Empirical Processes with Applications in Statistics

3. Glivenko–Cantelli and Donsker Theorems

Johan Segers

UCLouvain (Belgium)

Conférence Universitaire de Suisse Occidentale

Programme Doctoral en Statistique et Probabilités Appliquées

Les Diablerets, February 4–5, 2020

Set-up

- i.i.d. random variables $X_1, X_2, \ldots$ on some probability space Ω taking values in some measurable space $(\mathcal{X}, \mathcal{A})$ with common law P

$$\mathrm{P}(X_i \in B) = P(B), \qquad \forall B \in \mathcal{A}$$

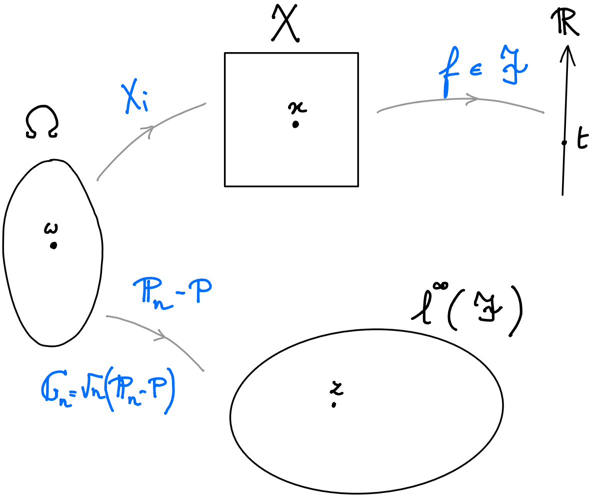
- Family $\mathcal{F} \subset L_1(P)$ of P-integrable functions $f : \mathcal{X} \to \mathbb{R}$
- *Empirical and population measures*: for $f \in \mathcal{F}$,

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

$$Pf = \int_{\mathcal{X}} f(x)\, dP(x) = \mathrm{E}[f(X_i)]$$

- *Empirical process*:

$$\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - P)$$


$\mathcal{X}$
$\mathbb{R}$
X_i
x
$f \in \mathcal{F}$
t
Ω
ω
$\mathbb{P}_n - P$
$\ell^\infty(\mathcal{F})$
$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$
z

Bounded stochastic processes

Assume that the following map is bounded almost surely:

$$\mathcal{F} \to \mathbb{R} : f \mapsto \mathbb{P}_n f - Pf$$

Then we can view $\mathbb{P}_n - P$ and $\mathbb{G}_n$ as (possibly non-measurable) maps

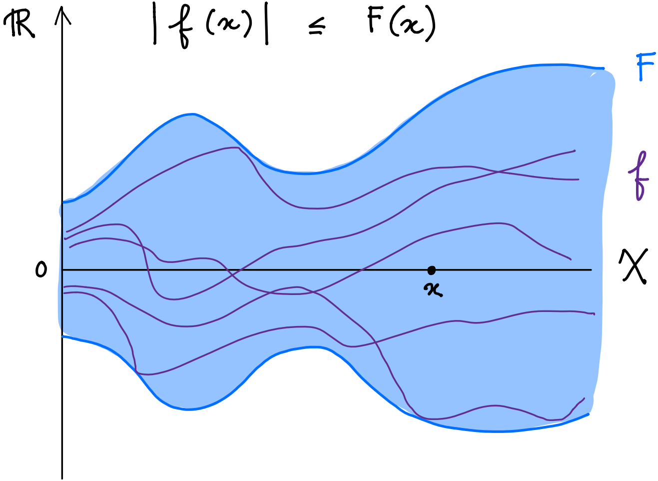
$$\Omega \to \ell^\infty(\mathcal{F})$$

- Assume that $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for P-almost every $x \in \mathcal{X}$ and $\sup_{f \in \mathcal{F}} |Pf| < \infty$
- Often, we will assume that $\mathcal{F}$ has a P-integrable *envelope* $F : \mathcal{X} \to [0, \infty)$, i.e.,

$$\forall x \in \mathcal{X}, \forall f \in \mathcal{F}, \qquad |f(x)| \leqslant F(x)$$

for all $x \in X$ and $f \in \mathcal{F}$:

$$|f(x)| \leq F(x)$$



Uniform law of large numbers: Glivenko–Cantelli

Strong law of large numbers. For all $f \in \mathcal{F}$,

$$\mathbb{P}_n f \to Pf \quad \text{a.s.}, \quad n \to \infty$$

Trivially, for every $f_1, \ldots, f_k \in \mathcal{F}$,

$$\max_{j=1,\ldots,k} |\mathbb{P}_n f - Pf| \to 0 \quad \text{a.s.}, \quad n \to \infty$$

Can we make this statement uniform in $\mathcal{F}$?

Definition: Glivenko–Cantelli. $\mathcal{F}$ is a *P-Glivenko–Cantelli class* if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\text{as}*}{\to} 0, \qquad n \to \infty$$

Uniform central limit theorem: Donsker

Multivariate central limit theorem
Finite-dimensional distributions of $\mathbb{G}_n$ converge to those of a centered Gaussian process $f \mapsto \mathbb{G}f$ on $\mathcal{F}$ with covariance function

$$E[\mathbb{G}f_1 \, \mathbb{G}f_2] = Pf_1 f_2 - Pf_1 \, Pf_2 = \operatorname{cov}(f_1(X), f_2(X))$$

for $f_1, f_2 \in \mathcal{F}$, where $X \sim P$.

Can we make this statement uniform in $f \in \mathcal{F}$?

Definition: Donsker class.
$\mathcal{F}$ is a *P-Donsker class* if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ as $n \to \infty$ in $\ell^\infty(\mathcal{F})$.

To show: *asymptotic tightness*

A counterexample

Let $\mathcal{X} = [0, 1]$ with the Borel σ-field and P the uniform distribution. Consider

$$\mathcal{F} = \{\text{all continuous functions } f : \mathcal{X} \to [0, 1]\}$$

For any $x_1, \ldots, x_n \in \mathcal{X}$ and any $\varepsilon > 0$, we can find $f \in \mathcal{F}$ such that

- $f(x_1) = \ldots = f(x_n) = 1$
- $Pf = \int_0^1 f(x)\, dx \leqslant \varepsilon$.

