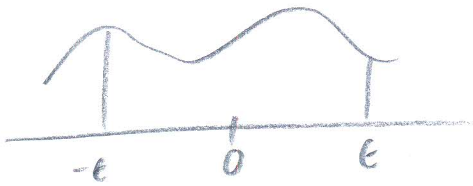
week 1 36-705 24/04/2020

· Youtube 02/09/2016

· lecture incomplete [illegible]

· looks like lemma 4 was proved : go over what is missed.

· $P(|\bar{Y}_n| > \epsilon)$



· Bound $P(\bar{Y}_n > \epsilon)$

· $P(\bar{Y}_n > \epsilon) = P(\Sigma Y_i \geq n\epsilon)$ $\boxed{3:02}$ ↓ $e^{(\cdot)}$ is monotone

$= P(e^{\Sigma Y_i} \geq e^{n\epsilon})$ -

$= P(e^{t\Sigma Y_i} \geq e^{tn\epsilon})$ - variational trick
- extra parameter

- via markov inequality

$\leq e^{-tn\epsilon}\, \mathbb{E}\left[e^{t\sum_{i=1}^{n} Y_i}\right]$

$= e^{-tn\epsilon} \prod_i \mathbb{E}[e^{tY_i}]$ via independence

$= e^{-tn\epsilon}\left(\underbrace{\mathbb{E}[e^{tY_i}]}\right)^n$ → we use a bound on this

$\leq e^{-tn\epsilon}\, e^{\frac{nt^2(b-a)^2}{8}}$ (*) - notes

· Advantage of including t - variational trick

· choose: $t = \frac{4\epsilon}{(b-a)^2}$ (can be found with derivatives)

Then $P(\bar{Y}_n > \epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}}$

• can apply to Bernoulli r.v.s. (coin flips) :-
$a=0, b=1$

· Gives a very tight bound :- ↗ goes to 0 very quickly (Hoeffding inequality gives a very, very tight bound)

· intuitively: probability fraction of heads is going to depart far from θ (parameter) is very small.

LW: If this doesn't go to 0 exponentially quickly, then 95% of stats and ML would not work.

Remarks
- For the case X_i with mean μ; then define $Y_i = (X_i - \mu)$ → ? $\boxed{8:21}$
- Prove whole theorem in terms of Y_i; then substitute
- choice of $t = \frac{4\epsilon}{(b-a)^2}$; (*) holds for any $t > 0$, we want to make it as small possible (W): check this (A1)
  - use deriv, solve; plot on wolfram.

Section 3 (optional)
- Extends Hoeffding to functions other than sums/averages of r.vs.
- Imagine arbitrary function of n r.vs.
- Same result (Hoeffding/variant) holds as long as function is smooth; in the sense that changing one of the function arguments by a small amount does not change the overall value of function much

- There for interest

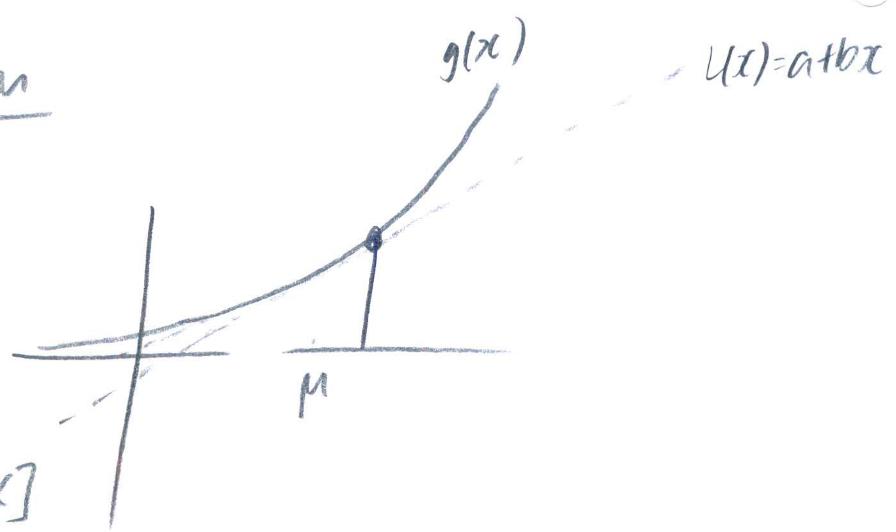
- Previously, we found bounds on probabilities (Markov, Chebyshev, Hoeffding)
- Bounds on expectations are slightly different
- Theorem 1 - Cauchy-Schwartz inequality
- will be used a lot; also

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

- LW: no details (A) (W): (A2) fill in

Jensen's inequality
- If g is convex then:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$



useful due to presence of $\mathbb{E}[X]$

Proof: Okay.

useful facts

- Require ideas of 'distance' between distributions (there are many)

Kullback-Leibler

- This arises very 'naturally' from information theory

LW. not really a 'distance': not symmetric; but has distance-like prop.

