CS 57800 – Statistical Machine Learning

Lecture 0 (08/24)

https://ruizhezhang.com/course_fall_2026.html

Course Staff

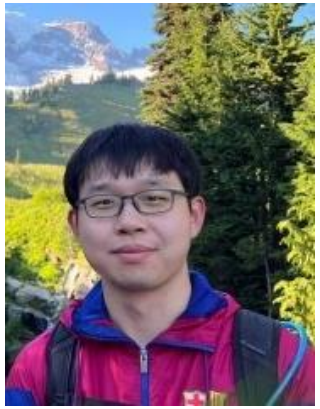
Ruizhe Zhang

- Assistant Professor, Purdue CS
- Research: Theory—Quantum Computing, Optimization, ML Theory
- **Office hours: Monday 10:30 AM - 11:30 AM at DSAI 1119B**

Teaching assistants

Shreya Nasa

Office hours:
Thu 9:00 AM – 10:00 AM
at DSAI B061



Yucheng Zhang

Office hours:
TBD

Outline

- Logistics
- Motivations
- Intro to Learning Theory

Course Instruction

- Course website: https://ruizhezhang.com/course_fall_2026.html
  - Slides and additional references

- Brightspace: https://purdue.brightspace.com/d2l/home/1631642
  - Announcements and problem sets submissions

- Ed Discussion for discussions
  - Access via the link in Brightspace

Grading

- **30%** Problem sets
  - 8 problem sets in total and 4 points for each
  - Must be typed in LaTeX and submit the PDFs
  - I may randomly choose problems for grading. Full solutions will be posted at Brightspace
- **30%** Two midterms
  - 15% each; in-class, closed-book, and pen-and-paper
  - Dates will be announced later
- **40%** Final project

Final Project

- **Goal:** Gain first-hand experience with (theoretical) machine learning research and develop basic scientific research skills

- **Theme:** Investigate a research question in statistical machine learning not answered by the literature, and make a focused, technically substantive, verifiable contribution

- **Collaboration:** Work in teams of two or three students

- **Timeline:**
  Proposal ⟶ Mid-semester check-in (ungraded) ⟶ Final report ⟶ Presentation

AI Policy for This Course

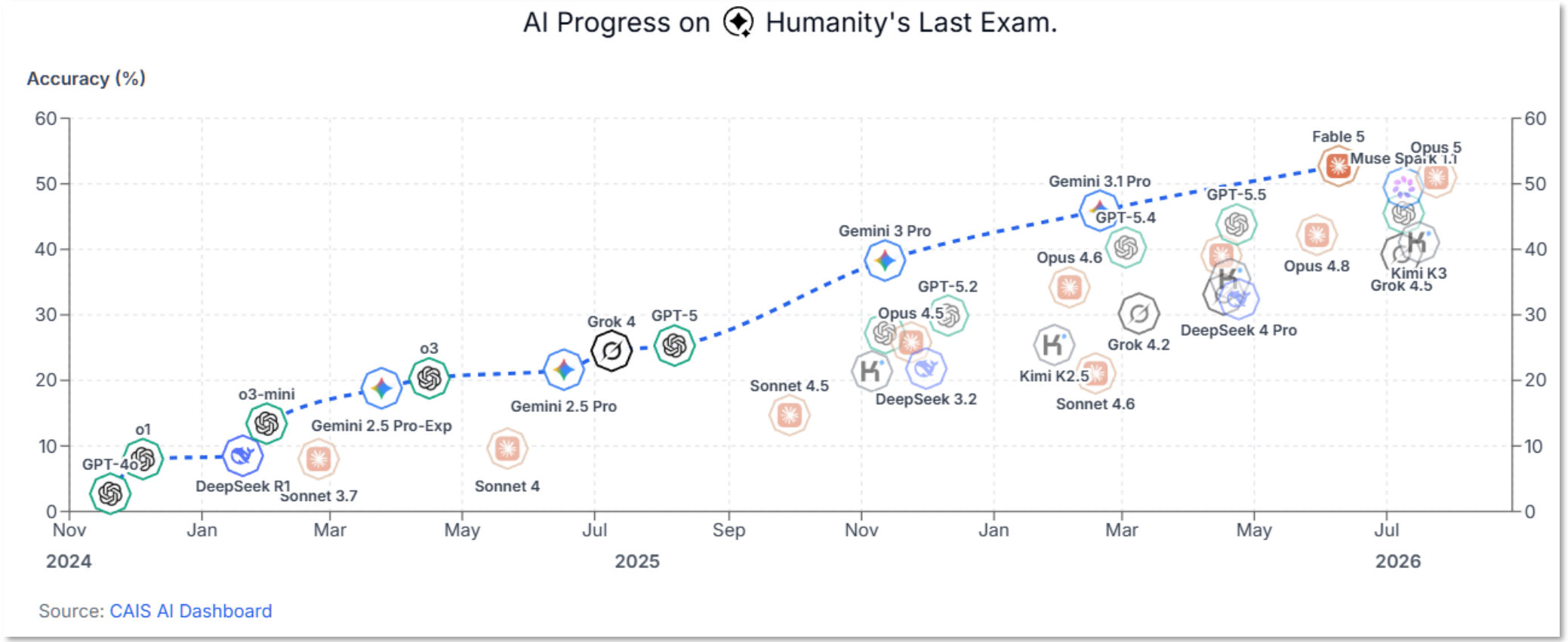
- Generative AI is **welcome as a research tool** throughout this course
- You remain responsible for **every claim, citation, proof, computation, figure, and line of code** you submit
- Your submissions should include a brief statement of **how AI was used** and **how its outputs were checked**

Outline

- Logistics
- Motivations
- Intro to Learning Theory

Why Statistical Machine Learning in the LLM Era?

