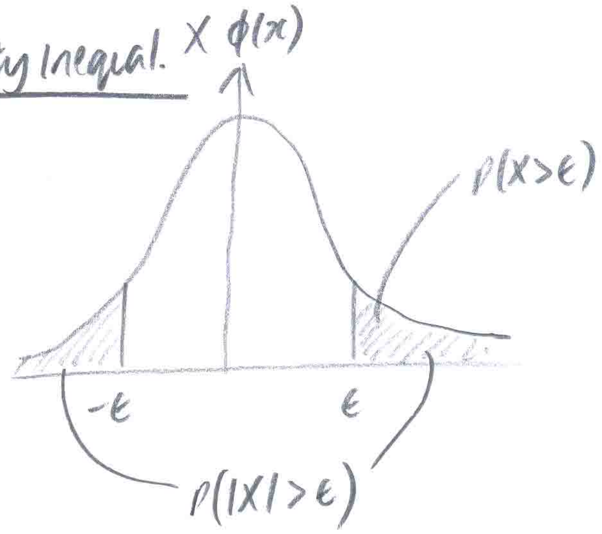
36-705 — 26/04/2020

- Lecture Notes 2 - supplementary
- Go over details you've highlighted

Theorem 1 - Gaussian Tail Inequality — **1. Probability Inequal.**



A (single r.v.)

- $X \sim N(0,1)$
- $P(|X|>\epsilon) \le \dfrac{2e^{-\epsilon^2/2}}{\epsilon}$

- **B Multiple r.vs.**

- $X_1,\dots,X_n \sim N(0,1)$

$$P(|\bar{X}_n|>\epsilon) \le \frac{2}{\sqrt{n}\,\epsilon} e^{-n^2\epsilon/2} \overset{\text{large } n}{\le} e^{-n\epsilon^2/2}$$

Proof (A)

- 2 parts:- i) Prove $P(X>\epsilon) \le \dfrac{e^{-\epsilon^2/2}}{\epsilon}$ · Density of X: $\phi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (std Normal)
  ii) use symmetry argument

i)

$$P(X>\epsilon) = \int_\epsilon^\infty \phi(s)\,ds \underset{(i)}{=} \int_\epsilon^\infty \frac{s}{s}\phi(s)\,ds \underset{(ii)}{\le} \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)\,ds$$

$$\text{And } \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)\,ds \underset{(iii)}{=} -\frac{1}{\epsilon}\int_\epsilon^\infty \phi'(s)\,ds = \frac{\phi(\epsilon)}{\epsilon} \underset{(iv)}{\le} \frac{e^{-\epsilon^2/2}}{\epsilon}$$

(i) multiply/divide by s
(ii) over interval $[\epsilon,\infty)$, $s \ge \epsilon$
(iii) $\phi'(s) = -s\phi(s) \Rightarrow s\phi(s) = -\phi'(s)$
(iv) note:- $\dfrac{1}{\epsilon\sqrt{2\pi}} e^{-\epsilon^2/2} \le \dfrac{1}{\epsilon} e^{-\epsilon^2/2}$

$$P(|X|>\epsilon) = P(X>\epsilon) + P(-X>\epsilon)$$
$$= P(X>\epsilon) + P(X<-\epsilon)$$
$$= 2P(X>\epsilon) \text{ via appeal to symmetry}$$

· So we have shown (i): $P(X>\epsilon) \le \dfrac{e^{-\epsilon^2/2}}{\epsilon}$

???

