{\theta}$ and we are assuming that $\hat{\theta}$ is consistent (minor).

Step 2: Study the process indexed by t.

$$\mathbb{G}_n(Y_i - f(X_i, \hat{\theta}))e^{t' X_i} = \underbrace{\mathbb{G}_n(Y_i - f(X_i, \theta_0))e^{t' X_i}}_{A} + \underbrace{\mathbb{G}_n(f(X_i, \theta_0) - f(X_i, \hat{\theta}))e^{t' X_i}}_{B}.$$

The first term on the right hand side here looks manageable, but what about the second? We will apply the delta method. To do so note that by first order Taylor expansion in (S1):

$$\left(\mathbb{P}_n \frac{\partial f(x_i, \bar{\theta})}{\partial \theta} e^{t' x_i}\right) \sqrt{n}(\theta_0 - \hat{\theta}) \approx \left(\mathbb{P}_n \frac{\partial f(X_i, \bar{\theta})}{\partial \theta} (Y_i - f(X_i, \theta_0))\right)' A^{-1} \mathbb{G}_n \frac{\partial f(X_i, \theta_0)}{\partial \theta} (Y_i - f(X_i, \theta_0)).$$

If we have

$$\sup_{t \in T, \theta \in \Theta} \left| \mathbb{P}_n \frac{\partial f(X_i, \theta)}{\partial \theta} e^{t' X_i} - \underbrace{\mathbb{E} \frac{\partial f(X_i, \theta)}{\partial \theta} e^{t' X_i}}_{:= b(t, \theta)} \right| = o_p(1).$$

and an appropriate continuity condition, then:

$$\mathbb{G}_n(f(X_i, \theta_0) - f(X_i, \hat{\theta}))e^{t' X_i} = b'(t, \theta) A^{-1} \mathbb{G}_n \frac{\partial f(X_i, \theta_0)}{\partial \theta} (f(X_i, \theta_0) - Y_i) + o_p(1).$$

where the $o_p(1)$ is uniform in t, i.e.

$$\sup_{T} \left| \mathbb{G}_n(f(X_i, \theta_0) - f(X_i, \hat{\theta}))e^{t' X_i} - b'(t, \theta_0) A^{-1} \mathbb{G}_n \frac{\partial f(X_i, \theta_0)}{\partial \theta} \right| = o_p(1).$$

Then, putting $A + B$ together, we have that

$$\mathbb{G}_n(Y_i - f(X_i, \hat{\theta})e^{t' X_i} = \mathbb{G}_n(Y_i - f(X_i, \hat{\theta})) \left[e^{t' X_i} - b'(t, \theta_0) A^{-1} \frac{\partial f(X_i, \theta_0)}{\partial \theta_0}\right] + o_p(1). \tag{S2}$$

Step 3: Find the limiting distribution in $L_\infty(T)$: Let

$$\mathcal{F} = \left\{ (y - f(x, \theta_0) \left[e^{t' x} - b'(t, \theta_0) A^{-1} \frac{\partial f(x, \theta_0)}{\partial \theta}\right] : t \in T \right\}.$$

we want to show that $\mathcal{F}$ is Donsker. Let $f = f_t, \tilde{f} = f_{\tilde{t}}$ for $t, \tilde{t} \in T$ and $f, \tilde{f} \in \mathcal{F}$. Then

$$\begin{aligned}
|f(x) - \tilde{f}(x)| &\leq |y - f(x,\theta_0)||e^{t'x} - e^{\tilde{t}'x}| + |y - f(x,\theta_0|\|A^{-1}\frac{\partial f(x,\theta_0)}{\partial \theta}\|\|b(t,\theta_0) - b(\tilde{t},\theta_0)\| \\
&\leq |y - f(x,\theta_0)|e^{\tilde{t}'x}|x'(t-\tilde{t})| + |y - f(x,\theta_0)|\|A^{-1}\frac{\partial f(x,\theta_0)}{\partial \theta}\|\|\mathbb{E}\frac{\partial f(x_i,\theta_0)}{\partial \theta}(e^{t'x} - e^{\tilde{t}'x})\| \\
&\lesssim |y - f(x,\theta_0)|\|t - \tilde{t}\| + |y - f(x,\theta_0)|\|\mathbb{E}|\frac{\partial f(x_i,\theta_0)}{\partial \theta_0}|\|\|t - \tilde{t}\| \\
&\lesssim \underbrace{|y - f(x,\theta_0)|\left[1 + \left\|\mathbb{E}\left|\frac{\partial f(x_i,\theta}{\partial \theta}\right|\right\|\right]}_{F(x,y)}\|t - \tilde{t}\|
\end{aligned}$$

so $\mathcal{F}$ is Lipschitz in $t \in T$. We can show that

$$\mathcal{N}_{[]}(2\epsilon\|F\|_{P,2}, \mathcal{F}, L_2(P)) \leq \mathcal{N}(\epsilon, T, \|\cdot\|) \leq \left(\frac{d \cdot \text{diam}(T)}{\epsilon}\right)^d.$$

so the uniform entropy condition needed for the bracketing Donsker Theorem, Theorem 3.5 is easily satisfied

$$\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon < \infty.$$

We conclude that

$$\mathbb{G}_n(Y_i - f(X_i,\hat{\theta}))e^{t'X_i} \rightsquigarrow \mathbb{G}(T). \tag{S3}$$

for a tight Gaussian process on $L_\infty(T)$.

Step 4: Find the asymptotic distribution of the test statistic T_n. By continuous mapping theorem

$$T_n \rightsquigarrow \max_T \|\mathbb{G}(t)\|.$$

However, using the result in (S3) we can generate any number of test statistics. For example, consider

$$\tilde{T}_n = \int_T \mathbb{P}_n(Y_i - f(X_i,\hat{\theta})e^{t'X_i})^2\, dt \rightsquigarrow \int \mathbb{G}^2(t)\, dt = \|\mathbb{G}\|_{P,2}^2.$$

Step 5: Consider the behavior under the alternative.

If the process is Glivenko-Cantelli even under the alternative, then

$$\begin{aligned}
\sup_T \left|\mathbb{P}_n(Y_i - f(X_i,\hat{\theta}))e^{t'X_i} - \mathbb{E}(Y_i - f(X_i,\theta_0))e^{t'X_i}\right| &\leq \sup_T \left|\mathbb{E}(f(X_i,\theta_0) - f(X_i,\hat{\theta}))e^{t'X_i}\right| + o_p(1) \\
&\lesssim \left|\mathbb{E}|f(X_i,\theta_0) - f(X_i,\hat{\theta}))\right| = o_p(1)
\end{aligned}$$

So the function of t will converge uniformly in $L_\infty(T)$

$$\mathbb{P}_n(Y_i - f(X_i,\hat{\theta}))e^{t'X_i} \to_p \mathbb{E}(Y_i - f(X_i,\theta_0)e^{t'X_i}.$$

By continous mapping theorem,

$$T_n = \sqrt{n}\max_T \|\mathbb{P}_n(Y_i - f(X_i,\hat{\theta}))e^{t'X_i}\| \to_p \infty.$$

so the test is consistent.

4 Delta Method and Applications to Statistics

We now look to apply the results of Section 3 to the Delta Method and other statistical problems.

4.1 Multiplier Central Limit Theorems

This section follows Chapter 2.9 in Van der Vaart and Wellner.

