# PAC learning
# + uniform convergence

CPSC 532D: Modern Statistical Learning Theory
14 September 2022
cs.ubc.ca/~dsuth/532D/22w1/

# Admin

- Everyone should be registered now; if not, talk to me
  - If you want to audit, email me a form

- A1 is up
  - Work in pairs if you want
  - Cite **any sources** you use other than the course books (SSBD, MRT, Tel)
    - Including talking to people not in your group: **say so** + what extent
  - Gradescope link to submit will be up soon

- UBC is **closed next Monday** for the Queen's funeral
  - So, class is canceled again…sorry
  - Assignment deadline **likely** to become Tuesday – will update on Piazza

- Final is scheduled: Wednesday Dec 14, 2-4:30pm, ICCS 246
  - Let me know if there's a serious problem and we can maybe adapt

# Last time: definitions

- $(x, y) \sim \mathscr{D}$, a distribution over $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$
- Training "set" $S = (z_1, \ldots, z_n) = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \sim \mathscr{D}^n$
- Loss function $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$, e.g. $\ell_{0-1}(h, (x, y)) = \mathbb{I}(h(x) \neq y)$
- Want to find $h$ minimizing $L_{\mathscr{D}}(h) = \mathbb{E}_{z \sim \mathscr{D}}[\ell(h, z)]$, e.g. error rate = 1-accuracy for 0-1
  - name ∈ {"true", "population"} × {"risk", "loss"}
- Have $L_S(h) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \ell(h, z_i)$;     name ∈ {"empirical", "training"} × {"risk", "loss"}
- Empirical risk minimization (ERM): choose $h$ minimizing $L_s(h)$
  from a ***hypothesis class*** $\mathscr{H}$ of functions $h : \mathscr{X} \to \mathscr{Y}$
- To start with something simple, assume **realizability** for a nonnegative loss:

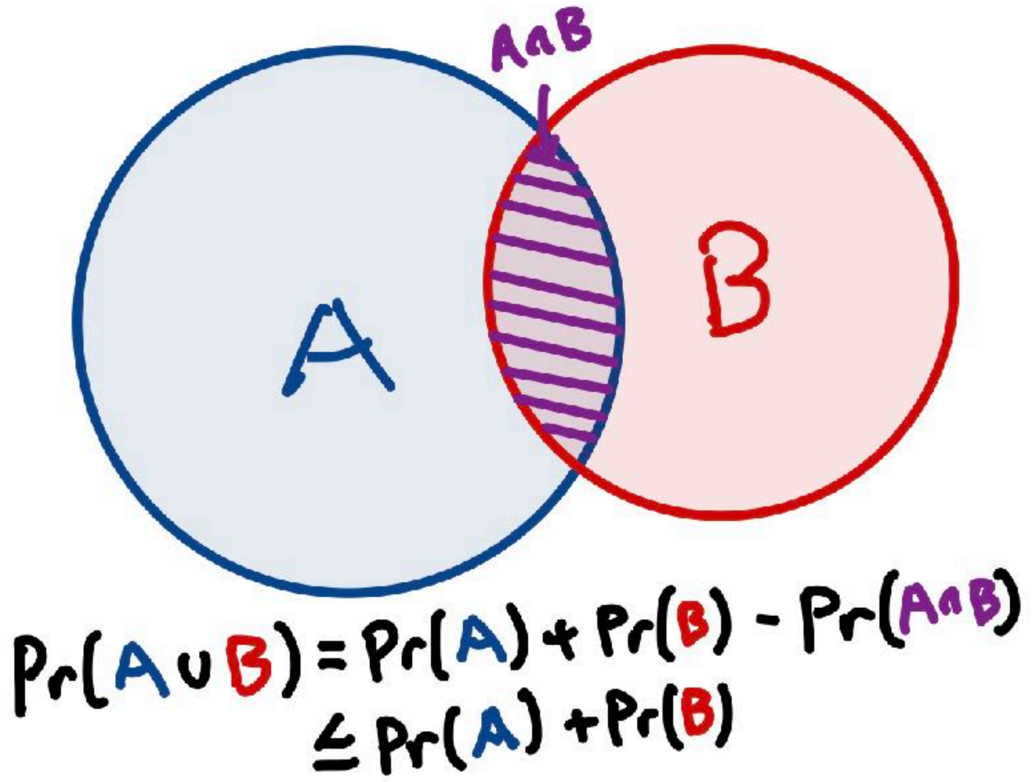
  **there is an $h^* \in \mathscr{H}$ with $L_{\mathscr{D}}(h^*) = 0$**