"Frontier models can implement, tune, and explain every algorithm—why spend a semester learning the ML theory & algorithms?"



Why Statistical Machine Learning in the LLM Era?

"Frontier models can implement, tune, and explain every algorithm—why spend a semester learning the ML theory & algorithms?"

Start by Conceding the Premise

Frontier models can already …

- implement standard learning algorithms
- generate experiments and visualizations
- tune models and hyperparameters
- explain concepts and proofs in textbooks

So only learning how to **implement** the algorithms is no longer enough

Start by Conceding the Premise

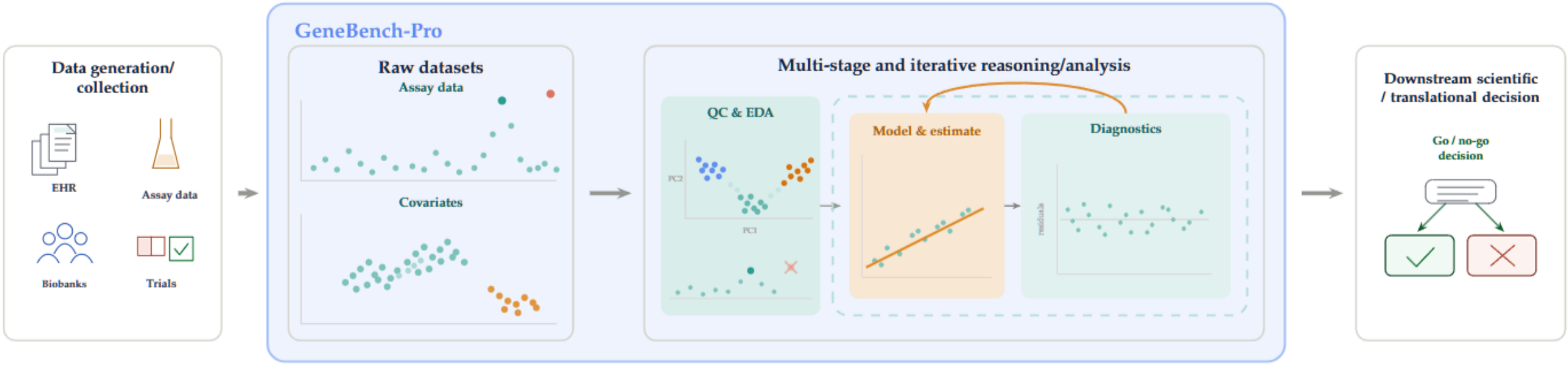
Frontier models can already …

- implement standard learning algorithms
- generate experiments and visualizations
- tune models and hyperparameters
- explain concepts and proofs in textbooks

GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine

Jeremy Li†, Andrew Ho†
OpenAI

June 30, 2026



Start by Conceding the Premise

Frontier models can already …

- implement standard learning algorithms
- generate experiments and visualizations
- tune models and hyperparameters
- explain concepts and proofs in textbooks

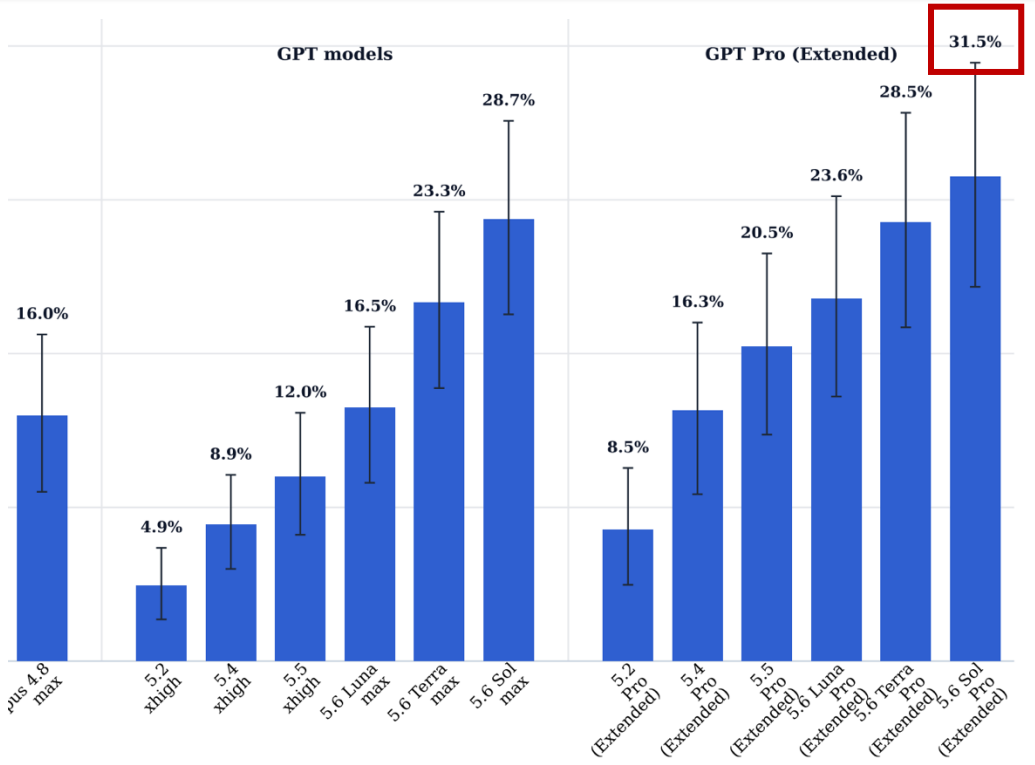
So only learning how to **implement** the algorithms is no longer enough