As a consequence,

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| = 1$$

The class $\mathcal{F}$ is therefore *not* P-Glivenko–Cantelli nor P-Donsker.

By the Stone–Weierstrass theorem, the example extends to the subfamily

$$\mathcal{F}_0 = \{\text{all polynomials } f : \mathcal{X} \to [0, 1] \text{ with rational coefficients}\}$$

See lecture 1.

Complexity of a class of functions

In general, how to put conditions on $\mathcal{F}$ limiting its complexity?

The previous counterexample shows that it is *not* sufficient to impose that

- functions f in $\mathcal{F}$ are bounded
- functions f in $\mathcal{F}$ are smooth
- the cardinality of $\mathcal{F}$ is not too large
- ...

Two techniques: controlling the

- *bracketing* numbers
- *covering* numbers

These will provide sufficient but not necessary conditions

Main sources for this lecture: van der Vaart and Wellner (1996), van der Vaart (1998)

Glivenko–Cantelli and Donsker Theorems

Bracketing entropy

Uniform entropy

VC-classes

Extension: Changing function classes

Definition: Bracketing numbers.

- *Bracket* induced by functions $l, u : \mathcal{X} \to \mathbb{R}$:

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

An ε-bracket $[l, u]$: if $\|u - l\| \leqslant \varepsilon$, with $\|\cdot\|$ some norm

Norm is usually $L_1(P)$ or $L_2(P)$

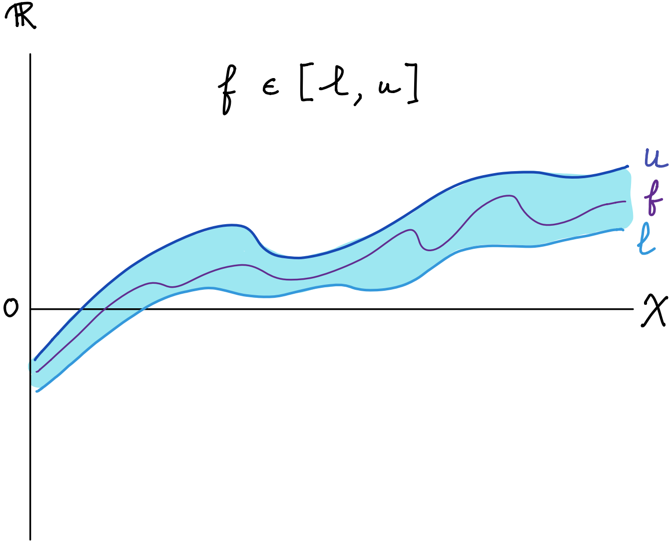
- *Bracketing number*:

$$N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|) = \text{the minimum number, } N\text{, of } \varepsilon\text{-brackets needed to cover } \mathcal{F}\text{, i.e.,} \quad \mathcal{F} \subset \bigcup_{i=1}^N [l_i, u_i] \text{ and } \|u_i - l_i\| \leqslant \varepsilon \text{ for all } i$$

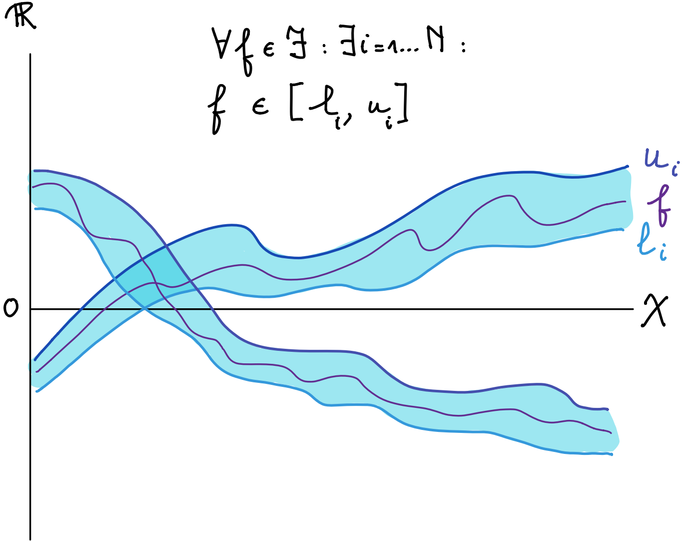
l_i and u_i need not belong to $\mathcal{F}$, but $\|u_i\|$ and $\|l_i\|$ need to be finite.

- *Entropy with bracketing*:

$$\log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$$


ℝ
f ∈ [ℓ, u]
u
f
ℓ
0
X

$$\forall f \in \mathcal{F} : \exists i = 1 \dots N : f \in [l_i, u_i]$$



Glivenko–Cantelli with bracketing

- i.i.d. random elements $X_1, X_2, \ldots$ in $(\mathcal{X}, \mathcal{A})$ with common law P
- Family $\mathcal{F} \subset L_1(P)$ of P-integrable functions $f : \mathcal{X} \to \mathbb{R}$
- Law of large numbers:

$$\forall f \in \mathcal{F}, \qquad \mathbb{P}_n f \to Pf \text{ a.s.}, \quad n \to \infty$$

- Uniformly in $f \in \mathcal{F}$?

Glivenko–Cantelli theorem: bracketing. If

$$\forall \varepsilon > 0, \qquad N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$$

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

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\text{as}*}{\to} 0, \qquad n \to \infty.$$

Proof.

Fix $\varepsilon > 0$.

- Consider a covering $\mathcal{F} \subset \bigcup_{i=1}^{N} [l_i, u_i] \subset L_1(P)$ with ε-brackets:

$$\forall i = 1, \ldots, N, \qquad 0 \leqslant P(u_i - l_i) \leqslant \varepsilon$$

- For every $f \in \mathcal{F}$, find $i \in \{1, \ldots, N\}$ such that $f \in [l_i, u_i]$ and then

$$\begin{array}{ccccccc} P_n l_i & \leqslant & P_n f & \leqslant & P_n u_i & & \\ P l_i & \leqslant & P f & \leqslant & P u_i & \leqslant & P l_i + \varepsilon \end{array}$$

- Bound the supremum over $f \in \mathcal{F}$ by a maximum over $i = 1, \ldots, N$:

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \leqslant \max_{i=1,\ldots,N} |\mathbb{P}_n l_i - P l_i| \vee |\mathbb{P}_n u_i - P u_i| + \varepsilon$$

- Apply the law of large numbers to $l_1, \ldots, l_N$ and $u_1, \ldots, u_N$.

