Study Group

on

Empirical Processes.

Contact: pasupath@purdue.edu

- Apology for handwritten notes!
- How did we get here?
- Background and preparation
- Frequency and Other Logistics
- Agenda

Why Study Empirical Processes?

Four motivating contexts

I. Estimate a cdf or a quantile function.

II. Stochastic Optimization & Stochastic ~~Fixed Point~~ Problem in Euclidean space.

III. Stochastic Optimization & Stochastic ~~Fixed Point~~ Problem in "non standard" spaces.

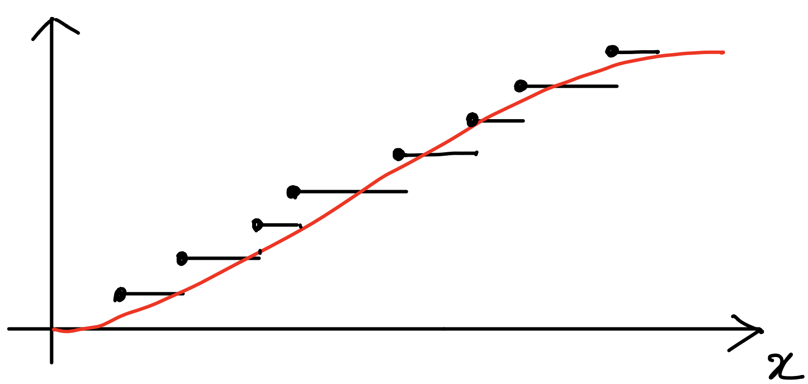
IV. Analyzing the bootstrap. X

I. Estimate a CDF, Quantile

Let $X_1, X_2, \ldots$ be iid random variables with distribution function F.

The natural estimator of F is:

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$$



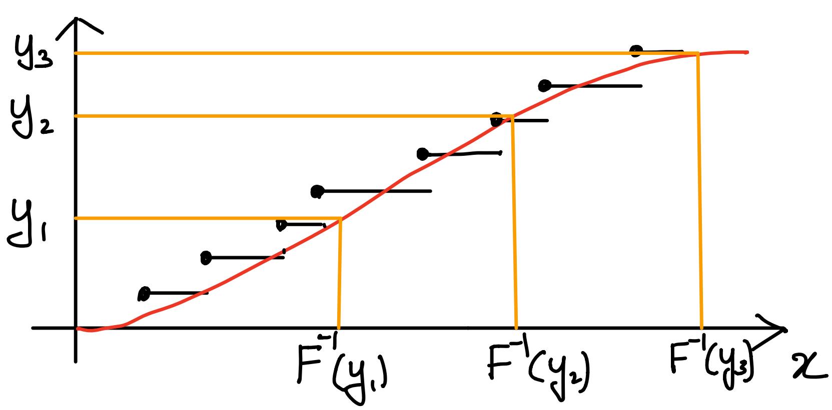
I. Estimate a CDF, Quantile

The natural estimator of the quantile function

$$F^{-1}(y) := \inf\{x : F(x) \geq y\}$$

is the sample quantile function

$$Q_n(y) := \inf\{x : F_n(x) \geq y\}.$$



I. Estimate a CDF, Quantile

What can we say about F_n as an estimator of F?

Specifically:

(i) $\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$?

(ii) How fast?

Analogous questions about Q_n as an estimator of F^{-1}.

I. Estimate a CDF, Quantile

The principal route to answering such questions analyzes the empirical process:

$$n^{\beta}\left(P_n(x) - P(x)\right)$$

$$\beta_n(x) = \sqrt{n}\left(F_n(x) - F(x)\right), \quad -\infty < x < \infty$$

$$q_n(y) = \sqrt{n}\left(Q_n(y) - F^{-1}(y)\right), \quad 0 < y < 1$$

I. Estimate a CDF, Quantile

For example, if $X_j \in [0,1]$, $j=1,2,\dots,n$ then,

$$B_n \Longrightarrow B \qquad (1)$$

where B is the Gaussian random element of $D[0,1]$ specified by

$$\mathbb{E}[B(t)] = 0, \quad \mathbb{E}[B(s)\,B(t)] = F(s \wedge t) - F(s)\,F(t).$$

(Theorem 14.3 in Billingsley, 1999.)

Beware: (1) is weak convergence in metric space.

I. Estimate a CDF, Quantile

So what?

<u>Weak</u> <u>Invariance</u> <u>Principle</u>

Because of (1), it may be expected that "operations" on B_n will "behave like operations" of B.

Hence the behavior of F_n "can be approximated" by passing to the limit.

I. Estimate a CDF, Quantile

Two Classic Examples (DasGupta, 2011)

(i) Kolmogorov-Smirnov Statistic

$$\sup_{0<x<1} |B_n(x)| \Longrightarrow \sup_{0<x<1} |B(x)|$$

where $\sup_{0<x<1} |B(x)|$ has the

known cdf $1 - 2\sum_{k=1}^{\infty}(-1)^{k-1} e^{-2k^2y^2}, \; y>0$

(ii) Cramér-von Mises

$$\int_0^1 B_n^2(x)\, dF \Longrightarrow \int_0^1 B^2(x)\, dx$$

I. Estimate a CDF, Quantile

The analogous setting for the quantile process

Suppose F is absolutely continuous with density $f>0$, f differentiable and $F(x)(1-F(x))\dfrac{|f'(x)|}{f^2(x)}$ is uniformly bounded on supp. (f). Then, for any $c\in(0,1)$

$$\sup_{c\le y\le 1-c}\left|q_n(y)\,f\left(F^{-1}(y)\right)\right| \Longrightarrow \sup_{c\le y\le 1-c}|B(y)|$$

I. Estimate a CDF, Quantile

Stepping back . . .

$$\mathbb{B}_n(x) = \underbrace{\sqrt{n}}_{\text{"scaling"}} \Big(\underbrace{F_n(x) - \overbrace{F(x)}^{\text{"parameter"}}}_{\text{"est. error"}} \Big), \quad x \in \mathbb{R}$$

$$\Longrightarrow \underbrace{\{ \mathbb{B}(x),\ x \in \mathbb{R} \}}_{\text{Gaussian process}}$$

Notice: We can also write the above as

$$\mathbb{B}_n(x) = \sqrt{n}\Big(\mathbb{P}_n\big((-\infty, x]\big) - \mathbb{P}\big((-\infty, x]\big)\Big), \quad x \in \mathbb{R}$$

$$\Longrightarrow \Big\{ \mathbb{P}_{\mathbb{B}}\big((-\infty, x]\big),\ x \in \mathbb{R} \Big\}$$

II Stochastic Optimization

Example 1

Relate employment with education.

Data (Y_i, Z_i), $i = 1, 2, \dots n$

where

$$Y_i = \begin{cases} 1 & \text{if } i \text{ has a job.} \\ 0 & \text{otherwise} \end{cases}$$

Z_i is the education of individual i.

Assume:

$$P(Y=1 \mid Z=z) = G(\theta^* z), \quad \theta^* \in \mathbb{R}^d$$

where $G(\xi) = (1+e^{-\xi})^{-1}$, $\xi \in \mathbb{R}$.

II Stochastic Optimization

Optimization Problem (Classic MLE)

Find $\theta_n \in \operatorname{argmax} H_n(\theta)$ ——— (P_n)

where

$$H_n(\theta) := \frac{1}{n} \sum_{j=1}^{n} H(\theta, Y_j)$$

$$H(\theta, Y) := \log G(\theta Z)^{Y} \left(1 - G(\theta Z)\right)^{1-Y}$$

$$\theta^* := \operatorname{argmax} h(\theta) = \mathbb{E}\left[H(\theta, Y)\right]$$

II Stochastic Optimization

Key Questions

(i) Does θ_n converge to θ in any sense, e.g., $\|\theta_n - \theta^*\| \xrightarrow{a.s.} 0$? $H(\theta_n) - H(\theta^*) \xrightarrow{a.s.} 0$?

(ii) How fast?