The scarce skill is moving from implementation to **formulation**, **evaluation**, **diagnosis**, and **judgment**

GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine

Jeremy Li†, Andrew Ho†
OpenAI

June 30, 2026



There Is No Task-Independent "Best Classifier"

"Best" depends on the decision problem:

- deployment population and sampling process
- meaning of the label and available information
- loss function and asymmetric error costs
- privacy, latency, and computational constraints

Cost-sensitivity Bayes rule:

$$f_\star(x) = \mathbf{1}\left\{\eta(x) \geq \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}\right\}$$

Rare-event example

Prevalence 0.1%; sensitivity = specificity = 99%; 100,000 cases

	Predict +	**Predict –**
Actual +	99 True Positive	1 False Negative
Actual –	999 False Positive	98,901 True Negative

- The classifier achieves 99% accuracy on positive and negative examples
- Only $99/(99 + 999) \approx 9\%$ of positive predictions are correct

Models optimize specified objectives; **theory** asks whether we specified the right one.

AI Makes Search Cheap; Validation Data Remains Finite

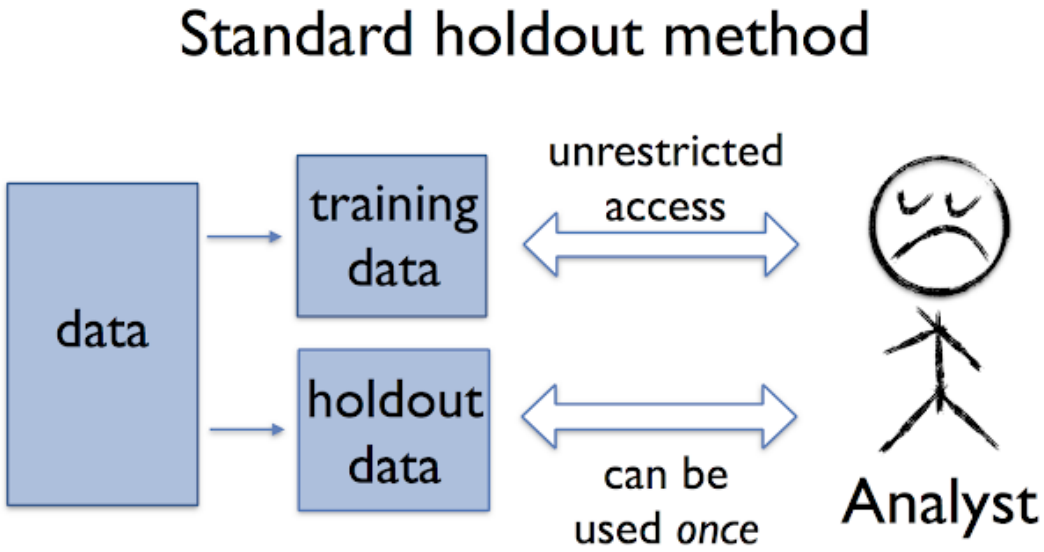
Suppose an AI generates M candidate models and then choose the best using n holdout examples

We can prove:

$$\max_{1 \le i \le M} \left| \mathcal{R}(f_i) - \hat{\mathcal{R}}_{\text{val}}(f_i) \right| \le O\left(\sqrt{\log(M)/n}\right)$$

