Last updated October 3, 2025

Empirical processes

Contents

1 Introduction

2 Glivenko-Cantelli theorems with the help of bracketing numbers

The problem we need to tackle is that as the set $\mathcal{F}$ is typically uncountable there is in general no guarantee that $\sup_{f\in\mathcal{F}} |P_n(f) - P(f)|$ is a measurable random variable. To overcome this difficulty we generalize the convergence in probability and almost sure convergence.

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be the probability space. Then for $B \subset \Omega$ we define the **outer probability** as

$$\mathsf{P}^*(B) = \inf\big\{\,\mathsf{P}(A) : B \subset A,\ A \in \mathcal{A}\big\}. \tag{2.1}$$

In what follows let $\mathbb{D}$ be a metric space with a metric d.

Definition 1. Let $X_1, X_2, \ldots, X$ be (random) maps from Ω to $\mathbb{D}$.

(i) We say that X_n converges **in outer probability** to X if $\mathsf{P}^*\big(d(X_n, X) > \eta\big) \xrightarrow[n\to\infty]{} 0$ for each $\eta > 0$. This convergence is denoted as $X_n \xrightarrow[n\to\infty]{P^*} X$.

(ii) We say that X_n converges **outer almost surely** to X if there exists a sequence of measurable random variable $\{\Delta_n\}$ such that $d(X_n, X) \le \Delta_n$ and $\Delta_n \xrightarrow[n\to\infty]{\text{alm. surely}} 0$. This convergence is denoted as $X_n \xrightarrow[n\to\infty]{\text{alm. surely}^*} X$.

Bracketing numbers

Let $\mathcal{F}$ be the set of real functions defined on the (sample) space $\mathcal{X}$ that is equipped with the norm $\|\cdot\|$. The first concept how to measure the size of $\mathcal{F}$ is based on the *bracketing numbers.* The second concept based on *covering numbers* will be introduced in Chapter 5.

Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f that are between l and u, i.e.

$$[l, u] = \big\{f : \mathcal{X} \to \mathbb{R} :\ l(x) \le f(x) \le u(x),\ \forall x \in \mathcal{X}\big\}.$$

Further an ϵ-*bracket* is a bracket $[l, u]$ such that $\|u - l\| < \epsilon$.

Definition 2. The **bracketing number** $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of ϵ-brackets needed to cover the set $\mathcal{F}$. The upper and lower bounds u and l of these brackets do not have to belong to $\mathcal{F}$ but are assumed to be measurable and to have finite norms.

In what follows we will be interested in the norms that posses the (Riesz) property, i.e. if $|f(x)| \le |g(x)|$ for all $x \in \mathcal{X}$ then $\|f\| \le \|g\|$.

Theorem 1. ***(Glivenko-Cantelli)***
Let $\mathcal{F}$ be a class measurable functions, such that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for each $\epsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli, i.e.

$$\sup_{f\in\mathcal{F}} \big|P_n(f) - P(f)\big| \xrightarrow[n\to\infty]{\textit{alm. surely}^*} 0.$$

Proof. Let $\epsilon > 0$ be given. Then there exists a finite bracketing number K_ε and brackets $[l_1, u_1], \dots, [l_{K_\varepsilon}, u_{K_\varepsilon}]$ such that

$$\mathcal{F} \subset \bigcup_{j=1}^{K_\epsilon} [l_j, u_j] \quad \text{and} \quad \|u_j - l_j\|_{L_1(P)} < \epsilon \quad \forall j \in \{1, \dots, K_\epsilon\}.$$

Thus for $f \in \mathcal{F}$ there exists $j \in \{1, \dots, K_\epsilon\}$ such that $f \in [l_j, u_j]$ and so one can bound

$$P_n(f) - P(f) \le P_n(u_j) - P(l_j) = P_n(u_j) - P(u_j) + P(u_j) - P(l_j) \le P_n(u_j) - P(u_j) + \epsilon.$$

Thus

$$\sup_{f \in \mathcal{F}} \big(P_n(f) - P(f)\big) \le \max_{j \in \{1,\dots,K_\epsilon\}} \big|P_n(u_j) - P(u_j)\big| + \epsilon. \tag{2.2}$$

Analogously

$$\inf_{f \in \mathcal{F}} \big(P_n(f) - P(f)\big) \ge - \max_{j \in \{1,\dots,K_\epsilon\}} \big|P_n(l_j) - P(l_j)\big| - \epsilon,$$

which together with (2.2) implies that

$$\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f)\big| \le \Delta_n(\epsilon) + \epsilon,$$

where

$$\Delta_n(\epsilon) = \max_{g \in \mathcal{G}_\varepsilon} \big\{|P_n(g) - P(g)|\big\}, \quad \text{where} \quad \mathcal{G}_\varepsilon = \big\{l_j, u_j : j \in \{1, \dots, K_\epsilon\}\big\}.$$

As $\epsilon > 0$ is arbitrary one gets that for each $m \in \mathbb{N}$ there exists a finite set of functions $\mathcal{G}_{\frac{1}{m}}$ such that

$$\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f)\big| \le \Delta_n(\tfrac{1}{m}) + \tfrac{1}{m}$$

and thus also

$$\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f)\big| \le \Delta_n, \quad \text{where} \quad \Delta_n = \inf_{m \in \mathbb{N}} \big(\Delta_n(\tfrac{1}{m}) + \tfrac{1}{m}\big).$$

Note that Δ_n is a measurable random variable thus from the definition of outer almost sure convergence (see Definition 1(ii)) it remains to show that $\Delta_n \xrightarrow[n\to\infty]{\text{alm. surely}} 0$. This will be verified provided that one can show that

$$\mathsf{P}\Big(\bigcup_{k=1}^{\infty} \Big[\limsup_{n\to\infty} \Delta_n > \tfrac{1}{k}\Big]\Big) = 0. \tag{2.3}$$

Note that

$$\mathsf{P}\Big(\bigcup_{k=1}^{\infty} \Big[\limsup_{n\to\infty} \Delta_n > \tfrac{1}{k}\Big]\Big) \le \sum_{k=1}^{\infty} \mathsf{P}\Big(\limsup_{n\to\infty} \Delta_n > \tfrac{1}{k}\Big) \tag{2.4}$$

and that for each $k \in \mathbb{N}$

$$\mathsf{P}\Big(\limsup_{n\to\infty} \Delta_n > \tfrac{1}{k}\Big) \le \mathsf{P}\Big(\limsup_{n\to\infty} \Delta_n(\tfrac{1}{2k}) + \tfrac{1}{2k} > \tfrac{1}{k}\Big) = \mathsf{P}\Big(\limsup_{n\to\infty} \Delta_n(\tfrac{1}{2k}) > \tfrac{1}{2k}\Big) = 0, \tag{2.5}$$

as $\Delta_n(\frac{1}{2k}) \xrightarrow[n\to\infty]{\text{alm. surely}} 0$. Combining (2.4) and (2.5) verifies (2.3) which further implies that $\Delta_n \xrightarrow[n\to\infty]{\text{alm. surely}} 0$ and finishes the proof of the theorem. □

The approach using bracketing numbers gives a very strict control of the size of $\mathcal{F}$. The advantage is that the above theorem can be easily generalized to not i.i.d. situation. The only thing that is needed is that the strong law of large numbers is available.

On the other hand the disadvantage of this approach is that requiring $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ may be rather strict in i.i.d. settings. Sometimes it is advantageous to use a different approach how to control the size of $\mathcal{F}$ (see Chapter 5).

Remark 1. The assumption $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ is sufficient but not necessary. Consider the class of constant functions.

Example 1. Consider the class of real functions

$$\mathcal{F} = \big\{x \to \rho(x;\theta) \,:\, \theta \in \Theta\big\},$$

where the space (Θ, d) is compact and $\rho(x;\theta)$ is continuous in θ for each $x \in \mathcal{X}$. Further there exists a dominating function R such that

$$\sup_{\theta\in\Theta} |\rho(x;\theta)| \leq R(x), \quad \forall x \in \mathcal{X} \qquad \text{and} \qquad \mathsf{E}\ R(X_i) < \infty.$$

Show that $\mathcal{F}$ is P-Glivenko-Cantelli.

Application to consistency of M-estimators

Let $\rho(x;\theta) : \mathcal{X} \times \Theta \to \mathbb{R}$ be a 'loss function' such that the parameter of interest can be identified as

$$\theta_X = \arg\min_{\theta\in\Theta} \mathsf{E}\ \rho(X_i;\theta).$$

Then the M-estimator of θ_X is defined as

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \rho(X_i;\theta).$$

Suppose that the parameter of interest is 'well-separated', i.e. for each $\eta > 0$ there exists $\delta > 0$ such that

$$\inf_{\theta\in\Theta_\eta^c} \mathsf{E}\ \rho(X_i;\theta) \geq \mathsf{E}\ \rho(X_i;\theta_X) + \delta, \tag{2.6}$$

where

$$\Theta_\eta^c = \big\{\theta \in \Theta : d(\theta, \theta_X) \geq \eta\big\}$$

with d being a metric on the parameter space Θ.

The following theorem illustrates how the uniform law of large numbers can be used to show the consistency of M-estimator.

Theorem 2. *Suppose that the above identifiability assumption* (2.6) *holds and that the class of functions*

$$\mathcal{F} = \big\{x \to \rho(x;\theta) \,:\, \theta \in \Theta\big\}$$

is P-Glivenko-Cantelli. Then

$$d(\widehat{\theta}_n, \theta_X) \xrightarrow[n\to\infty]{P} 0.$$

Proof. For simplicity of notation introduce

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(X_i;\theta) \quad \text{and} \quad M(\theta) = \mathsf{E}\ \rho(X_i;\theta).$$

Let $\eta > 0$ be fixed. Then there exists $\delta > 0$ such that (2.6) holds. Now one can bound

$$\begin{aligned}
\mathsf{P}\left(d(\widehat{\theta}_n,\theta) > \eta\right) &\leq \mathsf{P}^*\Big(\inf_{\theta\in\Theta_\eta^c} M_n(\theta) \leq M_n(\theta_X)\Big) \\
&\leq \mathsf{P}^*\Big(\inf_{\theta\in\Theta_\eta^c} \big[M_n(\theta) - M(\theta)\big] + \inf_{\theta\in\Theta_\eta^c} M(\theta) \leq M(\theta_X) + M_n(\theta_X) - M(\theta_X)\Big) \\
&\leq \mathsf{P}^*\Big(\inf_{\theta\in\Theta_\eta^c} \big[M_n(\theta) - M(\theta)\big] \leq M(\theta_X) - \inf_{\theta\in\Theta_\eta^c} M(\theta) + M_n(\theta_X) - M(\theta_X)\Big) \\
&\leq \mathsf{P}^*\Big(\inf_{\theta\in\Theta} \big[M_n(\theta) - M(\theta)\big] \leq -\delta + M_n(\theta_X) - M(\theta_X)\Big) \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

as

$$\inf_{\theta\in\Theta}\big[M_n(\theta) - M(\theta)\big] \geq -\sup_{\theta\in\Theta}\big|M_n(\theta) - M(\theta)\big| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0$$

and also $M_n(\theta_X) - M(\theta_X) \xrightarrow[n\to\infty]{\text{alm. surely}} 0.$ □

Unfortunately in applications it is typically not at all straightforward to use this theorem. The problem is that using of the result derived in Example 1 requires that the parameter space is compact which is usually not the case. Thus the first step of the proof of the consistency of the M-estimator is showing that $\mathsf{P}(\widehat{\theta}_n \in \Theta_K) \xrightarrow[n\to\infty]{} 1$, where Θ_K is a compact subset of Θ that contains the true value of the parameter. The other problem that may be rather difficult is to show that the identifiability assumption (2.6) holds.

Consistency of M-estimators is in more detail discussed in Chapter 5.2 of van der Vaart (2000). In particular consistency of maximum likelihood estimators and least squares estimators are discussed in Chapter 4 of van de Geer (2000).

Exercise 1. Let $X_1,\ldots,X_n$ be a random sample from the uniform distribution on $[0,2\pi]$. For $a \in [0,10]$ define

$$Y_n(a) = \frac{1}{n}\sum_{i=1}^{n} \cos(a\,X_i).$$

Show that $\sup_{a\in[0,1]}\big|Y_n(a) - \frac{\sin(2a\pi)}{2a\pi}\big| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0.$

Now, consider that $\widehat{a}_n = \frac{\pi}{\overline{X}_n}$. Show that $Y_n(\widehat{a}_n) \xrightarrow[n\to\infty]{P^*} 0.$

Hint. Denote $Y(a) = \mathsf{E}\ \cos(a\,X_i)$. Note that with probability going to one one can bound

$$\begin{aligned}
\big|Y_n(\widehat{a}_n) - Y(1)\big| &\leq \big|Y_n(\widehat{a}_n) - Y(\widehat{a}_n)\big| + \big|Y(\widehat{a}_n) - Y(1)\big| \\
&\leq \sup_{a\in[0,2\pi]}\big|Y_n(a) - Y(a)\big| + \big|Y(\widehat{a}_n) - Y(1)\big|.
\end{aligned}$$

Exercise 2. Let $X_1,\ldots,X_n$ be a random sample such that X_i has an exponential distribution with the mean equal to 1. Consider the process

$$Z_n(a) = \frac{1}{n}\sum_{i=1}^{n} |X_i - a|, \quad a \in [0,2].$$

Show that $\sup_{a\in[0,2]} \left|Z_n(a) - \mathsf{E}\ Z_n(a)\right| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0$.

With the help of the previous result show that

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n| \xrightarrow[n\to\infty]{P^*} \mathsf{E}\ |X_i - \mathsf{E}\ X_i|.$$

The end of the self study for the week (1.10.-7.10.)

3 Convergence in distribution in metric spaces

Let $\mathbb{D}$ be a metric space with a metric d.

Definition 3. Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and $Y : \Omega \to \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be a (possible non-measurable) mapping. Then the **outer expectation** of Y is defined as

$$\mathsf{E}^* [Y] = \inf \big\{ \mathsf{E} [U] \mid U \geq Y,\ U : \Omega \to \bar{\mathbb{R}}\ (\mathcal{A}, \bar{\mathcal{B}}^1)\text{-measurable},\ \mathsf{E} [U] \text{ exists} \big\}.$$

Analogously the **inner expectation** is defined as

$$\mathsf{E}_* [Y] = \sup \big\{ \mathsf{E} [U] \mid U \leq Y,\ U : \Omega \to \bar{\mathbb{R}}\ (\mathcal{A}, \bar{\mathcal{B}}^1)\text{-measurable},\ \mathsf{E} [U] \text{ exists} \big\}.$$

Remark 2. (a) The inner expectation can be also defined as

$$\mathsf{E}_* [Y] = -\mathsf{E}^* [-Y].$$

(b) If the function Y is measurable then

$$\mathsf{E}^* [Y] = \mathsf{E}_* [Y] = \mathsf{E} [Y].$$

(c) One has to be careful that in general for the outer expectation the Fubini theorem does not hold. Be also careful that in general the outer expectation is only subadditive, i.e. $\mathsf{E}^* [Y + Z] \leq \mathsf{E}^* [Y] + \mathsf{E}^* [Z]$.

(d) It can be proved (see e.g. Lemma 1.2.1 of van der Vaart and Wellner, 1996) that the infimum in the definition of the outer function is attained provided that the outer expectation exists. That is there exists '*a minimal measurable cover function*' Y^* such that $Y^*(\omega) \geq Y(\omega)$ for each $\omega \in \Omega$ and $\mathsf{E}^* [Y] = \mathsf{E} [Y^*]$. Further for any other $\widetilde{Y}$ such that $\widetilde{Y} \geq Y$ almost surely it holds that $Y^* \leq \widetilde{Y}$ almost surely.

(e) For each $B \subset \Omega$ it holds that $\mathsf{E}^* [\mathbb{I}_B] = \mathsf{P}^*(B)$. Similarly for each $B \subset \Omega$ it holds that $\mathsf{E}_* [\mathbb{I}_B] = \mathsf{P}_*(B)$, where $\mathsf{P}_*(B)$ is the inner probability defined as

$$\mathsf{P}_*(B) = \sup\{\mathsf{P}(A) \mid A \subset B,\ A \in \mathcal{A}\}.$$

For a formal proof see see e.g. Lemma 1.2.3.(i) of van der Vaart and Wellner (1996). Note also that the inner probability can be defined as

$$\mathsf{P}_*(B) = 1 - \mathsf{P}^*(\Omega \setminus B).$$

Definition 4. Let $\mathbb{D}$ be a metric space with the metric d and Borel-σ-algebra $\mathcal{D}$, $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space, $X_n : \Omega \to \mathbb{D}$ be a (possible non-measurable random) mapping and $X : (\Omega, \mathcal{A}, \mathsf{P}) \to (\mathbb{D}, \mathcal{D})$ be a measurable mapping. Then we say that X_n **converges weakly (in distribution)** to X (notation: $X_n \rightsquigarrow X$), provided that

$$\mathsf{E}^* [f(X_n)] \xrightarrow[n\to\infty]{} \mathsf{E} [f(X)]$$

for every bounded continuous function $f : \mathbb{D} \mapsto \mathbb{R}$.

