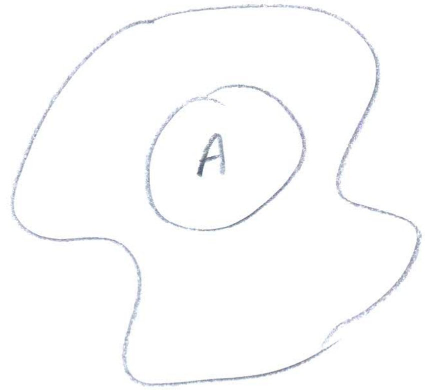
36-705
YouTube 12/09/2016

04/05/2020

- LW: HW question (?)
- LW: continue talking about uniform bounds.

- Suppose sample space:
- Draw r.v.



$P(A) = P(X \in A)$ (unknown prob)

$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in A)$ (empirical prob.)

- Q: How far is $P_n(A)$ from $P(A)$?
  - A: Hoeffding

$$P\left(|P_n(A) - P(A)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

- LW: Guaranteeing this is small one 'thing' at a time does not mean that it holds uniformly (ow a class/set)

- LW: Instead; we want:-

$$P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq \text{small}$$

- this is a uniform bound over set $\mathcal{A}$

- Ultimately; we will have a bound like Hoeffding; but with a complex term

2. Finite classes

- Let $\mathcal{A} = \{A_1, \dots, A_N\}$ - A finite set of events

notes give $P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq c_1 K(\mathcal{A}) e^{-c_2 n\epsilon^2}$

(I)

$$P\left(\sup_{?} |P_?(A_?) - P(A)| > \epsilon\right)$$

- Define B_j be the event that $B_j = \{|P_n(A_j) - P(A_j)| > \epsilon\}$
- event that probability of absolute difference of empirical and population probability of event A_j is greater than ϵ.
(true)

(*)(*)

LW: An alternative to (I) is ~~from~~ observing that it's identical to saying at least one of the events B_j is true.

(*) events B_1, B_2, B_3 are NOT disjoint. Hence $P(\bigcup B_j) \neq \sum_j P(B_j)$

- However we can use the union bound (LW)(Ai) - Recall that? - I don't recall (Q)

$$P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) = P\left(\bigcup_{j=1}^{N} B_j\right) \leq \sum_{j=1}^{N} P(B_j)$$

- Note: introduction of B_j yields $\sum_{j=1}^{N} P(B_j) = \sum_{j=1}^{N} P\left(|P_n(A_j) - P(A_j)| > \epsilon\right) = 2Ne^{-2n\epsilon^2}$

- Hence apply Hoeffding:-

$$P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2Ne^{-2n\epsilon^2}$$

Hoeffding (with extra term denoting no. of events B_j or A_j)

- LW: $2N$ is a 'measure' of complexity of the class/set of events $\mathcal{A} = \{A_1, \dots, A_N\}$ or $\mathcal{B} = \{B_1, \dots, B_N\}$.
  - yields simplest way to control probability (uniform bounds)

- LW: A limitation to this; what if class/set $\mathcal{A}$ is infinite i.e. infinite cardinality; bound then goes $\to \infty$

- LW: Need more sophisticated method than counting

Consider $Q = \{\text{discs on } \mathbb{R}^2\}$ - infinite no. of discs (how to improve on counting?)

$$Pr\left(\sup_{A} |P_n(A) - P(A)| > \epsilon\right)$$

LW: Invoke shattering/VC dimension

- Give me a class of sets Q

(class is possibly infinite set (of) sets)

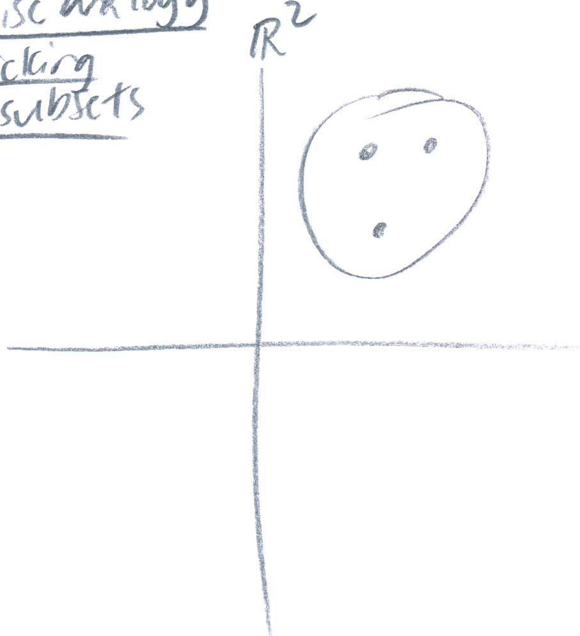
- Examples of infinite classes: $\mathcal{Q}$ or $\mathcal{A}$ (collection of sets)