Population risk

Empirical risk



More candidates increase the statistical cost of selection

AI Makes Search Cheap; Validation Data Remains Finite

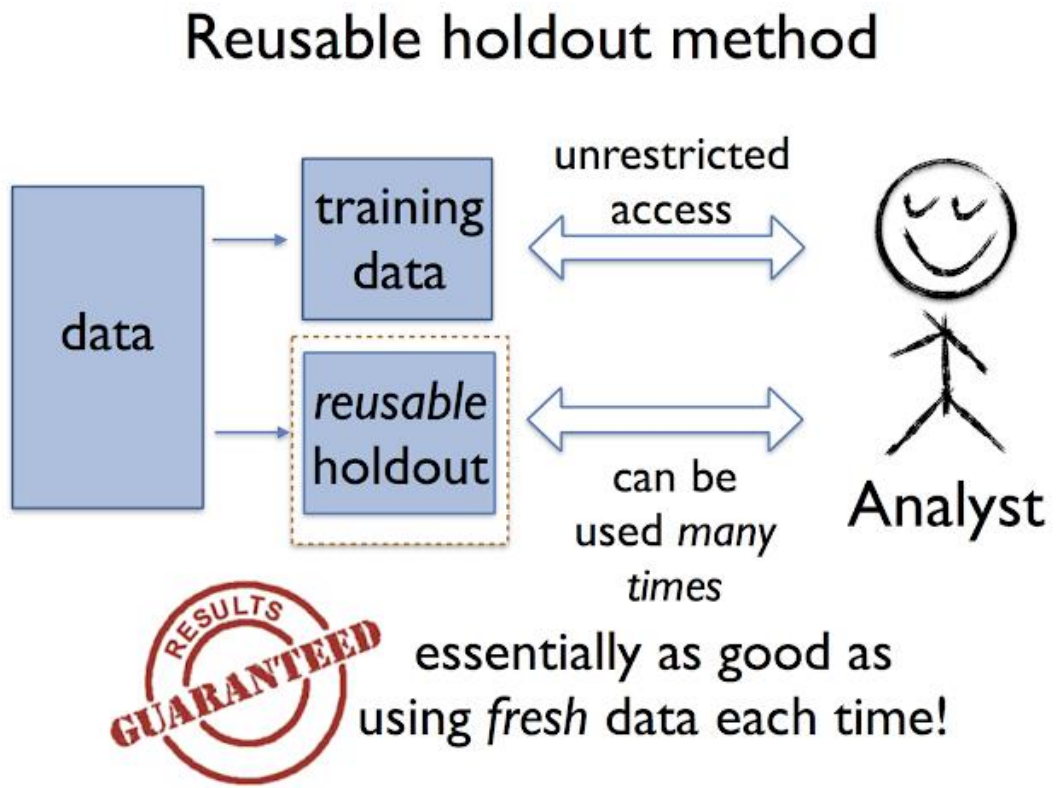
Suppose an AI generates M candidate models and then choose the best using n holdout examples

We can prove:

$$\max_{1 \le i \le M} \left| \mathcal{R}(f_i) - \hat{\mathcal{R}}_{\text{val}}(f_i) \right| \le O\left(\sqrt{\log(M)/n}\right)$$

Population risk — **Empirical risk**

More candidates increase the statistical cost of selection



Statistical theory can diagnose failure and supply a mechanism that changes the pipeline

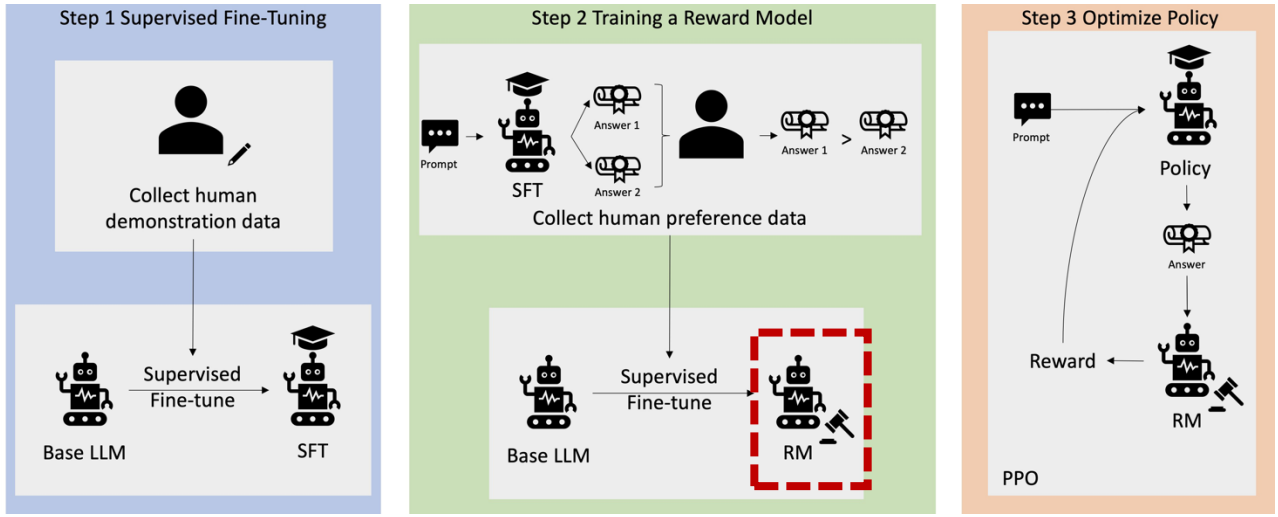
The reusable holdout: Preserving validity in adaptive data analysis

