36-705 - (L3) 27/04/2020

Youtube - 07-09-2016 (Asymptotics Lecture Notes 2

- $a_n = o(1)$ $a_n \to 0$
- $a_n = O(1)$ $\exists$ $|a_n| \leq C$ for all large $n$ (i.e. beyond some $n$; it is always the case that beyond an; $a_n$ is bounded by a finite const.)
- make a distinction between deterministic and probabilistic little-o and Big-O notation
- $a_n = o(b_n)$ i.e. divide both sides - same with $O(1)$
- $\frac{a_n}{b_n} = o(1)$

---

$a_n \sim b_n$ - both sequences are increasing at the same rate - (H) in CS.

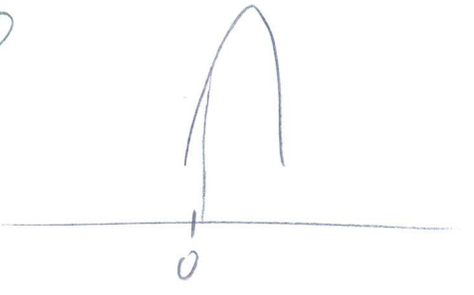
$\frac{a_n}{b_n} \sim \frac{b_n}{a_n}$ (?)

- distinction between statistics and compute science language
- 4:41 (AI) - check not.

---

Probabilistic versions:- (W): $o_p$ and $O_p$ have very specific probabilistic/analysis formalised definitions

- imagine a sequence of r.v.s $Y_1, Y_2, Y_3, \ldots Y_n$
- $Y_n$ is an arbitrary sequence of r.v.s ; but has some kind of feature. (not necessarily IID)
- To say $Y_n$ is $o_p(1)$ means that distribution (not r.v) is concentrated around 0.
- That is probability outside interval is going to 0




---

- $Y_n = o_p(1)$ means:- (*)

$$P(|Y_n| > \epsilon) \xrightarrow{n \to \infty} 0 \quad \forall \epsilon > 0$$

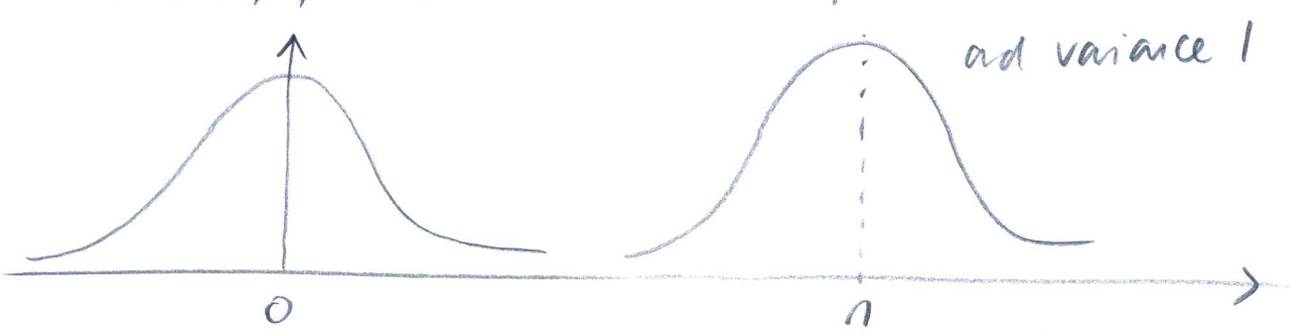
- LW: so if we drew a distribution - piling up near 0
- LW: make sure you understand distinction

---

$O_p(1)$ is trickier - we want to capture idea of boundedness, but what does this mean in probabilistic sense?