Let $\varepsilon \downarrow 0$. □

Donsker theorem with bracketing

Recall $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ seen as map $\Omega \to \ell^\infty(\mathcal{F})$, for $\mathcal{F} \subset L_2(P)$

Donsker theorem: bracketing.
If for some (and then for all) $\delta > 0$ we have a finite *bracketing integral*

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) := \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}\, d\varepsilon < \infty$$

then $\mathcal{F}$ is P-Donsker, i.e.,

$$\mathbb{G}_n \rightsquigarrow \mathbb{G}, \qquad n \to \infty, \qquad \text{in } \ell^\infty(\mathcal{F})$$

- $L_2(P)$-brackets $[l_i, u_i]$ satisfy $P[(u_i - l_i)^2] \leqslant \varepsilon^2$ and thus $P(u_i - l_i) \leqslant \varepsilon$
  $\Longrightarrow$ $L_2(P)$ brackets are *smaller* than $L_1(P)$ brackets
  $\Longrightarrow$ *higher* bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$
- Finite bracketing integral
  $\Longrightarrow$ $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$ cannot go to ∞ too quickly as $\varepsilon \downarrow 0$

Proof.

Show *asymptotic tightness* of $\mathbb{G}_n$ via finite-partition criterion (part 2, p. 20).

Construct partition $\mathcal{F} = \bigcup_{i=1}^{N} \mathcal{F}_i$ from δ-brackets $[l_i, u_i]$.

Use *maximal inequality*: there exists finite $a(\delta) > 0$ such that

$$\mathrm{E}^* \left[\max_{i=1,\dots,N} \sup_{f,g \in \mathcal{F}_i} |\mathbb{G}_n f - \mathbb{G}_n g| \right] \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + a(\delta)^{-1} P \left[F^2 \, \mathbb{1}\left\{ F > a(\delta) \sqrt{n} \right\} \right]$$

with $F \in L_2(P)$ an envelope of $\mathcal{F}$, constructed via 1-brackets.

Let first $n \to \infty$, then $\delta \to 0$, and apply Markov's inequality. □

Proof of the maximal inequality is difficult. Techniques: Bernstein's inequality, Orlicz norms, chaining. See Lemmas 19.32–34 in van der Vaart (1998).

Example: weighted distribution function

Suppose $\mathcal{X} = [0, 1]$ and P the uniform distribution.

Weighted empirical distribution function: for $w : [0, 1] \to [0, \infty]$, let

$$
\begin{aligned}
f_t(x) &= w(t)\mathbb{1}_{[0,t]}(x), \qquad x, t \in [0, 1] \\
\mathbb{G}_n f_t &= \sqrt{n}\,\{\mathbb{F}_n(t) - t\}\, w(t)
\end{aligned}
$$

Identify $\mathcal{F} = \{f_t \mid t \in [0, 1]\}$ with $[0, 1]$.

For a Brownian bridge B, do we have weak convergence in $\ell^\infty([0, 1])$

$$
\left(\sqrt{n}\,\{\mathbb{F}_n(t) - t\}\, w(t)\right)_{t\in[0,1]} \overset{?}{\rightsquigarrow} (B(t)w(t))_{t\in[0,1]}\,, \qquad n \to \infty
$$

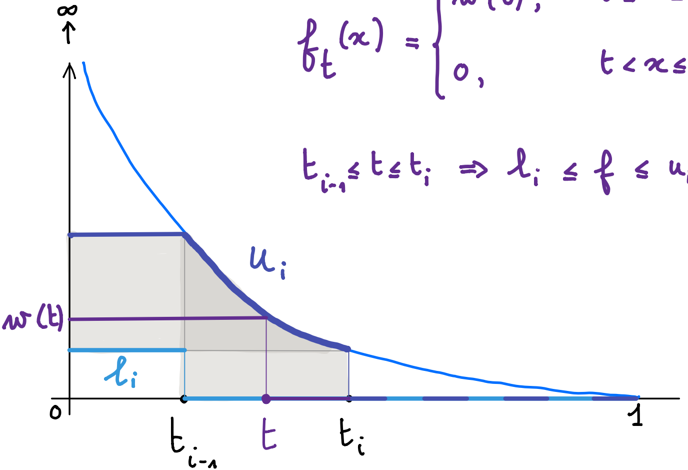
If yes, then construct tail-sensitive Kolmogorov–Smirnov test statistics

$$
\sup_{t\in[0,1]} \left| \sqrt{n}\,\{\mathbb{F}_n(t) - t\}\, w(t) \right|
$$

via w such that $w(t) \to \infty$ as $t \to 0$ or $t \to 1$

$$f_t(x) = \begin{cases} w(t), & 0 \le x \le t \\ 0, & t < x \le 1 \end{cases}$$

$$t_{i-1} \le t \le t_i \Rightarrow \ell_i \le f \le u_i$$



Assume $w : [0,1] \to [0,\infty]$ is decreasing, $w(0) = \infty$, $w(1) = 0$, and $\int_0^1 w^2(t)\,dt < \infty$.