Part (ii)

- By symmetry, i.e. $P(|X|>\epsilon) = P(X>\epsilon) + P(-X<-\epsilon) = 2P(X>\epsilon)$ of std. Normal
  - we have $P(|X|>\epsilon) \le \dfrac{2e^{-\epsilon^2/2}}{\epsilon}$

Proof (B)

- $X_1, \dots, X_n \sim N(0,1)$
- Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{X}_n \sim N(0, \frac{1}{n})$ (recall $Var(\bar{X}_n) = \frac{\sigma^2}{n}$, $\sigma^2 = 1$)

$$\Rightarrow \bar{X}_n \overset{d}{=} \frac{1}{\sqrt{n}} Z = n^{-1/2} Z \quad \text{where } Z \sim N(0,1)$$

$$\Rightarrow P(|\bar{X}_n| > \epsilon) = P(n^{-1/2}|Z| > \epsilon) = P(|Z| > \sqrt{n}\epsilon) \leq \frac{2}{\sqrt{n}\epsilon} e^{-\frac{n\epsilon^2}{2}}$$

(i) - ? - clarify

Intuition + plots -

Theorem 2 - Markov's inequality

- Let X be a non-negative random variable and suppose $\mathbb{E}[X]$ exists
- That is $X > 0$, $\mathbb{E}[|X|] < \infty$
- For any $t > 0$,

$$P(X > t) \leq \frac{\mathbb{E}[X]}{t} \qquad (1)$$

Proof

- As $X > 0$,

$$\mathbb{E}[X] = \int_0^\infty x p(x)\,dx \overset{(i)}{=} \int_0^t x p(x)\,dx + \int_t^\infty x p(x)\,dx$$

$$\geq \int_t^\infty x p(x)\,dx \geq t \int_t^\infty p(x)\,dx = t P(X > t)$$

(ii)

Hence $P(X > t) \leq \frac{\mathbb{E}[X]}{t}$ ☒

Proof steps

i) Partition expectation/integral over $(0,\infty)$ into sum of integrals over $[0,t)$ and (t,∞)

ii) Over interval $[t,\infty)$, $x \geq t \Rightarrow \int_t^\infty x p(x)\,dx \geq t \int_t^\infty p(x)\,dt$

Theorem 3 - Chebyshev's inequality

- Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var(X)$; then:-

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t} \quad \text{and} \quad P(|Z| \geq k) \leq \frac{1}{k^2}$$

- where $Z = \frac{(X-\mu)}{\sigma}$

- Note that $P(|Z|>2) \leq \frac{1}{4}$ $P(|Z|>3) \leq \frac{1}{9}$

- Assuming variance exists:- i.e. $\mathbb{V}(|X|) = \int |x-\mu|^2 p(x)\,dx$ exists ?

Proof:

- via Markov's inequality;

$$P(|X-\mu| \geq t) = P(|X-\mu|^2 \geq t^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

- setting $t = k\sigma$:- (explicitly)

$$P(|X-\mu| \geq t) = P(|X-\mu| \geq k\sigma) = P\left(\frac{|X-\mu|}{\sigma} \geq k\right) = P(|Z| \geq k) \leq \frac{\sigma^2}{(k\sigma)^2}$$

$$\Rightarrow P(|Z| \geq k) \leq \frac{1}{k^2}$$

$\blacksquare$

- Application to Bernoulli r.v.s with exp. (from Wasserman)

- If $X_1, \dots, X_n \sim Bernoulli(p)$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Then $Var(\bar{X}_n) = \frac{Var(X_i)}{n} = \frac{p(1-p)}{n}$

- And via Chebyshev:

$$P(|\bar{X}_n - p| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

- As $p(1-p) \leq \frac{1}{4}$ $\forall p$ ✓

(*)
- Note how the bound is used here

2. Hoeffding's Inequality

- note the proof strategy m notes

Lemma 4 (Hoeffding's Lemma)

- Suppose $a \leq X \leq b$
- Then
$$\mathbb{E}[e^{tX}] \leq e^{t\mu} e^{\frac{t^2(b-a)^2}{8}} \quad \text{where } \mu = \mathbb{E}[X]$$