- LW: example of something not bounded probabilistically

- consider $Y_n = N(n, 1)$ - that is $n^{th}$ r.v. in sequence has mean $n$ and variance 1



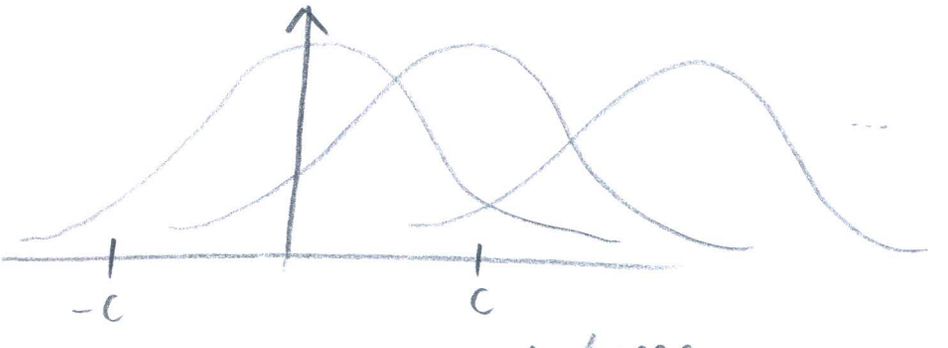

- As $n \to \infty$; of $n^{th}$ distribution, random variables still being drawn, but distribution as $n$ gets large 'flies off to $\infty$'
- more concretely, it means that probability mass is being shifted in $+\infty$ direction
- This is an example of an unbounded sequence of random variables
- Alternatively; if I pick an interval, and ask how much probability mass is trapped in a fixed interval; then no matter how big the interval, probability mass will 'escape' interval as $n \to \infty$.

---

- so the opposite of probabilistic unboundedness i.e. probabilistic boundedness heuristically means I can trap the probability mass in a large interval (as $n \to \infty$?)
- Definition: (you give me) analysis (and I can find) (9)(12) - analysis reasoning crucial

$Y_n = O_p(1)$ if $\forall \epsilon > 0 \quad \exists C : P(|Y_n| > C) \leq \epsilon$ for all large $n$ i.e. $n > n_0$

- Example (in context of above)
- You tell me you want $1 - \epsilon = 0.9$; can I find an interval $[-C, C]$ that traps 90% of probability
- In earlier example, no that is not possible as $n \to \infty$
- So for $Y_n = O_p(1)$ distribution of $Y_n$ can't move around too much, probabilistic notion of boundedness




distribution 'settling down'

- Bounded probability and going to 0 probability - 2 key notions

- $Y_n = o_p(a_n)$ $a_n$ - sequence of deterministic functions } little op

- $\frac{Y_n}{a_n} = o_p(1)$ - divide both sides by $a_n$

- $Y_n = O_p(a_n)$ } Big $O_p$

$\frac{Y_n}{a_n} = O_p(1)$ }

(W): will become 2nd nature

Examples of op and Op

- Define sequence of IID Bernoulli r.v.s $Y_1, \ldots, Y_n$ (coin flips)
- $Y_i \in \{0,1\}$ $p = P(Y_i = 1)$
- Claim: For $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $|\hat{p}_n - p| = O_p\left(\frac{1}{\sqrt{n}}\right) = o_p(1)$
- via Hoeffding: (applied to Bern.) (01)

$$P(|\hat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} \xrightarrow{n\to\infty} 0$$

(Note this is the definition of op)

- Note RHS $\to 0$ as $n \to \infty$, so we have already shown that

$\hat{p}_n - p = o_p(1)$ also written $\hat{p}_n \underset{(ii)}{=} p + o_p(1)$

(ii) Read as random variable $\hat{p}_n$ is equal to a constant plus a term $o_p(1)$ that approaches 0 as $n \to \infty$

- LW: Little $o_p$ is a little cruder; only tells us something is converging to 0
  Big $O_p$ have more info; tells us information about size of the deviation (inside brackets)