CYNTHIA DWORK, VITALY FELDMAN, MORITZ HARDT, TONIANN PITASSI, OMER REINGOLD, AND AARON ROTH Authors Info & Affiliations

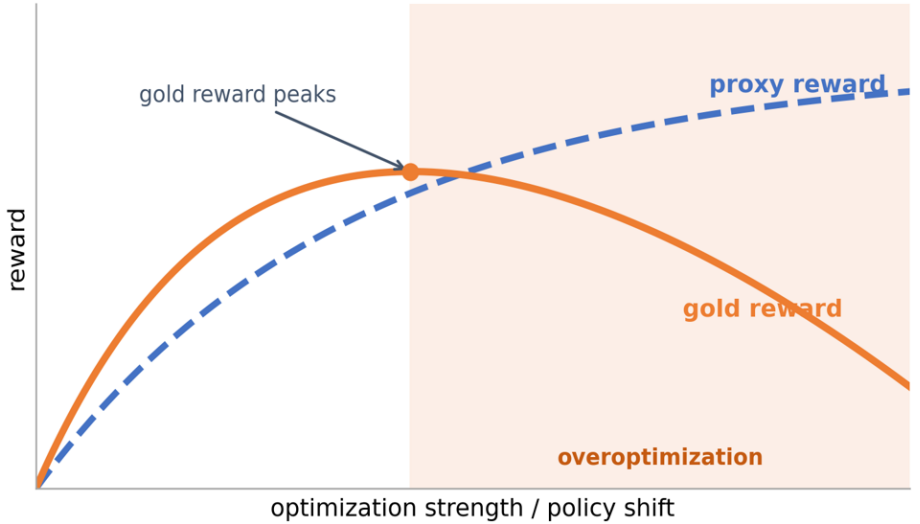
SCIENCE · 7 Aug 2015 · Vol 349, Issue 6248 · pp. 636-638 · DOI: 10.1126/science.aaa9375

Optimization Can Exploit Errors in a Learned Objective

RLHF workflow



Reward model overoptimization



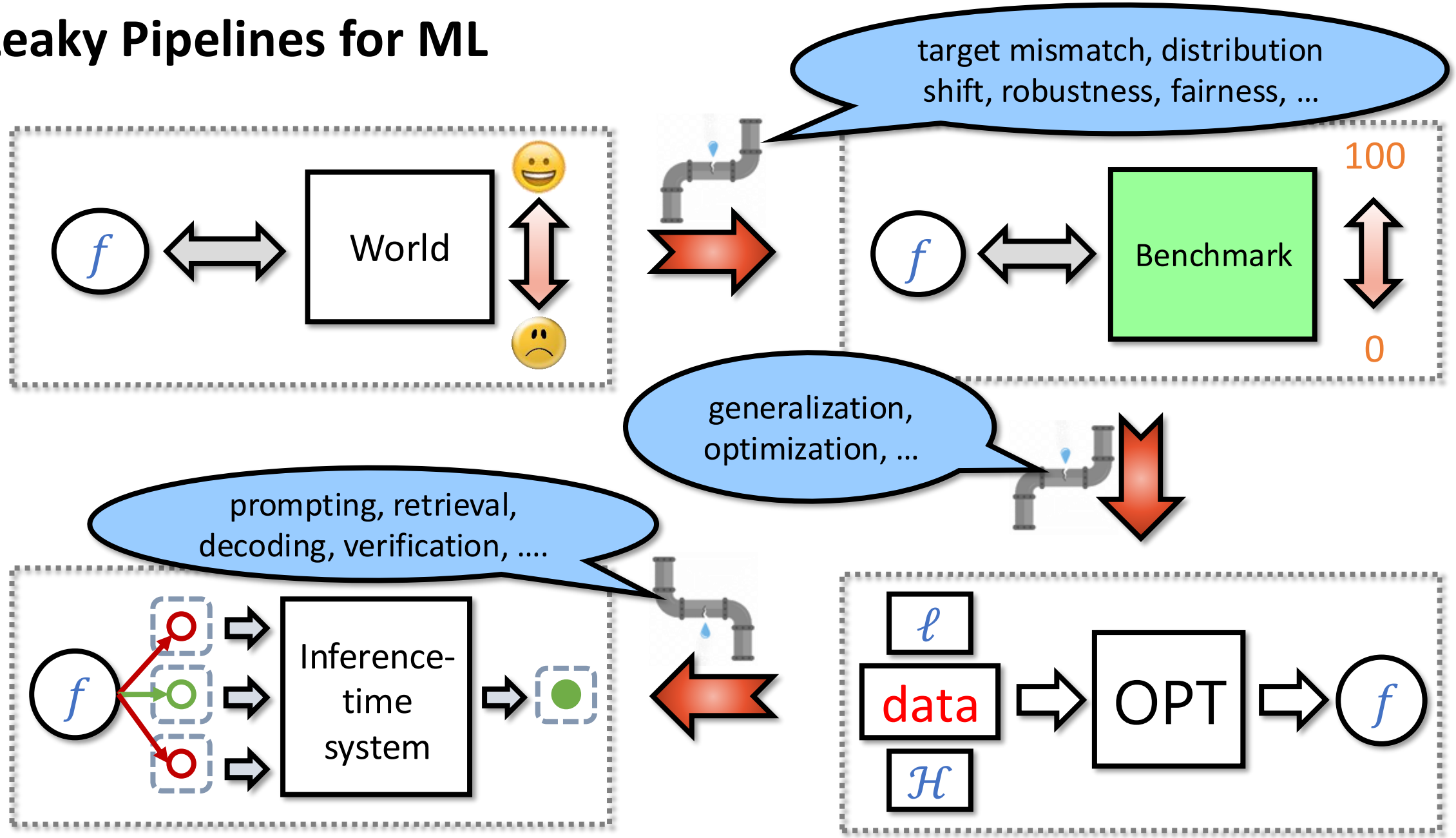
A reward model (RM) is only a proxy for true human values