With the notation $Z_i = \delta_{X_i} - P$, the empirical central limit theorem can be written

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \rightsquigarrow \mathbb{G}.$$

where $\mathbb{G}$ is a tight stochastic process on $\ell^\infty(\mathcal{F})$. In contrast, given i.i.d real-valued random variables $\xi_1, \dots, \xi_n$ which are independent of $Z_1, \dots, Z_n$, a multiplier central limit theorem would assert that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i Z_i \rightsquigarrow \mathbb{G}.$$

A deeper result is a *conditional multiplier central limit theorem*, which asserts that for almost every sequence $Z_1, Z_2, \dots$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i Z_i \rightsquigarrow \mathbb{G}.$$

To establish these results we will make use of symmetrization results of the sort seen in Lemma 3.6. First define the "2-1" norm:[1]

Definition 4.1 (2-1 Norm)**.** For a random variable $X : \Omega \to \mathbb{R}$ define

$$\|X\|_{2,1} = \int_0^\infty \sqrt{\Pr\left(|X| > x\right)}\, dx.$$

Lemma 4.1 (Generalized Symmetrization)**.** *Let $Z_1, \dots, Z_n$ be independent stochastic processes with mean zero and let $\epsilon_1, \epsilon_2, \dots$ be independently generated Rademacher random variables.*[2] *Then:*

$$\mathbb{E}^\star \Phi\left(\frac{1}{2}\left\|\sum_{i=1}^{n} \epsilon_i Z_i\right\|_{\mathcal{F}}\right) \le \mathbb{E}^\star\left(\left\|\sum_{i=1}^{n} Z_i\right\|_{\mathcal{F}}\right) \le \mathbb{E}^\star\left(2\left\|\sum_{i=1}^{n} \epsilon_i (Z_i - \ell_i)\right\|_{\mathcal{F}}\right).$$

for every nondecreasing, convex $\phi : \mathbb{R} \to \mathbb{R}$ and arbitrary functions $\ell_i : \mathcal{F} \to \mathbb{R}$.

Lemma 4.2 (Donkser Implication)**.** *Let $Z_1, Z_2, \dots$ be i.i.d stochastic processes such that $\sqrt{n}\sum_{i=1}^n Z_i$ converges weakly in $\ell^\infty(\mathcal{F})$ to a tight Gaussian process. Then*

$$\lim_{x \to \infty} x^2 \sup_n \mathbb{P}^\star\left(\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n} Z_i\right\|_{\mathcal{F}} > x\right) = 0.$$

In particular, the random variable $\|Z_1\|_{\mathcal{F}}^\star$ possesses a weak second moment.

Lemma 4.3 (Alternative Weak Convergence Characterizations)**.** *Let $Z_1, Z_2, \dots$ be i.i.d stochastic processes, linear in f. Set $\rho_Z(f,g) = \mathrm{Var}_Z(f-g)$ and $\mathcal{F}_\delta = \{f - g : \rho_Z(f,g) < \delta\}$. Then the following statements are equivalent and imply that the sequence $\mathbb{E}^\star\|n^{-1/2}\sum_{i=1}^n Z_i\|_{\mathcal{F}}^r$ converges to $\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}}^r$ for every $0 < r < 2$:*

1. *$n^{-1/2}\sum_{i=1}^n Z_i$ converges weakly to a tight limit in $\ell^\infty(\mathcal{F})$;*

[1] I have no idea what this is actually called.

[2] Can basically just think of an independent stochastic process as independent data. Each data point represents a random functional on $\mathcal{F}$ (evaluate each function $f \in \mathcal{F}$ at Z_i).

2. $(\mathcal{F}, \rho_Z)$ *is totally bounded*[3] *and* $\|n^{-1/2}\sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \to_{p^\star} 0$ *for every* $\delta_n \downarrow 0$*;*

3. $(\mathcal{F}, \rho_Z)$ *is totally bounded and* $\mathbb{E}^\star\|n^{-1/2}\sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \to 0$.

Lemma 4.4 (Multiplier Inequalities)**.** *Let* $Z_1, \ldots, Z_n$ *be i.i.d stochastic processes with* $\mathbb{E}^\star\|Z_i\|_\mathcal{F} < \infty$ *independent of the Rademacher random variables* $\epsilon_1, \ldots, \epsilon_n$*. Then, for every i.i.d sample* $\xi_1, \ldots, \xi_n$ *of mean-zero symmetric random variables independent of* $Z_1, \ldots, Z_n$ *and any* $1 \le n_0 \le n$*:*

$$\begin{aligned}\|\xi\|_1\mathbb{E}^\star\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i Z_i\right\|_\mathcal{F} &\le \mathbb{E}^\star\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i Z_i\right\|_\mathcal{F} \\ &\le (n_0-1)\mathbb{E}^\star\|Z_1\|_\mathcal{F}\mathbb{E}\max_{1\le i\le n}\frac{|\xi_i|}{\sqrt{n}} \\ &\quad + \|\xi\|_{2,1}\max_{n_0\le k\le n}\mathbb{E}^\star\left\|\frac{1}{\sqrt{k}}\sum_{i=n_0}^k \epsilon_i Z_i\right\|_\mathcal{F}\end{aligned}$$

These lemmas are used to show the following theorem:

Theorem 4.1 (Unconditional Multiplier Central Limit Theorem)**.** *Let* $\mathcal{F}$ *be a class of measurable functions. Let* $\xi_1, \ldots, \xi_n$ *be i.i.d symmetric random variables with mean zero, variance one, and* $\|\xi\|_{2,1} < \infty$*, independent of* $X_1, \ldots, X_n$*. Then the sequence* $n^{-1/2}\sum_{i=1}^n \xi_1(\delta_{X_i} - P)$ *converges to a tight limiting process in* $\ell^\infty(\mathcal{F})$ *if and only if* $\mathcal{F}$ *is Donsker.*

Proof. Since we can replace any $f \in \mathcal{F}$ with $f - Pf$ without changing the value of either the original or multiplier empirical process, it can be assumed without loss of generality that $Pf = 0$ for every f. Marginal convergence of both sequences is equivalent to $\mathcal{F} \subset L_2(P)$. In light of Theorem 2.5 and Theorem 2.6, it suffices to show that the asymptotic equicontinuity results for the empirical and multiplier processes are equivalent.

If $\mathcal{F}$ is Donsker then $\Pr^\star(F > x) = o(x^{-2})$ by Lemma 4.2. By the same lemma, convergence of the multiplier process to a tight limit implies that $\Pr^\star(|\xi F| > x) = o(x^{-2})$. In particular, $\mathbb{P}^\star F < \infty$ in both cases.

Since $\|\xi\|_{2,1} < \infty$ implies the existence of a second moment, we have that $\mathbb{E}^\star \max_{1\le i\le n}|\xi_i|/\sqrt{n} \to 0$ by Markov's Inequality. Applying Lemma 4.4 gives

$$\begin{aligned}\|\xi\|_1 \limsup_{n\to\infty}\mathbb{E}^\star\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i Z_i\right\|_{F_\delta} &\le \limsup_{n\to\infty}\mathbb{E}^\star\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i Z_i\right\|_{\mathcal{F}_\delta} \\ &\le \|\xi\|_{2,1}\sup_{n_0\le k}\mathbb{E}^\star\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i Z_i\right\|_{\mathcal{F}_\delta}\end{aligned}$$

for every n_0 and $\delta > 0$. By Lemma 4.1 we can remove the Rademacher variables ϵ_i in this statement at the cost of changing the constants. This gives us that $\mathbb{E}^\star\|n^{-1/2}\sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \to 0$ if and only if $\mathbb{E}^\star\|n^{-1/2}\sum_{i=1}^n \xi_i Z_i\|_{\mathcal{F}_{\delta_n}} \to 0$. By Lemma 4.3 this is equivalent to asymptotic equicontuity and weak convergence. □

Corollary 4.1 (Unconditional Multiplier Central Limit Theorem)**.** *Let* $\mathcal{F}$ *be Donsker with* $\|P\|_\mathcal{F} < \infty$*. Let* $\xi_1, \ldots, \xi_n$ *be i.i.d random variables with mean* μ*, variance* σ^2 *and* $\|\xi_i\|_{1,2} < \infty$ *generated independently of* $X_1, \ldots, X_n$*. Then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (\xi_i\delta_{X_i} - \mu P) \rightsquigarrow \mu\mathbb{G} + \sigma\mathbb{G}' + \sigma ZP.$$

where $\mathbb{G}$ *and* $\mathbb{G}'$ *are independent (tight) processes on* $\ell^\infty(\mathcal{F})$ *and are both independent of* $Z \sim N(0,1)$*. The limiting process* $\mu\mathbb{G} + \sigma\mathbb{G}' + \sigma ZP$ *is a mean zero Gaussian process*[4] *and covariance function* $(\sigma^2 + \mu^2)Pfg - \mu^2(Pf)(Pg)$.