Fix $\varepsilon > 0$. Find a sufficiently fine grid

$$0 = t_0 < t_1 < \ldots < t_m = 1$$

such that the brackets $[l_i, u_i]$ defined by

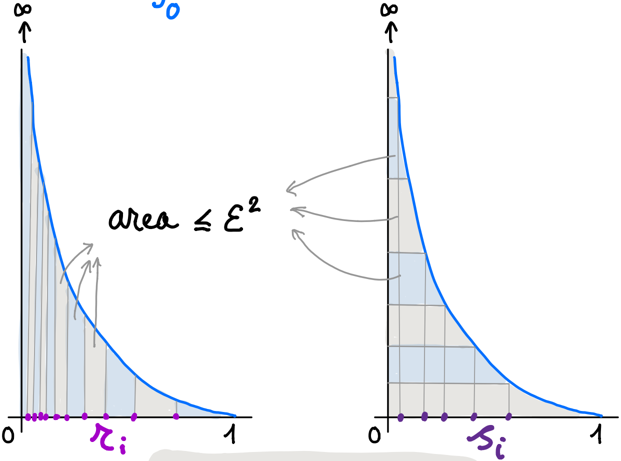
$$\begin{aligned} l_i(x) &= w(t_i)\mathbb{1}_{[0,t_{i-1}]}(x) \\ u_i(x) &= w(t_{i-1})\mathbb{1}_{[0,t_{i-1}]}(x) + w(x)\mathbb{1}_{(t_{i-1},t_i]}(x) \end{aligned}$$

have $L_2(P)$-size $[\int_{[0,1]}(u_i - l_i)^2]^{1/2} \leqslant \varepsilon$. The number m of points t_i needed is

$$N_{[]}(\varepsilon, \mathcal{F}, L_2(P)) = \mathrm{O}(1/\varepsilon^2)$$

As $\int_0^\delta \sqrt{\log(1/\varepsilon)}\,d\varepsilon < \infty$, the bracketing integral $J_{[]}(\delta, \mathcal{F}, L_2(P))$ is finite. □

$$\int_0^1 w^2(t)\,dt < \infty$$



$$\{t_i\} = \{r_i\} \cup \{s_i\}$$

Example: Parametric class

On general $\mathcal{X}$, consider $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ for bounded $\Theta \subset \mathbb{R}^d$.

Suppose there exists $m \in L_2(P)$ such that

$$\forall x \in \mathcal{X}, \forall \theta_1, \theta_2 \in \Theta, \qquad \left| f_{\theta_1}(x) - f_{\theta_2}(x) \right| \leqslant m(x) \, \|\theta_1 - \theta_2\|$$

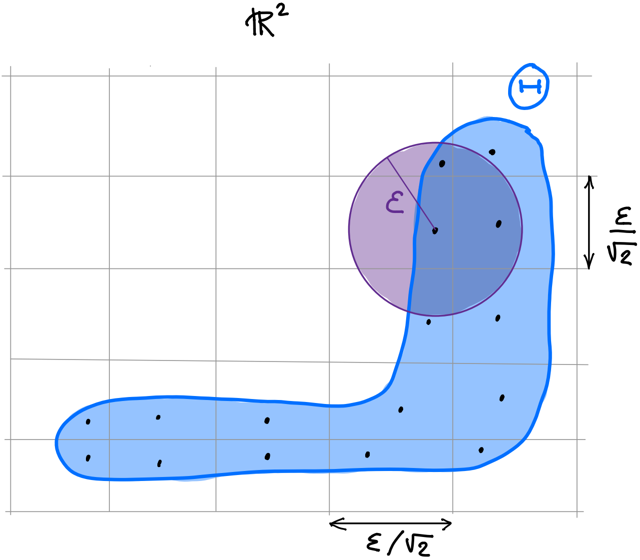
Claim: There exists $K > 0$ depending only on Θ, d, and m such that

$$\forall 0 < \varepsilon < \operatorname{diam} \Theta, \qquad N_{[]}\left(\varepsilon, \mathcal{F}, L_2(P)\right) \leqslant K \left(\frac{\operatorname{diam} \Theta}{\varepsilon} \right)^d$$

Proof: Find grid $\{\theta_i\}_{i=1}^N \subset \Theta$ such that ε-balls with centers θ_i cover Θ.

For $\varepsilon > 0$, cover $\mathcal{F}$ by brackets $[l_i, u_i] = [f_{\theta_i} - \varepsilon m, f_{\theta_i} + \varepsilon m]$.

Grid size N can be bounded by $O\left((\operatorname{diam} \Theta / \varepsilon)^d\right)$ as $\varepsilon \downarrow 0$. □



$\mathbb{R}^2$
Θ
ε
$\frac{\varepsilon}{\sqrt{2}}$
$\varepsilon/\sqrt{2}$

Mean absolute deviation

Let $X_1, X_2, \ldots$ be iid P over $\mathcal{X} = \mathbb{R}$; assume $\mathrm{E}[X^2] < \infty$.
Mean absolute deviation:

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

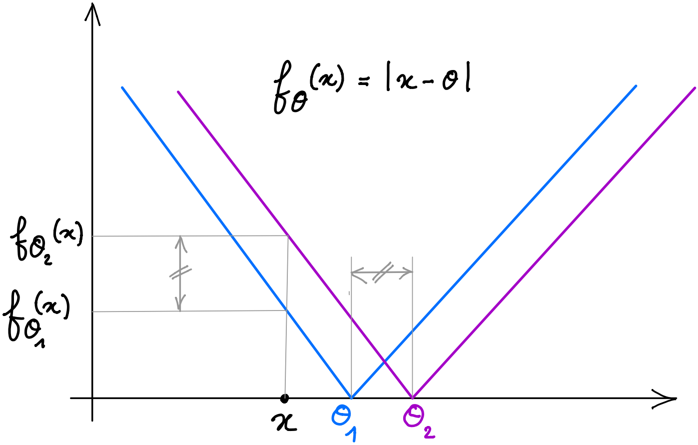
Asymptotic distribution of $\sqrt{n}(M_n - \mathrm{E}[|X - \mu|])$? Define

$$\forall \theta, x \in \mathbb{R}, \qquad f_\theta(x) = |x - \theta|$$

Writing $\mu = \mathrm{E}[X]$, we have

$$\begin{aligned} \sqrt{n}\,(M_n - \mathrm{E}[|X - \mu|]) &= \sqrt{n}\left(\mathbb{P}_n f_{\overline{X}_n} - P f_\mu\right) \\ &= \mathbb{G}_n f_{\overline{X}_n} + \sqrt{n}\left(P f_{\overline{X}_n} - P f_\mu\right) \end{aligned}$$

How to handle the two terms?