II Stochastic Optimization

In answering questions in A. & B.

EP is the natural viewpoint. Why?

Define the EP

$$\mathcal{E}_n(\theta) = \sqrt{n}\,\big(\underbrace{H_n(\theta) - h(\theta)}_{\text{error}}\big), \quad \theta \in \mathbb{R}^d.$$

parameter

scaling

Principle: If $\mathcal{E}_n$ can be shown to converge to a Gaussian process, then, answers to (i) and (ii) become possible.

II Stochastic Optimization

There is nothing special about the MLE example.

General Stochastic Optimization:

$$\boxed{\begin{aligned} &\text{Find } \operatorname*{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} h(\theta) := \mathbb{E}\left[H(\theta, Y)\right] \\ &\text{where } Y \text{ in } (\mathcal{Y}, \mathcal{A}), \; H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}. \end{aligned}}$$

(SO)

II Stochastic Optimization

$$\text{Find } \underset{\theta \in \Theta \subseteq \mathbb{R}^d}{\operatorname{argmin}} h(\theta) := \mathbb{E}\left[H(\theta, Y)\right]$$

where Y in $(\mathcal{Y}, \mathcal{A})$, $H: \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$.

"Sample Version" of (SO).

$$\text{Find } \underset{\theta \in \Theta \subseteq \mathbb{R}^d}{\operatorname{argmin}} H_n(\theta) := \frac{1}{n} \sum_{j=1}^{n} H(\theta, Y_j)$$

where Y_j, $j = 1, 2, \ldots \in (\mathcal{Y}, \mathcal{A})$, i.i.d.

II Stochastic Optimization

Example Settings for SO.

- Virtually all of regression
- A large fraction of optimization problems in Operations Research, Computer Science, & Engineering.

III Stochastic Optimization

The non-Euclidean context emerges...

Example 2. Recall that in Example 1,

$$P(Y=1 \mid Z=z) = G(\theta^{*} z), \quad \theta^{*} \in \mathbb{R}^d$$

$$\boxed{\underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \; h(\theta) := \mathbb{E}\left[H(\theta, Y)\right]}$$

What if, instead:

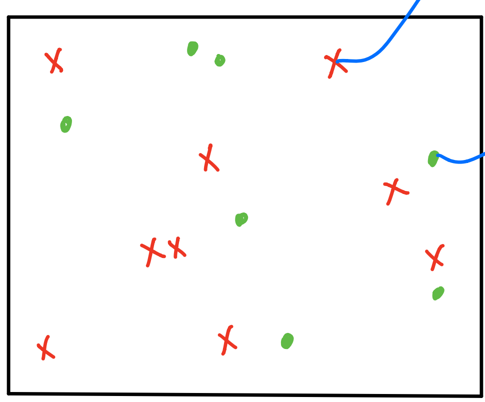
$$P(Y=1 \mid Z=z) = G_0(z)$$

$$\boxed{\underset{G \in \mathcal{G} = \{\mathbb{R} \to [0,1], \text{ increasing}\}}{\operatorname{argmax}} \; h(G) := \mathbb{E}\left[H(G, Y)\right]}$$

III Stochastic Optimization

Example 3

S



"event" occurs according to prob. measure Λ.

"good guys" sprinkled according to Poisson measure θ

$$h(\theta) = \int_S \Lambda(dx) \overbrace{\int_0^1 P\big(f(R_{x,\theta}) \geq u\big)\, du}^{\text{exp. prob. of death}}$$

$$= \int_S \Lambda(dx) \int_0^1 \left[\exp\left\{-\theta\left(B(x, v f^{-1}(u))\right)\right\}\right]$$

$$\boxed{\text{Find } \operatorname{argmin} h(\theta), \quad \int \theta(dx) = b}$$

III Stochastic Optimization

Define the EP

$$\mathcal{E}_n(\theta) = \underbrace{\sqrt{n}}_{\text{scaling}} \Big(\underbrace{H_n(\overset{\text{parameter}}{\theta}) - h(\theta)}_{\text{error}} \Big), \quad \theta \in \Theta$$

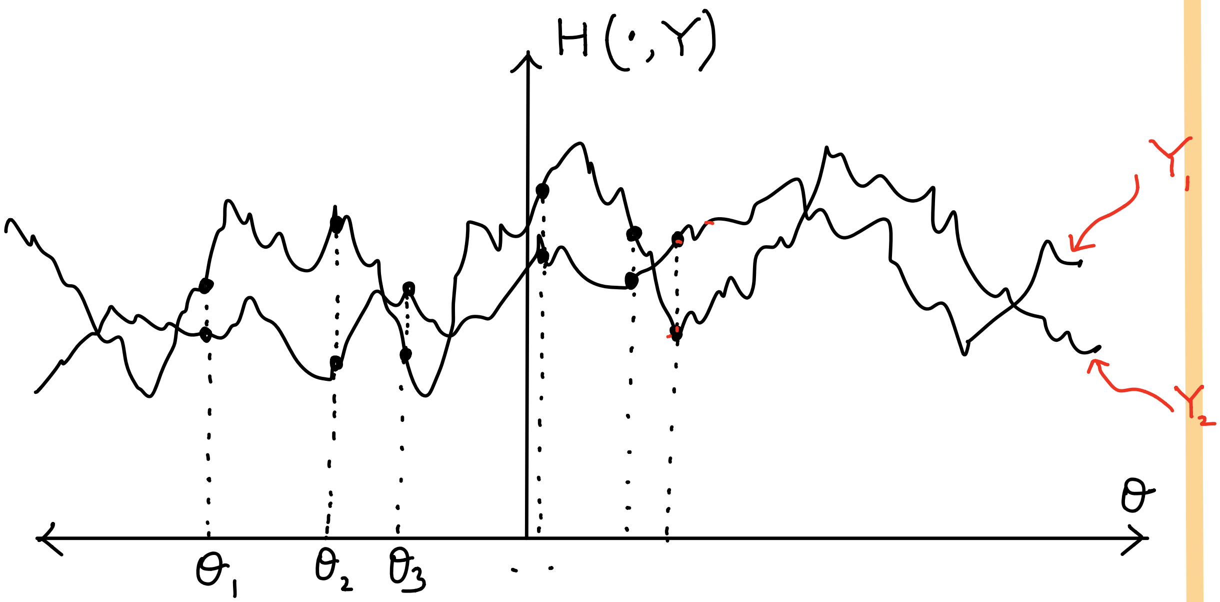
Q.1 What are conditions for

$$\sup_{\theta \in \Theta} \left| H_n(\theta) - h(\theta) \right| \xrightarrow{a.s.} 0 \, ?$$

Q.2. At what rate?

III Stochastic Optimization

Answers to Q.1 and Q.2 will be stated in terms of entropy. or "diversity" or "richness" of the class $\{ H(\theta, Y_j), \theta \in \Theta, j = 1, 2, \ldots, n \}$.



III Stochastic Optimization

AGENDA

I. Basic Results

(i) Donsker's thm for partial sums

(ii) EP for CDFs

(iii) Quantile process

II. M-Estimation

(i) Entropy

(ii) Uniform convergence

(iii) Asymptotic equicontinuity.

WEAK CONVERGENCE ESSENTIALS

Today's Agenda:

Demonstrate the first example of weak convergence in "non-Euclidean" space.

- Weak convergence essentials
- Specialize to $C[0,1]$, main theorem
- Wiener measure, Donsker's Thm.

WEAK CONVERGENCE ESSENTIALS

Our treatment is on a <u>metric space</u> $S = (S, \rho)$, where S is a set and ρ is a metric on S.