Convexity

- recall - A function g is convex iff for each x, y and each $\alpha \in [0,1]$

$$g(\alpha x + (1-\alpha) y) \leq \alpha g(x) + (1-\alpha) g(y)$$

Proof of Lemma 4

- Assume $\mu = 0$

Since $a \leq X \leq b$, we write X as a convex combination of a and b:-

$$X = \alpha b + (1-\alpha) a$$

where $\alpha = \frac{(X-a)}{(b-a)}$ and $(1-\alpha) = \frac{b-X}{b-a}$

Note that the function $g(y): y \to e^{ty}$ is convex, meaning that

$$e^{tX} \leq \alpha e^{tb} + (1-\alpha) e^{ta}$$

more explicitly; note that we are setting $X = \alpha b + (1-\alpha) a$ $x = b$ $y = a$ in the general definition of a convex function

$$g(X) = g(\alpha b + (1-\alpha) a) \leq \alpha g(b) + (1-\alpha) g(a)$$

$$e^{tX} \leq \alpha e^{tb} + (1-\alpha) e^{ta} = \frac{X-a}{b-a} e^{tb} + \frac{b-X}{b-a} e^{ta}$$

$\mathbb{E}[\cdot]$ both sides; $\mathbb{E}[X] = 0 \Rightarrow$

$$\mathbb{E}[e^{tX}] \leq \frac{e^{tb}}{b-a} \mathbb{E}[X-a] \quad \frac{e^{ta}}{b-a} \mathbb{E}[b-X]$$

Giving

$$\mathbb{E}[e^{tX}] \leq \frac{-ae^{tb}}{b-a} + \frac{be^{ta}}{b-a}$$

- At this stage we will express RHS in a form $e^{g(u)}$, using properties of $g(u)$ and Taylor's theorem for 1st three terms (up to quadratic)

- Define: $u = t(b-a)$

$$g(u) = -\gamma u + \log(1-\gamma+\gamma e^u)$$

$$\gamma = \frac{-a}{b-a}$$

- Note: $g(0) = g'(0) = 0$, $g''(u) \leq \frac{1}{4}\ \forall u > 0$ (*)

- Taylor expansion: ? W ~~X~~

$\exists\, \xi \in (0,u)$ such that

$$g(u) = \underbrace{g(0)}_{=0} + u\underbrace{g'(0)}_{=0} + \frac{u^2}{2}g''(\xi) = \frac{u^2}{2}\underbrace{g''(\xi)}_{\leq \frac{1}{4}} \leq \frac{u^2}{8} = \frac{t^2(b-a)^2}{8}$$

?- What is the 'point' about which Taylor expansion is being carried out about?

(*) check this:

$$g(0) = -\gamma(0) + \log(1-\gamma+\gamma e^0) = 0$$

$$g'(u) = -\gamma + \frac{\gamma e^u}{1-\gamma+\gamma e^u} \qquad g'(0) = -\gamma + \frac{\gamma e^0}{1-\gamma+\gamma e^0} = -\gamma+\gamma = 0$$

$$g''(u) = \frac{\gamma e^u(1-\gamma+\gamma e^u) - (\gamma e^u)^2}{(1-\gamma+\gamma e^u)^2} = \frac{\gamma e^u}{1-\gamma+\gamma e^u}\left(1 - \frac{\gamma e^u}{1-\gamma+\gamma e^u}\right) = \underbrace{s(1-s)}_{\leq \frac{1}{4}},$$

$s > 0$;

Hence

$$\mathbb{E}[e^{tX}] \leq e^{g(u)} \leq e^{\frac{t^2(b-a)^2}{8}}$$

Remark: Lemma 4 is known as Hoeffding's lemma

- uses Taylor's theorem and Jensen's inequality
- It is an inequality that bounds the moment generating function of any bounded random variable (above and below)

Lemma 5 - Chernoff's method

- Let X be a random variable. Then

$$P(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon} \mathbb{E}[e^{tX}]$$

where 'inf' can be understood as 'min'

Proof

for any $t > 0$

$$P(X > \epsilon) \underset{(i)}{=} P(e^X > e^\epsilon) \underset{(ii)}{=} P(e^{tX} > e^{t\epsilon}) \leq e^{-t\epsilon} \mathbb{E}[e^{tX}]$$

- since this is true for any $t \geq 0$, the result follows

(i) Raise/manipulate inequality within probability

(ii) introduce a variational parameter t

Theorem 6 (Hoeffding's inequality)