$f_\theta(x) = |x - \theta|$
$f_{\theta_2}(x)$
$f_{\theta_1}(x)$
x
θ_1
θ_2

1. Put $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ with $\Theta = (\mu - 1, \mu + 1)$. Then
   - $\overline{X}_n \xrightarrow{\mathrm{P}} \mu \in \Theta$ with large probability
   - $\mathcal{F}$ is P-Donsker by preceding example with $m(x) \equiv 1$
   - the map $\Theta \to L_2(P) : \theta \mapsto f_\theta$ is continuous

   $\mathbb{G}_n$ is asymptotically uniformly $L_2(P)$-equicontinuous in probability (lecture 2, slide 20):

$$\begin{aligned} \mathbb{G}_n f_{\overline{X}_n} &= \mathbb{G}_n f_\mu + o_{\mathrm{P}}(1) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n |X_i - \mu| - \mathrm{E}[|X - \mu|]\right) + o_{\mathrm{P}}(1), \qquad n \to \infty \end{aligned}$$

2. If the cdf F of X is continuous at μ, the map $\theta \mapsto Pf_\theta = \mathrm{E}[|X - \theta|]$ is differentiable at $\theta = \mu$ with derivative $2F(\mu) - 1$.

   Delta method:

$$\sqrt{n}\left(Pf_{\overline{X}_n} - Pf_\mu\right) = \sqrt{n}\,(2F(\mu) - 1)\left(\overline{X}_n - \mu\right) + o_{\mathrm{P}}(1)$$

Summing up the two terms:

$$\sqrt{n}\,(M_n - \mathrm{E}[|X-\mu|]) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathrm{E}[h(X)]\right) + o_{\mathrm{P}}(1)$$
$$\xrightarrow{d} \mathcal{N}\left(0, \mathrm{var}(h(X))\right)$$

where

$$h(x) = |x-\mu| + (2F(\mu) - 1)\,x$$

□

If $F(\mu) = 1/2$, i.e., $\mathrm{E}[X] = \mu$ is also the median, the second term vanishes and it is as if we used μ rather than $\overline{X}_n$ in the definition of M_n.

Glivenko–Cantelli and Donsker Theorems

Bracketing entropy

Uniform entropy

VC-classes

Extension: Changing function classes

Covering numbers

- family $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$
- norm $\|\cdot\|$ on such functions

  usually $L_r(Q)$ for some $r \geqslant 1$ and some probability measure Q on $\mathcal{X}$

Definition: Covering number.

- *Covering number* of $\mathcal{F}$ with respect to $\|\cdot\|$:

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) = \text{minimum number of } \varepsilon\text{-balls needed to cover } \mathcal{F},$$
$$\text{i.e., such that } \mathcal{F} = \bigcup_{i=1}^{N} \{g \in \mathcal{F} : \|g - f_i\| < \varepsilon\}$$

  Centers f_i need not belong to $\mathcal{F}$ but should have $\|f_i\| < \infty$.

- *Entropy (without bracketing)*:

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|)$$

Covering vs bracketing: if norm is such that $|f| \leqslant |g| \implies \|f\| \leqslant \|g\|$, then

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) \leqslant N_{[]}(2\varepsilon, \mathcal{F}, \|\cdot\|)$$

Glivenko–Cantelli with covering

Glivenko–Cantelli theorem with random covering numbers

Suppose $\mathcal{F} \subset L_1(P)$ satisfies:

- $\mathcal{F}$ is P-measurable (see below)
- $\mathcal{F}$ has a P-integrable envelope $F : \mathcal{X} \to \mathbb{R}$

  $|f| \leqslant F$ for all $f \in \mathcal{F}$ and $PF < \infty$
- *random entropy condition*:

$$\frac{1}{n} \log N(\varepsilon \, \mathbb{P}_n F, \mathcal{F}, L_1(\mathbb{P}_n)) \xrightarrow{\mathrm{P}} 0, \qquad n \to \infty$$

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

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \xrightarrow{\mathrm{as}*} 0, \qquad n \to \infty$$

Sufficient that $\mathcal{F}$ is *pointwise measurable*, i.e., there exists countable $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$ there exists $g_m \in \mathcal{G}$ such that $\lim_{m\to\infty} g_m(x) = f(x)$ for all $x \in \mathcal{X}$ (van der Vaart and Wellner, 1996, p. 110).

Uniform covering numbers

- Recall covering number $N(\varepsilon, \mathcal{F}, \|\cdot\|)$
- Let $F : \mathcal{X} \to [0, \infty)$ be a cover of $\mathcal{F}$

Definition: Uniform covering.
For $r \geqslant 1$ and $\varepsilon > 0$, the *uniform covering number* is

$$\sup_Q N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q))$$

- $\|F\|_{Q,r} = \|F\|_{L_r(Q)} = (QF^r)^{1/r} = (\int F^r \, dQ)^{1/r}$
- Supremum over all finitely discrete probability measures Q on $\mathcal{X}$ with $\|F\|_{Q,r} > 0$

The discrete measure $\mathbb{P}_n$ is among the Q in the supremum
$\Longrightarrow$ Random covering numbers are bounded by the uniform ones

Donsker with entropy

Donsker theorem with uniform entropy

Suppose $\mathcal{F} \subset L_2(P)$ satisfies

- $\mathcal{F}$ is "suitably measurable"
- $\mathcal{F}$ has a cover $F \in L_2(P)$
- $\mathcal{F}$ satisfies the *uniform entropy condition*:
  for some (and then for all) $\delta > 0$,

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

then $\mathcal{F}$ is P-Donsker, i.e., $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ as $n \to \infty$ in $\ell^\infty(\mathcal{F})$.

For "suitably measurable": see Theorem 2.5.2 in van der Vaart and Wellner (1996).

Vapnik–Červonenkis index

Let $\mathcal{C}$ be a collection of subsets $C \subset \mathcal{X}$.

