A whirlwind tour of probability + agnostic PAC / uniform convergence + no free lunch

CPSC 532S: Modern Statistical Learning Theory
17 January 2022
cs.ubc.ca/~dsuth/532S/22/

Admin

- FYI, I've been updating slides after class to stop where we actually stop + minor clarifications

- **Office hours:** Tuesdays 10-11am; Thursdays 4-5pm
  - Online for now (same Zoom link); at least one hybrid when we go to hybrid mode
  - Feel free to ask to schedule another time on Piazza
    - My potential available calendar is on my.cs.ubc.ca (if you have a CS account)

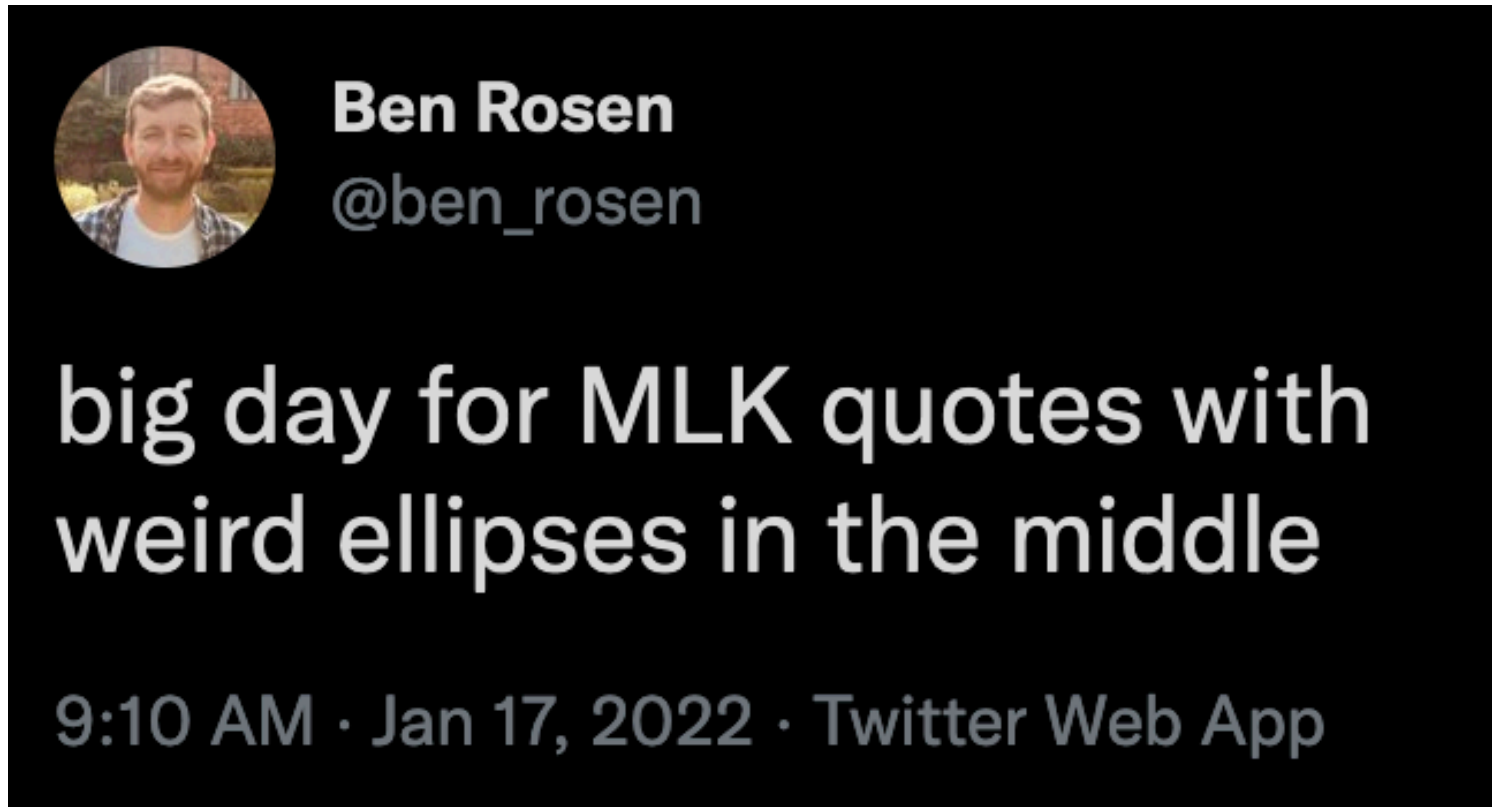
- **A1 due Thursday 11:59pm**
  - Do alone; cite sources in the question for anything you look up
  - Submit on Gradescope; if there's an issue, email your PDF to me
  - It's not short; make sure you've started! Might require brushing up on linear algebra