[3] See Definition 1.17.

[4] Mean Zero means that $(\mu\mathbb{G} + \sigma\mathbb{G} + \sigma ZP)f = 0$ for all $f \in P$, Gaussian means that the marginals are normally distributed.

We now want to show a conditional version of this result. That is, we want to show that we have weak convergence of the multiplier central limit theorem for almost all sequences $Z_1, \ldots, Z_n$.[5] For finite $\mathcal{F}$, this sort of result is a simple consequence of the Lindeberg central limit theorem. Before getting into it, it is useful to recall the central limit theorem as it applies to independent, but not necessarily identically distributed, data.

Theorem 4.2 (Lindeberg Central Limit Theorem)**.** *Suppose that $X_1, X_2, \ldots$ is a sequence of independent random vectors in $\mathbb{R}^k$ with $\mathbb{E}[X_i] = \mu_i$. Define $V_n = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i)$ and suppose the following* Lindeberg Condition *is satisfied:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|X_i - \mu_i\|^2 \mathbb{1}\{\|X_i - \mu_i\| > \epsilon\sqrt{n}\}\right] = 0, \qquad \forall \epsilon > 0 \tag{LC}$$

Then

$$V_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu_i) \rightsquigarrow N\left(0, I_k\right).$$

Lemma 4.5 (Conditional Multiplier CLT for Finite Classes)**.** *Let $Z_1, Z_2, \ldots$ be i.i.d random vectors with $\mathbb{E}Z_i = 0$ and $\mathbb{E}\|Z_i\|^2 < \infty$ independent of the i.i.d sequence $\xi_1, \xi_2, \ldots$ with $\mathbb{E}\xi_i = 0$ and $\mathbb{E}\xi_i^2 = 1$. Then, conditionally on $Z_1, Z_2, \ldots$,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i Z_i \rightsquigarrow N\left(0, \mathrm{Var}(Z_1)\right),$$

for almost every sequence $Z_1, Z_2, \ldots$

Proof. Treating $Z_1, Z_2, \ldots$ as just a stream of constant vectors, by the Lindeberg Central Limit Theorem, the statement is true for every sequence $Z_1, Z_2, \ldots$ such that, for every $\epsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^n Z_i Z_i' \to \mathrm{Var}(Z_1) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \|Z_i\|^2 \mathbb{E}_\xi\left[\xi_i^2\left\{|\xi_i|\|Z_i\| > \epsilon\sqrt{n}\right\}\right].$$

By Kolmogrov Strong Law of Large Numbers, the first statement is true for almost all all sequences $\{Z_i\}_{i=1}^\infty$. A finite second moment $\mathbb{E}\|Z_i\|^2 < \infty$, implies that $\max_{1\le i\le n}\|Z_i\|/\sqrt{n} \to 0$ for almost all sequences, which gives that the second statement holds for almost all sequences $\{Z_i\}_{i=1}^\infty$. Under a probability measure, the intersection of two (measurable) sets with measure 1 also has measure 1. □

Lemma 4.5 provides the weak convergence of marginals in the multiplier processes. What remains is to show some version of asymptotic equicontinuity (Theorem 2.6 or 2.7) to show weak convergence in $\ell^\infty(\mathcal{F})$.

Let BL_1 be the set of all bounded Lipschitz functions. That is the set of all functions $h : \ell^\infty(\mathcal{F}) \to [0,1]$ such that $|h(z_1) - h(z_2) \le \|z_1 - z_2\|_\mathcal{F}$ for every z_1, z_2. Using the bounded Lipschitz functions, we can define a metric (the bounded Lipschitz metric) between two distribtutions on a space $\mathbb{D}$.

$$d_{\mathrm{BL}}(L_1, L_2) = \sup_{f\in \mathrm{BL}_1}\left|\inf f\, dL_1 - \int f\, dL_2\right|.$$

It turns out that weak convergence is equivalent to convergence in the bounded Lipschitz metric.

Theorem 4.3 (Weak Convergence and the Bounded Lipschitz Metric)**.** *Weak convergence of separable (Definition 3.10) Borel probability measures on a metric space $\mathbb{D}$ corresponds to convergence in a topology that is metrizable by the bounded Lipschitz metric.*

The following theorems (presented without proof) gives conditions for convergence of the multiplier central limit theorem. First the other probability version is given under the same conditions as Theorem 4.1 then the almost sure version is given under only slightly stronger conditions.

[5] And for this to be useful for bootstrap, we'd like the weak limit to be the same as the (non multiplier) empirical process.

Theorem 4.4 (Conditional Multiplier Central Limit Theorem)**.** *Let $\mathcal{F}$ be a class of measurable functions. Let $\xi_1, \ldots, \xi_n$ be i.i.d random variables with mean zero, variance one and $\|\xi\|_{2,1} < \infty$, independent of $X_1, \ldots, X_n$. Let $\mathbb{G}'_n = n^{1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - p)$. Then the following assertions are equivalent:*

1. *$\mathcal{F}$ is Donsker;*
2. *$\sup_{h \in BL_1} \left| \mathbb{E}_\xi h(\mathbb{G}'_n) - \mathbb{E}h(\mathbb{G}) \right| \to 0$ in outer probability and the sequence $\mathbb{G}'_n$ is asymptotically measurable.*

Theorem 4.5 (Conditional Mulitplier Central Limit Theorem)**.** *Let $\mathcal{F}$ be a class of measurable functions. Let $\xi_1, \ldots, \xi_n$ be i,i.d random variables with mean zero, variance 1, and $\|\xi\|_{2,1} < \infty$, independent of $X_1, \ldots, X_n$. Define the multiplier process $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$. Then the following assertions are equivalent:*

1. *$\mathcal{F}$ is Donsker with $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ and $P^\star \|f - Pf\|_{\mathcal{F}}^2 < \infty$;*
2. *$\sup_{h \in BL_1} \left| \mathbb{E}_\xi h(\mathbb{G}'_n) - \mathbb{E}h(\mathbb{G}) \right| \to 0$ outer almost surely and the sequence $\mathbb{E}_\xi h(\mathbb{G}'_n)^\star - \mathbb{E}_\xi h(\mathbb{G})_\star$ converges to zero almost surely for every $h \in BL_1$.*

Here $h(\mathbb{G}'_n)^\star$ and $h(\mathbb{G}'_n)_\star$ denote measurable majorants and minorants with respect to $(\xi_1, \ldots, \xi_n, X_1, \ldots, X_n)$ jointly.

4.2 The Emprirical Bootstrap

Section 4.1 gives results that are useful for establishing the consistency of the multiplier bootstrap. We now quickly describe the empirical bootstrap and give some results. Let $\mathbb{P}_n$ be the empirical measure of an i.i.d sample $X_1, \ldots, X_n$ from a probability measure P. Given the sample values, let $\hat{X}_1, \ldots, \hat{X}_n$ be an i.i.d sample from $\mathbb{P}_n$. The *bootstrap empirical distribution* is the empirical measure $\hat{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n \delta_{\hat{X}_i}$ and the *bootstrap empirical process* $\hat{\mathbb{G}}_n$ is given

$$\hat{\mathbb{G}}_n = \sqrt{n} \left(\hat{\mathbb{P}}_n - \mathbb{P}_n \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1) \, \delta_{X_i}.$$

where M_{ni} is the number of times that X_i is redrawn from the original samples. We can also define a bootstrap empirical process where we draw k bootstrap values, $\hat{X}_1, \ldots, \hat{X}_k$. The corresponding bootstrap empirical process is

$$\hat{\mathbb{G}}_{n,k} = \sqrt{k} \left(\hat{\mathbb{P}}_k - \mathbb{P}_n \right) = \frac{1}{\sqrt{k}} \sum_{i=1}^n \left(M_{ki} - \frac{k}{n} \right) \delta_{X_i}.$$

In either case, it is important that the vector $(M_{k1}, \ldots, M_{kn})$ is independent of $X_1, \ldots, X_n$, that is multinomially distributed with parameters k and probabilities $\frac{1}{n}, \ldots, \frac{1}{n}$ for any random sample of size n.