The following theorem is an analogy of the standard Portmanteau-Theorem (see e.g. Lemma 2.2 in van der Vaart, 2000).

Theorem 3. ***(Portmanteau-Theorem)***
The following statements are equivalent.

(i) $X_n \rightsquigarrow X$.

(ii) $\mathsf{E}^*\left[f(X_n)\right] \xrightarrow[n\to\infty]{} \mathsf{E}\left[f(X)\right]$ *for all bounded and Lipschitz functions* $f:\mathbb{D}\to\mathbb{R}$.

(iii) $\limsup_{n\to\infty} \mathsf{P}^*(X_n \in F) \leq \mathsf{P}(X \in F)$ $\forall F \subset \mathbb{D}$ *closed.*

(iv) $\liminf_{n\to\infty} \mathsf{P}_*(X_n \in G) \geq \mathsf{P}(X \in G)$ $\forall G \subset \mathbb{D}$ *open.*

(v) $\mathsf{P}^*(X_n \in B) \xrightarrow[n\to\infty]{} \mathsf{P}(X \in B)$ $\forall B \in \mathcal{D}$ *with* $\mathsf{P}(X \in \partial B) = 0$.

Proof.

(i) ⇒ (ii) ✓

(ii) ⇒ (iii) Let $F \subset \mathbb{D}$ be closed, d be the metric in $\mathbb{D}$ and $d(x,F) := \inf\{d(x,y) \,|\, y \in F\}$. We approximate $\mathbb{I}_F$ with the help of the function $f:\mathbb{D}\to\mathbb{R}$ given by

$$f(x) = \big(1 - \tfrac{d(x,F)}{\epsilon}\big)_+.$$

Note that f is bounded and Lipschitz-continuous. Further let $F^\epsilon := \{x \in \mathbb{D} \,|\, d(x,F) < \epsilon\}$. Then it holds that $\mathbb{I}_F(x) \leq f(x) \leq \mathbb{I}_{F^\epsilon}(x)$ and thus for each $\epsilon > 0$

$$\begin{aligned}\limsup_{n\to\infty} \mathsf{P}^*(X_n \in F) &= \limsup_{n\to\infty} \mathsf{E}^*\left[\mathbb{I}_F(X_n)\right] \leq \limsup_{n\to\infty} \mathsf{E}^*\left[f(X_n)\right] \\ &= \mathsf{E}\left[f(X)\right] \leq \mathsf{P}(X \in F^\epsilon).\end{aligned}$$

Now the statement follows from the fact that $\mathsf{P}(X \in F^\epsilon) \to \mathsf{P}(X \in F)$ as $\epsilon \searrow 0$ (by the continuity of the probability measure).

(iii) ⇔ (iv) Homework exercise.

(iii)+(iv) ⇒ (v) Let $B \in \mathcal{D}$ with $\mathsf{P}(X \in \partial B) = 0$. Then it holds that

$$\mathsf{P}(X \in B) = \mathsf{P}(X \in \mathrm{int}(B)) \leq \liminf_{n\to\infty} \mathsf{P}_*\big(X_n \in \mathrm{int}(B)\big) \leq \liminf_{n\to\infty} \mathsf{P}^*(X_n \in B)$$

and at the same time also

$$\mathsf{P}(X \in B) = \mathsf{P}(X \in \bar{B}) \geq \limsup_{n\to\infty} \mathsf{P}^*(X_n \in \bar{B}) \geq \limsup_{n\to\infty} \mathsf{P}^*(X_n \in B).$$

Thus $\lim_{n\to\infty} \mathsf{P}^*(X_n \in B) = \mathsf{P}(X \in B)$.

(v) ⇒ (i) Without loss of generality it sufficient to consider only functions such that $0 < f < 1$. Then it can be shown that

$$\mathsf{E}^*\left[f(X_n)\right] \overset{\text{Exerc.}}{=} \int_0^1 \mathsf{P}^*(f(X_n) > t)\,dt, \quad \mathsf{E}\left[f(X)\right] = \int_0^1 \mathsf{P}(f(X) > t)\,dt.$$

Denote $B_t := \{x \in \mathbb{D} \,|\, f(x) > t\}$. Then by the continuity of f it holds that $\partial B_t \subset \{x \in \mathbb{D} \,|\, f(x) = t\}$. Further the probability $\mathsf{P}(f(X) > t)$ is a complement of a distribution function of the random variable $Y = f(X)$. Thus for all except for countable many t it holds that

$$0 = \mathsf{P}(f(X) = t) = \mathsf{P}(X \in \partial B_t).$$

For such t it holds (by (v)) that

$$\mathsf{P}^*(f(X_n) > t) = \mathsf{P}^*(X_n \in B_t) \xrightarrow[n\to\infty]{} \mathsf{P}(X \in B_t) = \mathsf{P}(f(X) > t).$$

Now wit the help of Lebesgue (dominated convergence) theorem it holds that

$$\mathsf{E}^*\left[f(X_n)\right] = \int_0^1 \mathsf{P}^*(f(X_n) > t)\,dt \xrightarrow[n\to\infty]{} \int_0^1 \mathsf{P}(f(X) > t)\,dt = \mathsf{E}\left[f(X)\right],$$

which was to be proved.

□

Exercise 3. Suppose that X is measurable. Show that

$$X_n \xrightarrow[n\to\infty]{P^*} X \implies X_n \rightsquigarrow X.$$

In applications we are often not interested in the empirical process itself but rather in a function of this process. The following theorem guarantees that if this function is continuous then the function of the empirical process also converges in distribution.

Theorem 4. ***(Continuous-Mapping-Theorem, CMT)***
Let X be measurable and $X_n \rightsquigarrow X$. Further let $\mathbb{D}'$ be a metric space with Borel-σ-algebra $\mathcal{D}'$ and $\psi : \mathbb{D} \to \mathbb{D}'$ be a mapping that is continuous on the set $C(\psi) \subset \mathbb{D}$ and $\mathsf{P}\big(X \in C(\psi)\big) = 1$. Then

$$\psi(X_n) \rightsquigarrow \psi(X).$$

Proof. Let $F' \in \mathcal{D}'$ be closed. Then it holds

$$\begin{aligned}\limsup_{n\to\infty} \mathsf{P}^*\left(\psi(X_n) \in F'\right) &= \limsup_{n\to\infty} \mathsf{P}^*\left(X_n \in \psi^{-1}(F')\right) \le \limsup_{n\to\infty} \mathsf{P}^*\left(X_n \in \overline{\psi^{-1}(F')}\right)\\ &\overset{Th.\ 3(iii)}{\le} \mathsf{P}\left(X \in \overline{\psi^{-1}(F')}\right) = \mathsf{P}\left(X \in \overline{\psi^{-1}(F')} \cap C(\psi)\right).\end{aligned}$$

Now show that (homework exercise)

$$\mathsf{P}\left(X \in \overline{\psi^{-1}(F')} \cap C(\psi)\right) \le \mathsf{P}\left(X \in \psi^{-1}(F')\right) = \mathsf{P}\left(\psi(X) \in F'\right).$$

Now the statement follows by Theorem 3(iii). □

Similarly as in the standard definition of weak convergence, also the generalized weak convergence introduced in Definition 4 is closely connected with tightness.

Definition 5. (a) Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space, $\mathbb{D}$ a metric space, $\mathcal{D}$ Borel-σ-algebra. We say that a $(\mathcal{A}, \mathcal{D})$-measurable random variable X is **tight**, when for all $\epsilon > 0$ there exist a compact set $K \subset \mathbb{D}$ such that $\mathsf{P}(X \notin K) < \epsilon$.

(b) Let $X_n : \Omega \to \mathbb{D}$ $(n \in \mathbb{N})$ be a sequence of possible non-measurable random elements. Then we say that a sequence of maps $\{X_n\}_{n=1}^{\infty}$ is **asymptotically tight** if $\forall \epsilon > 0$ there exists a compact set $K \subset \mathbb{D}$ such that for $\forall \delta > 0$

$$\limsup_{n \to \infty} \mathsf{P}^*(X_n \notin K^\delta) < \epsilon,$$

where $K^\delta = \{x \in \mathbb{D} \mid d(x, K) < \delta\}$ is a δ-enlargement of K.

Remark 3. Note that if $\mathbb{D}$ is separable and complete then the random variable is (always) tight (Theorem 1.3 Billingsley, 1999).

To formulate the (generalized) Prohorov's theorem we need to guarantee the 'measurability' in the limit. More precisely we say that a sequence of maps $\{X_n\}_{n=1}^{\infty}$ is *asymptotically measurable* if for each bounded continuous (real) function

$$\mathsf{E}^* f(X_n) - \mathsf{E}_* f(X_n) \xrightarrow[n \to \infty]{} 0.$$

Theorem 5. ***(Prohorov's theorem)*** *Let $X_n : \Omega_n \to \mathbb{D}$ be a sequence of (possibly non-measurable) random elements. Then*

- *If $X_n \rightsquigarrow X$ for some tight random element X, then $\{X_n\}_{n=1}^{\infty}$ is asymptotically tight and measurable.*
- *If $\{X_n\}_{n=1}^{\infty}$ is asymptotically tight and measurable, then there exists a subsequence which converges in distribution to a tight random element X (in the sense of Definition 4).*

For the proof of (i) see Lemma 1.3.8 of van der Vaart and Wellner (1996). The proof of (ii) corresponds to the proof of Theorem 1.3.9(ii) of van der Vaart and Wellner (1996).

Bounded stochastic processes

Let $X_1, \dots, X_n$ be independent identically distributed random variables with values in space $\mathcal{X}$ (think for instance of $\mathbb{R}$ or $\mathbb{R}^k$). Recall that the empirical process $\mathbb{G}_n$ indexed by the set of (real measurable) functions $\mathcal{F}$ (on $\mathcal{X}$) is a collection of the random variables

$$\mathbb{G}_n(f) = \sqrt{n}\,\big(P_n(f) - P(f)\big), \quad f \in \mathcal{F}, \tag{3.1}$$

where

$$P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \text{and} \quad P(f) = \mathsf{E}\, f(X_i).$$

Thus $\mathbb{G}_n$ can be viewed as a stochastic process $X = \{X(t) \mid t \in T\}$ where the role of t and the index set T is played by the function f and the set of functions $\mathcal{F}$.

Further in the subsequent chapter we will consider the sets of functions $\mathcal{F}$ such that

$$\sup_{f \in \mathcal{F}} \big|f(x) - P(f)\big| < \infty, \qquad \text{for every } x \in \mathcal{X}. \tag{3.2}$$

Thus for a given $\omega \in \Omega$ the empirical process has a sample path $f \mapsto \mathbb{G}_n(\omega, f)$ which is a bounded function. Thus, more generally consider the stochastic processes $\{X_n\}$ and X indexed by T whose values (sample paths) are in the space of bounded functions $\ell^\infty(T)$. This space is formally defined as

$$\ell^\infty(T) = \big\{f : T \to \mathbb{R} \mid \|f\|_T < \infty\big\}, \quad \text{where} \quad \|f\|_T = \sup_{t\in T} |f(t)|.$$

The advantage of considering $\ell^\infty(T)$ is that it covers all the commonly considered spaces. On the other hand one often knows that the empirical process lives in a smaller space. For instance the trajectories of the classical empirical process $\{\sqrt{n}(F_n(t) - F(t)),\ t \in \mathbb{R}\}$ are contained in the space of right-continuous functions with left limits (i.e. $\mathbb{D}(-\infty, \infty)$). This raises the following question. Is it possible that the empirical process does not converge in distribution in the space $(\ell^\infty(T), \|\cdot\|_T)$ but at the same time it converges in a subset of $\ell^\infty(T)$ (e.g. in $\mathbb{D}(-\infty, \infty)$)? The following lemma says that this cannot happen provided that the supremum metric $\|\cdot\|_T$ is also used in the smaller space.

Lemma 1. *Let $(\mathbb{D}, \rho)$ be a metric space and $\mathbb{D}_0 \subset \mathbb{D}$. Let $X_n(\omega) \in \mathbb{D}_0$ for all sufficiently large $n \in \mathbb{N}$ and also $X(\omega) \in \mathbb{D}_0$ for each $\omega \in \Omega$. Then $X_n \rightsquigarrow X$ in $(\mathbb{D}, \rho)$ if and only if $X_n \rightsquigarrow X$ in $(\mathbb{D}_0, \rho)$.*

Proof. The proof follows from the fact that the set G_0 is open in $\mathbb{D}_0$ if and only if $G_0 = G \cap \mathbb{D}_0$, where G is an open set in $\mathbb{D}$. □

By Prohorov's theorem (Theorem 5) the convergence in distribution implies the asymptotic tightness and asymptotic measurability. The question of interest is whether one can be more specific for processes with values in $\ell^\infty(T)$. The one of the possible answers is given in the following theorem. Roughly speaking this theorem says that convergence in distribution in $\ell^\infty(T)$ can be characterized by *finite approximation.* More precisely for each $\epsilon > 0$ the index set T can be partitioned into finitely many $T_1, \dots, T_k$ such that on each T_j with high probability the process X_n asymptotically oscillates less than ϵ uniformly in j. When this holds, then the behaviour of the process can be approximated by the marginal vectors $(X_n(t_1), \dots, X_n(t_k))^\mathsf{T}$ where $t_1, \dots, t_k$ are fixed points from $T_1, \dots, T_k$ respectively. If these marginal vectors converge then the process converges.

Theorem 6. *Let $X_n = \{X_n(t) \mid t \in T\}$ be a sequence of stochastic processes in $\ell^\infty(T)$. Then $X_n \rightsquigarrow X$, where X is a tight random variable if and only if both of the following assumptions hold:*

(i) For each $k \in \mathbb{N}$ and each $t_1, \dots, t_k \in T$ the random vector $(X_n(t_1), \dots, X_n(t_k))^\mathsf{T}$ converges in distribution in $\mathbb{R}^k$;

(ii) $\forall \epsilon > 0, \forall \eta > 0$ there exists a partition $T_1, \dots, T_k$ of T, such that

$$\limsup_{n\to\infty} \mathsf{P}^* \left(\sup_{j\in\{1,\dots,k\}} \sup_{s,t\in T_j} |X_n(s) - X_n(t)| \geq \epsilon \right) \leq \eta. \tag{3.3}$$

Remark 4. It can be proved (see Theorem 1.5.6 of van der Vaart and Wellner, 1996) that the assumption (ii) together with the asymptotic tightness of the sequence $\{X_n(t)\}$ for each $t \in T$ is in fact equivalent to the asymptotic tightness (Definition 5(ii)).

Proof. We give only a sketch of the proof of the sufficiency of the assumptions (i) and (ii). This part of the proof can be divided into four steps. More details can be found in the proof of Theorem 18.14 of van der Vaart (2000).

Step 1: With the help of assumption (ii) we find a semimetric[1] ρ on T such that (T, ρ) is totally bounded.

For $m \in \mathbb{N}$ denote $T_1^m, \dots, T_{k_m}^m$ the partition of T as in assumption (ii) with $\epsilon = \eta = 2^{-m}$. Then without loss of generality one can assume that the partitions are successive refinements as m increases. For $m \in \mathbb{N}$ define the semimetric ρ_m as

$$\rho_m(s,t) = \begin{cases} 0, & \exists j \in \{1, \dots, k_m\} : s, t \in T_j^m, \\ 1, & \text{otherwise.} \end{cases}$$

Note that (T, ρ_m) is totally bounded (exercise). As the partitions are successive refinements it holds that $\rho_1 \leq \rho_2 \leq \dots$. Now define

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

Then ρ is a semimetric. Further for each $\epsilon > 0$ one can find m such that $2^{-m} \leq \epsilon$. But then T_j^m is contained in a ball of diameter less than ϵ as for $\forall s, t \in T_j^m$

$$\rho(s,t) \leq \sum_{l=m+1}^{\infty} \frac{1}{2^l} = \frac{1}{2^m}.$$

Thus (T, ρ) is totally bounded (which is crucial to guarantee that the limiting process is tight, see **Step 4**).

Step 2: The construction of the limit process.

First, for each $j \in \{1, \dots, k_m\}$ and $m \in \mathbb{N}$ choose an arbitrary point $t_j^m \in T_j^m$. Then the set

$$T_0 := \{t_j^m, \, j \in \{1, \dots, k_m\}, \, m \in \mathbb{N}\}$$

is countable and dense in (T, ρ).

Now with the help of Daniel-Kolmogorov theorem there exists a process $\big\{X(t), t \in T_0\big\}$ so that assumption (i) holds. Next, it can be shown that for almost all $\omega \in \Omega$ the sample paths $X(\omega, t)$ of the limiting process are uniformly continuous in (T_0, ρ). Thus the process can be extended to T so that almost all paths are uniformly continuous in (T, ρ).

Step 3: We show that $X_n \rightsquigarrow X$.

To do that we make use of Theorem 3(ii). Let $f : \ell^\infty(T) \to \mathbb{R}$ be bounded and Lipschitz.