  - Implies (a.s.) that $L_S(h^*) = 0$

# Realizable, finite $\mathscr{H}$

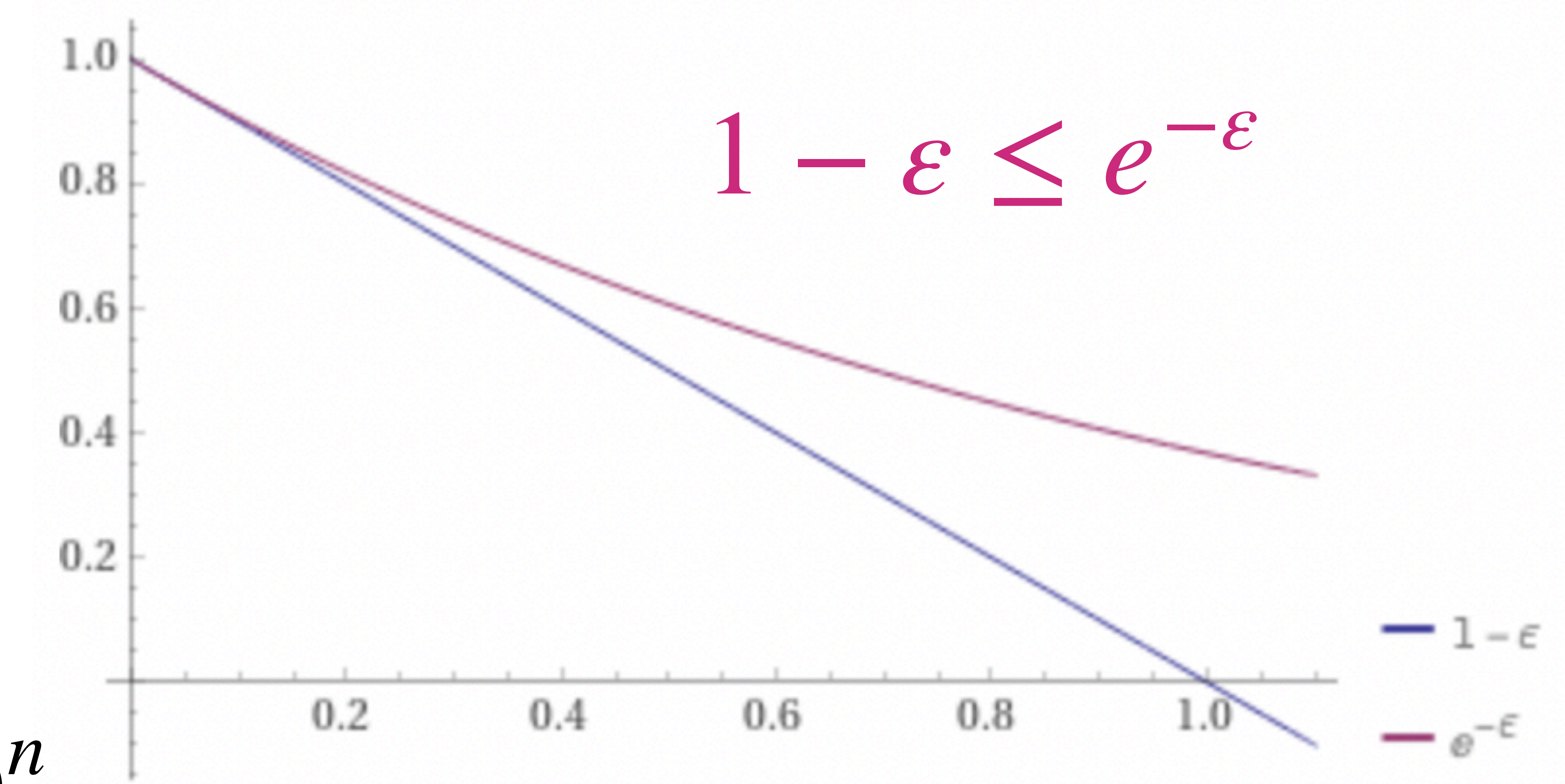
- Assume $0 \le \ell(h, z) \le 1$ for all $h, z$; also assume realizability
- $\hat{h}_S \in \arg\min_{h \in \mathscr{H}} L_S(h)$
  - Realizable means that $L_S(\hat{h}_S) = 0$, but maybe $L_{\mathscr{D}}(\hat{h}_S) > 0$
- Would like to show $\Pr_S \left( L_{\mathscr{D}}(\hat{h}_S) \le \varepsilon \right) \ge 1 - \delta$, i.e. $\Pr(L_{\mathscr{D}}(h_S) > \varepsilon) < \delta$
- Call $\mathscr{H}_\varepsilon$ the set of "bad" hypotheses, $\left\{ h \in \mathscr{H} : L_{\mathscr{D}}(h) > \varepsilon \right\}$
- If ERM failed, $S$ must be consistent with a bad hypothesis:

**Union bound**




$$\Pr(L_{\mathscr{D}}(\hat{h}_S) > \varepsilon) \le \Pr\left( S \in \bigcup_{h \in \mathscr{H}_\varepsilon} \{ S : L_S(h) = 0 \} \right) \le \sum_{h \in \mathscr{H}_\varepsilon} \Pr_{S \sim \mathscr{D}^n} \left( L_S(h) = 0 \right)$$

# Realizable, finite $\mathscr{H}$




$1 - \varepsilon \leq e^{-\varepsilon}$

- $\Pr(L_{\mathscr{D}}(\hat{h}_S) > \varepsilon) \leq \sum_{h \in \mathscr{H}_\varepsilon} \Pr\left(L_S(h) = 0\right)$
- $\Pr(L_S(h) = 0) = \Pr(\forall i \in [n]\,.\, \ell(h, z_i) = 0)$
- Because $S$ is iid, this is just $\prod_{i=1}^{n} \Pr_{z_i \sim \mathscr{D}} (\ell(h, z_i) = 0) = p_0(h)^n$

  where $p_0(h) = \Pr_{z \sim \mathscr{D}} (\ell(z, h) = 0)$

If a hypothesis is bad, we're likely to sample at least one data point where it's wrong

- Know that $L_{\mathscr{D}}(h) = p_0(h) \times 0 + (1 - p_0(h)) \times \mathbb{E}_z[\ell(z, h) \mid \ell(z, h) > 0]$
  - So, if $L_{\mathscr{D}}(h) > \varepsilon$, then must have $1 - p_0(h) > \varepsilon$, i.e. $p_0(h) < 1 - \varepsilon$
- $\Pr(L_{\mathscr{D}}(\hat{h}_S) > \varepsilon) < \sum_{h \in \mathscr{H}_\varepsilon} (1 - \varepsilon)^n$

Not too likely to get unlucky with *any* bad hypothesis

$$= |\mathscr{H}_\varepsilon|(1 - \varepsilon)^n < |\mathscr{H}|(1 - \varepsilon)^n \leq |\mathscr{H}|e^{-\varepsilon n}$$

# Finite $\mathscr{H}$ are (realizable) PAC-learnable

- We showed that $\Pr\left(L_{\mathscr{D}}(\hat{h}_S) < \varepsilon\right) \geq 1 - |\mathscr{H}|e^{-\varepsilon n}$
- Or: if we have $n \geq \dfrac{1}{\varepsilon}\left(\log|\mathscr{H}| + \log\frac{1}{\delta}\right)$, $L_{\mathscr{D}}(h) \leq \varepsilon$ with prob. at least $1 - \delta$.
- Or: error is at most $\dfrac{1}{n}\left(\log|\mathscr{H}| + \log\frac{1}{\delta}\right)$ with probability at least $1 - \delta$

- $\mathscr{H}$ is **PAC learnable** if there is a function $n_{\mathscr{H}} : (0,1)^2 \to \mathbb{N}$ and a learning alg. s.t.:
    - For every $\varepsilon, \delta \in (0,1)$, for every $\mathscr{D}$ over $\mathscr{X} \times \{0,1\}$ which is realizable by $\mathscr{H}$,
    - then running the algorithm on $n \geq n_{\mathscr{H}}(\varepsilon, \delta)$ i.i.d. examples from $\mathscr{D}$
    - will return a hypothesis $h$ with $L_{\mathscr{D}}(h) \leq \varepsilon$
    - with probability at least $1 - \delta$ over the choice of examples $S$