- We're making good progress towards fitting in the 40-person cap!
  - Will give instructions (on Piazza) to help prioritize waitlist soon, if still needed
  - If you're **not on the official waitlist but want to register**, or **want to officially audit**, follow instructions on Piazza

Briefly

- Obviously not a Canadian holiday,
  but want to acknowledge Martin Luther King, Jr day

- Letter from a Birmingham Jail (and other writings/speeches)
  still extremely relevant today, including in Canada (and around the world)



First: Probability overview

- A quick overview of probability as we'll mostly talk about it in this class
  - "measure-theoretic probability from someone who audited
    one measure-theoretic probability course in grad school
    (but got busy and and mostly stopped going halfway through)"

- We won't need to know "real" measure theory in this course
  - But the way I (and the Shais, and lots of work in the field) talk about probability is apparently more unintuitive than I thought to people who haven't learned it!

- There are links on the course page to sources to learn it "for real"

Why measure theoretic probability?

- Can handle discrete and continuous distributions in the same framework
- Can handle things that are neither discrete nor continuous
  - e.g. “spike-and-slab” prior: exactly 0 60% of the time, $\mathcal{N}(1, \sigma^2)$ o.w.
  - Joint distribution of (x, y) if $y = f(x)$ for a deterministic f
- Easier to handle things like *random functions* rigorously