For each $m \in \mathbb{N}$ define a mapping $\pi_m : T \to T$ such that $\pi_m(t) = t_j^m$ when $t \in T_j^m$. Then one can bound

$$\begin{aligned} \limsup_{n\to\infty} \left| \mathsf{E}^* \left[f(X_n) \right] - \mathsf{E} \left[f(X) \right] \right| \leq & \limsup_{n\to\infty} \left| \mathsf{E}^* \left[f(X_n) \right] - \mathsf{E}^* \left[(f(X_n \circ \pi_m)) \right] \right| \\ & + \limsup_{n\to\infty} \left| \mathsf{E}^* \left[f(X_n \circ \pi_m) \right] - \mathsf{E} \left[f(X \circ \pi_m) \right] \right| \\ & + \left| \, \mathsf{E} \left[f(X \circ \pi_m) \right] - \mathsf{E} \left[f(X) \right] \right| \end{aligned}$$

Now by the assumption (ii) of the theorem and the fact that f is bounded and Lipschitz *the first term* term on the right-hand side of the inequality above can be made arbitrarily small by taking m large enough.

[1] In the context of empirical processes the semimetric is symmetric in its arguments and satisfies a triangular inequality. But compared to the metric the 'zero distance' (i.e. $\rho(x, y) = 0$) does not imply that $x = y$. Note that some authors would call ρ rather a pseudometric.

The second term is zero for each $m \in \mathbb{N}$ from the convergence of the finite-dimensional distributions (assumption (i) of the theorem).

Finally *the third term* can be made arbitrarily small by taking m large enough as by the uniform continuity of the sample paths of X one has $X \circ \pi_m \xrightarrow[m\to\infty]{\text{alm. surely}} X$.

Step 4: It remains to show that the limit process X is a **tight** random variable in $\ell^\infty(T)$. This can be seen as follows. Recall that almost all sample paths of X are uniformly ρ-continuous in T. Thus thanks to the fact that the semimetric space (T, ρ) is *totally bounded* one can deduce that X has (almost all) sample paths in $\ell^\infty(T)$. Further each ρ-continuous function has a unique continuous extension to the ρ-completion of T, say $\overline{T}$. As T is totally bounded then $(\overline{T}, \rho)$ is a compact semimetric space. Thus the unique extension of X to $\overline{T}$ has almost all sample paths in the set ρ-continuous functions on $\overline{T}$. But the space of continuous function on the compact semimetric space is a separable and complete subspace of $\ell^\infty(\overline{T})$ (see Example 1.5.1 van der Vaart and Wellner, 1996). Thus a measure on this space is tight.

□

Remark 5. **Step 4** of the proof can be also deduced from Remark 4 and Prohorov's Theorem (i.e. Theorem 5(ii)).

It is worth noting that during the proof of Theorem 6 a semimetric ρ is constructed in such a way that the limiting process X has uniformly ρ-continuous sample paths and the semimetric space (T, ρ) is totally bounded. This shows that if the processes $\{X_n\}$ converge in distribution in the space of bounded functions $\ell^\infty(T)$ (which is rather large), then the processes have to become more and more 'well-behaved' so that the limiting process is concentrated on a space of uniformly ρ-continuous (which is a much smaller space than $\ell^\infty(T)$).

The above ideas are formalized by the concept of asymptotic equicontinuity.

Definition 6. Let $X_n : \Omega_n \to \ell^\infty(T)$ and ρ be a semimetric on T. We say that the sequence of processes $\{X_n\}$ is **asymptotically uniformly ρ-equicontinuous in probability** if for every $\epsilon > 0$ and $\eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathsf{P}^* \left(\sup_{s,t \in T : \rho(s,t) < \delta} |X_n(s) - X_n(t)| > \epsilon \right) < \eta.$$

Theorem 7. *The sequence of stochastic processes satisfies assumption (ii) of Theorem 6 if and only if there exists a semimetric ρ such that (T, ρ) is totally bounded and $\{X_n\}$ is asymptotically uniformly ρ-equicontinuous in probability.*

Proof. '⇒' Assume that assumption (ii) of Theorem 6 is satisfied. Then the statement follows from **Step 1** of the proof of that theorem.

'⇐' Assume that there exists a semimetric ρ such that (T, ρ) is totally bounded and $\{X_n\}$ is asymptotically uniformly ρ-equicontinuous in probability. Then T can be covered with finitely many balls of radius δ (details left as an exercise). □

Note that so far we only know that if the $\{X_n\}$ converges in distribution in $\ell^\infty(T)$, then there exists a semimetric ρ for which the sequence $\{X_n\}$ is asymptotically uniformly ρ-equicontinuous in probability. Further, for the same semimetric ρ the limiting process X has uniformly continuous almost all trajectories. But the construction of the semimetric ρ in the proof of Theorem 6 is rather implicit as it depends on the existence of an appropriate partion of the set T.

Nevertheless, recall that we have in mind the application to the empirical process $\mathbb{G}_n$ based on the i.i.d. random variables. In this case by the central limit theorem we know that the finite-dimensional distributions of the limiting process are Gaussian with zero expectations. Thus the limiting process is zero-mean Gaussian. And for Gaussian processes it is known that the continuity of the sample paths is tied with the continuity of the covariance function. That is why probably the following theorem may be not so surprising.

Theorem 8. *Suppose that the assumptions of Theorem 6 hold and that the limiting process X is zero-mean Gaussian. Then $\{X_n\}$ is asymptotically uniformly ρ-equicontinuous in probability for the semimetric ρ defined as*

$$\rho(t,s) = \sqrt{\mathsf{E}\,[X(t) - X(s)]^2} = \sqrt{\mathsf{var}(X(t) - X(s))}.$$

Proof. See the proof of Lemma 18.15 of van der Vaart (2000). □

4 Donsker-Theorem with the help of bracketing numbers

Recall the empirical process given in (3.1) and assume that the assumption for bounded sample paths (3.2) holds.

Our aim is to show the weak convergence of the process $\mathbb{G}_n$. By the central limit theorem for i.i.d. random vectors for each $k \in \mathbb{N}$ and each $f_1, \dots, f_k \in \mathcal{F}$

$$\big(\mathbb{G}_n(f_1), \dots, \mathbb{G}_n(f_k)\big)^{\mathsf{T}} \xrightarrow[n\to\infty]{d} \mathsf{N}_k\big(\mathbf{0}_k, \mathbb{V}\big),$$

where the (j, l) element of the variance matrix $\mathbb{V}$ is given by

$$v_{jl} = \mathsf{cov}\,\big(f_j(X_i), f_l(X_i)\big), \quad j, l \in \{1, \dots, k\} = P(f_j\, f_l) - P(f_j)P(f_l).$$

Thus provided that also the assumption (ii) of Theorem 6 (i.e. asymptotic tightness) is satisfied, then one gets that

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}), \tag{4.1}$$

where $\mathbb{G}$ is a zero mean Gaussian process with the covariance function given by

$$\mathsf{cov}\,\big(\mathbb{G}(f_1), \mathbb{G}(f_2)\big) = \mathsf{cov}\,\big(f_1(X_i), f_2(X_i)\big) = P(f_1\, f_2) - P(f_1)P(f_2).$$

If the weak convergence (4.1) holds, then we say that the class of functions $\mathcal{F}$ is *P-Donsker.*

Theoretical results

The following lemma will be helpful when proving assumption (ii) of Theorem 6. It states that the bracketing numbers can be used to bound the expectation of the supremum of the empirical process $\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|$. To formulate this lemma it convenient to introduce an envelope (function). We say that a function F is an **envelope** for the set (class) of functions $\mathcal{F}$ when it satisfies

$$\sup_{f\in\mathcal{F}} |f(x)| \leq F(x), \quad \forall x \in \mathcal{X}.$$

I.e. the envelop function is a dominating function for absolute values of functions in $\mathcal{F}$.

Lemma 2. *Let $\mathcal{F}$ be a class measurable functions $f : \mathcal{X} \to \mathbb{R}$, such that $P(f^2) < \delta^2$ for each $f \in \mathcal{F}$ and F be a measurable envelope of $\mathcal{F}$. Then there exists a constant C, such that*

$$\mathsf{E}^*\left[\sup_{f\in\mathcal{F}} |G_n(f)|\right] \;\leq\; C\left(\int_0^{2\delta} \sqrt{\operatorname{Log} N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))}\; d\epsilon + \sqrt{n}\, P\big(F\,\mathbb{I}\{F > \sqrt{n}\; a(\delta)\}\big)\right), \tag{4.2}$$

where $a(\delta) = \frac{\delta}{\sqrt{\operatorname{Log} N_{[\,]}(\delta, \mathcal{F}, L_2(P))}}$ and $\operatorname{Log} x = \max\{1, \log x\}$.

The proof of this lemma is rather technical and it is based on Bernstein inequality and chaining technique (see e.g. Lemma 19.34 of van der Vaart, 2000).

In what follows we will use the lemma for δ 'small'. Thus note that there is an important assumption that the $\mathcal{F}$ is class of functions that are 'small' in $L_2(P)$-norm, i.e. $P(f^2) = \mathsf{E}\, f^2(X_i) < \delta^2$.

Before we proceed let us introduce the *bracketing integral*, i.e.

$$\mathcal{J}_{[\,]}\big(\delta, \mathcal{F}, L_2(P)\big) = \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\epsilon, \mathcal{F}, L_2(P)\big)}\; d\epsilon. \tag{4.3}$$

The bracketing integral can be viewed as a measure of the size of the class $\mathcal{F}$ and it plays an important role in the inequality (4.2)[1]

Remark 6. Note that the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$ increases (typically to ∞) for ϵ approaching zero. Thus the bracketing integral is finite, if $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$ does not increase too quickly when $\epsilon \to 0_+$. The nice thing is that the rate of bracketing number is 'slowed down' first by the logarithm and then by the square root. Thus a sufficient condition for the finitiness of the bracketing integral is that for some $\theta > 0$

$$\sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))} = O\big(\tfrac{1}{\varepsilon^{1-\theta}}\big), \quad \text{as} \quad \epsilon \to 0_+.$$

This happens for instance when $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$ is bounded by a polynomial in $\frac{1}{\epsilon}$ or when

$$N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) \leq C \exp\{\tfrac{1}{\epsilon^{2-\theta}}\}, \quad \text{for some } C < \infty \text{ and } \theta > 0.$$

The following theorem says that the finiteness of the bracketing integral is a sufficient condition for $\mathcal{F}$ to be P-Donsker.

Theorem 9. *Suppose that $\mathcal{J}_{[\,]}(1, \mathcal{F}, L_2(P)) < \infty$. Then the class of measurable functions $\mathcal{F}$ is P-Donsker.*

Proof. As noted before, it is sufficient to show (3.3). That is it remains to show that for each $\epsilon > 0$, $\eta > 0$ there exists a finite partition $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of $\mathcal{F}$ such that

$$\limsup_{n\to\infty} \mathsf{P}^* \left(\sup_{j\in\{1,\ldots,k\}} \sup_{f,g\in\mathcal{F}_j} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| \geq \epsilon \right) \leq \eta. \tag{4.4}$$

Let $\delta > 0$ (later on δ will be taken sufficiently small). Then one can cover $\mathcal{F}$ with δ-brackets (with respect to $L_2(P)$ norm) $[l_1, u_1]$,...,$[l_k, u_k]$, where $k = N_{[\,]}(\delta, \mathcal{F}, L_2(P)) < \infty$. These brackets create partition $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of $\mathcal{F}$ and it holds that

$$\|f - g\|_{L_2(P)} < \delta \quad \forall f, g \in \mathcal{F}_j,\ \forall j \in \{1, \ldots, k\}.$$

Now denote $p_n = \mathsf{P}^* \Big(\sup_{j\in\{1,\ldots,k\}} \sup_{f,g\in\mathcal{F}_j} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| \geq \epsilon \Big)$. Then with the help of Markov's inequality one can bound

$$\begin{aligned} p_n &\overset{\text{Markov}}{\leq} \frac{1}{\epsilon}\, \mathsf{E}^* \Big[\max_{j\in\{1,\ldots,k\}} \sup_{f,g\in\mathcal{F}_j} |\underbrace{\mathbb{G}_n(f) - \mathbb{G}_n(g)}_{=\mathbb{G}_n(f-g)}| \Big] \\ &\leq \frac{1}{\epsilon}\, \mathsf{E}^* \Big[\sup_{h\in\mathcal{H}} |\mathbb{G}_n(h)| \Big], \quad \text{where} \quad \mathcal{H} = \big\{ f - g \,|\, f, g \in \mathcal{F},\ P(f-g)^2 \leq \delta^2 \big\}. \end{aligned} \tag{4.5}$$

[1] In fact there is Log instead of log in (4.2), but note that the difference typically disappears as $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) \geq 3$ for sufficiently small ϵ.

Now show that (*homework exercise*)

$$N_{[]}(\delta, \mathcal{H}, L_2(P)) \leq \big[N_{[]}\big(\delta/2, \mathcal{F}, L_2(P)\big)\big]^2$$

and thus

$$\log N_{[]}(\delta, \mathcal{H}, L_2(P)) \leq 2 \log N_{[]}(\delta/2, \mathcal{F}, L_2(P)).$$

Now we are almost ready to use Lemma 2. But to that we need to have a measurable envelope. For this reason put

$$F(x) = \max_{j \in \{1,\ldots,k\}} \max\{|u_j(x)|, |l_j(x)|\}$$

and note that $2\,F$ is an envelope for $\mathcal{H}$.

Now with the help of Lemma 2 there exists a finite constant C such that

$$\mathsf{E}^* \Big[\sup_{h \in \mathcal{H}} |\mathbb{G}_n(h)|\Big] \leq C \Bigg(\int_0^{2\delta} \sqrt{2 \operatorname{Log} N_{[]}(\epsilon/2, \mathcal{F}, L_2(P))}\, d\epsilon + \sqrt{n}\, P\big(2F\, \mathbb{I}\{2F > \sqrt{n} a(\delta)\}\big) \Bigg), \quad (4.6)$$

where $a(\delta) = \frac{\delta}{\sqrt{2 \operatorname{Log} N_{[]}(\delta/2, \mathcal{F}, L_2(P))}}$.

Now note that the envelope F has a finite second moment, as

$$P(F^2) \leq \sum_{j=1}^{k} \big[P(u_j^2) + P(l_j^2)\big] < \infty.$$

Thus the second term on the right-hand side of inequality (4.6) for each $\delta > 0$ satisfies

$$\sqrt{n}\, P\Big(2F\, \mathbb{I}\Big\{F > \tfrac{\sqrt{n}\, a(\delta)}{2}\Big\}\Big) \leq \frac{4}{a(\delta)} P\Big(F^2\, \mathbb{I}\Big\{F > \tfrac{\sqrt{n}\, a(\delta)}{2}\Big\}\Big) \xrightarrow[n \to \infty]{} 0. \quad (4.7)$$

Further the first term on the right-hand side of inequality (4.6) for δ sufficiently small satisfies

$$\begin{aligned} \int_0^{2\delta} \sqrt{2 \operatorname{Log} N_{[]}(\epsilon/2, \mathcal{F}, L_2(P))}\, d\epsilon &= 2\sqrt{2} \int_0^{\delta} \sqrt{\operatorname{Log} N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon \\ &= 2\sqrt{2}\, \mathcal{J}_{[]}(\delta, \mathcal{F}, L_2(P)), \end{aligned} \quad (4.8)$$

which goes to zero for $\delta \to 0_+$.

Now combining (4.5), (4.6), (4.7) and (4.8) implies (4.4), which was to be proved.

□

Remark 7. Note that in comparison with Theorem 1, it is not straightforward to generalize this theorem to *not i.i.d.* settings. The problem is that it is rather complicated to generalize Lemma 2 to not i.i.d. settings.

Applications

Theorem 9 can be very useful when showing that a class of functions $\mathcal{F}$ is P-Donsker. One only needs to check that the bracketing integral is finite. Some sufficient conditions to guarantee this are discussed in Remark 6. The nice thing is that the bracketing numbers for the commonly used classes of functions are already known. These include for instance classes smooth functions, monotone functions, convex functions and functions Lipschitz in a parameter. See e.g. Chapter 2.7 of van der Vaart and Wellner (1996) or Chapter 2 of van de Geer (2000).

In what follows we concentrate on the classical empirical process and then on the empirical copula estimation. In this second example it is worth noting how the asymptotic uniform ρ-equicontinuity (see Definition 6) is utilised. This technique is often useful when one wants derive asymptotic properties of the estimators that are based on estimated quantities (for instance in regression problems the diagnostic statistics are often based on estimated residuals as the true regression errors are not observed).

Classical empirical process

Let $X_1, \dots, X_n$ be a random sample of univariate observations. Then the classical empirical process $\mathbb{F}_n$ is given by

$$\mathbb{F}_n(t) = \sqrt{n}\,\big(F_n(t) - F(t)\big), \quad t \in \mathbb{R},$$

where $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\}$ is the empirical cumulative distribution function.