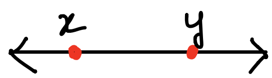
For $x, y, z \in S$

$$\boxed{\begin{array}{ll} \text{(M1)} & \rho(x,y) \in [0,\infty) \\ \text{(M2)} & \rho(x,y) = 0 \text{ iff } x = y \\ \text{(M3)} & \rho(x,y) = \rho(y,x) \\ \text{(M4)} & \rho(x,z) \le \rho(x,y) + \rho(y,z) \end{array}}$$

(See D2011 or K1978)

WEAK CONVERGENCE ESSENTIALS

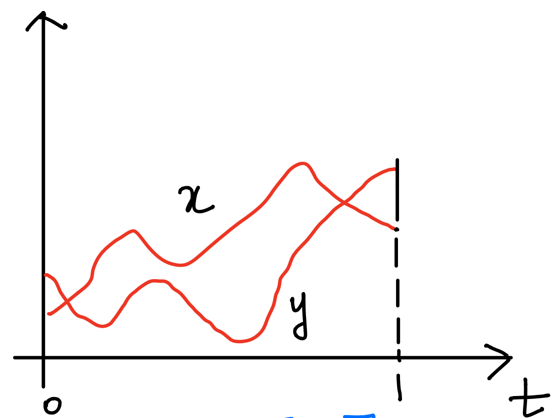
I Reals ($\mathbb{R}$)



$S = \mathbb{R}$

$\rho(x,y) = |x-y|$

II Function Space



$S = C[0,1]$

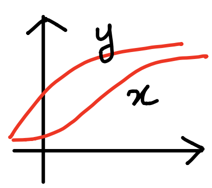
$$\rho(x,y) = \sup_{0 \le t \le 1} |x(t) - y(t)|$$

III Sequence Space, $S = \ell^\infty$.

$x = (\xi_1, \xi_2, \dots)$; $y = (\eta_1, \eta_2 \dots)$

$$\rho(x,y) = \sup_{j \in \mathbb{N}} |\xi_j - \eta_j|$$

IV Measure Space (M)



$x, y \in M$ with cdf F_x, F_y

$\rho(x,y)$

$$= \sup_{z \in \mathbb{R}^d} |F_x(z) - F_y(z)|$$

(See D 2011)

WEAK CONVERGENCE ESSENTIALS

The metric "does a lot of work."

$$\begin{tikzcd}
& \boxed{\text{Metric}} \arrow[d] & \\
& \boxed{\text{Open Sets}} \arrow[dl] \arrow[dr] & \\
\boxed{\text{Borel } \sigma\text{-algebra}} & & \boxed{\begin{array}{l}\text{convergence}\\ \text{continuity}\\ \text{separability}\\ \text{completeness}\end{array}}
\end{tikzcd}$$

(Sometimes, the Borel σ-algebra is "too big.")

So, defining the metric is key to defining the measurable space $(S, \mathcal{S})$.

Weak Convergence Essentials

A probability measure on $\mathcal{S}$ is a non-negative, countably additive set function with $P(S) = 1$.

P_n converges weakly to P means, for $A \in \mathcal{S}$

$$\boxed{P_n(A) \longrightarrow P(A) \text{ if } P(\partial A) = 0.}$$

Notation:

$$P_n \Rightarrow P$$

WEAK CONVERGENCE ESSENTIALS

What is the relevance of the P-continuity set condition $P(\partial A) = 0$?

Understand by Analogy

In $\mathbb{R}$, $X_n \Rightarrow X$ means

$F_{X_n}(t) \longrightarrow F_X(t)$ at points t where F_X is continous.

$F_{X_n}(t) = P_n((-\infty, t])$; $F_X(t) = P((-\infty, t])$

WEAK CONVERGENCE ESSENTIALS

Let $X : \Omega \longrightarrow S$ be a mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to the metric space S.

X is measurable $\mathcal{F}/\mathcal{S}$. $(X \in \mathcal{F}/\mathcal{S})$

Distribution of X is the probability measure induced by X.

$P = \mathbb{P}X^{-1}$, that is:

$$P(A) = (\mathbb{P}X^{-1})(A) = \mathbb{P}(X^{-1}(A))$$

$$= \mathbb{P}(\omega : X(\omega) \in A)$$

WEAK CONVERGENCE ESSENTIALS

(i) – (iv) mean the same.

(i) $X_n \Rightarrow X$

(ii) $P_n \Rightarrow P$

(iii) $X_n \Rightarrow P$

(iv) $P_n \Rightarrow X$

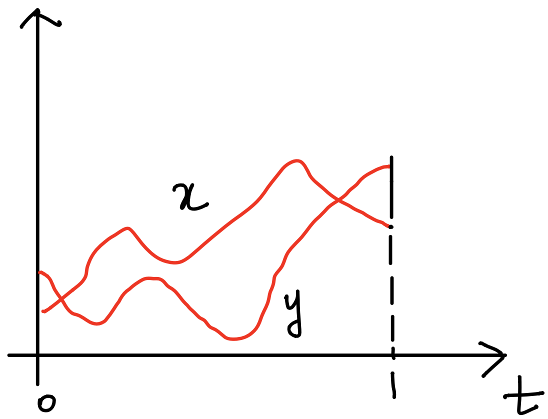
And, equivalent to:

$$\mathbb{E}\left[f(X_n)\right] \longrightarrow \mathbb{E}\left[f(X)\right]$$

for all bounded, uniformly continuous f.

WEAK CONVERGENCE ESSENTIALS

Let's live in $C[0,1]$ today.



$C = C[0,1]$ is the space of continuous functions endowed with the "uniform topology":

$$\rho(x, y) = \|x - y\| = \sup_t |x(t) - y(t)|$$

$(C, \mathcal{C})$: corresponding measurable space.

WEAK CONVERGENCE ESSENTIALS

Random Function; Random Vector; Random Variable.

$(\Omega, \mathcal{F}, \mathbb{P})$ and $X : \Omega \to C[0,1]$

If $X \in \mathcal{F}/\mathcal{C}$, then it is called a random function.

random variable

$\pi_t X(\omega) := X(t,\omega)$, $\pi_t X \in \mathcal{F}/\mathbb{R}$

$\pi_{t_1,t_2,\dots t_k} X(\omega) := \left(X(t_1,\omega), X(t_2,\omega), \dots, X(t_k,\omega)\right)$

$\pi_{t_1,t_2,\dots t_k} X \in \mathcal{F}/\mathbb{R}^k$

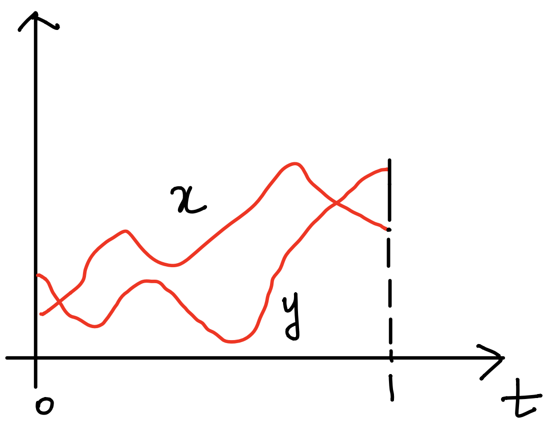
random vector

WEAK CONVERGENCE ESSENTIALS

Modulus of Continuity.

$$m_x(\delta) := \sup_{|s-t| \le \delta} |x(s) - x(t)|$$

$0 < \delta \le 1$



"$m_x(\cdot)$ quantifies how ε changes with δ in the ε-δ definition of continuity."

Examples

(i) If x is L-Lipschitz, $m_x(\delta) \le L\delta$

(ii) If x is Hölder, $m_x(\delta) \le L\delta^\alpha$

WEAK CONVERGENCE ESSENTIALS

Suppose $X, X^1, X^2, \ldots$ are random functions.

Theorem 7.5 If

$$\left(X^n_{t_1}, X^n_{t_2}, \ldots, X^n_{t_k}\right) \Rightarrow \left(X_{t_1}, X_{t_2}, \ldots, X_{t_k}\right)$$

$\forall\, t_1, t_2, \ldots, t_k \in [0,1]$

(FD)

and

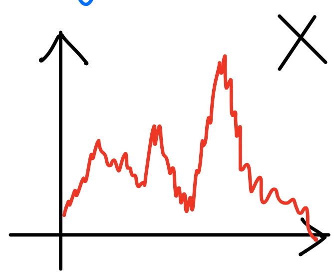
$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\left(m\left(X^n, \delta\right) > \varepsilon\right) = 0,$$

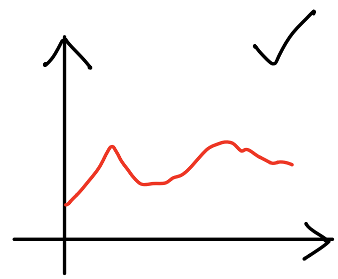
(KC)

then $X^n \Rightarrow X$.