- Let $Y_1, \dots, Y_n$ be iid observations such that $\mathbb{E}[Y_i] = \mu$ and $a \leq Y_i \leq b$
- Then for any $\epsilon > 0$,

$$P(|\bar{Y}_n - \mu| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}} \qquad (4)$$

Corollary 7

If $X_1, \dots, X_n$ are independent with $P(a \leq X_i \leq b) = 1$ and common mean μ then with probability at least $1-\delta$

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n} \log\left(\frac{2}{\delta}\right)} \qquad (5.)$$

Proof of Hoeffding

- without loss of generality considerations:

- Assume $\mu = 0$
- And observe $P(|\bar{Y}_n| \geq \epsilon) = P(\bar{Y}_n \geq \epsilon) + P(\bar{Y}_n \leq -\epsilon)$
$$= P(\bar{Y}_n \geq \epsilon) + P(-\bar{Y}_n \geq \epsilon)$$

- use Chernoff's method:-

$$P(\bar{Y}_n \geq \epsilon) = P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq \epsilon\right) = P\left(\sum_{i=1}^{n} Y_i \geq n\epsilon\right) \overset{(i)}{=} P\left(e^{\sum_{i=1}^{n} Y_i} \geq e^{n\epsilon}\right)$$

$$\overset{(ii)}{=} P\left(e^{t\sum_{i=1}^{n} Y_i} \geq e^{tn\epsilon}\right) \overset{(iii)}{\leq} e^{-tn\epsilon}\,\mathbb{E}\left(e^{t\sum_{i=1}^{n} Y_i}\right)$$

$$= e^{-tn\epsilon}\,\mathbb{E}\left[\prod_{i=1}^{n} e^{tY_i}\right] = e^{-tn\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{tY_i}]$$

$$= e^{-tn\epsilon}\left(\mathbb{E}[e^{tY_i}]\right)^n$$

(i) - e(.) monotone
(ii) - variational reparam.
(iii) - Markov's ineq.

- we bound $\mathbb{E}[e^{tY_i}]$ using lemma 4, Hoeffding's lemma:-

$$\mathbb{E}[e^{tY_i}] \leq e^{\frac{t^2(b-a)^2}{8}}$$

- So we have from above:-

$$P(\bar{Y}_n \geq \epsilon) \leq e^{-tn\epsilon} e^{\frac{-t^2 n(b-a)^2}{8}} \quad (*)$$ - holds for any $t > 0$

- Our variational tricks pays dividends as we can now minimise wrt t
- minimise RHS wrt t:
- select $t = \frac{4\epsilon}{(b-a)^2}$

- And we then have:-

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

- Applying the same argument to $P(-\bar{Y}_n \geq \epsilon)$ yields the same result ∎
- Extending to case with μ, define $Y_i(X_i - \mu)$, prove in terms of Y_i, then s.b.

3. Bounded Difference Inequality

- Hoeffding's inequality can be extended
- McDiarmid's inequality extends the general insight to more general functions $g(x_1, \dots, x_n)$ of Hoeffding.
- supplementary

4. Bounds on Expected Values

Theorem 11 - Cauchy-Schwartz Inequal.

- If X and Y have finite variance i.e. $Var(|X|)$ and $Var(|Y|) < \infty$ then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]} \qquad (9.)$$

- Some additional exposition on convex functions (Wasserman)
- $g(\cdot)$ is convex if for each x and y and each $\alpha \in [0,1]$

$$g(\alpha x + (1-\alpha) y) \leq \alpha g(x) + (1-\alpha) g(y)$$

- If g is twice differentiable and $g''(x) \geq 0 \; \forall x$ then g is convex (calculus def.)
- If g is convex then g lies above any line that g touches at that point (tangent line) (geometric)
- A function g is concave if $-g$ is convex
- convex examples:- $g(x) = x^2$, $g(x) = e^x$
- concave examples: $g(x) = -x^2$ $g(x) = \log x$

Q: Is inverse of a convex function concave?