Note that $\mathbb{F}_n$ can be viewed as the empirical process $\mathbb{G}_n$ introduced in (3.1) with

$$\mathcal{F} = \big\{x \mapsto \mathbb{I}\{x \leq t\}, t \in \mathbb{R}\big\}.$$

Now it is easy to check that

$$N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) \leq \frac{2}{\epsilon^2}.$$

Thus by Remark 6 the bracketing integral is finite and the class $\mathcal{F}$ is P-Donsker and so

$$\mathbb{F}_n \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathbb{R}), \tag{4.9}$$

where $\mathbb{G}$ is a zero mean Gaussian process with the covariance function given by

$$\begin{aligned}\mathsf{cov}\,\big(\mathbb{G}(t_1), \mathbb{G}(t_2)\big) &= \mathsf{E}\ \mathbb{I}\{X_i \leq t_1, X_i \leq t_2\} - \mathsf{E}\ \mathbb{I}\{X_i \leq t_1\}\ \mathsf{E}\ \mathbb{I}\{X_i \leq t_2\}\\ &= F(\min\{t_1, t_2\}) - F(t_1)\,F(t_2).\end{aligned}$$

Application to the Kolmogorov-Smirnov test

Let F_0 be a given cdf and suppose that one is interested in testing the hypotheses

$$H_0 : F(t) = F_0(t),\ \forall t \in \mathbb{R}, \quad H_1 : \exists t^* \in \mathbb{R}\ F(t^*) \neq F_0(t^*).$$

Then with the help of (4.9) and Theorem 4 (CMT) one gets that under the null hypothesis

$$\sqrt{n}\,\sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)| \rightsquigarrow \sup_{t\in\mathbb{R}} |\mathbb{G}(t)|, \tag{4.10}$$

as $\sup_{t\in\mathbb{R}} |\cdot|$ is a continuous functional in the space of bounded functions $\ell^\infty(\mathbb{R})$ equipped with the supremal norm.

Note that although Kolmogorov-Smirnov statistic is usually considered when F_0 is continuous, the above result (4.10) does not require the continuity of F_0. Nevertheless, if F_0 is continuous (and H_0 holds) then it is known that for $y > 0$:

$$\mathsf{P}\left(\sup_{t\in\mathbb{R}} |\mathbb{G}(t)| \leq y\right) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \mathrm{e}^{-2\,k^2 y^2}.$$

Exercise 4. Consider i.i.d. d-variate random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Show that then the corresponding classical empirical process

$$\mathbb{F}_n(\mathbf{t}) = \sqrt{n}\,\big(F_n(\mathbf{t}) - F(\mathbf{t})\big), \quad \mathbf{t} \in \mathbb{R}^d,$$

is P-Donsker. Note that it is not at all clear how the methods of the standard weak convergence theory presented in Billingsley (1999) (or in the course *NMTP434 Invariance Principles*) would be generalized to prove the weak convergence of the process $\mathbb{F}_n$.

Empirical copula process

Vaguely speaking a **copula** is a function that links univariate distributions together to create a multivariate distribution.

More precisely, let $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathsf{T}}$ be a random vector with the cumulative distribution function (cdf) $F_{\boldsymbol{X}}$. Further for $j \in \{1, \ldots, d\}$ denote F_j the corresponding marginal cdf of X_j (the j-th coordinate of $\boldsymbol{X}$). Then by Sklar's theorem (see e.g. Theorem 2.10.9. of Nelsen, 2006) there exists a function C defined on $[0,1]^d$ such that

$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big), \quad \forall (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{4.11}$$

Further if the cdfs $F_1, \ldots, F_d$ are continuous (which will be assumed in the sequel), then C is unique.

It can be shown that the copula function is a cdf on $[0,1]^d$ of the random vector

$$\big(U_1, \ldots, U_d\big)^{\mathsf{T}} = \big(F_1(X_1), \ldots, F_d(X_d)\big)^{\mathsf{T}}.$$

Thus the univariate margins of the copula function C are uniform on $[0,1]$.

Copula is sometimes called also a dependence function as it abstracts from the (univariate) marginal cdfs $F_1, \ldots, F_d$ and concentrates only on the dependence structure. Separating the margins from the dependence structure is also the reason why copulas became so popular in particular in financial applications.

Empirical estimation of C

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of from the distribution given by the generic random vector $\boldsymbol{X}$.

Estimation of the copula is not completely straightforward as the marginal cdfs $F_1, \ldots, F_d$ are typically unknown. Nevertheless with the help of (4.11) one can express the copula function C as

$$C(\mathbf{u}) = F_{\boldsymbol{X}}\big(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\big), \quad \text{where } \mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d.$$

Thus it is straightforward to estimate C as

$$C_n(\mathbf{u}) = F_n\big(F_{1n}^{-1}(u_1), \ldots, F_{dn}^{-1}(u_d)\big),$$

where

$$F_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \leq \mathbf{x}\} \quad \text{and} \quad F_{jn}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_{ji} \leq x\}.$$

In what follows we would like to show that the empirical copula process

$$\mathcal{C}_n(\mathbf{u}) = \sqrt{n}\,\big(C_n(\mathbf{u}) - C(\mathbf{u})\big), \quad \mathbf{u} \in [0,1]^d.$$

converges weakly as a function in $[0,1]^d$.

For this reason introduce the set of functions on $\mathbb{R}^d$

$$\mathcal{F} = \big\{\mathbf{x} \mapsto \mathbb{I}\{x_1 \leq F_1^{-1}(u_1), \dots, x_d \leq F_d^{-1}(u_d)\},\ \mathbf{u} \in [0,1]^d\big\}$$

and note that each of the functions from this can be identified with a unique $\mathbf{u} \in [0,1]^d$. Thus for simplicity of notation we will write $f_{\mathbf{u}}$ to distinguish functions from F. Further denote

$$f_{n\mathbf{u}}(\mathbf{x}) = \mathbb{I}\{x_1 \leq F_{1n}^{-1}(u_1), \dots, x_d \leq F_{dn}^{-1}(u_d)\}.$$

Note that $f_{n\mathbf{u}}$ is random but $P(f_{n\mathbf{u}} \in \mathcal{F}) = 1$. Using the empirical process notation one can write

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big\{X_{1i} \leq F_{1n}^{-1}(u_1), \dots, X_{di} \leq F_{dn}^{-1}(u_d)\big\} = P_n(f_{n\mathbf{u}}).$$

As for large n one can expect that $f_{n\mathbf{u}}$ is 'close' to $f_{\mathbf{u}}$ it seems to rewrite the process $\mathcal{C}_n$ as

$$\begin{aligned}\mathcal{C}_n(\mathbf{u}) &= \sqrt{n}\big(P_n(f_{n\mathbf{u}})) - P(f_{\mathbf{u}})\big)\\ &= \sqrt{n}\big[P_n(f_{n\mathbf{u}}) - P_n(f_{\mathbf{u}})\big] + \sqrt{n}\big[P_n(f_{\mathbf{u}}) - P_n(f_{\mathbf{u}})\big]\\ &\overset{Say}{=} \mathbb{A}_n(\mathbf{u}) + \mathbb{E}_n(\mathbf{u}) + \sqrt{n}\big[P_n(f_{\mathbf{u}})) - P(f_{\mathbf{u}})\big],\end{aligned} \tag{4.12}$$

where

$$\mathbb{A}_n(\mathbf{u}) = \sqrt{n}\,\big[P_n(f_{n\mathbf{u}}) - P_n(f_{\mathbf{u}})\big] - \sqrt{n}\,\big[P(f_{n\mathbf{u}}) - P(f_{\mathbf{u}})\big] \tag{4.13}$$

and

$$\mathbb{E}_n(\mathbf{u}) = \sqrt{n}\,\big[P(f_{n\mathbf{u}}) - P(f_{\mathbf{u}})\big]. \tag{4.14}$$

Note that in the expectation $P(f_{n\mathbf{u}})$ the function $f_{n\mathbf{u}}$ is considered as fixed, i.e.

$$P(f_{n\mathbf{u}}) = \mathsf{E}_{\boldsymbol{X}}\,\mathbb{I}\{X_1 \leq F_{1n}^{-1}(u_1), \dots, X_d \leq F_{dn}^{-1}(u_d)\} = F_{\boldsymbol{X}}\big(F_{1n}^{-1}(u_1), \dots, F_{dn}^{-1}(u_d)\big).$$

Dealing with $\mathbb{A}_n$

Note that with the help of (4.13) the process $\mathbb{A}_n$ can be rewritten as

$$\mathbb{A}_n(\mathbf{u}) = \mathbb{G}_n(f_{n\mathbf{u}} - f_{\mathbf{u}}). \tag{4.15}$$

Now with the help of Exercise 4 one knows that that the class $\mathcal{F}$ is P-Donsker. Thus with the help of Theorem 8 one knows that the process $\mathbb{G}_n$ is asymptotically uniformly ρ-equicontinuous for the semimetric ρ defined as

$$\rho(f_{\mathbf{u}_1}, f_{\mathbf{u}_2}) = \sqrt{\mathsf{E}\,\big[\mathbb{I}\{\boldsymbol{X} \leq \mathbf{F}^{-1}(\mathbf{u}_1)\} - \mathbb{I}\{\boldsymbol{X} \leq \mathbf{F}^{-1}(\mathbf{u}_2)\}\big]^2},$$

where for simplicity of notation we put $\mathbf{F}^{-1}(\mathbf{u}) = (F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))^{\mathsf{T}}$. Note that one can bound

$$\begin{aligned}\sup_{\mathbf{u}\in[0,1]^d} \rho^2(f_{n\mathbf{u}}, f_{\mathbf{u}}) &\leq \sup_{\mathbf{u}\in[0,1]^d} \sum_{j=1}^{d} \big|F_j\big(F_{jn}^{-1}(u_j)\big) - u_j\big|\\ &\leq \sup_{u\in[0,1]} \sum_{j=1}^{d} \big[\big|F_j(u) - F_{jn}(u)\big| + \tfrac{1}{n}\big] \xrightarrow[n\to\infty]{P} 0.\end{aligned}$$

Now with the help of (4.15) and from Definition 6 (of asymptotic ρ-equicontinuity) one can conclude that

$$\sup_{\mathbf{u}\in[0,1]^d} |\mathbb{A}_n(\mathbf{u})| \xrightarrow[n\to\infty]{P^*} 0. \tag{4.16}$$

Dealing with $\mathbb{E}_n$[2]

With the help of (4.14) the process $\mathbb{E}_n$ can be rewritten as

$$\begin{aligned}\mathbb{E}_n(\mathbf{u}) &= \sqrt{n}\,\big[F_{\mathbf{X}}\big(F_{1n}^{-1}(u_1),\ldots,F_{dn}^{-1}(u_d)\big) - F_{\mathbf{X}}\big(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d)\big)\big] \\ &= \sqrt{n}\,\Big[C\Big(F_1\big(F_{1n}^{-1}(u_1)\big),\ldots,F_d\big(F_{dn}^{-1}(u_d)\big)\Big) - C(u_1,\ldots,u_d)\Big]. \end{aligned}\tag{4.17}$$

Now the idea is to use the first order Taylor expansion. Provided that one can assume that for each $j \in \{1,\ldots,d\}$ the first order partial derivative of the copula function $C^{(j)}(\mathbf{u}) = \frac{\partial C(\mathbf{u})}{\partial u_j}$ is continuous on the set $V_j = \big\{\mathbf{u} \in [0,1]^d : 0 < u_j < 1\big\}$, then using (4.17) one can show

$$\sup_{\mathbf{u}\in[0,1]^d} \left| \mathbb{E}_n(\mathbf{u}) - \sum_{j=1}^d C^{(j)}(\mathbf{u})\sqrt{n}\big[F_j\big(F_{jn}^{-1}(u_j)\big) - u_j\big] \right| \xrightarrow[n\to\infty]{P^*} 0.$$

Now one needs

$$\sup_{u\in[0,1]} \left| \sqrt{n}\big[F_j\big(F_{jn}^{-1}(u)\big) - u\big] + \frac{1}{\sqrt{n}}\sum_{i=1}^n \big[\mathbb{I}\{F_j(X_{ji}) \le u\} - u\big] \right| \xrightarrow[n\to\infty]{P^*} 0$$

to show that

$$\sup_{\mathbf{u}\in[0,1]^d} \left| \mathbb{E}_n(\mathbf{u}) + \frac{1}{\sqrt{n}}\sum_{j=1}^d C^{(j)}(\mathbf{u})\big[\mathbb{I}\{U_{ji} \le u_j\} - u_j\big] \right| \xrightarrow[n\to\infty]{P^*} 0 \quad \text{where } U_{ji} = F_j(X_{ji}). \tag{4.18}$$

Dealing with $\mathcal{C}_n$

Now combining (4.12), (4.16) and (4.18) gives us the following i.i.d. representation of the copula process

$$\sup_{\mathbf{u}\in[0,1]^d} \left| \mathcal{C}_n(\mathbf{u}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \big[\mathbb{I}\{U_{1i} \le u_1,\ldots,U_{di} \le u_d\} - C(\mathbf{u})\big] \right. \\ \left. + \frac{1}{\sqrt{n}}\sum_{j=1}^d C^{(j)}(\mathbf{u})\big[\mathbb{I}\{U_{ji} \le u_j\} - u_j\big] \right| \xrightarrow[n\to\infty]{P^*} 0.$$

This can be further rewritten with the help of the empirical process as

$$\sup_{\mathbf{u}\in[0,1]^d} \left| \mathcal{C}_n(\mathbf{u}) - \mathbb{G}_n(f_{\mathbf{u}}) + \sum_{j=1}^d C^{(j)}(\mathbf{u})\mathbb{G}_n(f_{\mathbf{u}^{(j)}}) \right| \xrightarrow[n\to\infty]{P^*} 0,$$

where $\mathbf{u}^{(j)}$ denotes the vector whose jth entry is u_j and the $d-1$ others are 1. From this one can conclude that

$$\mathcal{C}_n \rightsquigarrow \mathbb{G}(f_{\mathbf{u}}) - \sum_{j=1}^d C^{(j)}(\mathbf{u})\,\mathbb{G}(f_{\mathbf{u}^{(j)}}),$$

[2] This step is only briefly sketched with missing derivations of some of the equations. It is included in the text just to give the idea how the proof proceeds.

where $\mathbb{G}$ is a Brownian bridge on $[0,1]^d$ whose covariance function is given, for all $\mathbf{u},\mathbf{v} \in [0,1]^d$, by

$$\begin{aligned}\mathsf{cov}\big(\mathbb{G}(f_{\mathbf{u}}),\mathbb{G}(f_{\mathbf{v}})\big) &= \mathsf{E}\,\big[\mathbb{I}\{U_{1i} \leq u_1,\ldots,U_{di} \leq u_d\}\,\mathbb{I}\{U_{1i} \leq v_1,\ldots,U_{di} \leq v_d\}\big]\\ &\quad - \mathsf{E}\,\big[\mathbb{I}\{U_{1i} \leq u_1,\ldots,U_{di} \leq u_d\}\big]\,\mathsf{E}\,\big[\mathbb{I}\{U_{1i} \leq v_1,\ldots,U_{di} \leq v_d\}\big]\\ &= C\big(u_1 \wedge v_1,\ldots,u_d \wedge v_d\big) - C(\mathbf{u})\,C(\mathbf{v}).\end{aligned}$$

Exercise 5. Let $X_1,\ldots,X_n$ be a random sample from the uniform distribution on $[0,2\pi]$. For $a \in [0,10]$ define

$$Y_n(a) = \frac{1}{n}\sum_{i=1}^{n} \cos(a\,X_i).$$

Show that the the appropriately standardized process $\{Y_n(a), a \in [0,10]\}$ converges in distribution. Describe the limiting process.

* Now, consider that $\widehat{a}_n = \frac{\pi}{\overline{X}_n}$. Show how one can derive the asymptotic distribution of $\sqrt{n}\,Y_n(\widehat{a}_n)$.

Hint. Denote $Y(a) = \mathsf{E}\,\cos(a\,X_i)$. *Note that*

$$\sqrt{n}\,\big[Y_n(\widehat{a}_n) - Y_n(1)\big] = \sqrt{n}\,\big[Y_n(\widehat{a}_n) - Y(\widehat{a}_n) - Y_n(1) + Y(1)\big] + \sqrt{n}\,\big[Y(\widehat{a}_n) - Y(1)\big].$$

Now with the help of uniform asymptotic ρ-equicontinuity of the process $\mathbb{G}_n(a) = \sqrt{n}\,\big[Y_n(a) - Y(a)\big]$ one can show that the first term on the right-hand side of the previous equation converges to zero in (outer) probability. Thus one gets

$$\sqrt{n}\,Y_n(\widehat{a}_n) = \sqrt{n}\,Y_n(1) + \sqrt{n}\,\big[Y(\widehat{a}_n) - Y(1)\big] + o_{P^*}(1).$$

Exercise 6. Let $X_1,\ldots,X_n$ be a random sample such that X_i has an exponential distribution with the mean equal to 1. Consider the process

$$Z_n(a) = \frac{1}{n}\sum_{i=1}^{n} |X_i - a|, \quad a \in [0,2].$$

Show that the appropriately standardized process $\{Z_n(a), a \in [0,2]\}$ converges in distribution. Describe the limiting process.