- $D(p,p)=0$; $D(p,q)\geq 0$ (some desirable properties of distance)
- can use Jensen to prove KL-divergence:
- Any time you see $\int p \ldots$ or $\Sigma p \ldots$ → think expected value
- note we only require specification of $X\sim p$; the rest i.e. in

$\mathbb{E}\left[\log\frac{p(x)}{q(x)}\right]$ is just a function/trans. $r(x)$ of X.

$$D(p,q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx = \mathbb{E}\left[\log\frac{p(x)}{q(x)}\right]$$

$$-D(p,q) = \mathbb{E}\left[\log\frac{q(x)}{p(x)}\right] \leq \log\left[\mathbb{E}\,\frac{q(x)}{p(x)}\right] = \log\int\frac{q(x)}{p(x)}\,p(x)\,dx = \log\int q(x)\,dx$$

via Jensen + concavity of log

$= 0$

(as $\int q(x)\,dx = 1$ (density))

- LW: Switching $\mathbb{E}$ and log: common use of Jensen
- so $-D(p,q)\leq 0$; hence $D(p,q)\geq 0$

(LW)(13): LW does not cover ex. 13

- LW: skip Theorem 15 and Proof

- useful: Bounding expected value of max. of r.v. (recurs a lot)
- $X_1,\ldots,X_n$ IID r.v.s.
- can impose an ordering $X_{(1)}\leq X_{(2)}\leq \ldots \leq X_{(n)}$; $X_{(n)} = \max\{X_1,\ldots,X_n\}$ by magnitude
- How to compute distr. of maximum? (conventionally find CDF)

LW: If we directly compute CDF $Pr(\max\{X_1,\ldots,X_n\}<t) \Leftrightarrow Pr(X_{(1)}<t, X_{(2)}\leq t \ldots X_{(n)}<t)$

· use independence ; 'multiply together' → (L) (A4)

· Get expression for CDF of max in terms of original CDF, get PDF, get $\mathbb{E}$

· in principle this is feasible ; but not in practice sometimes ; especially if we don't know exact distn, but some properties on shape

· $\mathbb{E}[\max\{X_1, \dots, X_n\}]$ (?"growing like n" again.) $\boxed{28:10}$

· If we know distn is thin-tailed → either from Normality or and/or Chernoff

· suppose we have a bond on MGF ← (L) (A5) (*)

$$\mathbb{E}[e^{tX_i}] \leq e^{\frac{t^2\sigma^2}{2}}$$

- σ - not necessarily variance

(mean grows "$\log n$")

- Theorem 16

- see notes $\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \sigma\sqrt{2\log n}$

Proof :-

- start with $\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right]$; apply transf :-

$$\exp\left\{t\,\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right]\right\} \leq \mathbb{E}\left(\exp\left\{t \max_{1\leq i\leq n} X_i\right\}\right)$$

via Jensen (L) (A6) - check

(property of max(·) ; think)

$$= \mathbb{E}\left(\max_{1\leq i\leq n} \exp\{tX_i\}\right)$$

$$\leq \sum_{i=1}^{n} \mathbb{E}\left[\exp\{tX_i\}\right]$$

$$\leq n e^{t^2\sigma^2/2}$$

Apply logs:

$$t\,\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \log n + \frac{t^2\sigma^2}{2}$$

$$\Rightarrow \mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}$$

• This is again a variational situation; set t to minimise

• i.e. set $t = \frac{\sqrt{2\log n}}{\sigma}$

Giving: $\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \sigma\sqrt{2\log n}$

LW: Recurring theme:

- If there exists a thin tail
- If we have an exponential bound on the MGF, we use that and Hoeffding's inequality style argument to say/make statements about $\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right]$

- Why thin-tailed distribution?

(TT) LW: only thin-tailed distributions have properties where MGF have exponential bounds → formally a Sub-Gaussian r.v.; informally thin tails

---

LW ☆

(TT). Hoeffding's inequality is most important; and related { bound MGF / Chernoff bound / 'tricks' }

- Next lecture: Asymptotic notation