# Example: Boolean conjunctions

| a | b | c | d | e | f | y |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 0 | 1 | 0 | 0 | 1 | + |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 1 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 1 | 0 | 0 | 1 | 0 | - |
| 1 | 0 | 1 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 1 | 0 | 1 | ? |

$\mathscr{H}$: conjunctions of the form

$$a \wedge \bar{c} \wedge f$$

Algorithm:

- Start with $a \wedge \bar{a} \wedge \cdots \wedge f \wedge \bar{f}$
- Cross out bits inconsistent with the positives

# Example: Boolean conjunctions

| a | b | c | d | e | f | y |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 0 | 1 | 0 | 0 | 1 | + |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 1 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 1 | 0 | 0 | 1 | 0 | - |
| 1 | 0 | 1 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 1 | 0 | 1 | ? |

$\mathscr{H}$: conjunctions of the form

$$a \wedge \bar{c} \wedge f$$

Algorithm:

- Start with $a \wedge \bar{a} \wedge \cdots \wedge f \wedge \bar{f}$
- Cross out bits inconsistent with the positives

# Example: Boolean conjunctions

| a | b | c | d | e | f | y |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 0 | 1 | 0 | 0 | 1 | + |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 1 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 1 | 0 | 0 | 1 | 0 | - |
| 1 | 0 | 1 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 1 | 0 | 1 | ? |

$\mathscr{H}$: conjunctions of the form

$$a \wedge \bar{c} \wedge f$$

Algorithm:

- Start with $a \wedge \bar{a} \wedge \cdots \wedge f \wedge \bar{f}$
- Cross out bits inconsistent with the positives

# Example: Boolean conjunctions

| a | b | c | d | e | f | y |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 0 | 1 | 0 | 0 | 1 | + |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 1 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 1 | 0 | 0 | 1 | 0 | - |
| 1 | 0 | 1 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 1 | 0 | 1 | ? |

$\mathscr{H}$: conjunctions of the form

$$a \wedge \bar{c} \wedge f$$

Algorithm:

- Start with $a \wedge \bar{a} \wedge \cdots \wedge f \wedge \bar{f}$
- Cross out bits inconsistent with the positives

# Example: Boolean conjunctions

$c \wedge \bar{d} \wedge f$

| a | b | c | d | e | f | y |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 0 | 1 | 0 | 0 | 1 | + |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 1 | 1 | 1 | 0 | 1 | 1 | + |
| 0 | 1 | 0 | 0 | 1 | 0 | - |
| 1 | 0 | 1 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 1 | 0 | 1 | ? |

$|\mathscr{H}| = 3^d$: $\left\lceil \frac{1}{\varepsilon}\left(d \log(3) + \log\frac{1}{\delta}\right)\right\rceil$ samples enough

$\mathscr{H}$: conjunctions of the form

$$a \wedge \bar{c} \wedge f$$

Algorithm:

- Start with $a \wedge \bar{a} \wedge \cdots \wedge f \wedge \bar{f}$
- Cross out bits inconsistent with the positives