- Number of subsets A of $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ *picked out* by $\mathcal{C}$:

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \#\Big\{A \subset \{x_1, \ldots, x_n\} \,\Big|\, \exists C \in \mathcal{C} : A = C \cap \{x_1, \ldots, x_n\}\Big\}$$

- $\mathcal{C}$ *shatters* $\{x_1, \ldots, x_n\}$ if *every* $A \subset \{x_1, \ldots, x_n\}$ is picked out by $\mathcal{C}$:

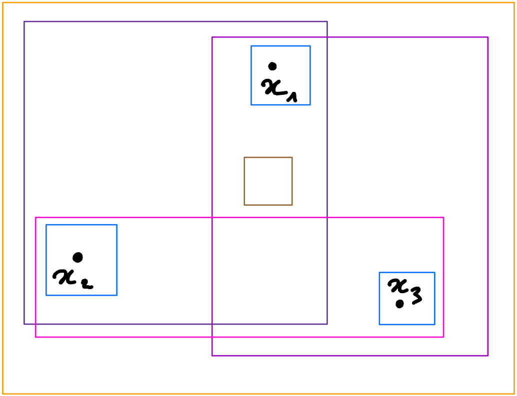
$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = 2^n$$

Definition: VC-index. The *VC-index* of $\mathcal{C}$ is the smallest n such that *no set* of size n is shattered by $\mathcal{C}$:

$$V(\mathcal{C}) = \inf\Big\{n \in \mathbb{N} \,\Big|\, \forall x_1, \ldots, x_n : \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n\Big\}$$

with $\inf \varnothing = \infty$

$\mathcal{C} = \{\text{rectangles in } \mathbb{R}^2, \text{ horizontal/vertical sides}\}$



All $2^3 = 8$ subsets of $\{x_1, x_2, x_3\}$ are picked out by $\mathcal{C}$

Example: rectangles in Euclidean space

In $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{C}$ be the class of all rectangles

$$\prod_{j=1}^{d} ([a_j, b_j] \cap \mathbb{R})$$

with $-\infty \leqslant a_j \leqslant b_j \leqslant \infty$. Then

$$V(\mathcal{C}) \leqslant 2d + 1$$

(Could take open/closed rectangles too.)

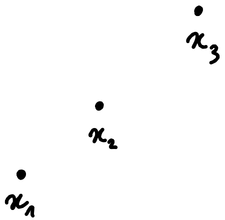
Proof.

For any $x_1, \ldots, x_n \in \mathbb{R}^d$ with $n = 2d + 1$, there exists $i \in \{1, \ldots, n\}$ such that x_i is "boxed in" by the other $n - 1$ points $\{x_k : k \neq i\}$.

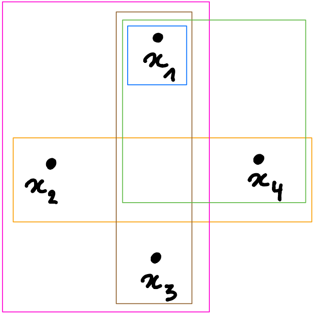
The set $A = \{x_i : k \neq i\}$ is *not* picked out by $\mathcal{C}$, i.e., there exists *no* rectangle $C \in \mathcal{C}$ such that

$$\{x_k : k \neq i\} = C \cap \{x_1, \ldots, x_n\}$$

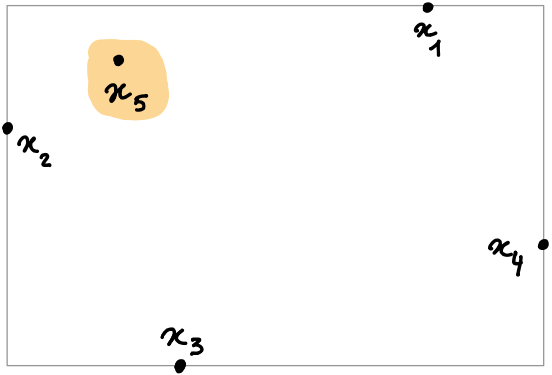
Hence, $\{x_1, \ldots, x_n\}$ is *not* shattered by $\mathcal{C}$. □



The set $A = \{x_1, x_3\}$ cannot be picked out.



All $2^4 = 16$ subsets of $\{x_1, \dots, x_4\}$ can be picked out.



With $n = 5$ points, there is always a point x_i such that $A = \{x_1, \dots, x_5\} \setminus \{x_i\}$ cannot be picked out $\Rightarrow$ $V(\mathcal{C}) = 5$

VC-classes of sets and functions

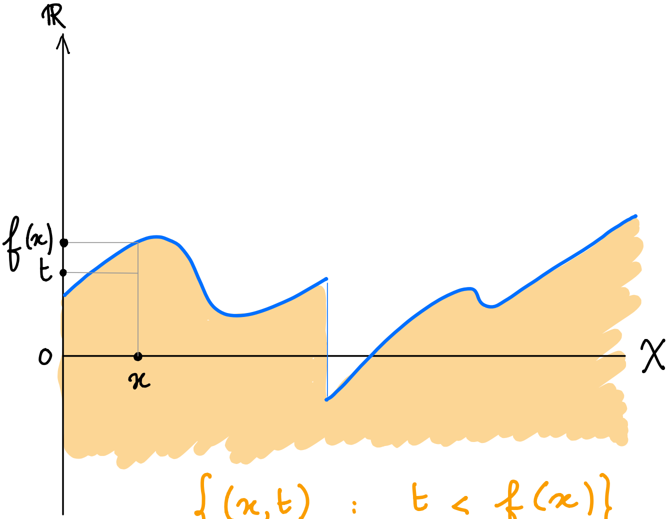
- Let $\mathcal{C}$ be a collection of subsets $C \subset \mathcal{X}$
- Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$

Definition: VC-class.

- $\mathcal{C}$ is a *VC-class* if $V(\mathcal{C}) < \infty$
  There exists $n \in \mathbb{N}$ such that no set of size n is shattered by $\mathcal{C}$
- $\mathcal{F}$ is a *VC-class* if the collection of *subgraphs*

$$\Big\{\{(x,t) \in \mathcal{X} \times \mathbb{R} \mid t < f(x)\} \,\Big|\, f \in \mathcal{F}\Big\}$$

  is a VC-class in $\mathcal{X} \times \mathbb{R}$



$$\{(x,t) \; : \; t < f(x)\}$$

VC-classes and uniform covering numbers

Theorem: uniform covering numbers of VC-classes
There exists a universal constant $K > 0$ such that for any VC-class $\mathcal{F}$ of functions with envelope F, any $r \geqslant 1$, any $0 < \varepsilon < 1$,

$$\sup_Q N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant K\, V(\mathcal{F})\, (16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}$$

the supremum being over all finitely discrete probability measures.