Let $\mathbb{E}_M$ denote the expectation with respect to the distribution of $(M_{k1}, \ldots M_{kn})$. In light of Theorem 4.3 which equates weak convergence of a sequence of probability measures to convergence in the bounded Lipschitz metric, the following theorem establishes the consistency of the bootstrap.

Theorem 4.6 (Consistency of the Empirical Bootstrap)**.** *Let $\mathcal{F}$ be a Donsker class of measurable functions such that $\mathcal{F}_\delta$ is measurable for every $\delta > 0$. Then*

$$\sup_{h \in BL_1} \left| \mathbb{E}_M h(\hat{\mathbb{G}}_{n,k_n}) - \mathbb{E}h(\mathbb{G}) \right| \overset{P^\star}{\to} 0.$$

as $n \to \infty$ for any sequence $k_n \to \infty$. Furthermore the sequence $\mathbb{E}_M h(\hat{\mathbb{G}}_{n,k_n})^\star - \mathbb{E}_M h(\hat{\mathbb{G}}_{n,k_n})_\star$ converges to zero in probability for every $h \in BL_1$. If $\mathbb{P}^\star \|f - Pf\|_{\mathcal{F}}^2 < \infty$, then the convergence is also outer almost surely.

4.3 Delta Method

We now cover the standard Delta Method (for "fully differentiable" functions). First we want to define some more general notions of differentiability.

4.3.1 Differentiability

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at a point x_0 if the limit

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x)}{h}.$$

exists. In this case we call the value of the limit the derivative of f at x_0 and denote this as $f'(x_0)$. The derivative is useful as it gives a linear approximation of f in a neighborhood of x_0, that is the derivative can instead be written as a scalar $f'(x_0)$ such that

$$\lim_{h \to 0} \frac{|f(x + h) - f(x) - f'(x)h|}{h} = 0.$$

or, more familiarly, $f(x + h) - f(x) = f'(x)h + o(h)$. This linear approximation property is the useful bit that we will use in econometrics, so we want notions of derivatives in general spaces to reflect this.

Example. Suppose $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{L} N(0, \sigma^2)$. We want to get the asymptotic distribution of $\hat{\theta}^2 - \theta_0^2$. We'll use the derivative of $f(x) = x^2; f'(x) = 2x$ evaluated at θ_0 to say that

$$\sqrt{n}\left(\hat{\theta}^2 - \theta_0^2\right) = 2\theta_0\sqrt{n}\left(\hat{\theta} - \theta_0\right) + o_p(1),$$

by taking $h = \hat{\theta} - \theta_0$. Now note that we know the distribution of $2\theta_0\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{L} N(0, 4\theta - 0^2\sigma^2)$. We have leveraged the linearity property.

Let's start by generalizing this property to functions in $\mathbb{R}^k$. This will allow us to differentiate between fully differentiable vs directionally differentiable functions, an important distinction later on. All the definitions below will reflect the linearity property that we desire.

Definition 4.2 (Differentiabiltiy in $\mathbb{R}^k$)**.** A function $f : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at $x_0 \in \mathbb{R}^k$ if there exists a linear transformation $f'(x_0) : \mathbb{R}^k \to \mathbb{R}^m$ such that[1]

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - f'(x)h\|}{\|h\|} = 0. \tag{4.1}$$

Note that this is very similar to our notion of differentiability from before. Let's see how this contrasts with our familiar notion of partial differentiability and consider a function that has a gradient (i.e partial derivatives in the standard unit vector directions) but that we would not consider fully differentiable according to Definition 4.2.

Example (Directionally but not Fully Differentiable)**.** Consider the function $f : \mathbb{R}^k \to \mathbb{R}$ given by:

$$f(x, y) = \begin{cases} \frac{x^2}{x+y} & \text{if } (x, y) \neq (, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

Let's consider the partial derivatives at $(0, 0)$. Note that $f(x, 0) = x$ and $f(0, y) = 0$ so that

$$\frac{\partial f}{\partial x}(0, 0) = 1 \quad \text{and} \quad \frac{\partial f}{\partial y}(0, 0) = 0 \implies \nabla f(0, 0) = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

We might be tempted then to say that the gradient $\nabla f(0, 0)$ satisfies the requirements of the linear map $f'(0, 0)$ in Definition 4.2. However, let's consider approaching $(0, 0)$ in the direction (h, h) for a real number $h \to 0$. Note that $f(h, h) = \frac{h}{2}$. Let's consider the limit in this direction

$$\lim_{h \to 0} \frac{|f(h, h) - f(0, 0) - \nabla f(0, 0) \cdot (h, h)|}{|(h, h)|} = \lim_{h \to 0} \frac{|h/2 - h|}{|\sqrt{2}h|} = \frac{1}{2\sqrt{2}} \neq 0.$$

So the gradient does not satisfy the conditions of Definition 4.2. In fact, no linear map will and so that function is not differentiable at $(0, 0)$.

[1] That is there is a $m \times k$ matrix A_{x_0}

Remark. The notion of differentiability that we want requires the *same* linear approximation to work uniformly (in all directions). But the partial derivatives that we are used to only look in one direction at a time. The existence of a gradient is necessary but not sufficient for Definition 4.2.

Let's generalize these definitions to general spaces. In general, let $(A, \|\cdot\|_A)$ and $(B, \|\cdot\|_B)$ be Banach spaces (complete normed vector spaces).[2]

Definition 4.3 (Fréchet Differential)**.** We say a function $f : A \to B$ is Fréchet differentiable at a point $x_0 \in A$ if there exists a continuous linear function $f'_{x_0} : A \to B$ such that

$$\lim_{\|h\|_A \to 0} \frac{\left\|f(x_0 + h) - f(x_0) - f'_{x_0}(h)\right\|_B}{\|h\|_A} = 0.$$

If this is the case we call the linear map f'_{x_0} the *Fréchet Differential* at x_0. We can alternatively formulate this

$$\lim_{\epsilon \to 0} \sup_{h \in S} \frac{\|f(x + \epsilon_h) - f(x) - \epsilon f'_{x_0}(h)\|_b}{\epsilon} = 0.$$

for all bounded (finite diameter) sets $S \subset A$.

Example (Fréchet Differential)**.** Let $(A, \|\cdot\|_\infty) = (L^\infty, \|\cdot\|_\infty)$ and $(B, \|\cdot\|_B) = (\mathbb{R}, |\cdot|)$. For a point $t_0 \in T$ and a function $x : T \to \mathbb{R} \in L^\infty$, let $f(x) = x(t_0)^2$. Fix x and consider the linear map on L^∞ into $\mathbb{R}$, $f'_x(h) = 2x(t_0)h(t_0)$. Then

$$\begin{aligned}
\lim_{\|h\|_A \to 0} \frac{\left\|f(x + h) - f(x) - f'_x(h)\right\|_B}{\|h\|_A} &= \lim_{\|h\| \to 0} \frac{\left|\left(x(t_0) + h(t_0)^2\right) - x(t_0)^2 - 2x(t_0)h(t_0)\right|}{\|h\|_\infty} \\
&= \lim_{\|h\|_\infty \to 0} \frac{h(t_0)^2}{\|h\|_\infty} \\
&\leq \lim_{\|h\|_\infty \to 0} \|h\|_\infty = 0
\end{aligned}$$

This generalizes the concept of a fully differentiable function, but what about directionally differentiable functions? For those we consider the Gateaux Differential.