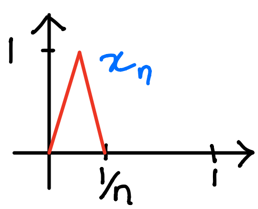
WEAK CONVERGENCE ESSENTIALS

(KC) means random function changes can be "controlled".





Ex. 1 (FD) is not enough:



$x_n \not\to x = 0$ and hence $\delta_{x_n} \not\Rightarrow \delta_x$.

Ex. 2 (FD) is enough: $S = \mathbb{R}^\infty$

WEAK CONVERGENCE ESSENTIALS

(Proof Sketch of 7.5)

- (FD) implies $P_n \pi_0^{-1}$ is tight.
- $\{P_n \pi_0^{-1}\}$ tight + KC $\iff$ $\{P_n\}$ tight.
- FD + $\{P_n\}$ tight is sufficient.

WEAK CONVERGENCE ESSENTIALS

(KC) looks "artificial."

Is it too much?

In short, because we are in $C[0,1]$, the answer is "No."

(KC) is fundamental and controls "richness" to just right extent.

WIENER MEASURE

Wiener measure, W, is a probability measure on $(C, \mathbf{C})$ having two properties.

(A) $W[x_t \leq \alpha] = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\alpha} \exp\left\{-\frac{1}{2}\frac{u^2}{t}\right\} du$

(B) for $0 \leq t_0 \leq t_1 \leq \cdots \leq t_k = 1$,

$x_{t_1} - x_{t_0},\ x_{t_2} - x_{t_1},\ \ldots,\ x_{t_k} - x_{t_{k-1}}$

are <u>independent</u> under W.

WIENER MEASURE

Two Crucial Things.

1. $W\left[x_{t_i} - x_{t_{i-1}} \le \alpha_i \,,\ i = 1, 2, \ldots, k\right]$

$$= \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \int_{-\infty}^{\alpha_i} \exp\left\{-\frac{1}{2}\frac{u^2}{t_i - t_{i-1}}\right\}$$

Thus, the finite-dimensional distributions are specified.

$$\left(f_{x_{t_1}, x_{t_2}, \ldots, x_{t_k}}(z_1, z_2, \ldots, z_k) = f_{x_{t_1}}(z_1)\, f_{x_{t_2} - x_{t_1}}(z_2 - z_1) \cdots f_{x_{t_k} - x_{t_{k-1}}}(z_k - z_{k-1}). \right)$$

2. The existence of W needs to be proved: at most one, sometimes none.

DONSKER'S THEOREM

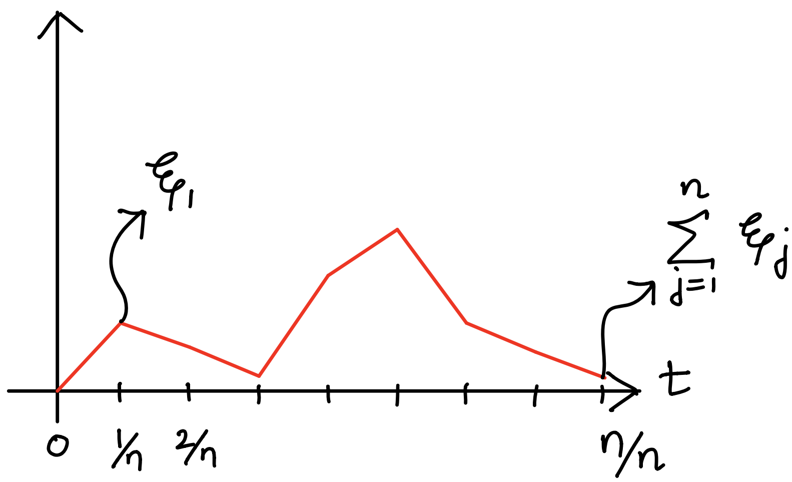
Lets assume W exists, and construct a sequence $\{X^n\}$ such that

$$X^n \Rightarrow W.$$

Let $\xi_1, \xi_2, \dots$ be i.i.d random variables such that $\mathbb{E}[\xi_1] = 0$ and $\mathrm{Var}(\xi_1) = \sigma^2 \in (0, \infty)$.

$$S_n = \begin{cases} 0 & n = 0 \\ \xi_1 + \xi_2 + \cdots + \xi_n, & n \geq 1 \end{cases}$$

DONSKER'S THEOREM



$$X_t^n(\omega) = \frac{1}{\sigma\sqrt{n}}\left(S_{\lfloor nt \rfloor}(\omega) + (nt - \lfloor nt \rfloor)\,\xi_{\lfloor nt \rfloor + 1}(\omega)\right)$$

$t \in [0, 1]$

Notice that $X^n \in C[0,1]$

Donsker's Theorem

Due to Theorem 7.5, we would have proved

$$X^n \Rightarrow W$$

if we can show (FD) and (KC) hold.

DONSKER'S THEOREM

Fix t :

$$X^n_t(\omega) = \underbrace{\frac{1}{\sigma\sqrt{n}} S_{\lfloor nt\rfloor}(\omega)}_{\substack{\Downarrow \\ \sqrt{t}\, N}} + \underbrace{\frac{(nt-\lfloor nt\rfloor)}{\sigma\sqrt{n}}\,\xi_{\lfloor nt\rfloor+1}(\omega)}_{\substack{\downarrow P \\ 0}}$$

(Lindeberg–Levy CLT) (by Chebyshev)

Fix s, t with $s \le t$:

$$\left(X^n_s(\omega),\ X^n_t(\omega) - X^n_s(\omega)\right)$$

$$= \frac{1}{\sigma\sqrt{n}}\left(S_{\lfloor ns\rfloor}(\omega),\ S_{\lfloor nt\rfloor}(\omega) - S_{\lfloor ns\rfloor}(\omega)\right) + o_p(1)$$

$$\Longrightarrow \left(\sqrt{s}\, N_1,\ \sqrt{t-s}\, N_2\right)$$

DONSKER'S THEOREM

Hence,

$$\left(X^n_s(\omega),\ X^n_t(\omega)\right) = \left(X^n_s(\omega),\ X^n_s(\omega) + X^n_t(\omega) - X^n_s(\omega)\right)$$

$$\Rightarrow \left(\sqrt{s}\, N_1,\ \sqrt{s}\, N_1 + \sqrt{t-s}\, N_2\right)$$

(by mapping thm.)

Fix $t_1, t_2, \ldots t_k$

$$\left(X^n_{t_1}(\omega),\ X^n_{t_2}(\omega) - X^n_{t_1}(\omega),\ \ldots,\ X^n_{t_k}(\omega) - X^n_{t_{k-1}}(\omega)\right)$$

$$\Rightarrow \left(\sqrt{t_1}\, N_1,\ \sqrt{t_2 - t_1}\, N_2,\ \ldots \sqrt{t_k - t_{k-1}}\, N_k\right)$$

This proves (FD).

DONSKER'S THEOREM

We will not prove (KC)... sorry but see pp. 88–90 in B99.

Theorem 8.2

If $\xi_1, \xi_2, \ldots$ are i.i.d with mean zero and variance $\sigma^2 \in (0, \infty)$, then

$$X^n \Rightarrow W.$$

DONSKER'S THEOREM

What have we skipped?

1. Existence of W
2. Proof that $X^n, n \geq 1$ satisfies (KC).

See pp. 88–90 in B99 for both.

WEAK CONVERGENCE ESSENTIALS

Is (KC) in Theorem 7.5 too much?