- The idea: we instead focus on probabilities of *events*

Probability spaces

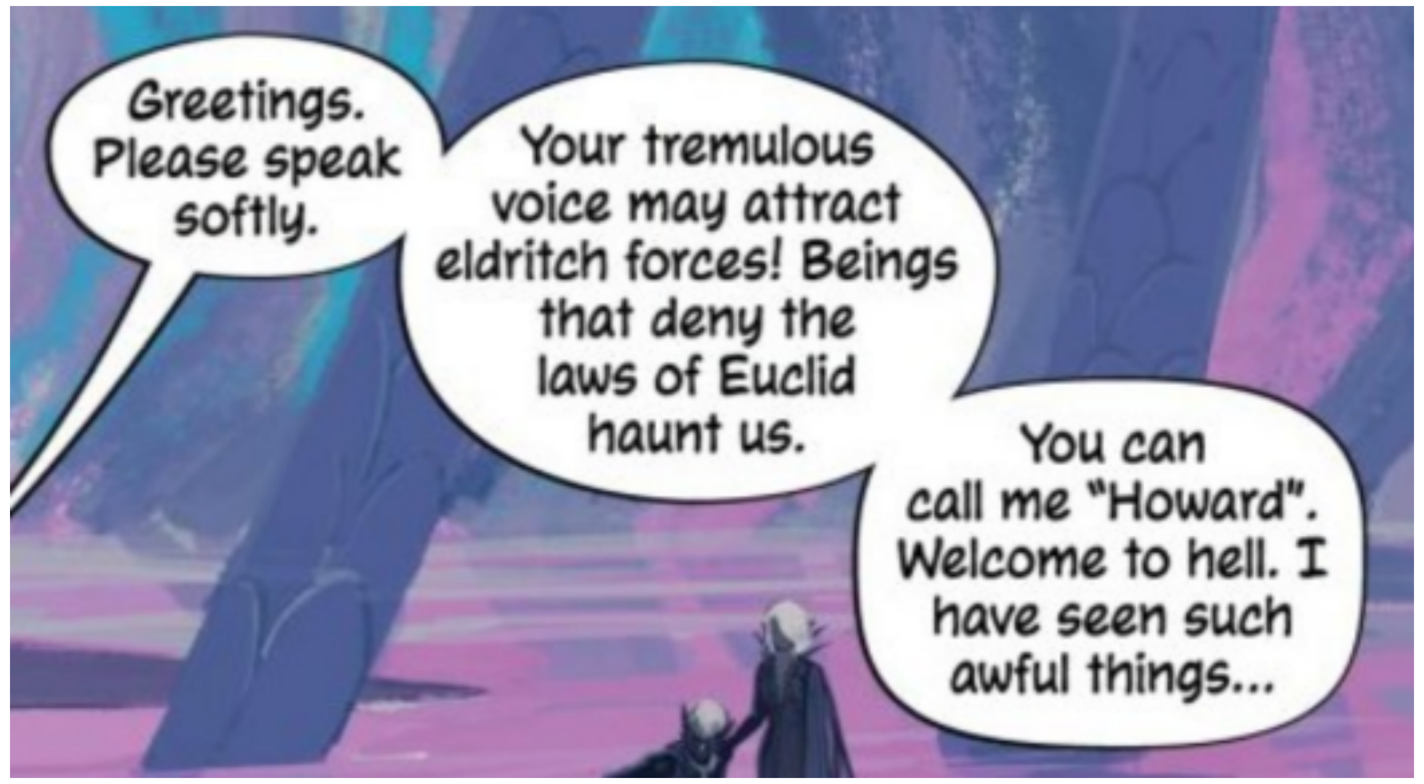
- Underlying **sample space** Ω – everything that *might happen*
  - If we roll a die once: $\Omega = \{1,2,3,4,5,6\} = [6]$
  - If we roll a die three times: $\Omega = [6] \times [6] \times [6]$
  - If we roll a die forever: $\Omega = [6]^\infty$

- An **event space** $\mathscr{F}$ containing possible events $E \subseteq \Omega$
  - "I rolled a 3": $\{3\}$
  - "My first two rolls were odd numbers": $\{(1,1,1), (1,1,2), \ldots, (5,5,6)\}$
  - "I didn't roll a four until my twenty-third roll"

- A **probability measure** $P : \mathscr{F} \to [0,1]$

Non-measurable sets (beware)

- We don’t allow ourselves to measure some things (avoid Banach-Tarski paradox)
- i.e. some $E \subset \Omega$ aren’t in $\mathscr{F}$
- Require $\mathscr{F}$ is a σ-algebra:
  - Contains Ω
  - Closed under complements
  - Closed under countable unions



Die #16 (Gillen/Hans/Cowles, 2021)

Probability axioms

- Kolmogorov axioms: a probability measure P needs

1. $P(E) \geq 0$ for all measurable events E

2. $P(\Omega) = 1$, where $\Omega = \cup_E E$: *something* happens with probability 1

3. If $E_1, E_2, \ldots$ is a countable sequence of *disjoint* sets,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Probability

- These axioms imply the kind of things you'd expect:
  - $P(\{\}) = 0$
  - Monotonicity: If $E_1 \subseteq E_2$ then $P(E_1) \leq P(E_2)$
  - $0 \leq P(E) \leq 1$
  - $P(E^c) = 1 - P(E)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Random variables

- Formally: a **random variable** is a function from $(\Omega, \mathscr{F})$ to some other measurable space, e.g. $(\mathbb{R}, \mathscr{R})$ – $\mathscr{R}$ is the Borel σ-algebra on $\mathbb{R}$