Scaling Laws for Reward Model Overoptimization

Leo Gao, John Schulman, Jacob Hilton *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:10835-10866, 2023.

A stronger optimizer can amplify objective misspecification.

Leaky Pipelines for ML



Adapted from Boaz Barak, "A blitz through classical statistical learning theory" (2021).

What This Course Is About

GENERALIZATION • INFERENCE • GENERATION • RELIABILITY

Statistical learning theory

PAC learning, ERM, Linear regression, …

Unsupervised and latent-variable learning

Clustering, method of moments, identifiability, …

Probabilistic inference and statistical physics

Energy-based models, variational methods, MCMC, mean-field, …

Diffusion-model theory

Langevin dynamics, score estimation, SDE discretization, …

Modern neural-network theory

NTK, edge of stability, in-context learning, …

Trustworthy ML

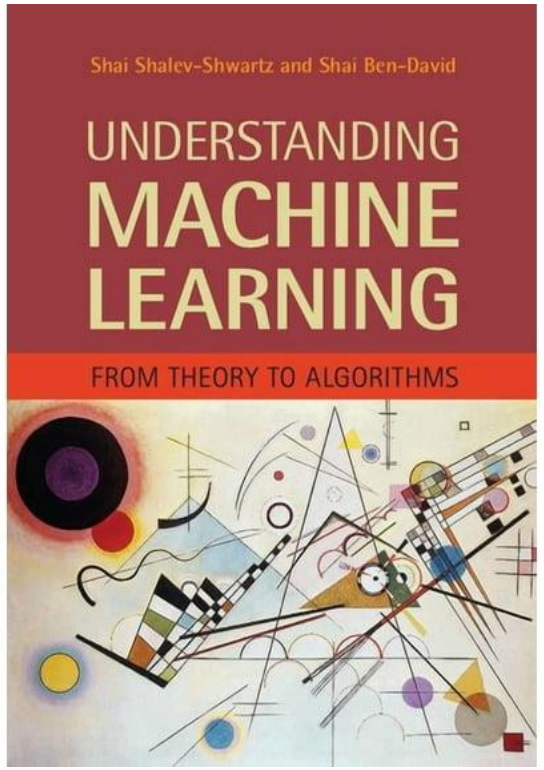
Watermarking, alignment, differential privacy, …

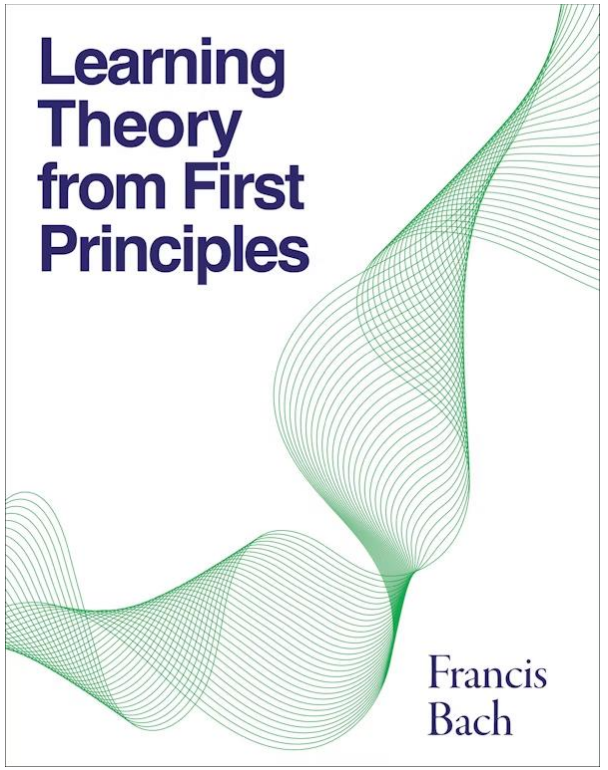
Concluding Remarks

- This course will be challenging; We are here to work together
- We will collaboratively learn from each other
- **Don’t hesitate to stop me at any point to ask questions**