Proof $|\hat{p}_n - p| = O_p\left(\frac{1}{\sqrt{n}}\right)$

- Rewrite $|\hat{p}_n - p| = O_p\left(\frac{1}{\sqrt{n}}\right)$

- $\sqrt{n}(\hat{p}_n - p) = O_p(1)$ - bounded in probability
  i.e. I can 'trap' LHS

$$P(\sqrt{n}|\hat{p}_n - p| > C) = P\left(|\hat{p}_n - p| > \frac{C}{\sqrt{n}}\right) \leq 2e^{\frac{-2nC^2}{n}} = 2e^{-2C^2} \leq \epsilon \text{ for } C \text{ large}$$ (2)

- Analysis narrative:
- You give me an arbitrary $\epsilon$, I want to trap all the probability except $\epsilon$, so if I choose $C$ large enough (?) then I will succeed.
- If you give me $\epsilon$, can I find a $C$ such that $P(|Y_n| > C) \leq \epsilon$
- ∴ we have shown that:-

  $\sqrt{n}(\hat{p}_n - p) = O_p(1)$ or equivalently $(\hat{p}_n - p) = O_p\left(\frac{1}{\sqrt{n}}\right)$

- Intuition: for $\hat{p}$ as an r.v., $p$ as a constant parameter, they are typically around $\frac{1}{\sqrt{n}}$ far apart, typically (at most)
  - $o_p$ - strictly less than
  - $O_p$ - less than or equal to

---

- calculus of $o_p$ and $O_p$
- $O_p(1)\,o_p(1) = o_p(1)$

- semantics

If $X_n = O_p(1)$ and $Y_n = o_p(1)$, then if $Z_n = X_n Y_n$, then $Z_n = o_p(1)$

- proof strategy:
- Assume $X_n = O_p(1)$, $Y_n = o_p(1)$, define $Z_n = X_n Y_n$
- Prove $Z_n = o_p(1)$ by going back to the definitions
  i.e. what property does $Z_n$ have to have to be $o_p(1)$
  use property of $X_n$ and $Y_n$ assumed to show

(W): should be able to see intuitively whether this is true

(A3) - learning objective - master it but will get further practice

- Asymptotics has a small diversion
- Key result is Hoeffding, and extensions → in depth in 10/36-702

- Hoeffding's inequality, Markov, Chebyshev bound one r.v.
- In stats/ML, need to bound multiple r.vs. → critical in ML

$$P(|\hat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

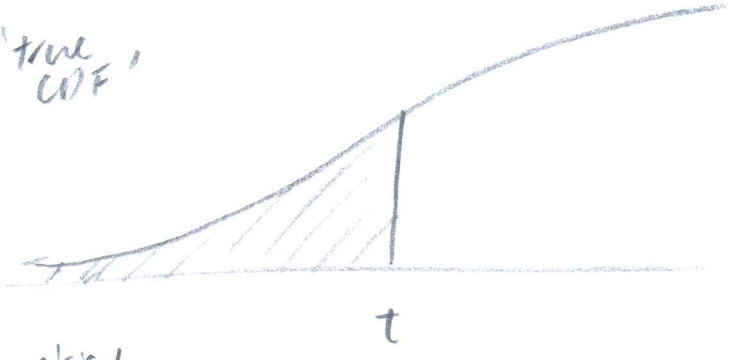
(LW): some examples of limitations of this (one r.v.)