- X induces a probability measure: $\mathbb{P}(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$
- Personally: usually don't talk about Ω; I use $\mathbb{P}$, or write $\Pr$ to mean roughly P

- Discrete probability distributions:
  - Probability mass function: $\Pr(X = a)$, if $X \sim \mathbb{P}$, is just $\mathbb{P}(\{a\})$
- Continuous probability distributions:
  - $\mathbb{P}(A) = \Pr(X \in A)$
  - Note that $\mathbb{P}(\{a\}) = 0$ for any a; we'll come back to densities in a minute
  - But the CDF is $\mathbb{P}\big((-\infty, a]\big) = \Pr(X \leq a)$

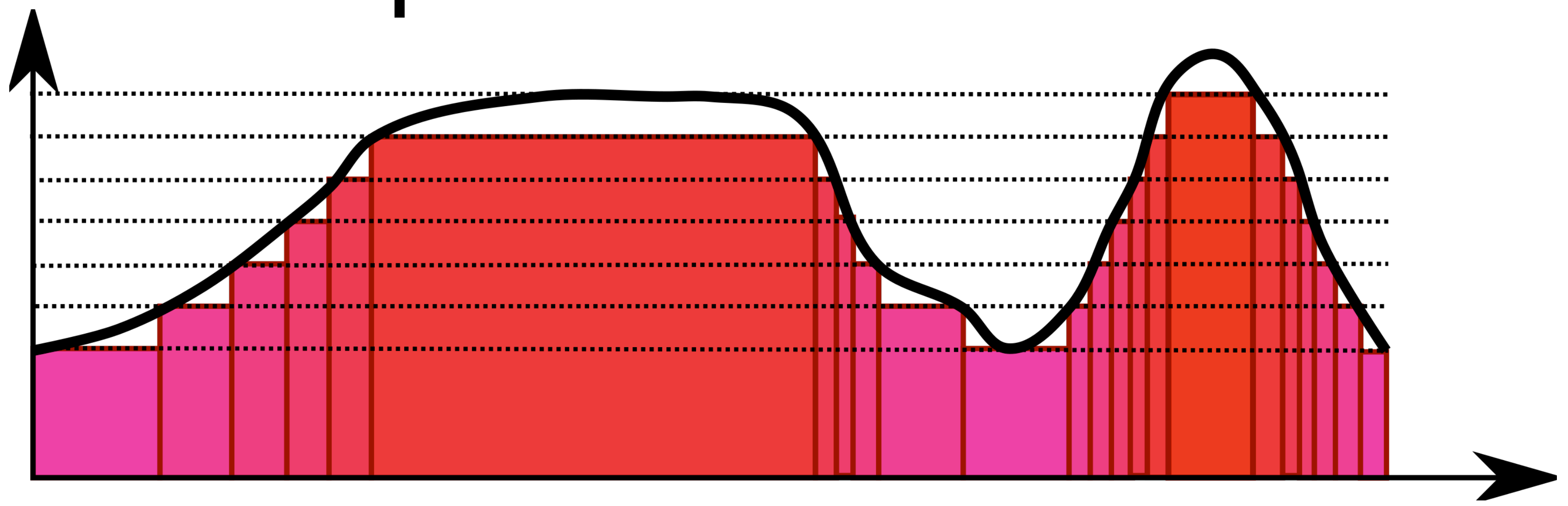
So what was $\mathscr{D}^n$ about?