Outline

- Logistics
- Motivations
- Intro to Learning Theory





The Statistical Learning Framework

- Domain set $\mathcal{X}$: images, sounds, text, DNA sequences, web pages, …
- Label set $\mathcal{Y}$: for binary classification, $\mathcal{Y} = \{0,1\}$ or $\{-1,1\}$; for multicategory classification, $\mathcal{Y} = \{1,2,\dots,k\}$; and for classical regression, $\mathcal{Y} = \mathbb{R}$
- Training data $S = \big((x_1, y_1), \dots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$
- Data-generation model: a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
  - Note that y_i may not be a deterministic function of x_i. E.g., $y = f(x) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, or $y = f(x, z)$ with unobserved z
  - $\mathcal{D}$ is unknown to the learner
- Machine learning algorithm A: a function that maps the training data S to a function (prediction rule/predictor/classifier/hypothesis) from $\mathcal{X}$ to $\mathcal{Y}$

The Statistical Learning Framework

- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - 0-1 loss $\ell(y, z) = \mathbf{1}\{y = z\}$
  - square loss $\ell(y, z) = (y - z)^2$

- Expected risk / Population loss / Generalization error / Test error of f:

$$\mathcal{R}(f) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$$

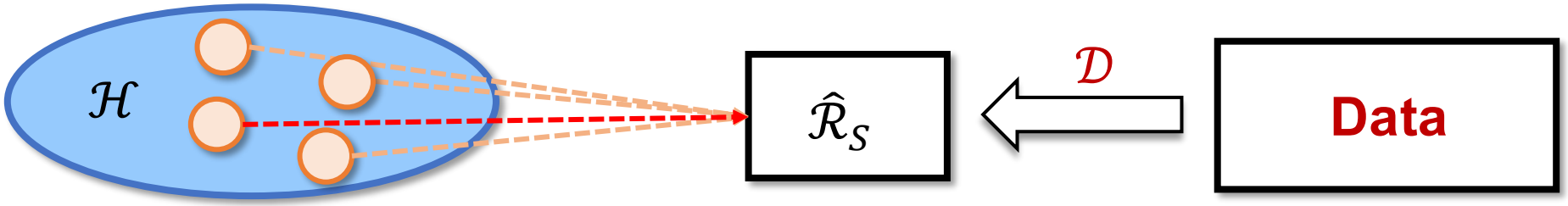
- Empirical risk / Empirical loss / Training error of f:

$$\hat{\mathcal{R}}_S(f) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)$$

Empirical Risk Minimization

- Let $\mathcal{H}$ be a hypothesis class $\{f : \mathcal{X} \to \mathcal{Y}\}$
- An ERM learner aims at finding a function $f \in \mathcal{H}$ that minimizes the empirical risk:

$$\mathrm{ERM}_{\mathcal{H}}(S) \in \arg\min_{f \in \mathcal{H}} \hat{R}_S(f)$$



Example: ordinary linear regression

- $\mathcal{H} = \{f_\theta(x) = \theta^\top \phi(x)\}$ where $\phi(\cdot)$ is a fixed feature map $\mathcal{X} \to \mathbb{R}^d$
- ERM is the least squares minimization:

$$\min_\theta \frac{1}{m} \sum_{i=1}^{m} (\theta^\top \phi(x_i) - y_i)^2$$

Empirical Risk Minimization

The Bayes optimal predicator:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x),y)] = \mathbb{E}\big[\mathbb{E}[\ell(f(x),y)|x]\big]$$
$$= \int_{\mathcal{X}} \underbrace{\mathbb{E}[\ell(f(x'),y)|x=x']}_{\text{conditional risk}}\, dp(x')\,,$$

where p is the density of the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$

- To minimize the expected risk $\mathcal{R}(f)$, we just need to minimize the conditional risk for each $x' \in \mathcal{X}$

$$f_\star(x') \coloneqq \arg\min_{z\in\mathcal{Y}} \mathbb{E}[\ell(z,y)|x=x']\,, \qquad \forall x' \in \mathcal{X}$$

- The Bayes risk is

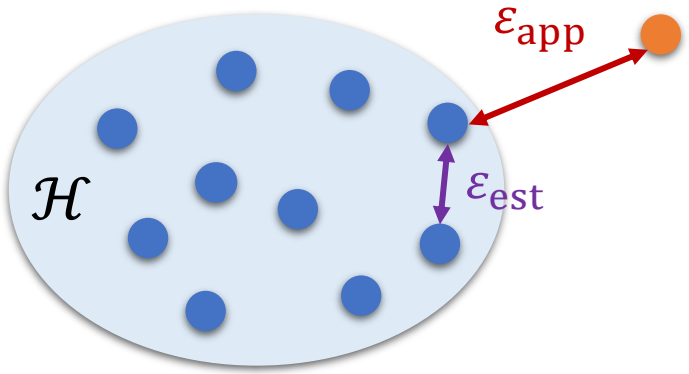
$$\mathcal{R}_\star \coloneqq \mathcal{R}(f_\star) = \int_{\mathcal{X}} \min_{z\in\mathcal{Y}} \mathbb{E}[\ell(z,y)|x=x']\, dp(x')$$