1. $\mathcal{A} = \{[-\infty, t] : t \in \mathbb{R}\}$ - set of all half intervals on real-line (CDF)

LW: what is sample space?
- Let F be an arbitrary finite set $F = \{x_1, \ldots, x_n\}$
- LW: So we have an arbitrary finite set F; but also a class of sets $\mathcal{Q}$ (*)
- LW: Take any set from the $\mathcal{Q}$ and intersect it with our finite set.

$A \cap F = G$ and $G \subset F$

- use disc analogy for picking out subsets

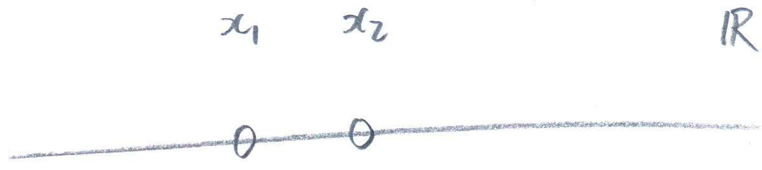


(*) we say $\mathcal{A}$ (collection) picks out G (subset) if $A \cap F = G$ for some $A \in \mathcal{A}$

(*) $S(\mathcal{A}, F)$ - no. of subsets that are picked out by $\mathcal{A}$.

simple example

$\mathcal{A} = \{[-\infty, t] \in \mathbb{R}\}$



- Finite set $F = [x_1, x_2]$
- Q How many possible sets can be extracted from F?
- possible subsets of F ((i.e. the Gs)) - $\emptyset, x_1, x_2, [x_1, x_2]$
- can we get these by considering infinite coll of sets $\mathcal{A}$?

- $S_n(\mathcal{A})$ is measure of complexity of class of sets $\mathcal{A}$

LW: This class of sets $\mathcal{A}$ is not as rich as we might think as $S_n(\mathcal{A}) \neq 2^n$ for many n.

- More complex sets $\longrightarrow S_n(\mathcal{A}) \simeq 2^n$ for many n.

- Vapnik-Chervonenkis (not proved) Theorem 5

$$P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8 S_n(\mathcal{A}) e^{\frac{-n\epsilon^2}{32}}$$

- "Extension of Hoeffding: Probability that maximum of absolute diff. between empirical prob. and probability over entire class of sets $\mathcal{A}$ is greater than ϵ is the probability bounded by a factor made up of an exponential term in n, ϵ and the shattering coefficient $S_n(\mathcal{A})$"

- $S_n(\mathcal{A})$ - shattering coefficient

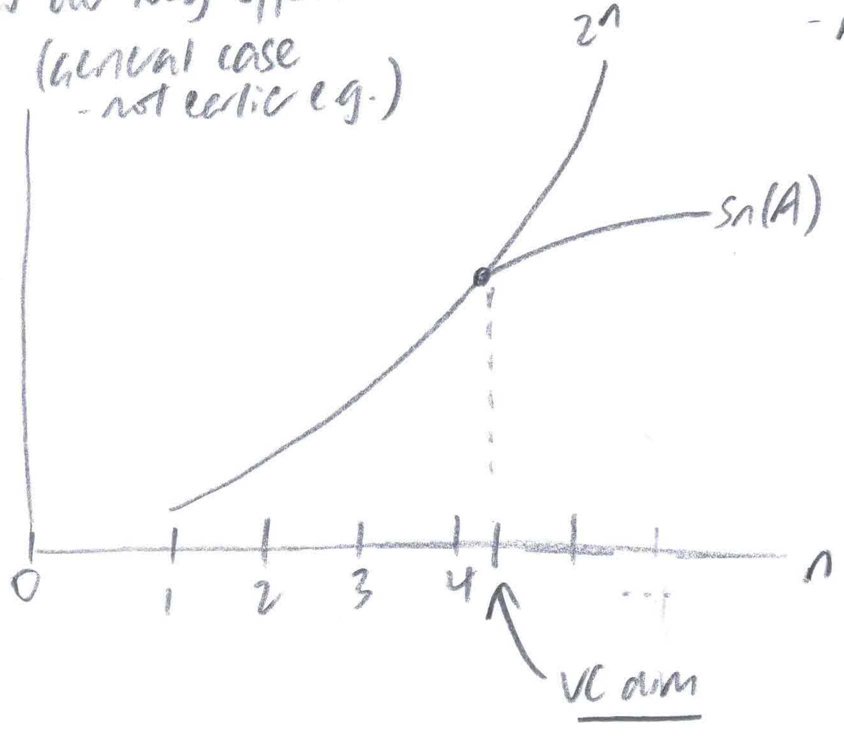
LW: - Bound is very tight over a complex class of sets $\mathcal{A}$ (?)