- If X and Y are **independent** random variables, then $\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$ by definition
- That is: $\mathbb{P}_{XY}(A \times B) = \mathbb{P}_X(A)\,\mathbb{P}_Y(B)$
- Write $\mathbb{P}_{XY} = \mathbb{P}_X \times \mathbb{P}_Y$: a **product measure**
- Also abbreviate $\mathbb{P}^2 = \mathbb{P} \times \mathbb{P}$ for an i.i.d. pair

- In the proof of (realizable) PAC learnability for finite $\mathscr{H}$, we had $x \sim \mathscr{D}_x$, $y = f(x)$, $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$
- Book: probability of $S|_x = (x_1, \ldots, x_n) \sim \mathscr{D}^n$ falling in set of "bad samples"
- Today: we'll use $(x, y) \sim \mathscr{D}$, so $S \sim \mathscr{D}^n$ (and do the same kind of thing)

Building the Lebesgue integral

- If μ is a measure (like a probability measure, but doesn't require $\mu(\Omega) = 1$) we can build up **Lebesgue integral** starting with
$$\int_A \mathrm{d}\mu(x) = \int \mathbb{1}_A(x)\,\mathrm{d}\mu(x) = \mu(A) \quad \text{where } \mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$
- Expand to **simple functions** $f = \sum_i a_i \mathbb{1}_{A_i}$ by $\int f\,\mathrm{d}\mu = \sum_i a_i \mu(A_i)$
- Nonnegative functions by taking supremum of smaller simple functions
- Signed functions by taking $f = f^+ - f^-$



- Agrees with Riemann integral when it exists, but Lebesgue is more general

Expectations

- Define $\mathbb{E} f(x) = \int f \,\mathrm{d}\mathbb{P}$
- For discrete X, f = simple func $\sum_a f(a)\, \mathbb{I}_{\{a\}}$ on its **support**, $\bigcup_{E \in \mathcal{F}:\, \mathbb{P}(E)>0} E$
  - Thus $\mathbb{E} f(X) = \sum_i f(x_i)\, \mathbb{P}(\{x_i\})$
- If f is zero **almost surely**, $\mathbb{P}(\{x : f(x) = 0\}) = 1$, then $\int f \,\mathrm{d}\mathbb{P} = 0$
  - Book example: $h(x) = \begin{cases} y_i & \text{if } x = x_i, (x, y) \in S \\ 0 & \text{otherwise} \end{cases}$ (pure memorization)
  - Continuous data distribution has $\mathscr{D}(S|_x) = 0$: $L_{\mathscr{D}_x, f}(h) = L_{\mathscr{D}_x, f}(x \mapsto 0)$
  - But **empirical distribution** has $\hat{\mathscr{D}}(S|_x) = 1$, so $L_S(h) = L_S(f)$

Probability densities

- **Lebesgue measure** (often λ) is the usual measure for volume on $\mathbb{R}^d$
  - e.g. $\lambda([a, b]) = b - a$ for $b \geq a \in \mathbb{R}$
  - If we just write $\displaystyle\int f(x)\,\mathrm{d}x$, we usually mean $\displaystyle\int f(x)\,\mathrm{d}\lambda(x)$
- Usual probability density exists only if $\mathbb{P}$ is **absolutely continuous** wrt λ, $\mathbb{P} \ll \lambda$
  - If $\lambda(A) = 0$, then we also have $\mathbb{P}(A) = 0$
  - Discrete distributions are *not* **dominated by** (absolutely continuous wrt) λ
    - Are dominated by **counting measure**, $\mu(A) = |A|$
- If $\mathbb{P} \ll \mu$, there is a measurable $p = \dfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}$ taking values in $[0,\infty)$ with $\mathbb{P}(A) = \displaystyle\int_A p(x)\,\mathrm{d}\mu(x)$

To learn this stuff for real

From the course site:

Resources on learning measure-theoretic probability (*not* required to know this stuff in detail, but you might find it helpful):

- A Measure Theory Tutorial (Measure Theory for Dummies) (Maya Gupta) – 5 pages, just the basics
- Measure Theory, 2010 (Greg Hjorth) – 110 pages but comes recommended as both thorough and readable
- A Probability Path (Sidney Resnick) – frequently recommended textbook aimed at non-mathematicians to learn it in detail, but it's a full-semester textbook scale of detail; available if you log in via UBC
- There are also lots of other places, of course; e.g. the probability textbooks by Billingsley, Klenke, and Williams are (I think) classics.