- CS inequality can be given a more explicit statistical context:-

$$\mathrm{cov}^2(X,Y) \leq \sigma_X^2 \sigma_Y^2$$

A (Q: why is this the case?)

Theorem 12 - Jensen's Inequality

- If g is convex, then

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \qquad (10)$$

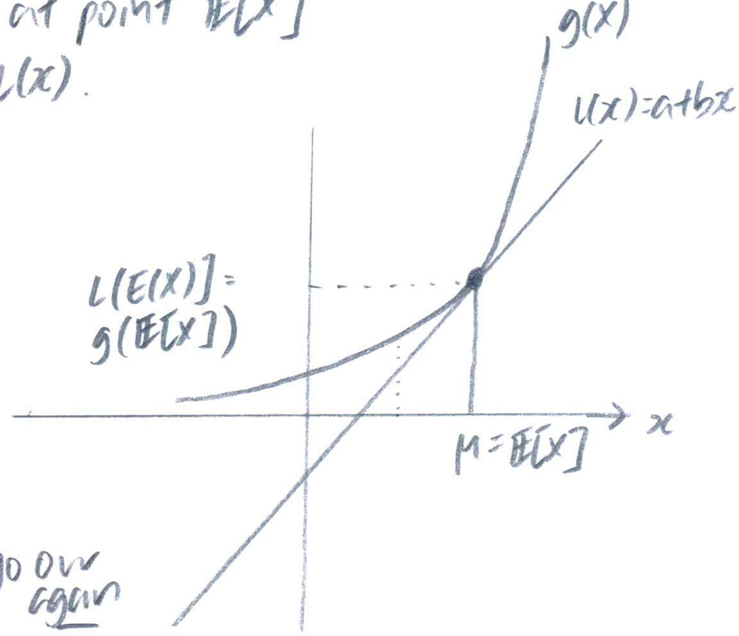
- If g is concave, then

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \qquad (11)$$

Proof:

- Let $L(x) = a + bx$ be a line tangent to $g(x)$ at point $\mathbb{E}[X]$
- Since g is convex, it lies above the line $L(x)$.
- $\mathbb{E}[g(X)] \geq \mathbb{E}[L(X)] = \mathbb{E}[a + bX] = a + b\mathbb{E}[X]$

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



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

(?) - For some reason the proof has become opaque to me again - Ⓦ Ⓐ - go over again

---

Example 13

As $g(x) = x^2$ is convex, we have $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$

i.e. the 2^{nd} moment is greater than the square of the 1^{st} moment (mean)

- constitutes a proof of $Var(X)$ being non-negative

Example 14

- Notes from lecture cover this fairly comprehensively

- Theorem 15 - skipped (spare time)

- Co... over bounding the maximum of a set of random variables

Theorem 16

- Let $X_1, \dots, X_n$ be random variables — sub-Gaussian
- Suppose there exists $\sigma > 0$ such that: -

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

Ⓐ Occurs for normal, bounded r.v.s. (thin-tailed r.v.s.)

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

Proof

- Start with statement; apply trans.

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

- (W): I couldn't initially see how proof invoked Jensen's inequality
- Define the convex function $g(y) = e^{ty}$, apply to (i) and invoke J.I.

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

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

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

$$\stackrel{(iv)}{\leq} n e^{\frac{t^2\sigma^2}{2}}$$

- Apply logs:-

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

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

- Variational situation, set t to minimise RHS.