Proof via *Sauer's lemma*: if $\mathcal{C}$ is a VC-class of sets, then

$$\max_{x_1,\ldots,x_n \in \mathcal{X}} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leqslant \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j} = \mathrm{O}\left(n^{V(\mathcal{C})-1}\right), \qquad n \to \infty$$

Empirical processes indexed by VC-classes

VC-classes $\mathcal{F}$ are "small":

- the *random entropy condition* (slide 30) is trivially satisfied
- the uniform entropy integral (slide 32) converges:

$$\int_0^\delta \sqrt{\log(1/\varepsilon)}\, d\varepsilon < \infty$$

$\implies$ A VC-class $\mathcal{F}$ is Glivenko–Cantelli and Donsker provided it is suitably measurable and possesses an appropriate envelope F

Extensions to:

VC-hull classes $\mathcal{F}$: There exists a VC-class $\mathcal{G}$ such that every $f \in \mathcal{F}$ is the pointwise limit of a sequence $f_m = \sum_{i=1}^m \alpha_{mi} g_{mi}$ with $\sum_{i=1}^m |\alpha_{mi}| \leqslant 1$ and $g_{mi} \in \mathcal{G}$

VC-major classes $\mathcal{F}$: The collection $\mathcal{C} = \{C_{f,t} \mid f \in \mathcal{F}, t \in \mathbb{R}\}$ with $C_{f,t} = \{x \in \mathcal{X} \mid f(x) > t\}$ is VC in $\mathcal{X}$

(van der Vaart and Wellner, 1996, Sections 2.6.3 and 2.6.4)

Example: finite-dimensional vector spaces

Suppose there exists $f_1, \ldots, f_k : \mathcal{X} \to \mathbb{R}$ such that

$$\mathcal{F} \subset \{\lambda_1 f_1 + \cdots + \lambda_k f_k \mid \lambda_1, \ldots, \lambda_k \in \mathbb{R}\}$$

Then $V(\mathcal{F}) \leqslant k + 2$.

Proof.

Let $n = k + 2$. Take any $(x_1, t_1), \ldots, (x_n, t_n) \in \mathcal{X} \times \mathbb{R}$. The vectors

$$v_f = (f(x_1) - t_1, \ldots, f(x_n) - t_n), \qquad f \in \mathcal{F}$$

lie in a $(k + 1)$-dimensional subspace $\mathbb{V}$ in $\mathbb{R}^{k+2}$. There exists $0 \neq a \in \mathbb{R}^{k+2}$ with at least one positive coordinate such that $a \perp \mathbb{V}$, i.e.,

$$\forall f \in \mathcal{F}, \qquad \sum_{i:a_i>0} a_i \,(f(x_i) - t_i) = \sum_{i:a_i<0} (-a_i)\,(f(x_i) - t_i)$$

The set $A = \{(x_i, t_i) \mid \text{those } i = 1, \ldots, n \text{ such that } a_i > 0\}$ is *not* picked out by a subgraph $\{(x, t) : t < f(x)\}$. □

Example: sets defined by polynomial equations

- If $\mathcal{F}$ is a VC-class of functions, then

$$\mathcal{C} = \Big\{ \{x \in \mathcal{X} \mid f(x) > 0\} \;\Big|\; f \in \mathcal{F} \Big\}$$

  is a VC-class of sets.

  Proof: Suppose $\mathcal{C}$ is not VC. For any n, there exists $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ shattered by $\mathcal{C}$. Then $\{(x_1, 0), \ldots, (x_n, 0)\} \subset \mathcal{X} \times \mathbb{R}$ is shattered by the subgraphs $\{(x,t) : t < f(x)\}$ for $f \in \mathcal{F}$. Hence $\mathcal{F}$ is not a VC-class of functions, a contradiction. □

- The set $\mathcal{F}$ of polynomials on $\mathbb{R}^d$ up to a fixed degree m is a VC-class of functions by slide 42.
- Hence, the collection of half-spaces

$$\mathcal{C} = \Big\{ \Big\{ x \in \mathbb{R}^d \mid a_1 x_1 + \cdots + a_d x_d \leqslant b \Big\} \;\Big|\; a_1, \ldots, a_d, b \in \mathbb{R} \Big\}$$

  is a VC-class of sets. So are the ellipsoids, ...

Generating more examples through stability properties

If $\mathcal{C}$ and $\mathcal{D}$ are VC-classes of sets, then so are

- $\{C \cap D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$
- $\{C \cup D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$

Proof.
By Sauer's lemma, the number of subsets of any $\{x_1, \ldots, x_n\}$ picked out by intersections $C \cap D$ is only $O(n^{V(\mathcal{C})+V(\mathcal{D})-2}) < 2^n$. □

If $\mathcal{F}$ and $\mathcal{G}$ are VC-classes of functions, then so are

- $\{f \wedge g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$
- $\{f \vee g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$

Proof.
The subgraph of $f \wedge g$ is the intersection of the subgraphs of f and g. □

Counterexample: convex sets

In $\mathcal{X} = [-1, 1]^2$, the following class is *not* VC:

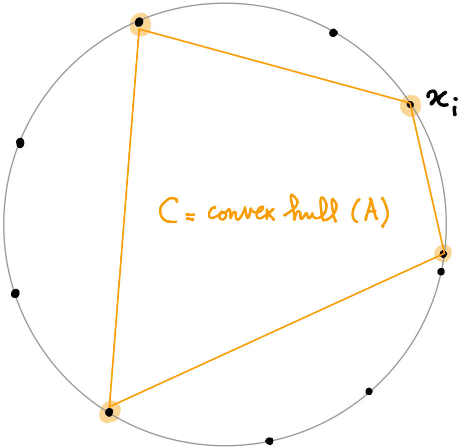
$$\mathcal{C} = \left\{\text{all convex polygons in } [-1, 1]^2\right\}$$

Proof.

Any set of n distinct points on the unit circle is shattered by $\mathcal{C}$. □

The class $\mathcal{F} = \{\mathbb{1}_C \mid C \in \mathcal{C}\}$ is *not* P-Glivenko–Cantelli or Donsker for the uniform distribution P on the unit circle: for all n,

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| = \sup_{C \in \mathcal{C}} \left|\mathbb{P}_n(C) - P(C)\right| = 1$$



Any $A \subset \{x_1, \ldots, x_n\}$ can be picked out

Glivenko–Cantelli and Donsker Theorems

Bracketing entropy

Uniform entropy

VC-classes

Extension: Changing function classes

Local empirical processes

Let $X_1, X_2, \ldots$ be iid uniform on $[0, 1]$.