(Prohorov's Thm., pp. 58)

FD + RC $\Longleftrightarrow$ Weak Convergence

And,

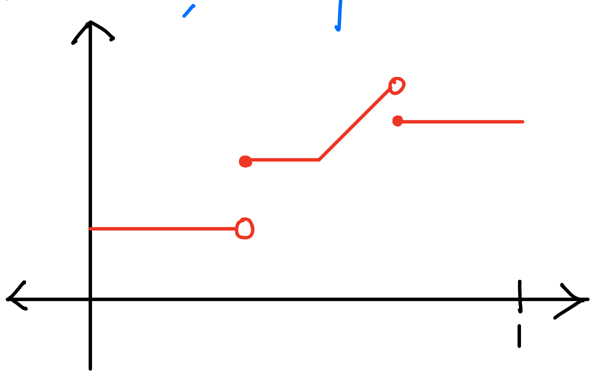
Under separability and completeness

FD + RC $\Longleftrightarrow$ FD + KC.

WEAK CONVERGENCE ESSENTIALS

Lets go to $D[0,1]$.

Recall that $D[0,1]$ is the class of cadlag functions, that is, right continuous functions with left limits, defined on $[0,1]$.



EMPIRICAL PROCESSES

AGENDA

(I) Step back . . .

(II) Generalization setup.

(III) Entropy & Examples.

(IV) ULLN (?)

I

Suppose $X_1, X_2, \ldots$ are i.i.d <u>real-valued</u> random variables with cdf F.

empirical cdf.

$$F_n(x) := n^{-1} \sum_{j=1}^{n} I_{(-\infty, x]}(X_j), \quad x \in \mathbb{R}$$

$$Z_n(x) := \sqrt{n}\left(F_n(x) - F(x)\right), \quad x \in \mathbb{R}$$

empirical process (indexed by x)

(I)

(Glivenko–Cantelli, 1933)

$$\|F_n - F\|_\infty := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

(Donsker, 1952)

std. Brownian Bridge on $[0,1]$.

$$Z_n \Rightarrow Z \equiv \{B(x),\ x \in [0,1]\}$$

in $D(\mathbb{R}, \|\cdot\|_\infty)$.

I

What are the analogues when $X_1, X_2, \ldots$ lie in a more general space $\mathcal{X}$, e.g., $C[0,1]$, or Riemanian manifold?

How to define $F(x) = P(X \leq x)$ when $X, x \in \mathcal{X}$?

Natural Idea : I

$$F_n(x) = P_n((-\infty, x])$$
$$F(x) = P((-\infty, x])$$

The above suggests considering $P_n(A)$ and $P(A)$ for an appropriate class of sets $\mathcal{C}$, $\mathcal{C} \ni A \subseteq \mathcal{X}$.

$\boxed{\text{Now ask}}$ $\sup_{A \in \mathcal{C}} |P_n(A) - P(A)| \xrightarrow{a.s.} 0$?

I

More conveniently, we can ask

$$\sup_{f \in \mathcal{F}} \left| P_n f - Pf \right| \xrightarrow{a.s.} 0 \,?$$

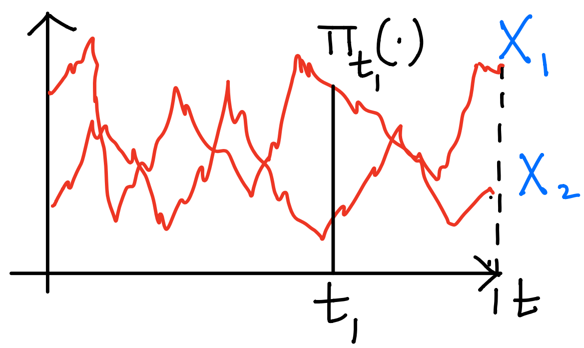
where $\mathcal{F}$ is a class of real-valued functions having domain $\mathcal{X}$.

$$\left(P_n f = n^{-1} \sum_{j=1}^{n} f(X_j) \,; \; Pf = \int f \, dP \right).$$

Two Examples.

I

(I)



$$\mathcal{X} \equiv C[0,1]$$

$\mathcal{F} :=$ "projection functions"

$$\left\{ \pi_t(x) := x(t) : t \in [0,1] \right\}$$

Two Comments (1)

(II) Suppose $\mathcal{X} \equiv \mathbb{R}$.

Then

$$\mathcal{F} := \text{"indicator functions"}$$

$$= \left\{ \mathbb{I}_{(-\infty, x]}(\cdot),\ x \in \mathbb{R} \right\}$$

I

The "richness" or "complexity" of $\mathcal{F}$ will determine the existence and nature of the GC and Donsker analogues.

II

Lets Setup...

$\boxed{\text{II}}$

$X_1, X_2, \ldots \sim P$ are i.i.d in $(\mathcal{X}, \mathcal{A})$

e.g.: $\mathcal{X} = \mathbb{R}^d$, $C[0,1]$, etc.

The Empirical Measure

$$P_n := n^{-1} \sum_{j=1}^{n} \delta_{X_j}$$

$$P_n(A) = n^{-1} \sum_{j=1}^{n} \mathbb{1}_A(X_j)$$

II

$\boxed{P_n f, \; Pf}$

For $f : X \longrightarrow \mathbb{R}$

$$Pf := \int f \, dP$$

$$P_n f := \int f \, dP_n = n^{-1} \sum_{j=1}^{n} f(X_i)$$

II

Empirical Process

Suppose $\mathcal{F}$ is a class of real-valued functions defined on $\mathcal{X}$.

$$\beta_n(f) := \sqrt{n}\left(P_n f - P f\right)$$

The stochastic process $\{\beta_n(f), f \in \mathcal{F}\}$ is called an empirical process.

II

We hope to identify <u>sufficient conditions</u> on $\mathcal{F}$ so that:

(Glivenko-Cantelli Analogue)

$$\|P_n - P\|_\infty^* = \Big(\sup_{f\in\mathcal{F}} |P_n f - Pf|\Big)^* \xrightarrow{a.s.} 0$$

$$\{x : \mathcal{F} \to \mathbb{R}, \|x\|_\infty < \infty\}$$

(Donsker Analogue)

$$\sqrt{n}\,(P_n - P) \Rightarrow G \text{ in } \ell^\infty(\mathcal{F})$$

Where G is a P-Brownian bridge.

II

The sufficient conditions will be phrased in terms of some notion of the "complexity" of $\mathcal{F}$.

So, we now set up toward entropy of $\mathcal{F}$.

II

$L_r(Q)$ norm

Q is a measure on $(\mathcal{X}, \mathcal{A})$.

For $1 \le r < \infty$,

$$\|f\|_{r,Q} := \left(\int |f|^r \, dQ \right)^{1/r}.$$

$\|f_1 - f_2\|_{r,Q}$: $L_r(Q)$ distance between f_1, f_2.

ε-cover

II

$\{f_1, f_2, \ldots, f_n\}$ is said to be an ε-cover for $\mathcal{F}$ if for any $f \in \mathcal{F}$, $\exists f_j$ such that

$$\|f - f_j\|_{r,Q} \leq \varepsilon.$$

($f_1, f_2, \ldots f_n$ need not live in $\mathcal{F}$)

II

ε-covering number.

$$N_r(\mathcal{F}, Q, \varepsilon) := \inf \left\{ n : \begin{array}{l} \exists \text{ an } L_r(Q)\ \varepsilon\text{-cover} \\ \{f_1, f_2, \ldots, f_n\} \text{ of } \mathcal{F} \end{array} \right\}$$

ε-entropy (or metric entropy)

$$H_r(\mathcal{F}, Q, \varepsilon) := \log N_r(\mathcal{F}, Q, \varepsilon)$$

II

ε-cover with bracketing.