Assuming realizability, this gives an ERM

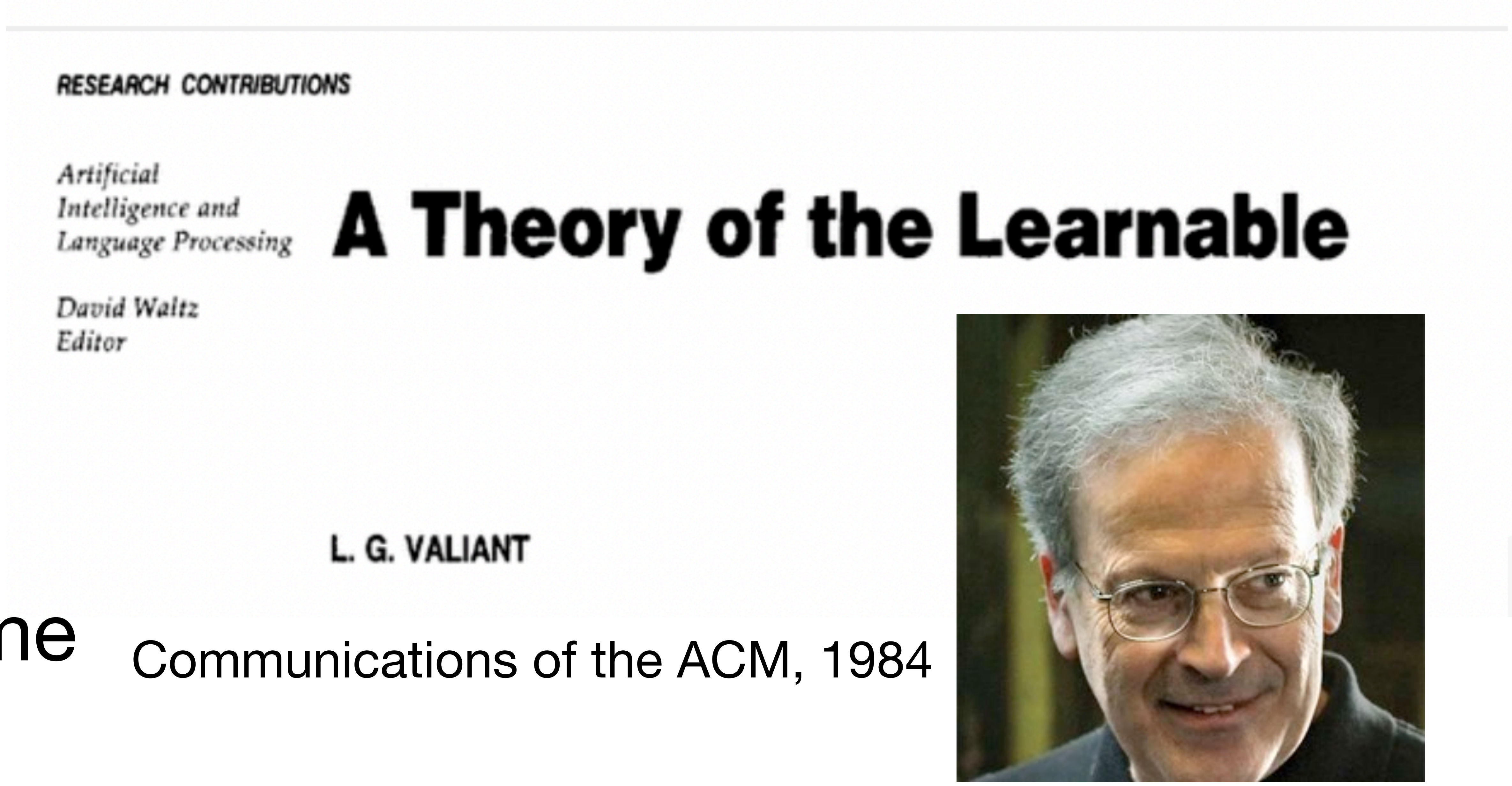
- Algorithm makes every + example a +
- True function f is only “less specific” than h: h(x) = - for anything truly -

# So, are we done with the course?

- Every practical $\mathscr{H}$ is finite if you put it on a computer
- Total size of weights in a big deep network is typically up to ~1GB
- Say 100MB, $8 * 100 * 2^{20}$ bits, so there are $2^{25 \cdot 2^{25}}$ possible networks
- $\log\left(2^{25 \cdot 2^{25}}\right) = 25\ 2^{25} \log(2) \approx$ 252 million
  - If we want, say, $\varepsilon = 0.1$ (90% accuracy): 2.5 billion training points

- (Plus, we don't actually do ERM with realizable, fixed hypothesis classes…)