Definition 4.4 (Gateaux Differential)**.** A function $f : A \to B$ is Gateaux differentiable at the point $x_0 \in A$ in the direction $h \in A$ if there exists a linear map $\Gamma_{x_0} : A \to B$ such that

$$\lim_{\epsilon \to 0} \frac{\left\|f(x + \epsilon h) - f(x) - \epsilon \Gamma_x(h)\right\|_B}{\epsilon} = 0.$$

In this case we call the map $\Gamma_x(\cdot)$ the *Gateaux Differential* in the direction h and denote $df(x_0; h) = \Gamma_x(\cdot)$.

Note that this is defined in each direction as opposed the Fréchet differential which is defined uniformly for all directions. Whenever the Fréchet differential exists the Gateaux differential will exist and coincide with the Fréchet.

The definitions of Fréchet differentiability is not quite what we need however. The following refinement becomes more useful to applications in econometrics.

Definition 4.5 (Hadamard Differential)**.** The function $f : A \to B$ is Hadamard differentiable at the point $x_0 \in A$ if there exists a continuous linear function $f'_{x_0} : A \to B$ with

$$\lim_{\epsilon \to 0} \sup_{h \in S} \frac{\left\|f(x + \epsilon h) - f(x) - \epsilon f'_{x_0}(h)\right\|_B}{\epsilon} = 0.$$

for all compact sets $S \subset A$. If this is the case the continuous linear function f'_{x_0} is called the *Hadamard Differential* at x_0.

The key idea here is that tight random variables concentrate on compact sets, so Hadamard is what we need.

[2] In general, I think we only need metrizable topological spaces. This is what is said in Andres' notes and in Van Der Vaart and Wellner. However, all the definitions below are given in terms of norms so to avoid confusion we'll just assume these are complete normed spaces.

4.3.2 Standard Delta Method

We are now ready to review the Delta Method. This discussion follows Andres' notes as well as Chapter 3.9 in Van der Vaart and Wellner.

First recall some useful theorems.

Theorem 4.7 (Continuous Mapping Theorem)**.** *Suppose $g_n : D \to E$ is a sequence of continuous maps with $g_n(x_n) \to g(x)$ for some continuous map g and every convergent sequence $x_n \to x$ with $x \in D_0$ and $\Pr(X \in D_0) = 1$. Then if $X_n \xrightarrow{L} X$ in D then $g_n(X_n) \xrightarrow{L} g(X)$ in E.*

This is a slight refinement of the continuous mapping theorem seen in Theorem 2.3, allowing for sequences of continuous maps.

Lemma 4.6 (Convergent Sequences are Compact)**.** *If a sequence $\{x_n\}$ converges to a point x, then the set $\{x, x_1, x_2, \dots\}$ is compact (in any topological space).*

Proof. Let $\{U_i\}_{i\in I}$ be an open cover of $S = \{x, x_1, x_2, \dots\}$. Pick a set U_x in $\{U_i\}_{i\in I}$ such that $x \in U_x$. This is an open neighborhood of x so there must exist a number N such that for all $n \geq N$, $x_n \in U_x$. For the finitely many points outside of U_x we can find sets in $\{U_i\}_{i\in I}$ that contain them. □

We are now ready to show the Delta Method.

Theorem 4.8 (Delta Method)**.** *Let D and E be Banach Spaces and $\phi : D \to E$ be Hadamard differentiable at θ_0 and suppose $\sqrt{n}(X_n - \theta_0) \xrightarrow{L} X$ in D. Then $\sqrt{n}\left(\phi(X_n) - \phi(\theta_0)\right) \xrightarrow{L} \phi'_{\theta_0}(X)$ where ϕ'_{θ_0} is the Hadamard Differential at θ_0.*

Proof. The goal will be to apply Theorem 4.7. Let $g_n(h) = \sqrt{n}\left(\phi(\theta_0 + \frac{h}{\sqrt{n}}) - \phi(\theta_0)\right)$ and let $g(h) = \phi'_{\theta_0}(h)$ and suppose $h_n \to h$. We want to show that $g_n(h_n) \to g(h)$.

$$\lim_{n\to\infty} |g_n(h_n) - g(h)| = \lim_{n\to\infty} \left|\sqrt{n}\left(\phi\left(\theta_0 + h_n/\sqrt{n}\right) - \phi(\theta_0)\right) - \phi'_{\theta_0}(h)\right|$$

Fix $\epsilon_n = 1/\sqrt{n}$ and rewrite

$$\begin{aligned}
&= \lim_{n\to\infty} \left|\frac{1}{\epsilon_n}\left[\phi(\theta_0 + h_n\epsilon_n) - \phi(\theta_0) - \epsilon_n\phi'_{\theta_0}(h)\right]\right| \\
&\leq \lim_{n\to\infty} \left|\frac{1}{\epsilon_n}\left[\phi\left(\theta_0 + h_n\epsilon_n\right) - \phi(\theta_0) - \epsilon_n\phi'_{\theta_0}(h_n)\right]\right| + \underbrace{\left|\phi'_{\theta_0}(h_n) - \phi'_{\theta_0}(h)\right|}_{\to 0 \text{ by continuity of } \phi'_{\theta_0}}
\end{aligned}$$

Since the last term goes to 0, consider only the first term and let $S = \{h, h_1, h_2, \dots\}$:

$$\leq \lim_{\epsilon_n\to 0} \sup_{\tilde{h}\in S} \left|\frac{1}{\epsilon_n}\left[\phi(\theta_0 + \tilde{h}\epsilon_n) - \phi(\theta_0) - \epsilon_n\phi'_{\theta_0}(\tilde{h})\right]\right|$$

By Lemma 4.6 the set S is compact and so this goes to zero because ϕ is Hadamard differentiable at θ_0. We can now apply Theorem 4.7 to show that $g_n\left(\sqrt{n}(X_n - \theta_0)\right) \xrightarrow{L} g(X)$. Plugging in we see that

$$\begin{aligned}
g_n\left(\sqrt{n}(X_n - \theta_0)\right) &= \sqrt{n}\left(\phi\left(\theta_0 + \frac{\sqrt{n}(X_n - \theta_0)}{\sqrt{n}}\right) - \phi(\theta_0)\right) \\
&= \sqrt{n}\left(\phi(X_n) - \phi(\theta_0)\right) \\
&\xrightarrow{L} \phi'_{\theta_0}(X)
\end{aligned}$$

□

4.3.3 Delta Method Examples

We now turn to some examples to show the usefulness of Theorem 4.8.

Example 4.1 (Regression Coeffecients)**.** Suppose $\beta_0 \in \mathbb{R}^k$ and $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{L} N(0, \Sigma)$. Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be differentiable.[3] Then $\sqrt{n}(\phi(\hat{\beta}) - \phi(\beta)) \xrightarrow{L} \phi'_{\beta_0}(Z)$ where $Z \sim N(0, \Sigma)$.

What does this mean? Recall that any continuous linear map from $\mathbb{R}^k$ to $\mathbb{R}^m$ can be expressed as a matrix (an element of $\mathbb{R}^{m\times k}$). Specifically we can write $\phi = (\phi_1, \ldots, \phi_m)'$, where each $\phi_i : \mathbb{R}^k \to \mathbb{R}$. Then $\phi'_{\beta_0} = \left(\nabla\phi_1(\beta_0), \ldots, \nabla\phi_m(\beta_0)\right)' \in \mathbb{R}^{m\times k}$. Then

$$\phi'_{\beta_0}(Z) \sim N\left(0, \phi'_{\beta_0}\Sigma\phi_{\beta_0}\right).$$

Example 4.2 (Uniform Semi-Parametric Inference)**.** Suppose $Y = m(X, \beta_0) + \epsilon$ where $\mathbb{E}[\epsilon|X] = 0$ and $\beta \in \mathbb{R}^K$ and $\{Y_i, X_i\}$ an i.i.d sample. Under usual assumptions we will get that $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{L} N(0, \Sigma)$. To forecast $\mathbb{E}[Y|X = x_0] = m(x_0, \beta_0)$ we want to use $m(x_0, \widehat{\beta})$. If $m(x_0, \cdot)$ is differentiable we can use Delta Method (Theorem 4.8) to show that $\sqrt{n}(m(x_0, \widehat{\beta}) - m(x_0, \beta_0)) \xrightarrow{L} \nabla_\beta m(x_0, \beta_0)Z$ where $Z \sim N(0, \Sigma)$.