$\left\{[f_j^L, f_j^U]_{j=1}^n\right\}$ is said to be an ε-cover with bracketing for $\mathcal{F}$ if for each $f \in \mathcal{F}$:

$$\| f_j^U - f_j^L \|_{r,Q} \leq \varepsilon \quad \forall j.$$

$$\exists j \text{ s.t } f_j^L \leq f \leq f_j^U$$

II

ε-covering with bracketing

$$N_{r,B}(\mathcal{F}, Q, \varepsilon)$$

$$:= \inf \left\{ n : \begin{array}{l} \exists \text{ an } \varepsilon\text{-cover with brac.} \\ \left\{ [f_j^L, f_j^U] \right\}_{j=1}^{n} \end{array} \right\}$$

ε-entropy with bracketing

$$H_{r,B}(\mathcal{F}, Q, \varepsilon) := \log N_{r,B}(\mathcal{F}, Q, \varepsilon)$$

II

ε-entropy for the sup. norm

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

$$H_\infty(\mathcal{F}, \varepsilon) := \log N_\infty(\mathcal{F}, \varepsilon)$$

where

$$N_\infty(\mathcal{F}, \varepsilon) :=$$

$$\inf \left\{ n : \begin{array}{l} \exists \{f_1, f_2, \dots, f_n\} \text{ s.t.} \\ \sup_{f \in \mathcal{F}} \min_{1 \le j \le n} \|f - f_j\|_\infty \le \varepsilon \end{array} \right\}$$

Notice: no dependence on Q

Comments

II

- Q need not be a prob. measure
- $\|f\|_{p,Q} \uparrow p$ but

$$\lim_{p\to\infty} \|f\|_{p,Q} \neq \underbrace{\|f\|_\infty}_{Q\text{-indep.}} = \sup_{x\in\mathcal{X}} |f(x)|$$

- $H_p(\mathcal{F}, Q, \varepsilon) \leq H_{p,B}(\mathcal{F}, Q, \varepsilon) \ \forall \varepsilon > 0$

$H_{p,B}(\mathcal{F}, Q, \varepsilon) \leq H_\infty(\mathcal{F}, \frac{\varepsilon}{2})$ if Q is a prob. measure.

(III)

Entropy calculation examples.

Example 1 (finite support)

$$\mathcal{F} := \left\{ \begin{array}{c} \text{increasing functions} \\ f : X \subset \mathbb{R} \to [0,1] \text{ where} \\ |X| = n < \infty . \end{array} \right\}.$$

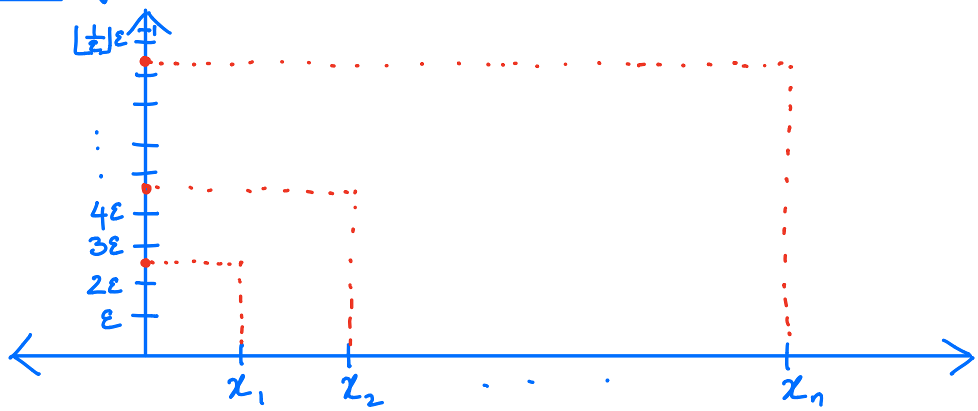
Then:

$$H_\infty(\mathcal{F}, \varepsilon) \leq \frac{1}{\varepsilon} \log\left(n + \frac{1}{\varepsilon}\right)$$

$\forall \varepsilon > 0.$

Proof Sketch

III



Suppose $X = \{x_1, x_2, \dots, x_n\}$

$$\bar{\mathcal{F}} := \left\{ \begin{array}{l} \text{non-decreasing functions } \bar{f} \\ \text{on } X \text{ such that:} \\ \bar{f}(x_j) \in \{ i\varepsilon, 0 \le i \le \lfloor 1/\varepsilon \rfloor \} \end{array} \right\}$$

$\bar{\mathcal{F}}$ is an ε-net of $\mathcal{F}$. Also:

$$|\bar{\mathcal{F}}| = \binom{n + \lfloor 1/\varepsilon \rfloor}{\lfloor 1/\varepsilon \rfloor}$$

$\boxed{\text{no. of non-negative int. solns to } y_1 + y_2 + \cdots + y_k = m \text{ is } \binom{m+k-1}{k}}$

III

<u>Example 2</u> (bounded derivatives)

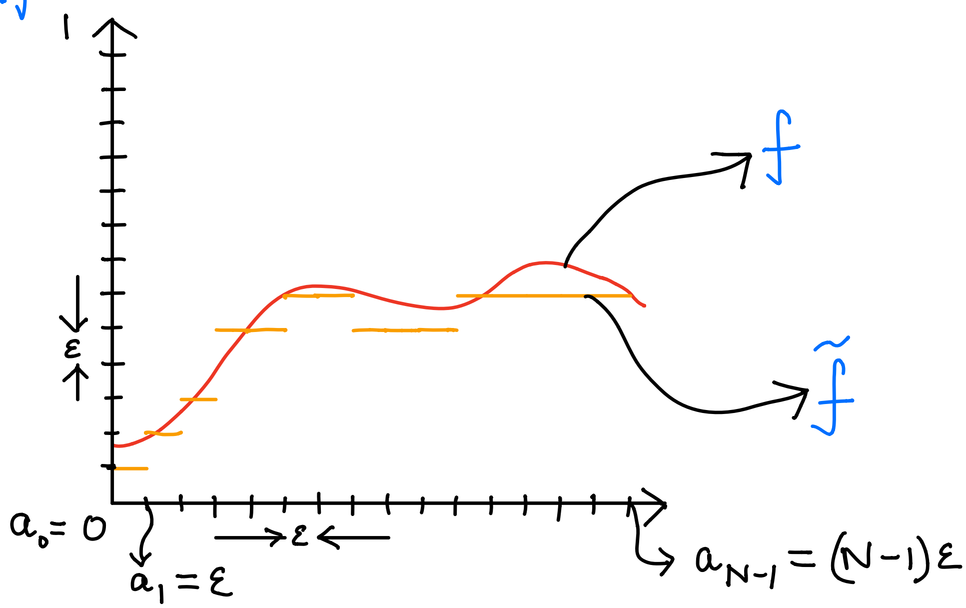
$$\mathcal{F} := \left\{ f : [0,1] \longrightarrow [0,1],\ |f'| \leq 1 \right\}.$$

Then, for some constant $A < \infty$,

$$H_\infty(\mathcal{F}, \varepsilon) \leq A \frac{1}{\varepsilon} \quad \forall\, \varepsilon > 0.$$

III

Proof Sketch.



$$\sup_{0 \le x \le 1} \left| f(x) - \tilde{f}(x) \right| \le 2\varepsilon .$$

$$\left| \tilde{f}(a_k) - \tilde{f}(a_{k-1}) \right| \le 3\varepsilon , \quad k = 1, 2, \ldots$$

$$N_\infty(\mathcal{F}, \varepsilon) \le \left(\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 \right) 7^{\lfloor 1/\varepsilon \rfloor}$$

III

Example 3 (Finite Dimensional Space)

Suppose $\psi_1, \psi_2, \ldots, \psi_d \in L_2(Q)$.

and

$$\mathcal{F} := \left\{ f = \sum_{k=1}^{d} \theta_k \psi_k : \theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^d \text{ and } \|f\|_{2,Q} \leq R \right\}$$

Then,

$$H_2(\mathcal{F}, \varepsilon, Q) \leq d \log\left(\frac{4R+\varepsilon}{\varepsilon}\right)$$

III

Example 3 (Finite Dimensional Space)

Suppose $\psi_1, \psi_2, \ldots, \psi_d \in L_2(Q)$.