Example 1 (A5) - IID? ecdf

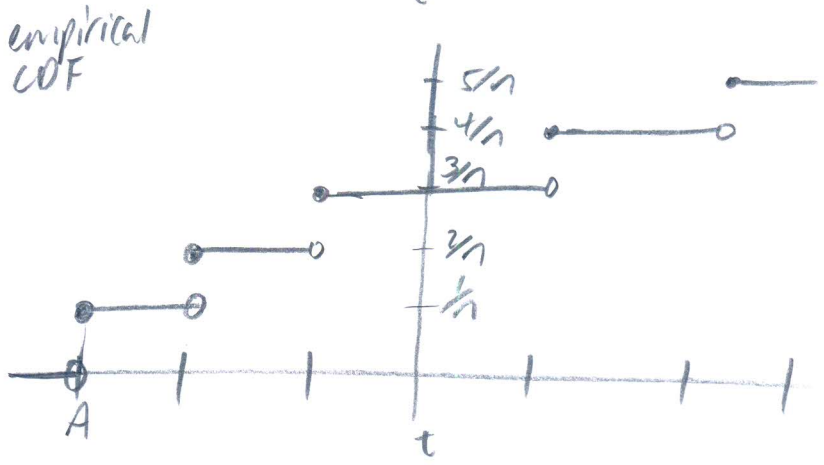
- Suppose we have samples from a distribution with $X_n \sim F$

Q: How do we go about estimating the CDF

What is true CDF?




- Empirical CDF; something/principle as CDF but using observed data i.e. actual rel frequencies rather than true probabilities
  - "fraction of points to the left is $\frac{1}{n}$"



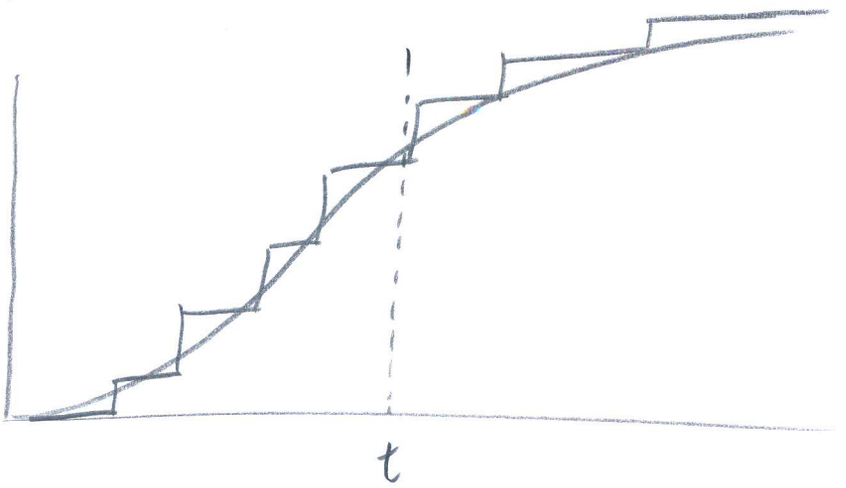

- $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$

'$\hat{F}_n$ is estimated from data, proportion of $\overset{obs}{n}$ data points less than t' → (A4)? intuit.

- $\hat{F}_n(t)$ is a CDF → check its properties e.g. right continuous, nondecreasing, normalised

Q: How close is $\hat{F}_n(t)$ to $F(t)$? i.e. empirical CDF and true CDF
- How much error am I making during estimation

(iv): reframe question through graphic:-




- fix both at a particular value of $t = (t_0)$; ask:

(Q) How close are the functions $F_n(t)\big|_{t_0}$ and $\hat{F}_n(t)\big|_{t_0}$

- $P\left(|\hat{F}_n(t) - F(t)| > \epsilon\right)$

(iv) - use Hoeffding inequality, but what is justification?

- $\hat{F}_n(t)$ can be seen as the average of Bernoulli r.v.s and $F_n(t)$ is true probability

(AS) will check this

- value of $\hat{F}_n(t)$ at each $t_i$; counting points - success if it falls to left
- failure if it falls to right
- so $\hat{F}_n(t)$ - average of Bernoullis
  $F(t)$ - true probability of 'heads'

- also, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$ (*)

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- where $Y_i = \mathbb{I}(X_i \leq t)$ and so $P(Y_i = 1) = P(X_i \leq t) = F(t) \quad \forall i$ (as IID)

- IID Bernoullis

$\therefore P\left(|\hat{F}_n(t) - F(t)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \quad n \to \infty \longrightarrow 0$

(iv): But there is a crucial subtlety here

- $P(|\hat{F_n}(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}$ is true for $\forall t$ ) (iv) AG int.