This is good for inference at a single point x_0, but what if we want a uniform confidence interval for $E[Y|X = x]$ for all points x in some set X. Let $\phi : \mathbb{R}^k \to L^\infty(X)$ be given by $\phi(\beta) = m(\cdot, \beta)$. That is, for each β we can give an $m(x, \beta)$ for each point $x \in X$. This defines a function on $L^\infty(X)$. Assume that $\sup m(\cdot, \cdot)$ and $\nabla_\beta m(\cdot, \cdot)$ are continuous and X is compact. Our guess for ϕ'_{β_0} is just $\phi'_{\beta_0}(\beta) = \nabla_\beta m(\cdot, \beta_0)\beta$.

1. $\phi'_{\beta_0} : \mathbb{R}^k \to L^\infty(X)$ is just like $\phi : \mathbb{R}^k \to L^\infty$, both take in a β and return a function on X.
2. Clearly ϕ'_{β_0} is linear (in β) and if $\beta_n \to \beta$ in $\mathbb{R}^k$ then $\phi'_{\beta_0}(\beta_n) \to \phi'_{\beta_0}(\beta)$ in $L^\infty(X)$. In other words, ϕ'_{β_0} is continuous.

$$\|\phi'_{\beta_0}(\beta_n) - \phi'_{\beta_0}(\beta)\|_\infty = \sup_{x\in X} \|\nabla_\beta m(x, \beta_0)(\beta_n - \beta)\| \leq \sup_{x\in X} \underbrace{\|\nabla_\beta m(x, \beta_0)\|}_{\text{finite by compactness}} \cdot \underbrace{\|\beta_n - \beta\|}_{\to 0}$$

Given this guess for ϕ'_{β_0} lets check Hadamard Differentiability (Definition 4.5) at β_0. Let B be an arbitrary compact set in $\mathbb{R}^k$.[4]

$$\lim_{\epsilon_n\to 0} \sup_{h\in B} \frac{\|\phi(\beta_0 + \epsilon_n h) - \phi(\beta_0) - \epsilon_n\phi'_{\beta_0}(h)\|_\infty}{\epsilon_n} = \lim_{\epsilon_n\to 0} \sup_{h\in B} \frac{\left\|m(x, \beta_0 + \epsilon_n h) - m(x, \beta_0) - \epsilon_n\nabla_\beta m(x, \beta_0)h\right\|_\infty}{\epsilon_n}$$

By mean value theorem, for some $\bar{\beta}(x) \in [\beta_0, \beta_0 + \epsilon_n h]$:

$$\begin{aligned} &= \lim_{\epsilon_n\to 0} \sup_{h\in B} \sup_{x\in X} \left|\nabla_\beta m(x, \bar{\beta}(x))h - \nabla_\beta m(x, \beta_0)h\right| \\ &\leq \lim_{\epsilon_n\to 0} \sup_{h\in B} \sup_{x\in X} \|\nabla_\beta m(x, \bar{\beta}(x)) - \nabla_\beta m(x, \beta_0)\| \cdot \|h\| \end{aligned}$$

The first term on the right goes to zero uniformly for all $x \in X$ by continuity of $m(\cdot, \cdot)$ and compactness of X. The second term is bounded by compactness of B and so the whole thing goes to 0. Since $\phi : \mathbb{R}^k \to L^\infty(X)$, $\phi(\beta) = m(\cdot, \beta)$ is Hadamard Differentiable at β_0 with $\phi'_{\beta_0}(\beta) = \nabla_\beta m(\cdot, \beta_0)\beta$, the Delta Method (Theorem 4.8) gives us that

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{L} Z \implies \sqrt{n}(\phi(\widehat{\beta}) - \phi(\beta_0)) \xrightarrow{L} \phi'_{\beta_0}(Z).$$

in other words,

$$\sqrt{n}\left(m(\cdot, \widehat{\beta}) - m(\cdot, \beta_0)\right) \xrightarrow{L} \nabla_\beta m(\cdot, \beta_0)Z.$$

where the convergence is in $L^\infty(X)$, that is uniformly over $x \in X$.

[3] In $\mathbb{R}^k$ Fréchet and Hadamard differentiability are equivalent.

[4] We should note that even if we restrict β_0 to be in some set $\Theta \subset \mathbb{R}^k$, compactness is invariant to superspaces. That is, let $A \subset \Theta$ be compact with respect to the subspace topology on Θ. Then A is compact with respect to $\mathbb{R}^k$.

Example 4.3 (Uniform Standard Deviation Estimation)**.** Suppose we have a class of function of square integrable functions, $\mathcal{F} = \{f : \mathbb{R}^k \to \mathbb{R}\}$ and $X \in \mathbb{R}^k$ is a random variable such that $\mathbb{E}[f(X)] = 0$ for all $f \in \mathcal{F}$. Suppose for each $f \in \mathcal{F}$ we want to study the limiting behavior of the empirical standard deviation

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} f^2(x_i)}$$

By the delta method, since the derivative of $\sqrt{x}$ is $\frac{1}{2\sqrt{x}}$, if $\sqrt{n}(\mathbb{E}_n f^2 - \mathbb{E}f^2) \to N(0, \sigma^2(f^2))$ then

$$\sqrt{n}\left(\sqrt{\mathbb{E}_n f^2} - \sqrt{\mathbb{E}f^2}\right) \xrightarrow{L} N\left(0, \frac{1}{4}\sqrt{\sigma^2(f^2)}\right).$$

This is all good for a since function $f \in \mathcal{F}$, but suppose that we want to conduct inference uniformly over $\mathcal{F}$. Let $\mathcal{F}^2 = \{f^2 : f \in \mathcal{F}\}$. Assume that $\mathcal{F}^2$ is Donsker so that $\mathbb{G}_n \to \mathbb{G}$ for a tight element $\mathbb{G}$ in $L^\infty(\mathcal{F}^2)$. Also assume that $0 < \inf_{f\in\mathcal{F}} \mathbb{E}f^2 < \sup_{f\in\mathcal{F}} \mathbb{E}f^2 < \infty$.

Let $\phi : L^\infty(\mathcal{F}^2) \to L^\infty(\mathcal{F}^2)$ be given by $\phi(G)(f^2) = \sqrt{G(f^2)}$. It is useful here to recall that $G \in L^\infty(\mathcal{F}^2)$ is a (bounded) function from $\mathcal{F}^2 \to \mathbb{R}$. Let's consider applying ϕ to the function $\theta_0(f) = \mathbb{E}f^2$ and guess that $\phi_0'(G)(f) = \frac{1}{2\sqrt{\mathbb{E}f^2}}G(f)$, which is clearly linear in G. It is also easy to verify continuity since $\mathbb{E}f^2$ is bounded from below uniformly for $f \in \mathcal{F}$. Let's verify that this function satisfies the property required of the Hadamard Differential (Definition 4.5). Let S be a compact set in $L^\infty(\mathcal{F})$:

$$\lim_{\epsilon_n\to 0}\sup_{h\in S}\frac{\|\phi(\theta_0 + h\epsilon_n) - \phi(\theta_0) - \phi_0'(h\epsilon_n)\|}{\epsilon_n} = \lim_{\epsilon_n\to 0}\sup_{h\in S}\sup_{f^2\in\mathcal{F}^2}\frac{\left|\sqrt{\theta_0(f^2) + \epsilon_n h(f^2)} - \sqrt{\theta_0(f^2)} - \epsilon_n\frac{h(f^2)}{2\sqrt{\theta_0(f^2)}}\right|}{\epsilon_n}$$