Or Math 418/544 (probability)
Math 420 (real analysis - includes some measure theory)

(pause)

ERM on finite $\mathscr{H}$

- Last time,
  we showed that ERM algorithms PAC-learn finite $\mathscr{H}$ in the realizable setting
  - Probability of a “bad” hypothesis (one with $L_{\mathscr{D}_x, f}(h) > \varepsilon$) being an ERM is low
  - Union bound over all “bad” hypotheses

- Today: do ERM algorithms PAC-learn finite $\mathscr{H}$ in the agnostic setting?

ERM with uniform convergence

- Want h_S to compete with best predictor in $\mathscr{H}$ with high probability

- First step: "good" S are **ε-representative**, $|L_S(h) - L_{\mathscr{D}}(h)| \leq \varepsilon$ for **all** h
  - The **generalization gap** is small, for all h
- Lemma: If S is $\varepsilon/2$-representative, then for *any* $h \in \mathscr{H}$,

  $L_{\mathscr{D}}(h_S) \leq L_S(h_S) + \frac{1}{2}\varepsilon \leq L_S(h) + \frac{1}{2}\varepsilon \leq L_{\mathscr{D}}(h) + \varepsilon$ and so $L_{\mathscr{D}}(h_S) \leq \inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) + \varepsilon$

- $\mathscr{H}$ has the **uniform convergence property** w.r.t. $\mathscr{Z}$ and ℓ if,
  with $n \geq n_{\mathscr{H}}^{\text{UC}}(\varepsilon, \delta)$ samples from *any* distribution $\mathscr{D}$ over $\mathscr{Z}$,
  $S \sim \mathscr{D}^n$ is ε representative with probability at least $1 - \delta$

- So: sufficient to show that finite $\mathscr{H}$ have the uniform convergence property

Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right)$$

(we want to show it's $< \delta$)

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

assume $A \leq \ell(h,z) \leq A + B$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\frac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\frac{2}{B^2}n\varepsilon^2\right)$$

Hoeffding Bound (1963)

Wassily Hoeffding

If $X_1, \ldots, X_n \in \mathbb{R}$ independent, $\mathbb{E}[X_i] = \mu$, $\Pr(a \leq X_i \leq b) = 1$,

then $\Pr\left(\left|\frac{1}{n}\sum X_i - \mu\right| > \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$

Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right)$$ (we want to show it's $< \delta$)

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

assume $A \leq \ell(h,z) \leq A + B$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\tfrac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\tfrac{2}{B^2}n\varepsilon^2\right)$$

$$2|\mathscr{H}|\exp\left(-\tfrac{2}{B^2}n\varepsilon^2\right) < \delta \text{ iff } -\tfrac{2}{B^2}n\varepsilon^2 < \log\frac{\delta}{2|\mathscr{H}|} \text{ iff } n > \frac{B^2}{2\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\tfrac{1}{\delta}\right]$$

ERM agnostically PAC-learns $\mathscr{H}$ with $n > \dfrac{2B^2}{\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\tfrac{1}{\delta}\right]$ samples

Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right)$$ (we want to show it's $< \delta$)

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\tfrac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\tfrac{2}{B^2}n\varepsilon^2\right)$$

assume $A \leq \ell(h,z) \leq A + B$

Equivalently: error of ERM over $\mathscr{H}$ is at most $\sqrt{\frac{2B^2}{n}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]}$

ERM agnostically PAC-learns $\mathscr{H}$ with $n > \frac{2B^2}{\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]$ samples

Summary

- Measure-theoretic probability
  - Hope that was helpful? But again, we won't need details.
- Finite classes are PAC learnable, both in realizable and agnostic settings
  - but rate is different
- Uniform convergence of $L_S(h)$ to $L_{\mathscr{D}}(h)$ over $\mathscr{H}$
  - Key tool: Hoeffding bound (a **concentration inequality**)
- Next time: choosing $\mathscr{H}$; what about infinite hypothesis classes?