L14 (?)

- $S_n(\mathcal{A})$ is potentially
- complex ~~set~~ class of sets $S_n(\mathcal{A}) \longrightarrow$ poor bound (?)
- Consider rates of increase of $S_n(\mathcal{A})$ and $e^{\frac{-n\epsilon^2}{32}}$ separately in n;

Q: How do they offset each other?

LW:



- Assume we can plot / have algorithm for $S_n(\mathcal{A})$ as fn of n.
- Plot $f = (2^n)$ $g = (S_n(\mathcal{A}))$
- VC dimension

- VC dimension :-

dfd $d = d(A) =$ largest $n : s_n(A) = 2^n$

- d is the size of the largest set that can be shattered (i.e. $s(A, F) = 2^n$)

LW: Sauer's lemma (*). As $s_n(A)$ increases in n before $n =$ VC dimension $= d$; rate of increase is exponential

As $s_n(A)$ increases in n after $n =$ VC dim $= d$; rate of increase is polynomial.

Theorem 7: Suppose A has a finite VC dimension d. Then for all $n \geq d$:-

$$s(A, n) \leq (n+1)^d$$

(*) At this particular polynomial rate

- After VC dimension d :- i.e. $n \geq d$, $n \to \infty$

$$8 s_n(A) e^{\frac{-n\varepsilon^2}{32}}$$

decreases exponentially

increases at polynomial rate

- i.e. an extension of Hoeffding; with some add. info

(LW): How to compute VC dimension? (*) VC Theory & Sauer proved in 36-702 / 10-702

- Sometimes easy / trivial

Assume we have a VC dimension dictionary (for course purposes)

WA(S) - Try for IR - VC dim of 2. (see Table 1) (?)

- LW (*) :- Know VC dim → know shattering → $s_n(A)$ ...

(*) - simple probability bounds ⟶ Hoeffding

collections of sets ⟶ use VC Theory/inequality (extension of Hoeffding).

36-705 04/05/2020

YouTube 12/09/2016

- convergence theory

· LW: Answers question of what happens ~~when~~ we get more data (asymptotics)

- 2 key results - LLN, CLT

- Deal with sequence of r.vs. $X_1, \ldots, X_n \sim F$
- Concerned with statistics of sequence of r.vs. $T_n = g(X_1, \ldots, X_n)$ - a function of data
- E.g. sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ - (*) sequence of sample means (as r.vs. are NOT iid) ✓

(*) A sequence of statistics may not be IID, whilst ~~its~~ underlying r.v.s. are IID

---

more abstractly

original sequence Y_1, Y_2, Y_3 (likely IID)

$X_n = g_n(Y_1, \ldots, Y_n)$ - compute sth from first n data points

- New ~~(arbitrary)~~ sequence of statistics; can be viewed as arbitrary seq of r.vs. :- (use/have $X_i = \bar{X}_n$ for concrete example).

$X_1, X_2, X_3, X_4, \ldots$ where $X_i = g(Y_1, \ldots, Y_i)$

Q: What does it mean for sequence $X_1, X_2, \ldots$ to converge??

- More complex than calculus for r.v.s.
- Find many different types of convergence.

LW: Will be confusing → will take a few lectues

~~A1~~

Almost sure convergence (to a constant)

$X_n \xrightarrow{a.s.} c$

$$P\Big(\underbrace{\lim_{n\to\infty} X_n = c}_{\text{conv. calc. definitn}}\Big) = 1$$

$\underbrace{\qquad\qquad}_{\text{probabilistic conv. defn.}}$

Convergence in probab. (to an r.v./constant) (*) essential

$X_n \xrightarrow{p} X$

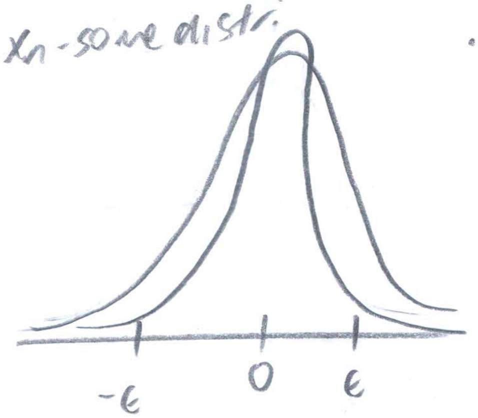
$\forall \epsilon > 0 \quad P(|X_n - X| > \epsilon) \to 0$

$X_n \xrightarrow{p} c$ (*) essential

$\forall \epsilon > 0 \quad P(|X_n - c| > \epsilon) \longrightarrow 0$

- As $n \to \infty$

LW: X_n - some distr.