and

$$\mathcal{F} := \left\{ f = \sum_{k=1}^{d} \theta_k \psi_k : \theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^d \text{ and } \|f\|_{2,Q} \leq R \right\}$$

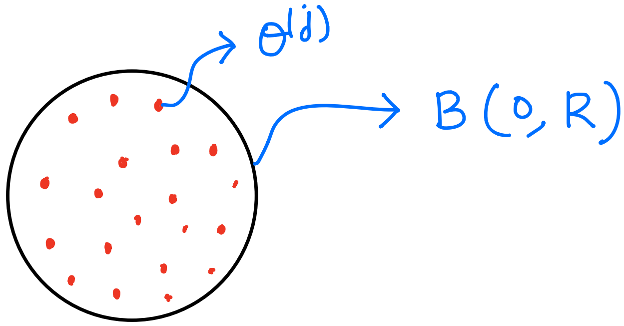
Then,

$$H_2(\mathcal{F}, \varepsilon, Q) \leq d \log\left(\frac{4R^2 + \varepsilon}{\varepsilon}\right)$$

Proof Sketch

III

Find balls $B\left(\theta^{(j)}, \frac{\varepsilon}{R}\right)$, $j = 1, 2, \ldots, N$ such that the $B(0,R)$ ball in $\mathbb{R}^d$ with Euclidean metric is covered.



$$\left\{ \tilde{f} : \tilde{f} = \sum_{k=1}^{d} \theta_k^{(j)} \psi_k , j = 1, 2, \ldots N \right\}$$

And,

$$N \leq \left(\frac{4R + \varepsilon/R}{\varepsilon/R} \right)^d .$$

III

Other Examples (Birman & Solomjak, 1967)

1. $\mathcal{F} := \left\{ f : \mathbb{R} \to [0,1], \underbrace{TV(f)}_{\int_X |df|} \leq 1 \right\}$

then, $\exists$ A s.t.

$$H_B(\mathcal{F}, \varepsilon, Q) \leq \frac{A}{\varepsilon} \quad \forall \varepsilon > 0.$$

2. $\mathcal{F} := \left\{ f : [0,1] \to [0,1], \int \left(f^{(m)}\right)^2 \leq 1 \right\}$

then, $\exists$ A s.t.

$$H_B(\mathcal{F}, \varepsilon, Q) \leq \frac{A}{\varepsilon^{1/m}} \quad \forall \varepsilon > 0.$$

IV

The Basic ULLN

Suppose $\mathcal{F}$ is such that

$$H_{1,B}(\mathcal{F}, \varepsilon, P) < \infty \quad \forall \varepsilon > 0.$$

Then

$$\| P_n - P \|_\infty = \sup_{f \in \mathcal{F}} | P_n f - Pf | \xrightarrow{a.s.} 0.$$

Proof Sketch

IV

We can find $\left\{ [f_j^L, f_j^U]_{j=1}^N \right\}$ so that for each $f \in \mathcal{F}$, we find i s.t

$$P_n f - Pf \leq (P_n - P) f_j^U + \varepsilon$$

$$P_n f - Pf \geq (P_n - P) f_j^L - \varepsilon \quad \text{--- (1)}$$

Since $N < \infty$, for large n,

$$\max_{1 \leq j \leq N} \left| (P_n - P) f_j^U \right| \leq \varepsilon$$

$$\max_{1 \leq j \leq N} \left| (P_n - P) f_j^L \right| \leq \varepsilon \quad \text{--- (2)}$$

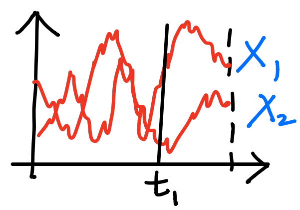
Use (1) and (2).

AGENDA

I. Strengthen Basic ULLN.

Suppose $\int \sup_{f \in \mathcal{F}} \|f\|_\infty \, dP < \infty$ and

$$\frac{1}{n} H_1(\mathcal{F}, \varepsilon, P_n) \xrightarrow{P} 0 \quad \forall \varepsilon > 0.$$



Then,

$$\|P_n - P\|_\infty := \sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{a.s.} 0$$

II. Example Classes (VC)

TWO KEY MACHINERY

~~(i)~~ Chaining

(ii) Symmetrization

SYMMETRIZATION.

means approximating

$$\| P_n - P \|_\infty \text{ using } \| P_n - P_n' \|_\infty$$

where:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} P$$

$$X_1', X_2', \ldots, X_n' \overset{iid}{\sim} P$$

$$(X_1, X_2, \ldots, X_n) \overset{ind}{\sim} (X_1', X_2', \ldots, X_n')$$

SYMMETRIZATION LEMMA

Suppose that $\forall f \in \mathcal{F}$, $\delta > 0$

$$P\left(\left| P_n f - Pf \right| > \delta/2 \right) \leq \frac{1}{2} .$$

Then,

$$P\left(\| P_n - P \|_\infty > \delta \right)$$

$$\leq 2 P\left(\| P_n - P_n' \|_\infty > \delta/2 \right)$$

Proof Sketch.

$$P\left(\sup_{f\in\mathcal{F}} \left|\int f\, d(P_n - P)\right| > \delta\right)$$

$$\leq \quad P\left(\left|\int f^* d(P_n - P)\right| > \delta\right)$$

$$\leq 2P\left(\left|\int f^* d(P_n - P)\right| > \delta \ \cap \ \left|\int f^* d(P_n' - P)\right| \leq \frac{\delta}{2}\right)$$

$$\leq 2P\left(\left|\int f^* d(P_n - P_n')\right| > \delta/2\right)$$

$$\leq 2P\left(\sup_{f\in\mathcal{F}} \left|\int f\, d(P_n - P_n')\right| > \frac{\delta}{2}\right)$$

Why symmetrization?

$$(W_1, W_2, \ldots, W_n) \underset{\sim}{\text{ind}} (X_1, X_2, \ldots X_n, X_1', \ldots X_n')$$

(Rademacher sequence)

$$W_j \underset{\sim}{\text{iid}} \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$$

Then, for each $f \in \mathcal{F}$,

$$\left\{ f(X_i) - f(X_i') : i = 1, 2, \ldots, n \right\}$$

$$\stackrel{d}{=} \left\{ W_i\big(f(X_i) - f(X_i')\big) : i = 1, 2, \ldots, n \right\}$$

Okay. So what?

Notice:

$$P\left(\|P_n - P_n'\|_\infty > \delta/2\right)$$

$$\stackrel{\text{def.}}{=} P\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(X_i')\right| > \frac{\delta}{2}\right)$$

$$= P\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} W_i\left(f(X_i) - f(X_i')\right)\right| > \frac{\delta}{2}\right)$$

$$\leqslant 2\, P\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} W_i\, f(X_i)\right| > \frac{\delta}{4}\right).$$

The tail probability

$$P\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}W_i\, f(X_i)\right| > \frac{\delta}{4}\right)$$

is easy to work with.

Okay, so what?

Suppose that we show:

$$P\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}W_i\, f(X_i)\right| > \frac{\delta}{4}\right) \xrightarrow{P} 0. \qquad \text{———} (1)$$

Then, $\|P_n - P\|_\infty \xrightarrow{P} 0.$ (symmetrization)

$$\|P_n - P\|_\infty \xrightarrow{a.s.} \text{something.}$$

(Pollard, 1984)

Therefore, $\|P_n - P\|_\infty \xrightarrow{a.s.} 0.$

Let's prove (1) for bounded r.v.s.

Assume $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq R < \infty$.