Tail empirical process: for bandwidths $h = h_n \to 0$ such that $nh \to \infty$,

$$0 \leqslant t \mapsto \sqrt{nh}\left(\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}\{X_i \leqslant ht\} - t\right)$$

Example of a *local empirical process*:

- Density estimation
- Extreme-value theory (usual notation: $nh = k$)

Lindeberg–Feller central limit theorem: since $h \to 0$, the finite-dimensional distributions converge to those of a Wiener process $(W(t))_{t \geqslant 0}$

Centered Gaussian process such that $\mathrm{E}[W(s)W(t)] = s \wedge t$ for $s, t \geqslant 0$

Weak convergence in $\ell^\infty([0, M])$ for fixed $M > 0$?

Sums of independent stochastic processes

Alternative representation of the usual empirical process:

$$\mathbb{G}_n f = \sum_{i=1}^{n} \frac{f(X_i) - Pf}{\sqrt{n}} = \sum_{i=1}^{n} \Big(Z_{n,i}(f) - \mathrm{E}[Z_{n,i}(f)] \Big), \qquad f \in \mathcal{F}$$

Representation also encompasses tail empirical process on slide 48:

$$Z_{n,i}(t) = \frac{1}{\sqrt{nh}} \mathbb{1}\{X_i \leqslant ht\}, \qquad t \in [0, M] = \mathcal{F}$$

Convergence in $\ell^\infty(\mathcal{F})$?

Asymptotic tightness

Theorem. For each n, let $Z_{n,1}, \ldots, Z_{n,n}$ be independent stochastic processes with finite second moments indexed by a set $\mathcal{F}$. Suppose

1. $\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}^* \left[\|Z_{n,i}\|_{\mathcal{F}} \mathbb{1}\{\|Z_{n,i}\|_{\mathcal{F}} > \lambda\} \right] = 0$ for all $\lambda > 0$
   Notation: $\|Z_{n,i}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Z_{n,i}(f)|$
2. There exists $c > 0$ and, for all sufficiently small $\varepsilon > 0$, a covering $\mathcal{F} = \bigcup_{j=1}^{N_\varepsilon} \mathcal{F}_{\varepsilon,j}$ such that, for every set $\mathcal{F}_{\varepsilon,j}$ and every n,

$$\sum_{i=1}^{n} \mathrm{E}^* \left[\sup_{f,g\in\mathcal{F}_{\varepsilon,j}} \left| Z_{n,i}(f) - Z_{n,i}(g) \right|^2 \right] \leqslant c\varepsilon^2$$

3. For some $\delta > 0$, we have $\int_0^\delta \sqrt{\log N_\varepsilon}\, d\varepsilon < \infty$

Then $\sum_{i=1}^{n} \{Z_{n,i} - \mathrm{E}(Z_{n,i})\}$ is asymptotically tight in $\ell^\infty(\mathcal{F})$.
It converges weakly provided finite-dimensional distributions do so.

Corollary to Theorem 2.11.9 in van der Vaart and Wellner (1996)

Example: bracketing

Empirical process $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$ indexed by $f \in \mathcal{F} \subset L_2(P)$ and

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) < \infty$$

Then $\mathcal{F}$ is P-Donsker by slide 16.
Alternatively, check conditions on slide 50 for $Z_{ni}(f) = n^{-1/2} f(X_i)$:

1. Cover $\mathcal{F}$ by N_1 brackets $[l_j, u_j]$ with $P[(u_j - l_j)^2] \leqslant 1$, then

$$\sup_{f \in \mathcal{F}} |f(x)| \leqslant \max_{j=1,\dots,N(1)} \max(|l_j(x)|, |u_j(x)|) =: F(x)$$

$$\|Z_{n,i}\|_{\mathcal{F}} \leqslant n^{-1/2} F(X_i)$$

   Since $F \in L_2(P)$, the condition follows from Markov's inequality

2. Cover $\mathcal{F}$ by N_ε brackets $[l_j, u_j] = \mathcal{F}_{\varepsilon,j}$ with $P[(u_j - l_j)^2] \leqslant \varepsilon^2$. Then

$$\sup_{f,g \in [l_j, u_j]} \left| Z_{n,i}(f) - Z_{n,i}(g) \right|^2 \leqslant \frac{1}{n} \left| u_j(X_i) - l_j(X_i) \right|^2$$

3. Since $J_{[]}(\delta, \mathcal{F}, L_2(P)) < \infty$.

Example: tail empirical process

Check three conditions on slide 50 for

$$Z_{n,i}(t) = (nh)^{-1/2}\mathbb{1}\{X_i \leqslant ht\}, \qquad t \in [0, M] = \mathcal{F}$$

1. Trivial, since $0 \leqslant \|Z_{n,i}\|_{\mathcal{F}} \leqslant (nh)^{-1/2} \to 0$
2. Construct a grid $0 = t_0 < t_1 < \ldots < t_{N_\varepsilon} = M$ with mesh $\varepsilon > 0$ and put

$$\mathcal{F}_{\varepsilon,j} = [t_{j-1}, t_j], \qquad j = 1, \ldots, N_\varepsilon$$

3. $N_\varepsilon = \mathrm{O}(1/\varepsilon)$ as $\varepsilon \downarrow 0$, hence $\int_0^\delta \sqrt{\log N_\varepsilon}\, d\varepsilon < \infty$

□

Summary

Empirical process indexed by family $\mathcal{F} \subset L_1(P)$ or $L_2(P)$ of functions $f : \mathcal{X} \to \mathbb{R}$

- P-Glivenko–Cantelli: uniform large of law numbers

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\text{as}*}{\to} 0$$

- P-Donsker: uniform central limit theorem

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{F})$$

Seek asymptotic tightness $\implies$ techniques to control the complexity of $\mathcal{F}$:

- bracketing numbers
- random and uniform covering numbers
  - VC-classes of sets and functions

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer Sciences+Business Media.