# PAC learnability and computational efficiency


RESEARCH CONTRIBUTIONS

*Artificial Intelligence and Language Processing*

*David Waltz Editor*

**A Theory of the Learnable**

L. G. VALIANT

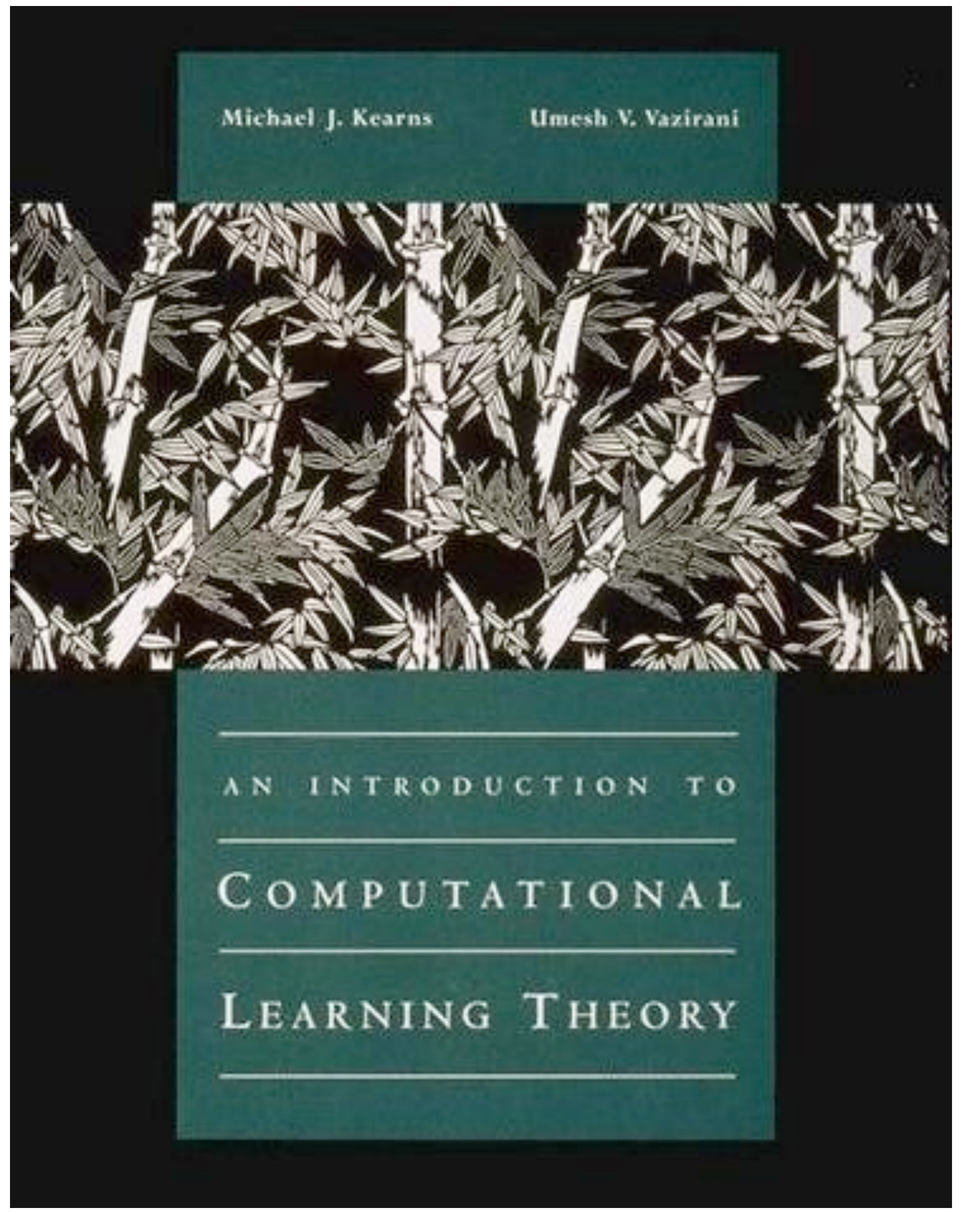
Communications of the ACM, 1984

- Valiant (1984)'s formulation
  required the algorithm
  to run in polynomial time
- We're going to mostly not care about runtime
  (call poly version "efficient PAC learning"),
  but be aware many authors keep that in the definition

- Independent(?), closely related development by Vapnik and Chervonenkis
  in the USSR; much more on their work soon

# PAC learnability and computational efficiency

(Sec 1.4-1.5
PDF through UBC: log in here)




- A class that can be PAC-learned but **not in polynomial time** (assuming P = BPP and P ≠ NP):
- 3-DNF: 3-term clauses in *disjunctive normal form*

  $T_1 \vee T_2 \vee T_3$

  terms are conjunctions: $T_1 = a \wedge \bar{c} \wedge \cdots$

  - Graph 3-coloring reduces to learning 3-DNFs

- But: 3-DNF ⊂ 3-CNF, $\bigwedge (a \vee b \vee c)$,
- $$T_1 \vee T_2 \vee T_3 = \bigwedge_{u \in T_1, v \in T_2, w \in T_3} (u \vee v \vee w)$$
- and 3-CNF **can** be efficiently PAC-learned

**Computational Limitations on Learning from Examples**

LEONARD PITT (1988)