* With the help of the previous result derive the asymptotic distribution of

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n| - \mathsf{E}\,|X_i - \mathsf{E}\,X_i|\right).$$

5 Covering numbers and VC-classes

Let $\mathcal{F}$ be a class of real functions ($f : \mathcal{X} \to \mathbb{R}$) that is equipped with the norm $\|\cdot\|$. As in Chapter 2 we are interested in the norms that posses the (Riesz) property, i.e. if $|f(x)| \leq |g(x)|$ for all $x \in \mathcal{X}$ then $\|f\| \leq \|g\|$.

Sometimes the approach using bracketing numbers is not useful, as the bracketing numbers are too big and do not satisfy assumptions of Theorem 1 or Theorem 9. Or simply the bounds on the bracketing numbers are not available and one does not know how to derive appropriate bounds. In such situations it may be useful to use an alternative approach how to measure the 'size' of a class of functions.

Let $g : \mathcal{X} \to \mathbb{R}$ be a given function. Denote $B_\epsilon(g) = \{f \in \mathcal{F} \mid \|f - g\| < \epsilon\}$ the *ball* of radius ϵ (of functions in $\mathcal{F}$) with the center g.

Definition 7. The **covering number** $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $B_\epsilon(g)$ of radius ϵ needed to cover the set $\mathcal{F}$. The centres of the balls need not belong to $\mathcal{F}$, but they should have finite norms.

Note that if $N(\epsilon, \mathcal{F}, \|\cdot\|) < \infty$ for each $\epsilon > 0$, then $(\mathcal{F}, \|\cdot\|)$ is totally bounded. Further note that (**exercise**)

$$N(\epsilon, \mathcal{F}, \|\cdot\|) \leq N_{[\,]}(2\epsilon, \mathcal{F}, \|\cdot\|). \tag{5.1}$$

But in general there si no converse inequality. Thus the bracketing numbers can in general be much bigger than the covering numbers.

On the other hand the nice thing about the bracketing numbers is that the sufficient conditions for the theorems of interests are stated in terms of *only one norm*. Note that only $L_1(P)$ is used for Glivenko-Cantelli property stated in Theorem 1 and $L_2(P)$ for Donsker property stated in Theorem 9. As we will see later, the sufficient conditions in terms of the covering numbers will involve a bound on the covering numbers with respect to *'many norms'*.[1]

Remark 8. The only general exception when the converse inequality to (5.1) holds is the case of *supremum norm*, i.e.

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

Then the the pair of functions $g(x) - \epsilon$ and $g(x) + \epsilon$ forms a 2ϵ-bracket and thus

$$N_{[\,]}(2\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq N(\epsilon, \mathcal{F}, \|\cdot\|_\infty),$$

which together with the general inequality (5.1) yields that

$$N_{[\,]}(2\epsilon, \mathcal{F}, \|\cdot\|_\infty) = N(\epsilon, \mathcal{F}, \|\cdot\|_\infty),$$

[1] To be more specific, later we will see that we need a bound on the *random* covering numbers $N(\epsilon, \mathcal{F}, L_1(P_n))$ for GC-property and on $N(\epsilon, \mathcal{F}, L_2(P_n))$ for Donsker property, where P_n is the empirical measure.

Example 2. **(Classes of smooth functions)**
Let $\mathcal{X}$ be an open, bounded and convex subset of $\mathbb{R}$. For $\alpha > 0$ put $\underline{\alpha} = \sup\{n \in \mathbb{N} \mid n < \alpha\}$. Further let $C_M^\alpha(\mathcal{X})$ be the class of functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$\max_{k \leq \underline{\alpha}} \sup_{x \in \mathcal{X}} |f^{(k)}(x)| + \sup_{x \neq y} \frac{|f^{(\underline{\alpha})}(x) - f^{(\underline{\alpha})}(y)|}{|x-y|^{\alpha - \underline{\alpha}}} \leq M$$

where $f^{(k)}$ denote the k-th derivative of f with $f^{(0)} = f$.
For instance for $\alpha = 1$ ($\underline{\alpha} = 0$) the set of functions $C_M^1(\mathcal{X})$ contain the functions whose absolute values are bounded by M and that are Lipschitz-continuous with the Lipschitz-constant bounded also by M.

Then (see e.g. Theorem 2.7.1 van der Vaart and Wellner, 1996) there exists a constant K such that

$$N(\epsilon, C_M^\alpha(\mathcal{X}), \|\cdot\|_\infty) = N_{[]}(2\epsilon, C_M^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K \exp\left\{\epsilon^{-\frac{1}{\alpha}}\right\}.$$

Note that a bigger α requires smoother functions which results in a smaller bracketing number.

VC-classes of sets

Very important classes of sets whose covering numbers are well controlled for our purposes (i.e. for stating the analogies of Theorems 1 and 9) are *Vapnik-Červonenkis* classes of functions, or simply *VC-classes.*[2] We start with VC-classes of sets and then we move to VC-classes of functions.

Let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$ and $\{x_1, \ldots, x_n\}$ be a finite subset of $\mathcal{X}$. We say that $\mathcal{C}$ *picks out* a certain subset of $\{x_1, \ldots, x_n\}$ (e.g. the subset $\{x_2, x_5\}$), if it can be written as $\{x_1, \ldots, x_n\} \cap C$ for some $C \in \mathcal{C}$ (e.g. there exists $C \in \mathcal{C}$ such that $\{x_2, x_5\} = \{x_1, \ldots, x_n\} \cap C$).

Further, the class $\mathcal{C}$ is said to *shatter* the set $\{x_1, \ldots, x_n\}$, if $\mathcal{C}$ picks out out each of its 2^n subsets.

Example 3. The class of sets $\mathcal{C} = \{(-\infty, t] \mid t \in \mathbb{R}\}$ (when $\mathcal{X} = \mathbb{R}$) shatters one-point sets, but it does not shatter two-point sets. To see this let $\{x_1, x_2\}$ be such that $x_1 < x_2$. Then $\mathcal{C}$ fails to pick out the set $\{x_2\}$.

The **VC-index** $V(\mathcal{C})$ of the class of sets $\mathcal{C}$ is the the smallest n for which no set of size n is shattered by $\mathcal{C}$. More formally, let $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ be the the number of subsets of $\{x_1, \ldots, x_n\}$ that can be chosen by $\mathcal{C}$, i.e.

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \#\Big\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\Big\}.$$

Then the VC-index is given by

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

Definition 8. A collection of measurable sets $\mathcal{C}$ is called **VC-class** if $V(\mathcal{C}) < \infty$.

Example 4. (a) Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \{(-\infty, t] \mid t \in \mathbb{R}\}$. Then from Example 3 we know that $V(\mathcal{C}) = 2$.

(b) Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \{(a, b] \mid a, b \in \mathbb{R}\}$. Then $V(\mathcal{C}) = 3$ (**exercise**).

[2] These sets were introduced in theoretical problems related to pattern recognition and later also in other machine learning problems.

(c) Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{C} = \{(-\infty, \mathbf{t}] \mid \mathbf{t} \in \mathbb{R}^d\}$. Then $V(\mathcal{C}) = d + 1$ (**exercise**).

(d) Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{C} = \{C \subset \mathbb{R}^2 \mid C \text{ is closed and convex}\}$. Then $V(\mathcal{C}) = \infty$ and thus this class of sets is not a VC-class.
To see this let the points $x_1, \dots, x_n$ be on the border of the unit circle. Then $\mathcal{C}$ shatters $\{x_1, \dots, x_n\}$ for each $n \in \mathbb{N}$.

The crucial property of VC-classes of sets is that from a set of size n it can pick out at most $O(n^{V(\mathcal{C})-1})$ subsets. This combinatorial result is known as Sauer's Lemma (see Corollary 2.6.3 of van der Vaart and Wellner, 1996).

Lemma 3. ***(Sauer's Lemma)***
Let $\mathcal{C}$ be a VC-class. Then

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

Consider now the class of functions $\mathcal{F}_{\mathcal{C}}$ given by the indicators of the sets $\mathcal{C}$, i.e.

$$\mathcal{F}_{\mathcal{C}} = \big\{x \to \mathbb{I}\{x \in C\} : \ C \in \mathcal{C}\big\}. \tag{5.2}$$

The proof of the following theorem can be found in van der Vaart and Wellner (1996), Theorem 2.6.4.

Theorem 10. *Let $\mathcal{C}$ be a VC-class of sets and $\mathcal{F}_{\mathcal{C}}$ be given by* (5.2). *Then there exists a* ***universal*** *constant K such that for* ***any*** *probability measure Q, any $r \geq 1$ and $0 < \epsilon < 1$*

$$N(\epsilon, \mathcal{F}_{\mathcal{C}}, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{C})-1)}.$$

There are several things worth noting. First, the covering numbers of indicators of VC-classes are bounded by a polynomial in $\left(\frac{1}{\epsilon}\right)$. Further the important thing is that the polynomial bound is uniform in the probability measure Q. Thus the result can be also restated in the form that there exist finite constants K and W such that

$$\sup_Q N\big(\epsilon, \mathcal{F}_{\mathcal{C}}, L_r(Q)\big) \leq K \left(\frac{1}{\epsilon}\right)^W,$$

where the supremum is taken with respect to all probability measures Q on the sample space $\mathcal{X}$.

The other nice properties of the VC-classes of sets is that it is closed with respect to standard operations. See Lemma 2.6.18 of van der Vaart and Wellner (1996) for a more general result.

Lemma 4. *Let $\mathcal{C}$ and $\mathcal{D}$ be VC-classes. Then also $\{C^c \mid C \in \mathcal{C}\}$, $\{C \cap D \mid C \in \mathcal{C},\ D \in \mathcal{D}\}$ and $\{C \cup D \mid C \in \mathcal{C},\ D \in \mathcal{D}\}$ are VC-classes.*

Proof. We will show the lemma only for $\mathcal{C} \cap \mathcal{D} = \{C \cap D \mid C \in \mathcal{C},\ D \in \mathcal{D}\}$. The remaining statements can be proved analogously (see Lemma 2.6.17 of van der Vaart and Wellner, 1996).

By Lemma 3 from each n points set $\{x_1, \dots, x_n\}$ the class $\mathcal{C}$ picks out at most $O(n^{V(\mathcal{C})-1})$ subsets. From each of this subsets the class $\mathcal{D}$ picks out at most $O(n^{V(\mathcal{D})-1})$ further subsets. Thus $\mathcal{C} \cap \mathcal{D}$ picks out at most $O(n^{V(\mathcal{C})+V(\mathcal{D})-2})$ subsets which is less than 2^n for n large enough. □

VC-classes of functions

Consider a function $f : \mathcal{X} \to \mathbb{R}$. Note that this function can be characterised by the subset

$$\big\{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\big\}. \tag{5.3}$$

The set of the form (5.3) is called the *subgraph* of the function f.

Definition 9. A collection $\mathcal{F}$ of measurable real functions on $\mathcal{X}$ is called a **VC-class of functions**, if the collections of all subgraphs of the functions in $\mathcal{F}$ form 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 $\mathcal{F}$. Further for simplicity of notation denote for a measure Q on $\mathcal{X}$ the L_r norm of a function $f : \mathcal{X} \to \mathbb{R}$ as $\|f\|_{Q,r}$, i.e.

$$\|f\|_{Q,r} = \|f\|_{L_r(Q)} = \bigg[\int_{\mathcal{X}} |f(x)|^r \, dQ(x)\bigg]^{1/r} = \big[Q(|f|^r)\big]^{1/r}.$$

The following theorem says, that similarly as in Theorem 10, also for the VC-classes of functions the covering numbers grow only at a polynomial rate a as $\epsilon \to 0_+$.

Theorem 11. *Let $\mathcal{F}$ be a VC-class of functions with measurable envelope function F. Then there exists $K < \infty$ such that for each $r \geq 1$, any probability measure Q on $\mathcal{X}$ such that $\|F\|_{Q,r} > 0$ and for each $\epsilon \in (0,1)$*

$$N\big(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big) \leq K \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Proof. For $f \in \mathcal{F}$ let C_f be the corresponding subgraph. Now let $\mathcal{C} = \{C_f : f \in \mathcal{F}\}$ be the set of all subgraphs.

Consider first $\underline{r=1}$. Let λ be a Lebesgue measure on $\mathbb{R}$. Note that by the Fubini's theorem for any measurable functions f and g (**exercise**)

$$\begin{aligned}\|f-g\|_{Q,1} = Q(|f-g|) &= \int_{\mathcal{X}} |f(x)-g(x)| \, dQ(x) \\ &= \int_{\mathcal{X}\times\mathbb{R}} \mathbb{I}\{0 < t < |f(x)-g(x)|\} \, d(Q\otimes\lambda)(x,t) = (Q\otimes\lambda)\,(C_f \triangle C_g),\end{aligned} \tag{5.4}$$

where $C_f \triangle C_g = (C_f \cup C_g) \setminus (C_f \cap C_g)$ stands for the symmetric difference of the sets C_f and C_g.

Define now the measure $P_\lambda = (Q\otimes\lambda)/(2Q(F))$, where F is the envelope function. Note that P_λ is a probability measure on the set $\big\{(x,t) \in \mathcal{X} \times \mathbb{R} : |t| \leq F(x)\big\}$.

Let $\mathbb{I}_{C_f}(\cdot)$ stands for the indicator function $(x,t) \to \mathbb{I}\{(x,t) \in C_f\}$. Now with the help of (5.4)

$$P_\lambda(C_f \triangle C_g) = \|\mathbb{I}_{C_f} - \mathbb{I}_{C_g}\|_{P_\lambda,1} < \epsilon \quad \text{implies that} \quad \|f-g\|_{Q,1} < 2\,Q(F)\,\epsilon = \epsilon\,2\,\|F\|_{Q,1}.$$

Let $\mathcal{F}_{\mathcal{C}}$ be the set of indicator functions defined of the form $\mathbb{I}_{C_f}$, where $f \in \mathcal{F}$. Then with the help of Theorem 10 one gets

$$N\big(\epsilon\,2\,\|F\|_{Q,1}, \mathcal{F}, L_1(Q)\big) = N\big(\epsilon, \mathcal{C}, L_1(P_\lambda)\big) \leq \widetilde{K}\,\big(\tfrac{1}{\epsilon}\big)^{V(\mathcal{F})-1}, \tag{5.5}$$

for a universal finite constant $\widetilde{K}$.

Now consider $r > 1$. Note that for each $f, g \in \mathcal{F}$

$$\begin{aligned} Q(|f-g|^r) &\le Q\big(|f-g|\,(2F)^{r-1}\big) = 2^{r-1}\int_{\mathcal{X}} |f(x)-g(x)|\,F^{r-1}(x)\,dQ(x) \\ &= 2^{r-1}\int_{\mathcal{X}} \tfrac{|f(x)-g(x)|\,F^{r-1}(x)}{Q(F^{r-1})}\,dQ(x)\,Q(F^{r-1}) \\ &= 2^{r-1}\widetilde{Q}(|f-g|)\,Q(F^{r-1}), \end{aligned}$$

where the measure $\widetilde{Q}$ has a density $F^{r-1}(x)/Q(F^{r-1})$ with respect to the measure Q. Thus

$$\|f-g\|_{Q,r} = \big[Q(|f-g|^r)\big]^{1/r} \le 2\,\big[Q(F^{r-1})\big]^{1/r}\,\|f-g\|_{\widetilde{Q},1}^{1/r}.$$

Suppose now that $\|f-g\|_{\widetilde{Q},1} \le \epsilon^r\,\|F\|_{\widetilde{Q},1} = \epsilon^r\,\widetilde{Q}(F)$. Then with help of the previous inequality

$$\|f-g\|_{Q,r} \le 2\,\big[Q(F^{r-1})\big]^{1/r}\,\epsilon\,\big[\widetilde{Q}(F)\big]^{1/r} \le 2\,\epsilon\,\big[Q(F^{r-1})\big]^{1/r}\,\tfrac{[Q(F^r)]^{1/r}}{[Q(F^{r-1})]^{1/r}} = \epsilon\,2\,\|F\|_{Q,r},$$

which together with (5.5) implies that

$$N\big(\epsilon\,2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big) \le N\big(\epsilon^r\,\|F\|_{\widetilde{Q},1}, \mathcal{F}, L_1(\widetilde{Q})\big) \le \widetilde{K}\big(\tfrac{2}{\epsilon^r}\big)^{V(\mathcal{F})-1}$$

and concludes the proof of the theorem (with $K = \widetilde{K}\,2^{V(\mathcal{F})-1}$). □

Similarly as in Lemma 4 one can show that the VC-classes of functions are closed under many standard operations/transformations. For a more general result see Lemma 2.6.18 of van der Vaart and Wellner (1996).

Lemma 5. *Let $\mathcal{F}$ and $\mathcal{G}$ be VC-classes of functions and $g : \mathcal{X} \to \mathbb{R}$ a given function. Then also the following classes are VC-classes of functions.*

(i) $\mathcal{F} \vee \mathcal{G} = \{f \vee g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$;

(ii) $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$;

(iii) $\mathcal{F} + g = \{f + g \mid f \in \mathcal{F}\}$.