· As n gets large, distr gets squished towards 0

- Already seen:-

$X_n - c = o_p(1)$

i.e.

LW: convergence in probability and almost sure convergence not the same

Convergence in quadratic mean

→ (W)(A1) E^2 or $E[\ ]^2$? ✓ $\mathbb{E}[(\ )^2]$

$X_n \xrightarrow{q.m.} X$: $\mathbb{E}[(X_n - X)^2] \longrightarrow 0$ - convergence in L_2

$X_n \xrightarrow{q.m} c$ $\quad \mathbb{E}[(X_n - c)^2] \longrightarrow 0$

Convergence in distri (*) - essential

LW: CDF of X_n → CDF of X

$X_n \rightsquigarrow X$ also $X_n \xrightarrow{d} X$ → $\lim_{n\to\infty} F_n(t) = F(t)$

(LW) - show by exampl.

$F_n(t) \to F(t)$ as $n \to \infty$; at all t for which F is continuous

(LW A2): Review and polish/check other definitions he has omitted

$$\text{a.s.} \Longrightarrow \overset{\circledast}{\text{prob}} \Longrightarrow \overset{\circledast}{\text{distribution}}$$
$$\Uparrow$$
$$\text{q.m.}$$

A taxonomy of convergence relationships

- LW: The ⊛ are what we are concerned with
  - a.s. → subtle (not used for our purposes)
  - q.m → stepping stone to proving prob.

LW: make clear through proofs, examples
- All above are general (no ass. about discreteness/continuous etc.)

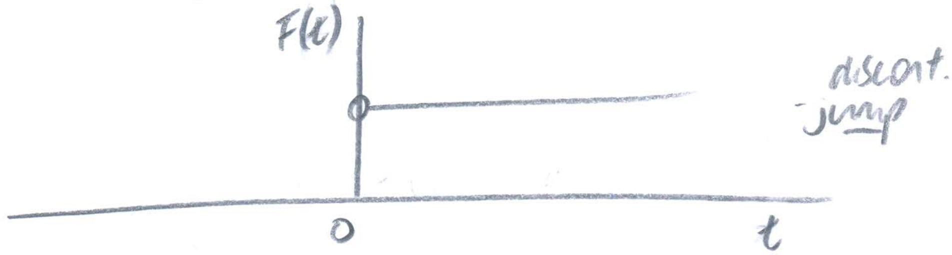
LW: Theorem 3: for curiosity

Example 4

- Consider following seq. of r.v.s. with $X_n \sim N(0, \frac{1}{n})$

  $X_1, X_2, \dots$

- Intuitively; concentrates around 0 as $n \to \infty$
- Consider r.v. whose value is always 0.



CDF $F(t) = \begin{cases} 0 & t<0 \\ 1 & t>0 \end{cases}$

· Note $\sqrt{n} X_n \sim N(0,1)$

- Showing $X_n \to 0$ $\Rightarrow$ show $F_n(t) \to F(t)$ as $n \to \infty$; ~~nota~~

$F_n(t) = P(X_n \leq t) \overset{(i)}{=} P(\sqrt{n} X_n \leq \sqrt{n} t) = P(Z \leq \sqrt{n} t)$

· consider two cases

· $t > 0$ $\quad P(Z \leq \sqrt{n}\,t) \to 1 = F(t)$

as $n \to \infty$

· $t < 0$ $\quad P(Z \leq \sqrt{n}\,t) \to 0 = F(t)$

as $n \to 0$

Q: How about at 0? ? W A4

i.e. $t = 0$

$F_n(0) = P(X_n \leq 0)$

$= \frac{1}{2}$

· F

· This does not converging

· convergence in distri only requires convergence at continuity points

· exclude discontinuity

$F_n(0) \neq F(\frac{1}{2}) = 1$

convergence fails

· we have proved :-

$X_n \xrightarrow{d} X$ where X is always equal to 0 i.e. $P(X=0) = 1$

$X_n \xrightarrow{d} 0$

In next lecture, more examples

formal proofs to develop intuition