*University of Illinois, Urbana-Champaign, Urbana, Illinois*

AND

LESLIE G. VALIANT

*Harvard University, Cambridge, Massachusetts*

(pause)

# Non-realizable (agnostic) learning

- What if we don't know that $\mathscr{H}$ can realize $\mathscr{D}$?
  - (Does the class of ResNet-101s realize ImageNet? 🤷‍♂️)
- What if we know that $\mathscr{H}$ *can't* realize $\mathscr{D}$?
  - If one $x$ can have two possible $y$s, no function can get zero loss*
    - *if there's a positive probability of getting such an $x$

# Agnostic PAC

- $\mathcal{H}$ is **agnostically PAC learnable** for a set $\mathcal{Z}$ and loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ if there is a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm such that:
  For every $\varepsilon, \delta \in (0,1)$ and every distribution $\mathcal{D}$ over $\mathcal{Z}$,
  then running the algorithm on $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples from $\mathcal{D}$
  will return a hypothesis $h \in \mathcal{H}$ with $L_{\mathcal{D}}(h) \leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$
  with probability at least $1 - \delta$ over the choice of examples

- We don't (necessarily) get error arbitrarily close to 0 anymore!
  - Realizable means $\inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') = 0$: then, this is same as realizable PAC
  - Otherwise, $\inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$ is the best loss achievable in $\mathcal{H}$

# Improper Agnostic PAC

- $\mathscr{H}$ is **improperly agnostically PAC learnable** in $\mathscr{H}'$ for $\mathscr{Z}$, loss $\ell : \mathscr{H}' \times \mathscr{Z} \to \mathbb{R}$
  if there is a function $n_{\mathscr{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm such that:
  For every $\varepsilon, \delta \in (0,1)$ and every distribution $\mathscr{D}$ over $\mathscr{Z}$,
  then running the algorithm on $n \geq n_{\mathscr{H}}(\varepsilon, \delta)$ i.i.d. examples from $\mathscr{D}$
  will return a hypothesis $h \in \mathscr{H}' \supset \mathscr{H}$ with $L_{\mathscr{D}}(h) \leq \inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') + \varepsilon$
  with probability at least $1 - \delta$ over the choice of examples

- e.g.: learn a polynomial classifier almost as good as the best linear classifier,
  or learn a 3-DNF function with a 3-CNF

- Shai+Shai: “there is nothing improper about representation-independent learning”

# Bayes error rate

- What can we say about $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$?
- It's at least as big as the **Bayes error**: error of the Bayes-optimal predictor

e.g. for 0-1 loss, $f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \Pr(y = 1 \mid x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

- The best predictor in $\mathcal{H}$ might be as good as this, or it might be worse
- Gap between Bayes error and $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ called **approximation error**

# ERM on finite classes, agnostic edition

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

- First step: "good" $S$ are **$\varepsilon$-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 $\varepsilon$-representative, then for *any* comparator $h' \in \mathscr{H}$,

$$L_{\mathscr{D}}(\hat{h}_S) \leq L_S(\hat{h}_S) + \varepsilon \leq L_S(h') + \varepsilon \leq L_{\mathscr{D}}(h') + 2\varepsilon \quad \text{and so } L_{\mathscr{D}}(\hat{h}_S) \leq \inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) + 2\varepsilon$$

- $\mathscr{H}$ has the **uniform convergence property** w.r.t. $\mathscr{Z}$ and $\ell$ 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 $\varepsilon$ 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 $< \delta$)

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

**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 $< \delta$)

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

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

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

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

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 $< \delta$)

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

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

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

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

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

# Realizable vs agnostic rates

- ERM for finite hypothesis classes, $n$ to get excess error $\varepsilon$ w/ prob. $1-\delta$, for a loss bounded in $[0,1]$:
  - Realizable: $n \geq \frac{1}{\varepsilon}\left(\log|\mathscr{H}| + \log\frac{1}{\delta}\right)$ "$\frac{1}{n}$ rate"
  - Agnostic: $n > \frac{2}{\varepsilon^2}\left[\log|\mathscr{H}| + \log\frac{2}{\delta}\right]$ "$\frac{1}{\sqrt{n}}$ rate"

- Late in the course, we'll (probably) see "optimistic rates": interpolate between the two regimes based on $\inf_{h\in\mathscr{H}} L_{\mathscr{D}}(h)$

# Summary

- PAC learnability: realizable, agnostic, improper
- 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?