Proof. (i) and (ii). Let C_f be a subgraph of the function f. Then $C_{f\vee g} = C_f \cup C_g$ and $C_{f\wedge g} = C_f \cap C_g$. Thus with the help of Lemma 4 also $\mathcal{F} \vee \mathcal{G}$ and $\mathcal{F} \wedge \mathcal{G}$ are VC-classes.

(iii). Note that the subgraphs of $\mathcal{F} + g$ shatter a given set of points $\{(x_1, t_1), \ldots, (x_n, t_n)\}$ if and only if the subgraphs of $\mathcal{F}$ shatter the set $\big\{\big(x_1, t_1 - g(x_1)\big), \ldots, \big(x_n, t_n - g(x_n)\big)\big\}$. □

Some concluding comments on bounding covering numbers

VC-classes are a very useful starting point for controlling covering numbers (uniformly in the measures on the sample space). In many applications the sets of functions of interest are already VC-classes or can be derived from the VC-classes (by an appropriate transformation or combination). It is for instance worth noting that there exist useful bounds (in terms of Glivenko-Cantelli and Donsker properties) for the convex hulls of VC-classes of functions. For further results and details see for instance Chapter 2.6 of van der Vaart and Wellner (1996).

6 Glivenko-Cantelli theorems with the help of covering numbers

The goal of this chapter to find an analogy of Theorem 1, where the sufficient assumptions on the size of the class $\mathcal{F}$ is given in terms of covering numbers.

Preliminary results

In what follows we make use of Hoeffding's inequality. This is a useful inequality for independent (but not necessarily identically distributed) random variables that are bounded.

Lemma 6. ***(Hoeffding's inequality)***
Let $Y_1, \dots, Y_n$ be independent random variables such that $\mathsf{E}\,[Y_i] = 0$ and $a_i \leq Y_i \leq b_i$, $i = 1, \dots, n$. Then for each $\eta > 0$ it holds that

$$\mathsf{P}\left(\Big|\sum_{i=1}^{n} Y_i\Big| > \eta\right) \;\leq\; 2\exp\left(\frac{-2\,\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

The important and useful trick that is used in the proofs of Theorems 12 and 13 is the *symmetrization* (below). This trick says that instead of of the empirical process

$$f \mapsto \big(P_n - P\big)(f) = \frac{1}{n}\sum_{i=1}^{n}\big[f(X_i) - \mathsf{E}\,f(X_i)\big]$$

one can consider the symmetrized process

$$f \mapsto P_n^0(f) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\, f(X_i),$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed random variables such that

$$\mathsf{P}(\varepsilon_i = 1) = \mathsf{P}(\varepsilon_i = -1) = \tfrac{1}{2}$$

that are independent of $(X_1, \dots, X_n)$. Random variables $\varepsilon_1, \dots, \varepsilon_n$ are often also called *Rademacher* random variables.[1]

Let P_ε stand for the probability calculated with respect to Rademacher random variables (conditionally on $X_1, \dots, X_n$). Then with the help of Hoeffding's lemma one gets the following exponential bound for the symmetrized empirical process

$$\begin{aligned}\mathsf{P}_\varepsilon\left(\big|P_n^0(f)\big| > \eta\right) &= \mathsf{P}_\varepsilon\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\, f(X_i)\Big| > \eta\right)\\ &\leq 2\exp\left(\tfrac{-2\,n^2\,\eta^2}{4\sum_{i=1}^{n} f^2(X_i)}\right) = 2\exp\left(\tfrac{-n\,\eta^2}{2P_n(f^2)}\right). \qquad (6.1)\end{aligned}$$

[1] Note that the cumulative sum of Rademacher random variables gives a standard symmetric random walk.

The important thing is that (for fixed $\eta > 0$) the right-hand side of the above inequality converges to zero at an exponential rate provided that the random variable $P_n(f^2)$ is bounded. This will be very helpful later.

The following lemma justifies the symmetrization (for large sample sizes).

Lemma 7. ***(Symmetrization)***
Let P_n be the empirical measure of i.i.d. random variables $X_1, \dots, X_n$ and $\mathcal{F}$ be a class of functions bounded by a finite constant M. Then for $n \geq \frac{8M^2}{\eta^2}$ it holds that

$$\mathsf{P}^* \big(\|P_n - P\|_{\mathcal{F}} > \eta \big) \;\leq\; 4\, \mathsf{P}^* \big(\|P_n^0\|_{\mathcal{F}} > \tfrac{\eta}{4} \big),$$

where the notation $\|Q\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |Q(f)|$ is used.

Note that while the probability on the left-hand side in the statement of the lemma is calculated with respect to $X_1, \dots, X_n$, the probability on the right-hand side is calculated with respect to the joint distribution of $X_1, \dots, X_n$ and $\varepsilon_1, \dots, \varepsilon_n$.

Proof of Lemma 7. Let $X_1', \dots, X_n'$ be independent copies of $X_1, \dots, X_n$. Denote the corresponding empirical measure of $X_1', \dots, X_n'$ by P_n'. Then for each $f \in \mathcal{F}$ by Chebyshev's inequality it holds that

$$\mathsf{P} \big(|P_n'(f) - P(f))| > \tfrac{\eta}{2} \big) \leq \frac{\mathsf{var}\big(f(X_i')\big)}{n\,(\eta/2)^2} \leq \frac{4\,M^2}{n\,\eta^2} \leq \frac{1}{2} \tag{6.2}$$

for $n \geq \frac{8M^2}{\eta^2}$.

Suppose that $\{\|P_n - P\|_{\mathcal{F}} > \eta\}$ holds. Then there exists $f \in \mathcal{F}$ such that $|P_n(f) - P(f))| \geq \eta$. Fix this function f for a moment. Then with the help of (6.2) for $n \geq \frac{8M^2}{\eta^2}$:

$$\tfrac{1}{2}\, \mathbb{I}\{\|P_n - P\|_{\mathcal{F}} > \eta\} \leq \mathsf{P} \big(|P_n'(f) - P(f))| \leq \tfrac{\eta}{2} \big)\, \mathbb{I}\{\|P_n - P\|_{\mathcal{F}} > \eta\}. \tag{6.3}$$

Note that on the intersection of the events $\big\{|P_n'(f) - P(f)| \leq \frac{\eta}{2}\big\}$ and $\big\{\|P_n - P\|_{\mathcal{F}} > \eta\big\}$ it holds that

$$|P_n'(f) - P_n(f)| \geq |P_n(f) - P(f)| - |P_n'(f) - P(f)| > \tfrac{\eta}{2}. \tag{6.4}$$

Combining (6.3) and (6.4) one gets

$$\begin{aligned} \tfrac{1}{2}\, \mathbb{I}\{\|P_n - P\|_{\mathcal{F}} > \eta\} &\leq \mathsf{P}_{X'} \big(|P_n'(f) - P_n(f)| > \tfrac{\eta}{2} \big)\, \mathbb{I}\{\|P_n - P\|_{\mathcal{F}} > \eta\} \\ &\leq \mathsf{P}^*_{X'} \big(\|P_n' - P_n\|_{\mathcal{F}} > \tfrac{\eta}{2} \big), \end{aligned} \tag{6.5}$$

where $\mathsf{P}_{X'}$ stands for the probability calculated (only) with respect to $X_1', \dots, X_n'$ (with $X_1, \dots, X_n$ in the event $\big\{\|P_n - P\|_{\mathcal{F}} > \eta\big\}$ being fixed). Now with the help of (6.5) one gets

$$\begin{aligned} \mathsf{P}^* \big(\|P_n - P\|_{\mathcal{F}} > \eta \big) &\leq 2\; \mathsf{E}^*_X \big[\mathsf{P}^*_{X'} \big(\|P_n' - P_n\|_{\mathcal{F}} > \tfrac{\eta}{2} \big) \big] \\ &\leq 2\; \mathsf{P}^* \big(\|P_n' - P_n\|_{\mathcal{F}} > \tfrac{\eta}{2} \big), \end{aligned} \tag{6.6}$$

where justifying the last inequality is technical and it will be postponed to the end of the proof.

Now suppose that Rademacher random variables $\varepsilon_1, \dots, \varepsilon_n$ are independent of $X_1, \dots, X_n$ as well as $X_1', \dots, X_n'$. The important thing is that due to symmetry for each measurable function f the distribution of the difference $f(X_i') - f(X_i)$ is the same as the distribution of $\varepsilon_i\big(f(X_i') - f(X_i)\big)$. Thus one can bound

$$\begin{aligned}\mathsf{P}^*\left(\|P_n' - P_n\|_{\mathcal{F}} > \tfrac{\eta}{2}\right) &= \mathsf{P}^*\left(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(f(X_i') - f(X_i)\big)\Big| > \tfrac{\eta}{2}\right)\\ &\le \mathsf{P}^*\left(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\, f(X_i')\Big| + \sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\, f(X_i)\Big| > \tfrac{\eta}{2}\right)\\ &\le 2\ \mathsf{P}^*\left(\|P_n^0\|_{\mathcal{F}} > \tfrac{\eta}{4}\right),\end{aligned}$$

which together with (6.6) verifies the statement of the lemma.

Finally, the inequality in (6.6) can be justified as follows. For simplicity of notation denote $A = \big\{\|P_n - P_n'\|_{\mathcal{F}} > \frac{\eta}{2}\big\}$. Then

$$\begin{aligned}\mathsf{P}^*(A) &= \inf\big\{\,\mathsf{P}(B)\mid A\subset B \text{ where } B \text{ is a measurable set}\big\}\\ &= \inf\big\{\,\mathsf{E}_X\big[\mathsf{E}_{X'}[\mathbb{I}_B]\big]\mid \mathbb{I}_A \le \mathbb{I}_B \text{ where } B \text{ is a measurable set}\big\}\\ &\ge \inf\big\{\,\mathsf{E}_X\big[\mathsf{E}_{X'}[V]\big]\mid \mathbb{I}_A \le V \text{ where } V \text{ is a measurable random variable}\big\}\\ &\ge \mathsf{E}_X^*\big[\inf\big\{\,\mathsf{E}_{X'}[V]\mid \mathbb{I}_A \le V \text{ where } V \text{ is a measurable random variable}\big\}\big]\\ &\ge \mathsf{E}_X^*\big[\mathsf{E}_{X'}^*[\mathbb{I}_A]\big],\end{aligned}$$

where we use the fact that for each V (jointly measurable in $X_1,\ldots,X_n$ and $X_1',\ldots,X_n'$) by the Fubini Theorem one has that $\mathsf{E}_{X'}[V]$ is a measurable majorant for $\inf\big\{\,\mathsf{E}_{X'}[V]\mid \mathbb{I}_A \le V$ where V is a measurable random variable$\big\}$. □

Unfortunately, one cannot in general proceed without any measurability assumptions. The reason is that in the proof of Theorems12 and 13 (below) we need to write the joint outer probability $\mathsf{P}^*(A)$ of an event A on the probability space where $X_1,\ldots,X_n$ and $\varepsilon_1,\ldots,\varepsilon_n$ are defined as $\mathsf{E}_X^*\,\mathsf{P}_\varepsilon(A)$. But Fubini's theorem is not valid for outer expectations.

To overcome this difficulty we will assume that the supremum $\|P_n^0\|_{\mathcal{F}}$ is a measurable random variable. Since the Rademacher variables are discrete this is the case if and only if the random elements $\big\|\sum_{i=1}^n e_i\, f(X_i)\big\|_{\mathcal{F}}$ are measurable for every n-tuple $(e_1,\ldots,e_n)\in\{-1,1\}^n$. For the intended application of Fubini's theorem it sufficies that this is the case for the completion of the probability space (as the subsets with the zero outer measure are irrelevant for the expectation).

Definition 10. Let $X_1,\ldots,X_n$ be independent random variables with values in $\mathcal{X}$ defined on the product probability space $(\Omega,\mathcal{A},\mathsf{P})$. A class of measurable function $\mathcal{F}$ (from $\mathcal{X}$ to $\mathbb{R}$) is called a ***P*-measurable class** if for each $n\in\mathbb{N}$ and each $(e_1,\ldots,e_n)\in\{-1,1\}^n$ the random element $\big\|\sum_{i=1}^n e_i\, f(X_i)\big\|_{\mathcal{F}}$ is measurable on the completion of the probability space $(\Omega,\mathcal{A},\mathsf{P})$.

In fact the restriction to P-measurable classes of functions is not necessary, but it is suggested by the method of the proof. The thing is that some kind of measurability is necessary, otherwise the following general theorems are not valid from the mathematical point of view.

On the other hand it is pretty tricky to create counter-examples for which either the Glivenko-Cantelli or Donsker theorem does not hold because of measurability issues. That is why in applications researchers usually do not bother with measurability.

Further, in many applications the measurability is justified as follows. Suppose there exists a countable subset $\mathcal{G}$ of $\mathcal{F}$ such that for each $f\in\mathcal{F}$ there existence a sequence of functions $\{g_m\}$ in $\mathcal{G}$ so that $g_m(x)\to f(x)$ as $m\to\infty$ for each $x\in\mathcal{X}$. Then

$$\Big\|\sum_{i=1}^{n} e_i\, f(X_i)\Big\|_{\mathcal{F}} = \Big\|\sum_{i=1}^{n} e_i\, g(X_i)\Big\|_{\mathcal{G}}.$$

Such classes $\mathcal{F}$ are called *pointwise measurable classes.*

Example 5. Suppose that $\mathcal{X} = \mathbb{R}^d$. For $y \in \mathbb{R}^d$ and $r > 0$ denote the open ball with the center y and diameter r as $B_r(y) = \{x \in \mathbb{R}^d \,:\, \|x - y\|_E < r\}$, where $\|\cdot\|_E$ stands for the Euclidean norm. Show that the set of indicator functions

$$\mathcal{F} = \big\{x \to \mathbb{I}\{x \in B_r(y)\} \,:\, y \in \mathbb{R}^d, r > 0\big\}$$

is a pointwise measurable class (and it is also a VC class of functions).

Glivenko-Cantelli theorem and its applications

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with values in $\mathcal{X}$ and $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$.

Theorem 12. *Let the class of functions $\mathcal{F}$ be P-measurable with the envelope F, such that $P^*(F) := \mathsf{E}^*[F(X_i)] < \infty$. Finally assume that for each $\epsilon > 0$ and each $M < \infty$*

$$\log N\big(\epsilon, \mathcal{F}_M, L_1(P_n)\big) = o_{P^*}(n). \tag{6.7}$$

Then $\mathcal{F}$ is P-Glivenko-Cantelli, i.e.

$$\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f)\big| \xrightarrow[n \to \infty]{\text{alm. surely}^*} 0.$$

Note that in comparison with Theorem 1 the covering numbers are allowed even to grow with the increasing sample size n, but at a lower than exponential rate. Further note that while for the bracketing numbers $L_1(P)$ norm is used, here we use (the empirical) $L_1(P_n)$ norm.

It can be proved that the condition (6.7) is not only sufficient but also necessary, (see e.g. Theorem 6.2 Wellner, 2005).

Proof of Theorem 12. We will show only that $\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f))\big| \xrightarrow[n \to \infty]{P^*} 0$. The almost sure convergence follows from the fact that the minimal measurable cover function of $\sup_{f \in \mathcal{F}} \big|P_n(f) - P(f))\big|$ (see Remark 2(f)) is a reverse martingale with respect to a suitable filtration (see Lemma 2.4.5 van der Vaart and Wellner, 1996).

Let $\eta > 0$ be given and for a given M denote

$$\mathcal{F}_M = \big\{x \to f(x)\, \mathbb{I}\{F(x) \leq M\}\big\}.$$

Then one can bound

$$\begin{aligned}
\|P_n - P\|_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \\
&\leq \sup_{f \in \mathcal{F}_M} |P_n(f) - P(f)| + \sup_{f \in \mathcal{F}} P_n\big(|f\, \mathbb{I}\{F > M\}|\big) + \sup_{f \in \mathcal{F}} P^*\big(|f\, \mathbb{I}\{F > M\}|\big) \\
&\leq \sup_{f \in \mathcal{F}_M} |P_n(f) - P(f)| + P_n\big(F\, \mathbb{I}\{F > M\}\big) + P^*\big(F\, \mathbb{I}\{F > M\}\big).
\end{aligned}$$

Now one can take M sufficiently large so that $P^*\big(F\, \mathbb{I}\{F > M\}\big) < \eta/2$. Further with the help of the law of large numbers one can show that $P_n\big(F\, \mathbb{I}\{F > M\}\big) < \eta/2$ with probability going to one as $n \to \infty$. Thus one needs to concentrate only on $\sup_{f \in \mathcal{F}_M} |P_n(f) - P(f)|$.