Again applying mean value theorem, for some $\bar{\theta} \in [\theta_0, \theta_0 + \epsilon_n h]$ (containment is pointwise):

$$\begin{aligned}
&= \lim_{\epsilon_n\to 0}\sup_{h\in S}\sup_{f^2\in\mathcal{F}^2}\left|\frac{h(f^2)}{2\sqrt{\bar{\theta}_n(f^2)}} - \frac{h(f^2)}{2\sqrt{\theta_0(f^2)}}\right| \\
&\le \lim_{\epsilon_n\to 0}\sup_{h\in S}\sup_{f^2\in\mathcal{F}^2}\left|\frac{1}{2\sqrt{\bar{\theta}_n(f^2)}} - \frac{1}{2\sqrt{\theta_0(f^2)}}\right|\left|h(f^2)\right|
\end{aligned}$$

The first term goes to zero uniformly and the second term is bounded uniformly because S is compact.[5] This verifies Hadamard differentiability. Applying the Delta Method then gives us that

$$\sqrt{n}\left(\sqrt{\mathbb{E}_n f^2} - \sqrt{\mathbb{E}f^2}\right) \xrightarrow{L} \frac{\mathbb{G}(f^2)}{2\sqrt{\mathbb{E}f^2}} \quad \text{uniformly for } f \in \mathcal{F}.$$

Remark. Comments on the delta method:

- Delta method is very powerful in infinite dimensions
- Lots of examples, e.j going from uniform inference on the empirical CDF to inference on the empirical quantile process.
- Domain and rules can be complicated, though have to be careful with norms.

4.4 Directionally Differentiable Functions

The Delta Method in Theorem 4.8 works if the function ϕ is Hadamard Differentiable. However, what if we only have *directional differentiability* of ϕ? Let's first review what directionally differentiable means.

[5] Recall that $h \in S \in L^\infty(\mathcal{F}^2)$. This means that h is a bounded function from $\mathcal{F}^2$ onto $\mathbb{R}$ so that $\epsilon_n h \to 0$. Compactness of S gives uniform boundedness which then gives uniform convergence of $\epsilon_n h$ over S.

Definition 4.6 (Directional Hadamard Differential)**.** Let A and B be Banach spaces and let $\phi : A_\phi \subseteq A \to B$. The map ϕ is Hadamard directionally differentiable at $\theta_0 \in A_\phi$ tangentially to a set $A_0 \subseteq A$ if there is a continuous map $\phi'_{\theta_0} : A_0 \to B$ such that

$$\lim_{n\to\infty} \frac{\left\|\phi(\theta_0 + \epsilon_n h_n) - \phi(\theta_0) - \epsilon_n \phi'_{\theta_0}(h)\right\|_B}{\epsilon_n} = 0 \tag{4.2}$$

for all $h_n \to h \in A_0$ and $\epsilon_n \downarrow 0$. In this case, the map ϕ'_{θ_0} is called the *Hadamard directional differential* at θ_0 tangent to A_0.

Example 4.4 (Directional Differentiability)**.** Let's consider the function $\phi : \mathbb{R} \to \mathbb{R}_+, x \mapsto \max\{x, 0\}$. We want to show that this function is Hadamard directionally differentiable tangent to $\mathbb{R}$ at any point $x_0 \in \mathbb{R}$ with directional differential given by

$$\phi'_{x_0}(h) = \begin{cases} h & \text{if } x_0 > 0 \\ \max\{h, 0\} & \text{if } x_0 = 0 \\ 0 & \text{if } x_0 < 0 \end{cases}.$$

Intuitively, we can see why this would be the case. If x_0 is above zero then locally in a neighborhood around x_0, $\phi(x)$ is equal to x, if x_0 equals zero then in any neighborhood around x_0 $\phi(x)$ is equal to x if $x > 0$ and 0 otherwise, and if x_0 is less than zero then $\phi(x)$ is equal to zero uniformly in a neighborhood around x_0.

We can clearly see that this is a continuous function in h for any x_0 so what remains is to verify eq. (4.2). Take any x_0 and any sequence $h_n \to h$ and $\epsilon_n \downarrow 0$.

$$\lim_{n\to\infty} \frac{\left|\max\{x_0 + \epsilon_n h_n, 0\} - \max\{x_0, 0\} - \epsilon_n \phi'_{\theta_0}(h)\right|}{\epsilon_n} \tag{Ex-1}$$

Formally apply the intuition above: if $x_0 > 0$ eventually we will have $x_0 + \epsilon_n h_n > 0$. In this case (Ex-1) will reduce to:

$$\lim_{n\to\infty} \frac{|x_0 + \epsilon_n h_n - x_0 - \epsilon_n h|}{\epsilon_n} = \lim_{n\to\infty} h_n - h = 0.$$

If $x_0 < 0$ eventually we will have $x_0 + \epsilon_n h_n < 0$. After this point (Ex-1) reduces to:

$$\lim_{n\to\infty} \frac{\left|-\epsilon_n \phi'_{\theta_0}(h)\right|}{\epsilon_n} = 0.$$

Finally, if $x_0 = 0$ then (Ex-1) reduces to:

$$\lim_{n\to\infty} \frac{\left|\max\{\epsilon_n h_n, 0\} - \epsilon_n \max\{h, 0\}\right|}{\epsilon_n} = \lim_{n\to\infty} \left|\max\{h_n, 0\} - \max\{h, 0\}\right| = 0.$$

where in the first equality we use that ϵ_n is decreasing towards zero and in the second we use continuity of $\max\{x, 0\}$.

Theorem 4.9 (Delta Method For Directionally Differentiable Functions)**.** *Suppose $\sqrt{n}(\theta_n - \theta_0) \xrightarrow{L} X$ for a random element X in A_0. If $\phi : A \to B$ is Hadamard directionally differentiable at θ_0 tangent to A_0 with directional differential ϕ'_{θ_0} then:*

$$\sqrt{n}\left(\phi(\theta_n) - \phi(\theta_0)\right) \xrightarrow{L} \phi'_{\theta_0}(X),$$

Proof. Proof follows the same steps as Theorem 4.8. Again take $g_n(h) = \sqrt{n}\left(\phi(\theta_0 + \frac{h}{\sqrt{n}}) - \phi(\theta_0)\right)$ and $g(h) = \phi'_{\theta_0}(h)$.We want to apply Theorem 4.7 so we want to show that if $h_n \to h$, $g_n(h_n) \to g(h)$ for any

sequence h_n converging to an $h \in A_0$. Letting $\epsilon_n = 1/\sqrt{n}$ and following verbatim the first two lines in the proof of Theorem 4.8 gives us that

$$\lim_{n\to\infty} \left|g_n(h_n) - g(h)\right| = \lim_{n\to\infty} \frac{\left|\phi(\theta_0 + \epsilon_n h_n) - \phi(\theta_0) - \epsilon_n \phi'_{\theta_0}(h)\right|}{\epsilon_n}$$

Applying the definition of Hadamard directional differentiability we see that this term goes to zero as $\epsilon_n \to 0$ and $h_n \to h$. Rest of the proof follows exactly that of Theorem 4.8. □

Example 4.5 (Delta Method for Directionally Differentiable Functions)**.** Suppose $\mathbb{E}[X^2] < \infty$ and we have an i.i.d sample from X, $X_1, X_2, \ldots$ with $X_i \sim X$. Let $\theta_0 = \mathbb{E}[X]$ and $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. By the Lindeberg Central Limit Theorem (Theorem 4.2) we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N(0, \sigma_X^2).$$

Suppose are interested in conducting inference on the quantity $\phi(\theta_0) = \max\{\theta_0, 0\}$. In Example 4.4 we verified that $\phi(\cdot)$ is directionally differentiable everywhere tangent to $\mathbb{R}$ with Hadamard directional differential at (arbitrary point) x_0:

$$\phi'_{x_0}(h) = \begin{cases} h & \text{if } x_0 > 0 \\ \max\{h, 0\} & \text{if } x_0 = 0 \\ 0 & \text{if } x_0 < 0 \end{cases}.$$

So we can apply Theorem 4.9 to get that

$$\sqrt{n}\left(\max\{\hat{\theta}_n, 0\} - \max\{\theta_0, 0\}\right) \xrightarrow{L} \phi'_{\theta_0}\left(\sigma_X Z\right).$$

where $Z \sim N(0, 1)$.