- But does that imply that :- (NO)

$$P\left(\sup_t |\hat{F_n}(t) - F(t)| > \epsilon\right) \leq \text{something small ?}$$

"sup" understood as max?

- i.e. it is true for all t, is it true for the maximum?

(vii): This is the distinction between pointwise and uniform convergence

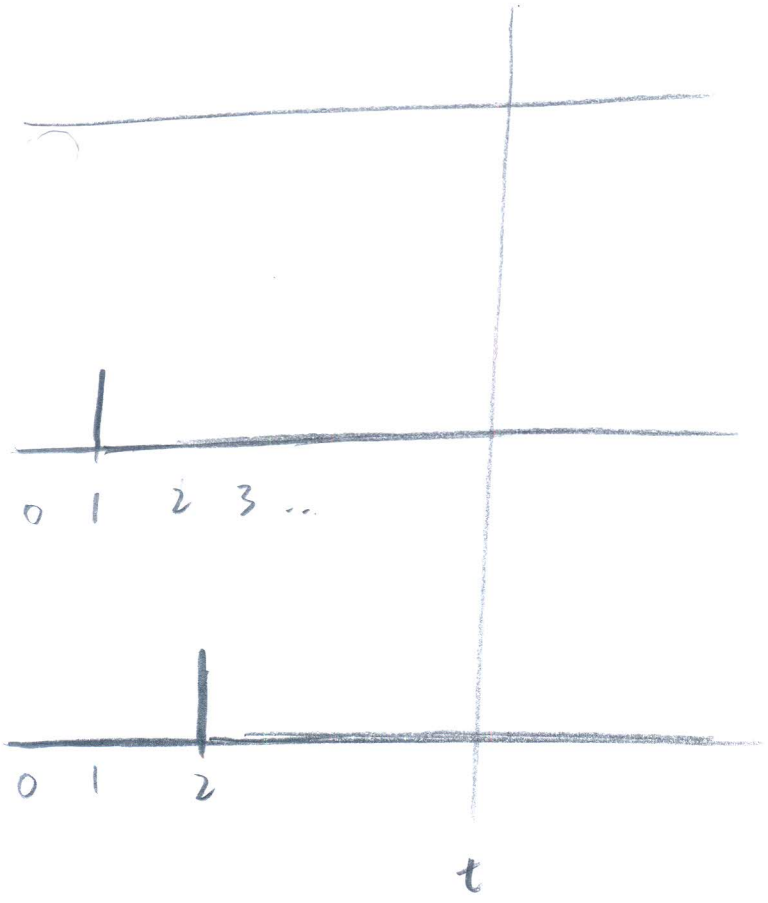
- selection bias (of selecting classifiers)

unclear not.

(?) 30:09

- To understand (vii): need some calculus
- Suppose we have a sequence of reals on the real line $\mathbb{R}$ indexed by $n$
- Define function

$$f_n(t) = \begin{cases} 1 & \text{if } t = n \\ 0 & \text{otherwise} \end{cases}$$




- look at fixed value of t
- and consider limit

$$\lim_{n \to \infty} f_n(t)$$

- At point t, ask what $f_n(t)$ will be for successive values of n
- so at t = 3; $f_0(3) = 0$, $f_1(3) = 0$, $f_2(3) = 0$ } $f_i(t) = 0$ for $i < t$

$f_3(3) = 1 \rightarrow f_i(t) = 1$ $i = t$

$f_4(3) = 0 \rightarrow f_i(t) = 0$ $i > t$

Hence $\lim_{n\to\infty} f_n(t) = 0 \quad \forall t$

- If I look at sequence of functions $f_1, f_2, f_3, \ldots$ at a fixed value of $t_0$, and I think of the outputs $f_1(t_0), f_2(t_0), \ldots$, it will eventually stay at 0 when $n > t$.