To do that use Lemma 7 to bound (for $n \geq \frac{2M^2}{\eta^2}$)

$$\mathsf{P}^* \big(\|P_n - P\|_{\mathcal{F}_M} > \eta \big) \leq 4\, \mathsf{P}^* \big(\|P_n^0\|_{\mathcal{F}_M} > \eta/4 \big). \tag{6.8}$$

Now one has to be carefull. Although by the assumption of the theorem the class $\mathcal{F}$ is P-measurable we do not know whether this holds also true for the class $\mathcal{F}_M$. That is why we need to proceed (not intuitively) as follows.

$$\begin{aligned} \mathsf{P}^* \big(\|P_n^0\|_{\mathcal{F}_M} > \eta/4 \big) &\leq \mathsf{P}^* \big(\|P_n^0\|_{\mathcal{F}} > \eta/4 \big) \overset{Fubini}{=} \mathsf{E}_X\, \mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}} > \eta/4 \big) \\ &\leq \mathsf{E}_X^*\, \mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}_M} > \eta/8 \big) + P^* \big(P_n(F\{F > M\}) > \eta/8 \big). \end{aligned} \tag{6.9}$$

Now the second term on the right hand of (6.9) can be handled analogously as above. Thus it is sufficient to concentrate only on $\mathsf{E}_X^*\, \mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}_M} > \eta/8 \big)$.

To proceed note that given $X_1, \ldots, X_n$ one can find $g_1, \ldots, g_{N_\eta}$, where $N_\eta = N\big(\eta/8, \mathcal{F}_M, L_1(P_n)\big)$, such that for each $f \in \mathcal{F}_M$ there exists g_j so that $P_n(|f - g_j|) \leq \eta/8$. Without loss of generality one can assume that also the absolute values of the functions $g_1, \ldots, g_N$ are bounded by M. Now using the Hoeffding's inequality as in (6.1) one can bound

$$\begin{aligned} \mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}_M} > \tfrac{\eta}{4} \big) &\leq \mathsf{P}_\epsilon \bigg(\max_{j \in \{1,\ldots,N_\eta\}} |P_n^0(g_j)| + \tfrac{\eta}{8} > \tfrac{\eta}{4} \bigg) \\ &\leq N_\eta \max_{j \in \{1,\ldots,N_\eta\}} \mathsf{P}_\epsilon \big(|P_n^0(g_j)| > \tfrac{\eta}{8} \big) = N_\eta \max_{j \in \{1,\ldots,N_\eta\}} \mathsf{P}_\varepsilon \bigg(\Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_j(X_i) \Big| > \tfrac{\eta}{8} \bigg) \\ &\overset{\text{Hoeff.}}{\leq} N_\eta \max_{j \in \{1,\ldots,N_\eta\}} 2 \exp \bigg(- \frac{n\, (\frac{\eta}{8})^2}{2\, P_n(g_j^2)} \bigg) \leq 2 N_\eta \exp \big(- \tfrac{n\eta^2}{128 M^2} \big) \xrightarrow[n\to\infty]{P^*} 0, \end{aligned} \tag{6.10}$$

as by the assumptions of the theorem $N_\eta = \exp\big(o_{P^*}(n)\big)$. Thus we have proved that for each $X_1, X_2, \ldots$ it holds that the conditionally probability $\mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}_M} > \frac{\eta}{4} \big) \xrightarrow[n\to\infty]{P^*} 0$. Further we know that this probability is bounded by one. Thus by the dominated convergence theorem one gets

$$\mathsf{E}_X^*\, \mathsf{P}_\varepsilon \big(\|P_n^0\|_{\mathcal{F}_M} > \eta/8 \big) \xrightarrow[n\to\infty]{} 0.$$

This combined with (6.8) and (6.9) yields

$$\mathsf{P}^* \big(\|P_n - P\|_{\mathcal{F}_M} > \eta \big) \xrightarrow[n\to\infty]{} 0,$$

which finishes the proof. □

Suppose that $\mathcal{F}$ is a P-measurable VC-class of functions with a measurable envelope function F such that $P(F) < \infty$. Then with the help of Theorem 11 there exist universal constants K and W such that for each $n \in \mathbb{N}$

$$N\big(\epsilon, \mathcal{F}, L_1(P_n)\big) \leq K \bigg(\frac{\|F\|_{P_n,1}}{\epsilon} \bigg)^W,$$

which with the help of the law of large numbers verifies the assumption (6.7) of Theorem 12. This implies than any VC class with a $L_1(P)$ integrable envelope is P-Glivenko-Cantelli.

Sometimes it is also possible to check assumption (6.7) even for classes that are not VC-classes. Consider for instance the class of closed convex subsets of the unit square $[0,1]^2$. Then from

Example 4(d) we know that this class of sets is not a VC-class and thus also the corresponding class of indicator functions is not a VC-class. In spite of that one can show that if the distribution of X_i is uniform on $[0,1]^2$ then the assumption (6.7) is satisfied (see Example 22 in Chapter II.4 of Pollard, 1984).

Remark 9. The question of interest is if this approach can be used also for not i.i.d. random variables, provided that one has an appropriate law of large numbers. The answer is hidden in the symmetrization Lemma 7. Let $(X_1',\dots,X_n')$ be an independent copy of $(X_1,\dots,X_n)$. Further let $\varepsilon_1,\dots,\varepsilon_n$ be Rademacher variables independent of $(X_1,\dots,X_n,X_1',\dots,X_n')$. Then it is needed that the distribution of $\sum_{i=1}^n \big(f(X_i)-f(X_i')\big)$ is the same as the distribution of $\sum_{i=1}^n \varepsilon_i\big(f(X_i)-f(X_i')\big)$. This is true for independent (not necessarily identically distributed) random variables $X_1,\dots,X_n$. But this fails to be generally true when $X_1,\dots,X_n$ are not independent.

Rates of convergence for bounded VC-classes of functions

Let $\mathcal{F}$ be a P-measurable VC-class of functions with the envelope function bounded by a finite constant M. Then the covering number satisfies

$$N\big(\epsilon,\mathcal{F},L_1(P_n)\big)\le K\left(\frac{M}{\epsilon}\right)^W,$$

from which one can conclude that $\mathcal{F}$ is P-Glivenko-Cantelli. Note that this bound on the covering number is much more than it is required by Theorem 12.

In fact for bounded VC-classes one can show that the rate of convergence in the uniform law of large numbers is at least $r_n=\sqrt{\frac{n}{q_n\log n}}$, where $q_n\to\infty$ (consider for instance $q_n=\log^\delta n$ for some $\delta>0$).

To prove that take $\eta_n=q_n^{-1/3}$ (and note that $\eta_n \xrightarrow[n\to\infty]{} 0$). Then for each fixed $f\in\mathcal{F}$ for all sufficiently large n

$$\mathsf{P}\left(r_n|P_n(f)-P(f)|>\tfrac{\eta_n}{2}\right)\le\frac{r_n^2\ \mathsf{var}\big(f(X_i')\big)}{n\,(\eta_n/2)^2}\le\frac{4\,M^2}{q_n^{1/3}\,\log n}\le\frac{1}{2}.$$

Thus similarly as in Lemma 7 one can show that for these sufficiently large sample sizes

$$\mathsf{P}^*\left(r_n\,\|P_n-P\|_{\mathcal{F}}>\eta_n\right)\le 4\ \mathsf{P}^*\left(r_n\,\|P_n^0\|_{\mathcal{F}}>\tfrac{\eta_n}{4}\right). \tag{6.11}$$

Further one can proceed completely analogously as in (6.10) to show that

$$\begin{aligned}\mathsf{P}_\varepsilon\left(r_n\,\|P_n^0\|_{\mathcal{F}}>\tfrac{\eta_n}{4}\right)&\le 2\,N\big(\tfrac{\eta_n}{r_n},\mathcal{F},L_1(P_n)\big)\exp\big(-\tfrac{n\eta_n^2}{r_n^2 128M^2}\big)\\&\le 2\,K\big(\tfrac{M\,\sqrt{n}}{q_n^{1/6}\sqrt{\log n}}\big)^W\exp\big(-\tfrac{q_n^{1/3}\log n}{128M^2}\big).\end{aligned} \tag{6.12}$$

Now combining (6.11) and (6.12) one gets that for all sufficiently large n one has

$$\mathsf{P}^*\left(r_n\,\|P_n-P\|_{\mathcal{F}}>\eta_n\right)\le 8\,K\big(\tfrac{M\,\sqrt{n}}{q_n^{1/6}\sqrt{\log n}}\big)^W\exp\big(-\tfrac{q_n^{1/3}\log n}{128M^2}\big), \tag{6.13}$$

which goes to zero as $n\to\infty$. Note that so far we have proved that $r_n\,\|P_n-P\|_{\mathcal{F}}\xrightarrow[n\to\infty]{P^*}0$ (see Definition 1(i)). Now we strengthen this convergence to outer almost sure convergence. By

Definition 1(ii) we need to find a dominating sequence $\{\Delta_n\}$ of measurable random variables that converges almost surely to zero.

Note that from (6.13) and the definition of the outer probability P^* given in (2.1) there exists a sequence of measurable events $A_n \in \mathcal{A}$ such that

$$\big[r_n \, \|P_n - P\|_{\mathcal{F}} > \eta_n\big] \subset A_n$$

and at the same time $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$. Thus by the Cantelli theorem for almost all $\omega \in \Omega$ one has $\lim_{n\to\infty} \mathbb{I}_{A_n}(\omega) = 0$. Now consider the (measurable) random variable defined for $\omega \in \Omega$ as

$$\Delta_n(\omega) = \eta_n \mathbb{I}_{A_n^c}(\omega) + 2\, r_n \, M \, \mathbb{I}_{A_n}(\omega).$$

Then $r_n \, \|P_n - P\|_{\mathcal{F}} \leq \Delta_n$ and at the same time $\Delta_n \xrightarrow[n\to\infty]{\text{alm. surely}} 0$, which finally implies that

$$r_n \, \|P_n - P\|_{\mathcal{F}} \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0.$$

Exercise 7. Show that

$$\|P_n - P\|_{\mathcal{F}} = O_{P^*}\Big(\sqrt{\tfrac{\log n}{n}}\Big).$$

Application to halfspace depth

Suppose that $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent and identically distributed random vectors with values in $\mathbb{R}^d$. Data depth is an attempt to characterize how much a given observation (or more generally a given point) is central with respect to the sample (or more generally with respect to a given probability measure on $\mathbb{R}^d$). Roughly speaking, high depth values indicate centrality of the given observations and low depth values indicate potential outlyingness.

The most widely used and famous depth is *Tukey's halfspace depth.* Let $\mathcal{H}$ be the class of all closed halfspaces in $\mathbb{R}^d$, i.e.

$$\mathcal{H} = \big\{\{\mathbf{x} \in \mathbb{R}^d : \mathbf{b}^\mathsf{T}\mathbf{x} \geq a\},\, \mathbf{b} \in \mathbb{R}^d, a \in \mathbb{R}\big\}.$$

Let P be a probability measure on $\mathbb{R}^d$ given by the distribution of $\boldsymbol{X}_i$ and $\mathbf{x}$ be a given point in $\mathbb{R}^d$. Then the halfspace depth is defined as the minimal probability of the halfspace that contains the point $\mathbf{x}$, i.e.

$$hD(\mathbf{x}) = \inf_{H \in \mathcal{H}: \mathbf{x} \in H} \mathsf{P}(\boldsymbol{X}_i \in H) = \inf_{H \in \mathcal{H}: \mathbf{x} \in H} P(\mathbb{I}_H).$$

The sample version of the halfspace depth is then given by

$$hD_n(x) = \inf_{H \in \mathcal{H}: \mathbf{x} \in H} \left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \in H\}\right\} = \inf_{H \in \mathcal{H}: \mathbf{x} \in H} P_n(\mathbb{I}_H).$$

Note that $\mathcal{H}$ does not shatter a set of $d+2$ points and thus it is a VC-class of sets. So by Theorem 10 the covering numbers of the corresponding set of indicator functions $\mathcal{F}_{\mathcal{H}}$ satisfy the assumption of Theorem 12. Further similarly as in Example 5 one can argue that $\mathcal{F}_{\mathcal{H}}$ is a pointwise measurable set and so also a P-measurable class. Thus one gets

$$\begin{aligned}\sup_{\mathbf{x} \in \mathbb{R}^d} \big|hD_n(\mathbf{x}) - hD(\mathbf{x})\big| &= \sup_{\mathbf{x} \in \mathbb{R}^d} \left|\inf_{H \in \mathcal{H}: \mathbf{x} \in H} P_n(\mathbb{I}_H) - \inf_{H \in \mathcal{H}: \mathbf{x} \in H} P(\mathbb{I}_H)\right| \\ &\leq \sup_{\mathbf{x} \in \mathbb{R}^d} \sup_{H \in \mathcal{H}: \mathbf{x} \in H} \big|P_n(\mathbb{I}_H) - P(\mathbb{I}_H)\big| \\ &= \sup_{H \in \mathcal{H}} \big|P_n(\mathbb{I}_H) - P(\mathbb{I}_H)\big| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0, \qquad (6.14)\end{aligned}$$

which implies that the sample halfspace depth hD_n is uniformly (in $\mathbf{x} \in \mathbb{R}^d$) consistent estimate of the population halfspace depth hD. Note also that similarly as in the Glivenko-Cantelli theorem for the empirical distribution function in Exercise 4, there is no assumption on the distribution of $\boldsymbol{X}_i$ in $\mathbb{R}^d$. Finally note that by the results of the previous section the rate of convergence is in fact $\sqrt{\frac{n}{q_n \log n}}$ for $q_n \xrightarrow[n\to\infty]{} \infty$.

Exercise 8. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be a random sample from the uniform distribution on the square $[-2, 2]^2$. Consider the following process

$$Z_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \in B(\mathbf{x})\}, \text{ where } \mathbf{x} \in \mathbb{R}^2$$

and $B(\mathbf{x})$ is a unit ball, i.e. $B(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{x} - \mathbf{y}\|_E \leq 1\}$ with $\|\cdot\|_E$ denoting the Euclidean norm.

Show that

$$\sup_{\mathbf{x}\in[-2,2]^2} \left| Z_n(\mathbf{x}) - \frac{\lambda\big(B(\mathbf{x}) \cap [-2,2]^2\big)}{16} \right| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0,$$

where λ is a Lebesgue measure (on $\mathbb{R}^2$).

Now, consider $\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$. Show that

$$Z_n(\overline{\boldsymbol{X}}_n) \xrightarrow[n\to\infty]{\text{alm. surely}^*} \frac{\pi}{16}.$$

Exercise 9. Let $\binom{Y_1}{X_1}, \dots, \binom{Y_n}{X_n}$ be a random sample such that X_i has a uniform distribution on $[-1, 1]$ and

$$Y_i = \beta\, X_i + \varepsilon_i,$$

where $\beta \in \mathbb{R}$, ε_i is independent of X_i and $\mathsf{E}\ \varepsilon_i = 0$. Consider the process

$$Z_n(a, b) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i - b\, X_i \leq a\}$$

Show that

$$\sup_{a,b} \left| Z_n(a, b) - \tfrac{1}{2} \int_{-1}^{1} F_\varepsilon\big(a + (b - \beta)x\big)\, dx \right| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0,$$

where F_ε is the distribution function of ε_i.

Further consider $\widehat{\beta}_n = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$ and the empirical distribution function of the residuals $\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_n X_i$ given by

$$\widehat{F}_\varepsilon(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\widehat{\varepsilon}_i \leq a\}$$

Show that $\sup_{a\in\mathbb{R}} \big|\widehat{F}_\varepsilon(a) - F_\varepsilon(a)\big| \xrightarrow[n\to\infty]{\text{alm. surely}^*} 0.$

Hint. Note that $\widehat{F}_\varepsilon(a) = Z_n(a, \widehat{\beta}_n)$ and $F_\varepsilon(a) = \mathsf{E}\ Z_n(a, \beta)$.

7 Donsker-Theorem with the help of covering numbers

The goal of this chapter to find an analogy of Theorem 9, where the sufficient assumptions on the size of the class $\mathcal{F}$ is given in terms of covering numbers.

Preliminary results

Similarly as in Chapter 6 denote P_n^0 the symmetrized empirical measure. The following lemma is an analogy of Lemma 7 for the expectations.