4.5 Inference on Directionally Differentiable Functions

In this section we'll briefly cover the results in Fang and Santos (2019, ReStud). Specifically, we'll want to cover how bootstrap methods differ when functions are only directionally (as opposed to fully) differentiable.

Throughout this discussion, let's let $\hat{\theta}_n^*$ denote a "bootstrapped version" of $\hat{\theta}_n$ and assume the limiting distribution of $r_n\{\hat{\theta}_n - \theta_0\}$ can be consistently estimated by the conditional law of

$$r_n\left\{\hat{\theta}_n^* - \hat{\theta}_n\right\}.$$

In order to allow for diverse resampling schemes, simply appose that $\hat{\theta}_n^*$ is a function of the data $\{X_i\}$ and some random weights $\{W_i\}$ that are independent of $\{X_i\}$. Recall that from Theorem 3.10 weak convergence of probability measures on a space X is equivalent to convergence in the bounded lipschitz metric

$$d_{\text{BL}}\left(P_1, P_2\right) = \sup_{f \in \text{BL}_1(A)} \left|\int f(a)\, dP_1 - \int f(a) dP_2\right|.$$

The above is summarized in the following assumption:

Assumption 4.1 (Assumption 3, Fang and Santos (2019))**.** Asumme that

1. $\hat{\theta}_n^* : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}_\phi$ with $\{W_i\}$ independent of X_i.
2. $\hat{\theta}_n^*$ satisfies $\sup_{f\in \text{BL}_1(\mathbb{D})} |\mathbb{E}[f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}|\{X_i\}] - \mathbb{E}[f(\mathbb{G}_0)]| = o_p(1)$.
3. $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ is asymptotically measurable (jointly in $\{X_i, W_i\}$).

4. $f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})$ is a measurable function of $\{W_i\}$, outer almost surely in $\{X_i\}$ for any continuous and bounded $f : \mathbb{D} \to \mathbb{R}$.

We will then be interested in bootstrap procedures to estimate the distribution of $r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\}$. Now give an important theorem characterizing when we can use a plug in bootstrap estimator, $\phi(\hat{\theta}_n^*)$.

Theorem 4.10 (Theorem 3.1, Fang and Santos (2019)). *Assume that $\mathbb{D}$ and $\mathbb{E}$ are Banach spaces with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$. Also assume that $\theta_0 \in \mathbb{D}_\phi$, the domain of ϕ, and that $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}_0$ for some tight gaussian element $\mathbb{G}_0$. Then under Assumption 4.1 it follows that ϕ is fully Hadamard differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to the support of $\mathbb{G}_0$* if and only if*:*

$$\sup_{f \in BL_1(\mathbb{E})} |\mathbb{E}[f(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\})|\{X_i\}_{i=1}^n] - \mathbb{E}[f(\phi'_{\theta_0}(\mathbb{G}_0))]| = o_p(1) \tag{4.3}$$

This is a positive result if your function ϕ is fully Hadamard differentiable (Definition 4.5), but a negative result otherwise. For example, the standard bootstrap would fail if the function ϕ is only Hadamard directionally differentiable tangent to the support of $\mathbb{G}_0$ (Definition 4.6), as in the case of Example 4.4.

The following supplemental theorem forms the backbone of the proof of Theorem 4.10.

Theorem 4.11 (Theorem S.3.1, Fang and Santos (2019)). *Assume that $\mathbb{D}$ and $\mathbb{E}$ are Banach spaces with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$. Also assume that $\theta_0 \in \mathbb{D}_\phi$, the domain of ϕ, and that $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}_0$ for some tight gaussian element $\mathbb{G}_0$ that contains zero in it's support. Finally assume that $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is Hadamard directionally differentiable at θ_0 tangent to the support of $\mathbb{G}_0$. Then under Assumption 4.1 the following statments are equivalent:*

1. $\mathbb{E}[f(\phi'_{\theta_0}(\mathbb{G}_0))] = \mathbb{E}[f(\phi'_{\theta_0}(\mathbb{G}_0 + a_0) - \phi'_{\theta_0}(a_0))]$ *for all* $a_0 \in \text{supp}(\mathbb{G}_0)$ *and* $f \in BL_1(\mathbb{E})$.
2. $\sup_{f \in BL_1(\mathbb{E})} |\mathbb{E}[f(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\})|\{X_i\}_{i=1}^n] - \mathbb{E}[f(\phi'_{\theta_0}(\mathbb{G}_0)]| = o_p(1)$

This is useful as, using Theorem 3.10, the second statment is equivalent to weak convergence of the bootstrap distribution, $r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\}$ to the limiting distribution of $r_n\{\phi(\hat{\theta}) - \phi(\theta_0)\}$ given by $\phi'_{\theta_0}(\mathbb{G}_0)$ via Theorem 4.9. The first statement can then be shown to be equivalent to full Hadamard differentiability of ϕ at θ_0.[1]

Given this negative result for naïve bootstrap inference on directionally diffentiable functions, Fang and Santos propose a modified bootstrap procedure that is consistent more generally. To use thie valid bootstrap procedure we will need a consistent estimator for the Hadamard directional differential $\hat{\theta}'_n$ in the following sense:

Assumption 4.2 (Assumption 4, Fang and Santos (2019)). The map $\hat{\theta}'_n : \mathbb{D} \to \mathbb{E}$ is a function $\{X_i\}_{i=1}^n$ such that for every compact set $K \subseteq \mathbb{D}_0$ and every $\epsilon > 0$:

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\left(\sup_{h \in K^\delta} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} > \epsilon \right) = 0 \tag{4.4}$$

where we recall that $K^\delta = \{x : d(x, K) < \delta\}$ is the δ-expansion of K.

Remark (Remark 3.3, Fang and Santos (2019)). In certain applications, for example if $\mathbb{D} = \mathbb{R}^d$ or if $\mathbb{D}$ is separable and $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ is Borel measurable as a function of $\{X_i, W_i\}$ then (4.4) can be relaxed to verifying the following

$$\sup_{h \in K} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = o_p(1) \tag{4.5}$$

for any compact set $K \subseteq \mathbb{D}$.

In either case, we get the following consistent bootstrap procedure that works even if the function ϕ is only Hadamard directionally differentiable.

[1] One direction is easy, only the returning direction is difficult.

Theorem 4.12 (Theorem 3.2, Fang and Santos (2019)). *Assume that $\mathbb{D}$ and $\mathbb{E}$ are Banach spaces with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$. Also assume that $\theta_0 \in \mathbb{D}_\phi$, the domain of ϕ, and that $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}_0$ for some tight gaussian element $\mathbb{G}_0$ that contains zero in it's support. Finally assume that $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is Hadamard directionally differentiable at θ_0 tangent to the support of $\mathbb{G}_0$. Then under Assumption 4.1 and Assumption 4.2*

$$\sup_{f \in BL_1(\mathbb{E})} |\mathbb{E}[f(\hat{\phi}'_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}))|\{X_i\}_{i=1}^n] - \mathbb{E}[f(\phi'_{\theta_0}(\mathbb{G}_0))]| = o_p(1).$$

Theorem 4.12 shows that the conditional distribution of $\hat{\phi}'_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\})$ given the data is a consistent estimator of the limiting distribution of $r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\}$.