---

- Define another zero-function $g(t) = 0$ everywhere
- Then $\forall t \quad |f_n(t) - g(t)| \overset{n\to\infty}{\to} 0$ pointwise

- Consider:

$$\max_t |f_n(t) - g(t)| = 1 \not\to 0 \text{ uniform}$$

(FII): UW: saying something is true at each $t$ does not guarantee maximum difference will converge to 0

(SII): Pointwise convergence $\not\Rightarrow$ uniform convergence

- Deep implications for overfitting

---

(LW): in classification in ML, we want uniform convergence and stats

i.e.

- (LW): we don't just want probability $P(|\hat{F}_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2} \to 0 \; \forall \epsilon$

we want maximum difference $P(\sup_t |\hat{F}_n(t) - F(t)| > \epsilon) \leq$ something

- (W): (A7) check you understood the narrative behind this subtlety
- Uniform convergence is a much stronger statement

Example - classification

- $R(h) = P(Y \neq h(X))$ classification risk / probability of an error
- Recall we do not have $R(h)$; but given $(X_n, Y_n) \sim P$

And therefore observed data $(X_n, Y_n)$, we have an estimate of the classification error, i.e. training data

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Y_n \neq h(X_n))$$

(Q) How close is $R(h)$ to $\hat{R}(h)$

- Have to be able to estimate $R(h)$ to yield $\hat{R}(h)$ well to choose good classifie

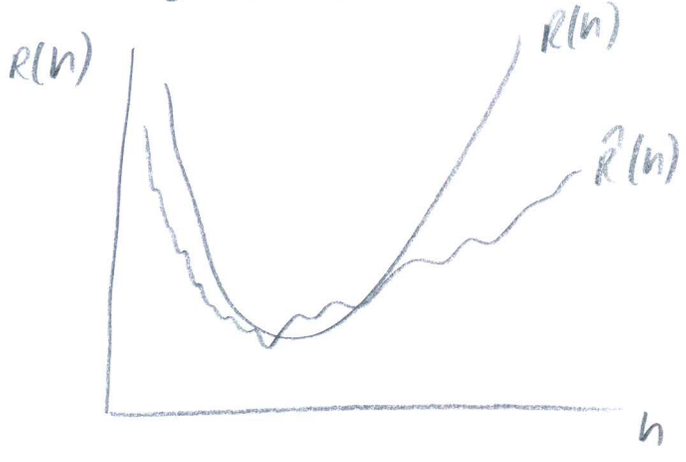
(W): coin flips + Hoeffding:-

$$P(|\hat{R}(h) - R(h)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

- This is fine, says that observed classification error is a good estimate of true class. error
- BUT (SI): we have a set of classifiers, which we have to choose from (i.e. a hypothesis space)
- So we want a classifie $h_i \in \mathcal{H}$: $R(h)$ is small, we don't know $R(h)$ so estimate with $\hat{R}(h)$
- Have to estmate $\hat{R}(h_i)$ for possibly infinite no. of classifiers in $\mathcal{H}$.

(SI):-

- If I don't estimate it 'uniformly well', I don't know when I minimise it, $\hat{R}(h)$ that I'll find a good classifier
- making assumption that you are estimating class. error well everywhere
- (W): searching whole space $\mathcal{H}$ for an $h$ that minimises $R$




- select $h$ with low $\hat{R}(h)$

(Q) we make implicit assumption that $\hat{R}(h)$ is close $R(h)$ everywhere

(A) (AB) - check and go through intuition here

- So we want :-

$$P\left(\sup_n \left|\hat{R}(h) - R(h)\right| > \epsilon\right) \leq \text{something small}$$

(LW): Take Hoeffding and not apply to r.v.; but to maximum of ... - a uniform bound

- the random variable here is $\hat{R}(h)$, indexed by the classifiers
- $R(h)$ is fixed but unknown to us.

- Next lecture: VC Theory