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

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

(ii) properties of $\max(\cdot)$ function

(iii) $\max(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} X_i$

(iv) by assumption

36-705 Lecture Notes 2 09/05/2020

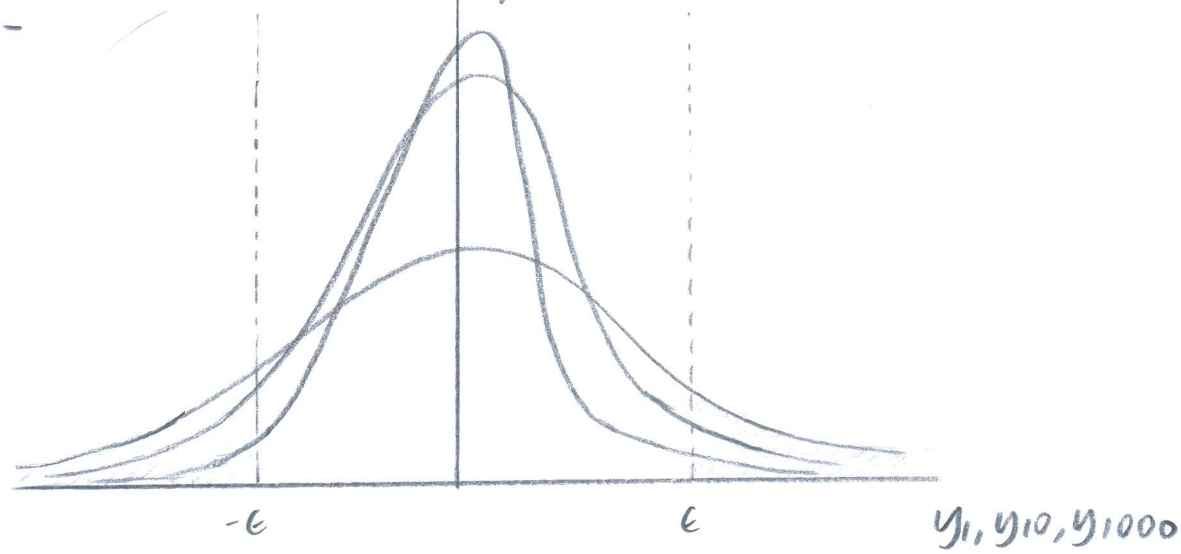
- Supplementary - review (Asymptotics)

- $Y_n = o_p(1)$:- (convergence in prob.) (sequence of r.v.s.)

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

$\to f_{Y_1}(y_1)$; $F_{Y_{10}}(y_{10})$; $F_{Y_{1000}}(y_{1000})$

- Diagram :-



- Note $P(|Y_n| > \epsilon)$ refers to tail probabilities
- A consequence of $Y_n = o_p(1)$ is that as $n \to \infty$; these tail probabilities (when you hold ϵ fixed at some arbitrary positive value) will approach 0.
- Note that for a fixed interval $[-\epsilon, \epsilon]$; the area under the PDF outside that interval, corresponding to tail probabilities, get smaller $(-\infty, -\epsilon]$; $[-\epsilon, \infty)$ and smaller, and approach 0.

$Y_n = O_p(1)$ (stochastic boundedness)

- Notes already cover this intuitively very well
- But I want to add a little more to capture some essential insight
- $Y_n = O_p(1)$
- If $\forall \epsilon > 0 \ \exists C_\epsilon : P(|Y_n| > C_\epsilon) \leq \epsilon \quad \forall n > n_0$ (for finite n_0, C_ϵ)
- Can we improve on this intuitively to better understand the differences in definition?
- (*) - A subtlety in definition $\longrightarrow$ PTO

For $o_p(1)$ (convergence in probability), we require the statement to hold not only for one, but for any arbitrarily small ϵ.
For $O_p(1)$ (stochastic boundedness), it suffices that there exists one arbitrarily large C_ϵ to satisfy the inequality, and C_ϵ is dependent on ϵ.

- This yields the analysis/adversarial way of thinking for $O_p(1)$ as a pedagogical tool for proofs.
- (*) If you give me an $\epsilon > 0$, can I find an arbitrarily large (finite) C_ϵ such that statement holds for large n greater than finite n_0?

(*) i.e.: If you give me $(1-\epsilon) = 0.9$, can I find an interval $[-C_\epsilon, C_\epsilon]$ to match that ϵ such that the interval traps 90% of the probability as $n \to \infty$?

(*) There are still some questions about this → place in overspill

- see stackexchange for proof examples; apply these insights to learn from them.

(*) Other interpretations (and formal def.), which can help understanding formalism → see add. notes