We will choose $A_n \subseteq \mathcal{X}^n$ s.t.

$$P\Big(\underbrace{\sup_{f \in \mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\Big| > \frac{\delta}{4}}_{E_n}\Big)$$

$$\leq P\Big(E_n \,\Big|\, (X_1, X_2, \dots X_n) \in A_n\Big) \times 1$$

$$+ 1 \times P\Big((X_1, X_2, \dots X_n) \in A_n^c\Big)$$

$$\longrightarrow 0.$$

Choose:

$$\boxed{A_n = \left\{ \sqrt{n}\,\frac{\delta}{8} \geq C\left(R\, H_2^{1/2}\left(\mathcal{F}, \frac{\delta}{32}, P_n\right) \vee R\right)\right\}}.$$

With some work, and Hoeffding (1963)

$$P\Big(E_n \,\Big|\, (X_1, X_2, \ldots X_n) \in A_n\Big) \leq C \exp\left\{-\frac{n\delta^2}{64 C^2 R^2}\right\}.$$

And:

$$P(A_n^c) \to 0 \text{ if } \frac{1}{n} H_2(\mathcal{F}, \delta, P_n) \xrightarrow{P} 0. \quad \forall \delta > 0.$$

Key Lemma

"envelope"

Suppose $F := \sup_{f \in \mathcal{F}} \|f\|_\infty \leq R < \infty$,

and that

$$\frac{1}{n} H_2(\mathcal{F}, \delta, P_n) \xrightarrow{P} 0, \quad \forall \delta > 0.$$

Then,

$$\|P_n - P\|_\infty \xrightarrow{a.s.} 0.$$

7.

Theorem.

Suppose $\int F \, dP < \infty$,

and suppose

$$\frac{1}{n} H_1(\mathcal{F}, \delta, P_n) \xrightarrow{P} 0, \quad \forall \delta > 0.$$

— (Ent)

Then,

$$\| P_n - P \|_\infty \xrightarrow{a.s.} 0.$$

7.

Proof Sketch. Choose R so that

$$\|P_n - P\|_\infty := \sup_{f \in \mathcal{F}} |P_n f - P f| .$$

$$\leq \sup_{f \in \mathcal{F}} \left| \int_{F \leq R} f \, d(P_n - P) \right| + \underbrace{\int_{F > R} F \, dP_n}_{\leq 2\delta \text{ a.s. for large } n.} + \underbrace{\int_{F > R} F \, dP}_{\leq \delta \text{ for large } n}$$

<u>Proof</u> <u>Sketch</u>. contd...

Let's deal with the first term.

Define the truncated class.

$$\mathcal{F}_R := \left\{ f\, \mathbb{I}\{f \leq R\} : f \in \mathcal{F} \right\}$$

For $f_1, f_2 \in \mathcal{F}$,

$$\int (f_1 - f_2)^2 \, dP_n \leq 2R \int |f_1 - f_2| \, dP_n$$

Proof Sketch. contd...

Under (Ent), this means

$$\frac{1}{n} H_2(\mathcal{F}_R, \delta, P_n) \xrightarrow{P} 0 \ \forall \delta > 0.$$

Therefore, Glivenko-Cantelli holds on $\mathcal{F}_R$.

⑦

Which $\mathcal{F}$s satisfy the condition

$$\frac{1}{n} H_1(\mathcal{F}, \delta, P_n) \xrightarrow{P} 0 \; ?$$

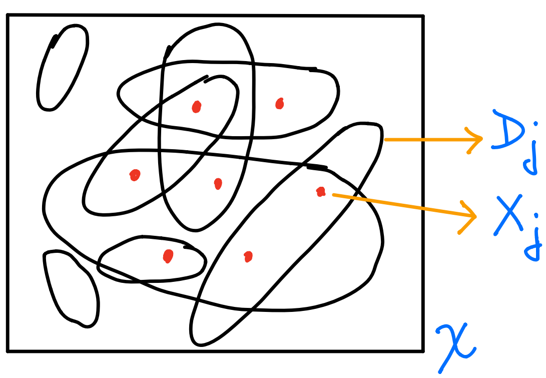
Many useful classes $\mathcal{F}$ satisfy this property. Let's see one such.

Vapnik - Chervonenkis (VC) Subgraph Classes.

$\mathcal{D}$:= collection of subsets of $\mathcal{X}$

(domain of $f \in \mathcal{F}$)

$$\Delta^{\mathcal{D}}(X_1, X_2, \ldots X_n) = \text{card}\left\{ D \cap \{X_1, X_2, \ldots, X_n\} : D \in \mathcal{D} \right\}$$



$$\Delta^{\mathcal{D}}(X_1, X_2, \ldots X_7) = 6.$$

$$m^{\mathcal{D}}(n) := \sup\left\{ \Delta^{\mathcal{D}}(x_1, x_2, \dots x_n) : (x_1, x_2, \dots x_n) \in \mathcal{X} \right\}.$$

$$V(\mathcal{D}) := \inf\left\{ n \geq 1 : m^{\mathcal{D}}(n) < 2^n \right\}$$

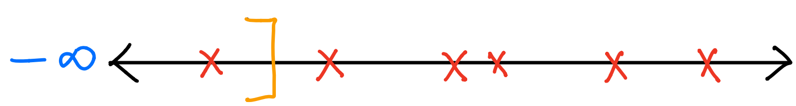
- $V(\mathcal{D})$ is called the index of the class $\mathcal{D}$.
- $\mathcal{D}$ is called a VC-class if $V(\mathcal{D}) < \infty$.

Example 1

$$\mathcal{D} := \left\{ (-\infty, t],\ t \in \mathbb{R} \right\}$$

Then,

$m^{\mathcal{D}}(n) = n+1$, so that $\mathcal{D}$ is a VC-class.



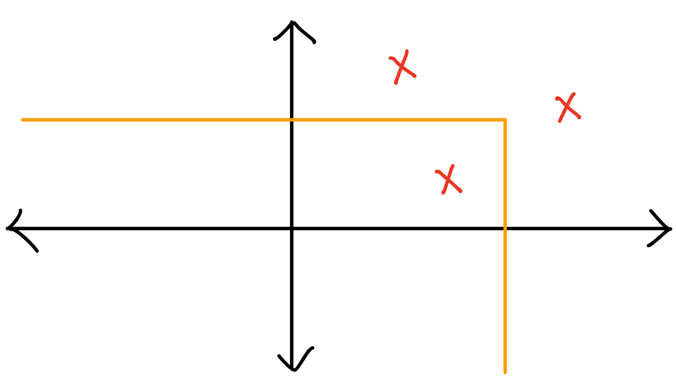
Example 2

$$\mathcal{D} := \left\{ (-\infty, t],\ t \in \mathbb{R}^d \right\}$$

Then,

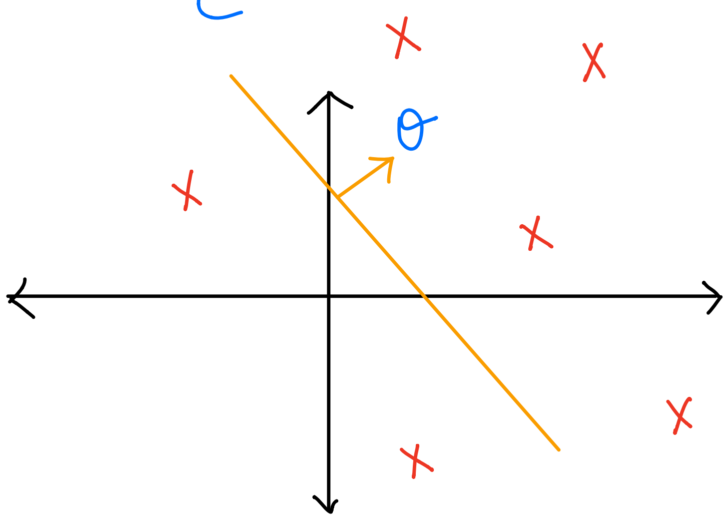
$$m^{\mathcal{D}}(n) = (n+1)^d, \quad \text{so that}$$

$\mathcal{D}$ is a VC-class.



Example 3

$$\mathcal{D} := \{ \{ x : \langle x, \theta \rangle > y \} : \theta \in \mathbb{R}^d, y \in \mathbb{R} \}.$$



$$m^{\mathcal{D}}(n) \leq 2^d \binom{n}{d}$$

$$V(\mathcal{D}) \leq d+2.$$