Lemma 8. ***(Symmetrization II)***
Let $\mathcal{F}$ be a class of measurable functions. Then it holds that

$$\mathsf{E}^*\left[\|P_n - P\|_{\mathcal{F}}\right] \leq 2\,\mathsf{E}^*\left[\,\|P_n^0\|_{\mathcal{F}}\,\right].$$

Proof. Let $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$ and P_n' be the corresponding empirical measure of $X_1', \ldots, X_n'$. Then (conditionally on $X_1, \ldots, X_n$):

$$\begin{aligned}\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathsf{E}_{X'}[f(X_i')])\Big| &\leq \sup_{f\in\mathcal{F}}\mathsf{E}_{X'}\left[\Big|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(X_i'))\Big|\right]\\ &\leq \mathsf{E}^*_{X'}\left[\|P_n - P_n'\|_{\mathcal{F}}\right].\end{aligned}$$

Now similarly as in the proof of Lemma 7

$$\mathsf{E}^*\left[\|P_n - P\|_{\mathcal{F}}\right] \leq \mathsf{E}^*_X\,\mathsf{E}^*_{X'}\left[\|P_n - P_n'\|_{\mathcal{F}}\right] \leq \mathsf{E}^*\left[\|P_n - P_n'\|_{\mathcal{F}}\right] \leq 2\,\mathsf{E}^*\left[\|P_n^0\|_{\mathcal{F}}\right].$$

□

Let $\{X(t), t \in T\}$ be a stochastic process. We say that this process is *sub-Gaussian* with respect to a semi-metric d if for each $\epsilon > 0$ and $s, t \in T$

$$\mathsf{P}\left(|X(s) - X(t)| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2d^2(s,t)}\right).$$

Let us now consider the symmetrized empirical process

$$\mathbb{G}_n^0(f) = \sqrt{n}\,P_n^0(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\,f(X_i),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are Rademacher variables independent of $X_1, \ldots, X_n$. Note that similarly as in (6.1) with the help of Hoeffding's inequality (Lemma 6) one gets

$$\mathsf{P}_\varepsilon\left(|\mathbb{G}_n^0(f) - \mathbb{G}_n^0(g)| > \epsilon\right) \leq 2\exp\left(-\tfrac{\epsilon^2}{2P_n((f-g)^2)}\right). \tag{7.1}$$

Thus (for fixed $X_1, \ldots, X_n$) the process $\mathbb{G}_n^0$ is sub-Gaussian in with respect to the empirical $L_2(P_n)$-norm. The nice thing about sub-Gaussian processes is that one can bound the expectations of their supremums with the help of covering numbers as stated in the following lemma.

Lemma 9. ***(Maximal inequality)*** *Let $\{X(t), t \in T\}$ be a stochastic process and the index set T is countable. Further suppose that for each $\epsilon > 0$ and $s, t \in T$*

$$\mathsf{P}\left(|X(s) - X(t)| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2d^2(s,t)}\right),$$

where d is a semi-metric on T. Then there exists a finite constant K such that for each $\delta > 0$

$$\mathsf{E}\left[\sup_{d(s,t)\leq\delta} |X(s) - X(t)|\right] \leq K \int_0^{\frac{\delta}{2}} \sqrt{\log N(\epsilon, T, d)}\; d\epsilon.$$

In particular for each $t_0 \in T$

$$\mathsf{E}\left[\sup_{t\in T} |X(t)|\right] \leq \mathsf{E}\left[|X(t_0)|\right] + K \int_0^{\infty} \sqrt{\log N(\epsilon, T, d)}\; d\epsilon.$$

Remark 10. The proof of this lemma can be found as the proof of Corollary 2.2.8 in van der Vaart and Wellner (1996). In this book the index set T is not assumed to be countable but the lemma is formulated for a separable sub-Gaussian process. This is equivalent as the separability means that the $\sup_{d(s,t)\leq\delta} |X(s) - X(t)|$ and $\sup_{t\in T} |X(t)|$ remain almost surely the same if the index set is replaced by a suitable countable subset.

Donsker theorem and its applications

Similarly as we introduced the bracketing integral $\mathcal{J}_{[]}(\delta, \mathcal{F}, L_2(P))$ in (4.3) in Chapter 4 we now define the *covering integral*

$$J(\delta, \mathcal{F}, L_2) := \int_0^{\delta} \sqrt{\log \sup_Q N\big(\epsilon\, \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\big)}\; d\epsilon,$$

where the supremum is taken with respect to all finitely discrete probability measures Q on $\mathcal{X}$ for which $\|F\|_{Q,2} = \int_{\mathcal{X}} F^2(x)\, dQ(x) > 0$.

Theorem 13. *Let $\mathcal{F}$ be a class of measurable functions with the envelope F, such that $P^*(F^2) < \infty$ and $J(1, \mathcal{F}, L_2) < \infty$. Further assume that the classes of functions $\mathcal{H}_\delta = \big\{f_1 - f_2 : f_1, f_2 \in \mathcal{F}, \|f_1 - f_2\|_{P,2} < \delta\big\}$ and $\mathcal{G} = \big\{(f_1 - f_2)^2 : f_1, f_2 \in \mathcal{F}\big\}$ are P-measurable for every δ. Then $\mathcal{F}$ is P-Donsker.*

Proof. By Theorem 7 it is sufficient to find a semimetric ρ such that $(\mathcal{F}, \rho)$ is *totally bounded* and the empirical process $\mathbb{G}_n = \sqrt{n}\,(P_n - P)$ is *asymptotically uniformly ρ-equicontinuous in probability* (see Definition 6). As suggested by Theorem 8 take ρ as $L_2(P)$-norm.

Let $\epsilon > 0$ be given. Then we need to show that for each sequence of positive constants $\{\delta_n\}$ going to zero

$$\limsup_{n\to\infty} \mathsf{P}^*\left(\sup_{h\in\mathcal{H}_{\delta_n}} |\mathbb{G}_n(h)| > \epsilon\right) = 0. \tag{7.2}$$

By Markov's inequality and the symmetrization Lemma 8 for each $\epsilon > 0$

$$\mathsf{P}^*\left(\sup_{h\in\mathcal{H}_{\delta_n}} |\mathbb{G}_n(h)| > \epsilon\right) \leq \frac{1}{\epsilon}\,\mathsf{E}^*\left[\sup_{h\in\mathcal{H}_{\delta_n}} |\mathbb{G}_n(h)|\right] \leq \frac{2}{\epsilon}\,\mathsf{E}^*\left[\sup_{h\in\mathcal{H}_{\delta_n}} \sqrt{n}\,\big|P_n^0(h)\big|\right] \tag{7.3}$$

Now by the assumption of the theorem the supremum on the right-hand side of the above inequality is measurable. Thus by Fubini's theorem the outer expectation can be calculated as

$\mathsf{E}_X\,\mathsf{E}_\varepsilon$, where E_ε is the expectation taken with respect to $\varepsilon_1,\ldots,\varepsilon_n$ (while $X_1,\ldots,X_n$ being fixed). In the same way as in (7.1) one can use Hoeffding's inequality to deduce that the process $\{\sqrt{n}\,P_n^0(h), h\in\mathcal{H}_{\delta_n}\}$ is sub-Gaussian for $L_2(P_n)$-seminorm. Furthermore (**exercise**)

$$N(\eta,\mathcal{H}_{\delta_n},L_2(P_n)) \leq N(\eta,\mathcal{H}_\infty,L_2(P_n)) \leq N^2(\tfrac{\eta}{2},\mathcal{F},L_2(P_n)), \tag{7.4}$$

which implies that the process $\{\sqrt{n}\,P_n^0(h), h\in\mathcal{H}_{\delta_n}\}$ is separable. Now using Remark 10 and the second part of Lemma 9 there exists a finite constant K such that

$$\mathsf{E}_\varepsilon\Big[\sup_{h\in\mathcal{H}_{\delta_n}}\sqrt{n}\,\big|P_n^0(h)\big|\Big] \leq K\int_0^\infty \sqrt{\log N\big(\eta,\mathcal{H}_{\delta_n},L_2(P_n)\big)}\,d\eta. \tag{7.5}$$

Now introduce $\nu_n := \sup_{h\in\mathcal{H}_{\delta_n}}\|h\|_{P_n,2}$. Note that for $\eta\geq\nu_n$ it holds $N\big(\eta,\mathcal{H}_{\delta_n},L_2(P_n)\big)=1$. Combining this with (7.4) one can bound

$$\begin{aligned}\int_0^\infty \sqrt{\log N(\eta,\mathcal{H}_{\delta_n},L_2(P_n))}\,d\eta &= \int_0^{\nu_n}\sqrt{\log N(\eta,\mathcal{H}_{\delta_n},L_2(P_n))}\,d\eta\\ &\leq \int_0^{\nu_n}\sqrt{2\,\log N(\tfrac{\eta}{2},\mathcal{F},L_2(P_n))}\,d\eta\\ &\leq 2\,\|F\|_{P_n,2}\int_0^{\nu_n/(2\|F\|_{P_n,2})}\sqrt{2\,\log N(\eta\|F\|_{P_n,2},\mathcal{F},L_2(P_n))}\,d\eta\\ &\leq 4\,\|F\|_{P_n,2}\int_0^{\nu_n/(\|F\|_{P_n,2})}\sqrt{\sup_Q\log N(\eta\|F\|_{Q,2},\mathcal{F},L_2(Q))}\,d\eta,\end{aligned} \tag{7.6}$$

where the integral exists by the assumptions of the theorem. Further by the law of large numbers $\|F\|_{P_n,2}=(\frac{1}{n}\sum_{i=1}^n F^2(X_i))^{1/2}$ is bounded in probability. Thus with the help of (7.3), (7.5) and (7.6) the convergence in (7.2) is verified provided that

$$\nu_n/(\|F\|_{P_n,2}) \xrightarrow[n\to\infty]{P^*} 0. \tag{7.7}$$

Note that without loss of generality one can assume that there exists $f\in\mathcal{F}$ such that $\mathsf{E}\,|f(X_i)|^2>0$. From this conclude that

$$\liminf_{n\to\infty}\|F\|_{P_n,2}>0 \quad \text{almost surely.}$$

Thus it sufficient to show that $\nu_n \xrightarrow[n\to\infty]{P^*} 0$. As $\lim_{n\to\infty}\sup_{h\in\mathcal{H}_{\delta_n}}\|h\|_{P,2}\to 0$, it remains to prove

$$\sup_{h\in\mathcal{H}_\infty}\big|P_n(h^2)-P(h^2)\big| = \sup_{g\in\mathcal{G}}\big|P_n(g)-P(g)\big| \xrightarrow[n\to\infty]{P^*} 0. \tag{7.8}$$

To prove that we use Theorem 12 for the class $\mathcal{G}$. First note that $\mathcal{G}$ has the integrable envelope $(2F)^2$ and is P-measurable by the assumption. Further for each $h_1,h_2\in\mathcal{H}_\infty$ one can with the help of Cauchy-Schwarz inequality bound

$$P_n\big(|h_1^2-h_2^2|\big)\leq P_n\big(|h_1-h_2|\,4F\big)\leq \|h_1-h_2\|_{P_n,2}\,\|4F\|_{P_n,2}.$$

With the help of this inequality and the equality $\|4F\|_{P_n,2}\,\|F\|_{P_n,2}=\|(2F)^2\|_{P_n,1}$ one can deduce that

$$N\big(\epsilon\|(2F)^2\|_{P_n,1},\mathcal{G},L_1(P_n)\big)\leq N\big(\epsilon\|F\|_{P_n,2},\mathcal{H}_\infty,L_2(P_n)\big).$$

Now the term on the right-hand side of the above inequality is bounded by the assumptions of the theorem and thus the assumptions of Theorem 12 are satisfied. This finishes the proof of (7.7) and thus also of (7.2).

It remains to show that $(\mathcal{F}, \rho)$ is totally bounded. Let $\epsilon > 0$ be given. For $\omega \in \Omega$ denote $\{P_n^\omega\}$ the sequence of the corresponding empirical measures. As we have shown that $\mathcal{G}$ is P-Glivenko-Canteli one has that for almost all $\omega \in \Omega$

$$\sup_{g \in \mathcal{G}} |P_n^\omega(g) - P(g)| = \sup_{h \in \mathcal{H}_\infty} |P_n^\omega(h^2) - P(h^2)| \xrightarrow[n \to \infty]{} 0.$$

Consider $\omega \in \Omega$ such that the above convergence holds. Further take n so large so that $\sup_{h \in \mathcal{H}_\infty} |P_n^\omega(h^2) - P(h^2)| < \frac{\epsilon^2}{2}$ and fix the measure P_n^ω. Then for each $f_1, f_2 \in \mathcal{F}$

$$\begin{aligned} \rho^2(f_1, f_2) = P\big((f_1 - f_2)^2\big) &\le \big|P\big((f_1 - f_2)^2\big) - P_n^\omega\big((f_1 - f_2)^2\big)\big| + P_n^\omega\big((f_1 - f_2)^2\big) \\ &\le \tfrac{\epsilon^2}{2} + P_n^\omega\big((f_1 - f_2)^2\big), \end{aligned}$$

which implies that

$$N(\epsilon, \mathcal{F}, \rho) \le N\big(\tfrac{\epsilon}{\sqrt{2}}, \mathcal{F}, L_2(P_n^\omega)\big).$$

Now the right-hand side of the last inequality is finite by the assumption of the theorem which finally implies that $(\mathcal{F}, \rho)$ is totally bounded.

□

Let $\mathcal{F}$ be a VC-class of functions with a measurable envelope function F. Then by Theorem 11 there exist finite constants K and W such that the covering integral is bounded as

$$J(1, \mathcal{F}, L_2) \le \int_0^1 \sqrt{\log\big[K\big(\tfrac{1}{\epsilon}\big)^W\big]}\, d\epsilon = \int_0^1 \sqrt{\log K - W \log \epsilon}\, d\epsilon < \infty.$$

If moreover the envelope function F is squared integrable (i.e. $P(F^2) < \infty$), then $\mathcal{F}$ is P-Donsker.

Some further classes of functions that have a finite covering integral can be found for instance in Chapter 2.6 of van der Vaart and Wellner (1996) and Chapter 9.1.2 of Kosorok (2007).

Example 6. Recall the halfspace depth introduced on page 6. Then the class of the indicator functions $\mathcal{F}_{\mathcal{H}}$ of halfspaces is P-Donsker class. Further with the help of (6.14) one can deduce that

$$\sqrt{n} \sup_{\mathbf{x} \in \mathbb{R}^d} \big|hD_n(\mathbf{x}) - hD(\mathbf{x})\big| \le \sup_{H \in \mathcal{H}} \big|\sqrt{n}\,\big(P_n(\mathbb{I}_H) - P(\mathbb{I}_H)\big)\big| \rightsquigarrow \sup_{H \in \mathcal{H}} |\mathbb{G}(\mathbb{I}_H)|,$$

where $\mathbb{G}$ is a centered Gaussian process with the covariance function

$$\begin{aligned} \mathsf{cov}\,\big(\mathbb{G}(\mathbb{I}_{H_1}), \mathbb{G}(\mathbb{I}_{H_2})\big) &= P\big(\mathbb{I}_{H_1}, \mathbb{I}_{H_2}\big) - P\big(\mathbb{I}_{H_1}\big)\, P\big(\mathbb{I}_{H_2}\big) \\ &= \mathsf{P}\,\big(\boldsymbol{X}_i \in H_1 \cap H_2\big) - \mathsf{P}\,\big(\boldsymbol{X}_i \in H_1\big)\ \mathsf{P}\,\big(\boldsymbol{X}_i \in H_2\big). \end{aligned}$$

Thus we have shown that the sample halfspace depths $hD_n(\mathbf{x})$ are uniformly (in $\mathbf{x} \in \mathbb{R}^d$) weakly $\sqrt{n}$-consistent estimators of the theoretical halfspace depths $hD(\mathbf{x})$, i.e.

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \big|hD_n(\mathbf{x}) - hD(\mathbf{x})\big| = O_P\big(\tfrac{1}{\sqrt{n}}\big).$$

This strengthen the result of Exercise 7, which implies ‘only’

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \big|hD_n(\mathbf{x}) - hD(\mathbf{x})\big| = O_P\Big(\sqrt{\tfrac{\log n}{n}}\Big).$$

It is worth noting that the rate $O_P\big(\frac{1}{\sqrt{n}}\big)$ cannot be further improved as for a given halfspace $H \in \mathcal{H}$ with $p_H = \mathsf{P}(\boldsymbol{X}_i \in H) > 0$ one has

$$\sqrt{n}\,\big(P_n(\mathbb{I}_H) - p_H\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, p_H(1-p_H)\big).$$

Exercise 10. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be a random sample from the uniform distribution on the square $[-2,2]^2$. Consider the following process

$$Z_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \in B(\mathbf{x})\}, \text{ where } \mathbf{x} \in \mathbb{R}^2$$

and $B(\mathbf{x})$ is a unit ball, i.e. $B(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{x}-\mathbf{y}\|_E \leq 1\}$ with $\|\cdot\|_E$ denoting the Euclidean norm.

Show that the appropriately standardized process $\{Z_n(\mathbf{x}), \mathbf{x} \in [-2,2]\}$ converges in distribution. Describe the limiting process.

* Now, consider $\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$. Derive the asymptotic distribution of $\sqrt{n}\,\big(Z_n(\overline{\boldsymbol{X}}_n) - Z(\mathbf{0})\big)$, where $Z(\mathbf{x}) = \mathsf{P}\big(\boldsymbol{X}_i \in B(\mathbf{x})\big)$.

Hint. See the hint for Exercise 5.

Bibliography

Billingsley, P. (1999). *Convergence of probability measures.* Wiley, New York.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference.* Springer, New York.

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer, New York. Second Edition.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press, New York.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Wellner, J. A. (2005). Empirical processes: Theory and applications. Special Topics Course, Spring 2005, Delft Technical University.

Classic theory of empirical processes:

G.R. Shorack, J.A. Wellner, Empirical processes with applications to statistics (1996), Wiley.