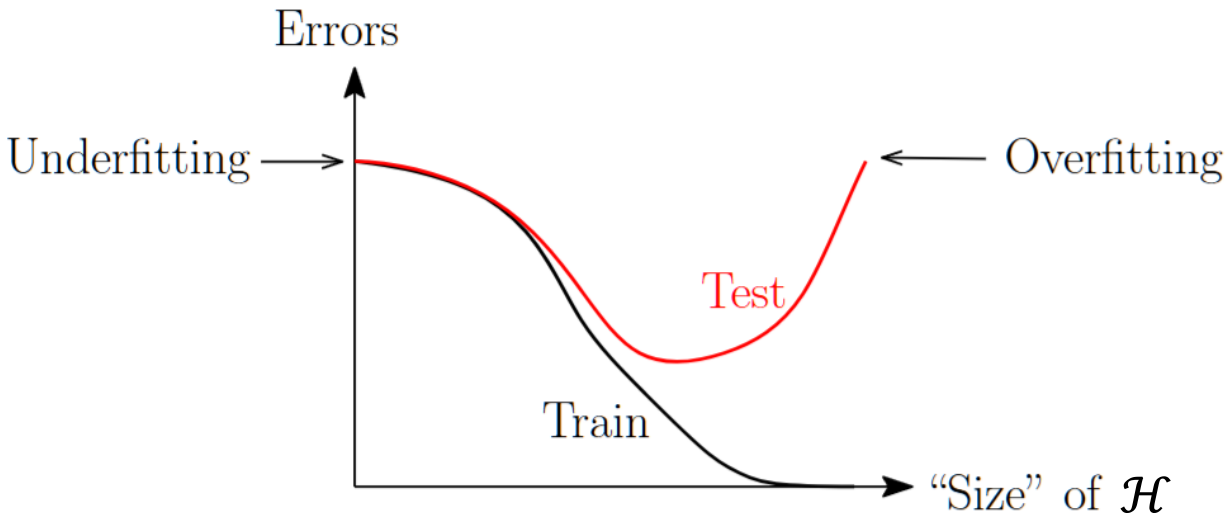
- **Exercise:** Show that the Bayes optimal predictor $f_\star$ is optimal, i.e., $\mathcal{R}_\star \le \mathcal{R}(f)$ for any $f\colon \mathcal{X}\to\mathcal{Y}$

Empirical Risk Minimization

Risk decomposition:

$$R(f) - \mathcal{R}_\star = \underbrace{\min_{h \in \mathcal{H}} R(h) - \mathcal{R}_\star}_{\text{approximation error}} + \underbrace{R(f) - \min_{h \in \mathcal{H}} R(h)}_{\text{estimation error}}$$





Toy Example: Polynomial Regression

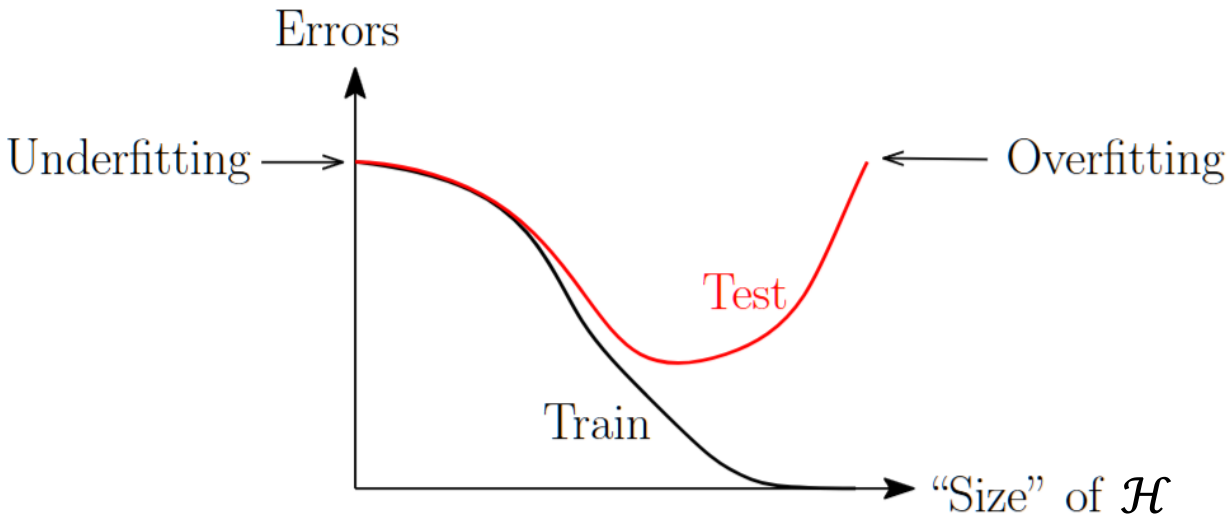
The goal is to find the best degree-P polynomial approximation of the function

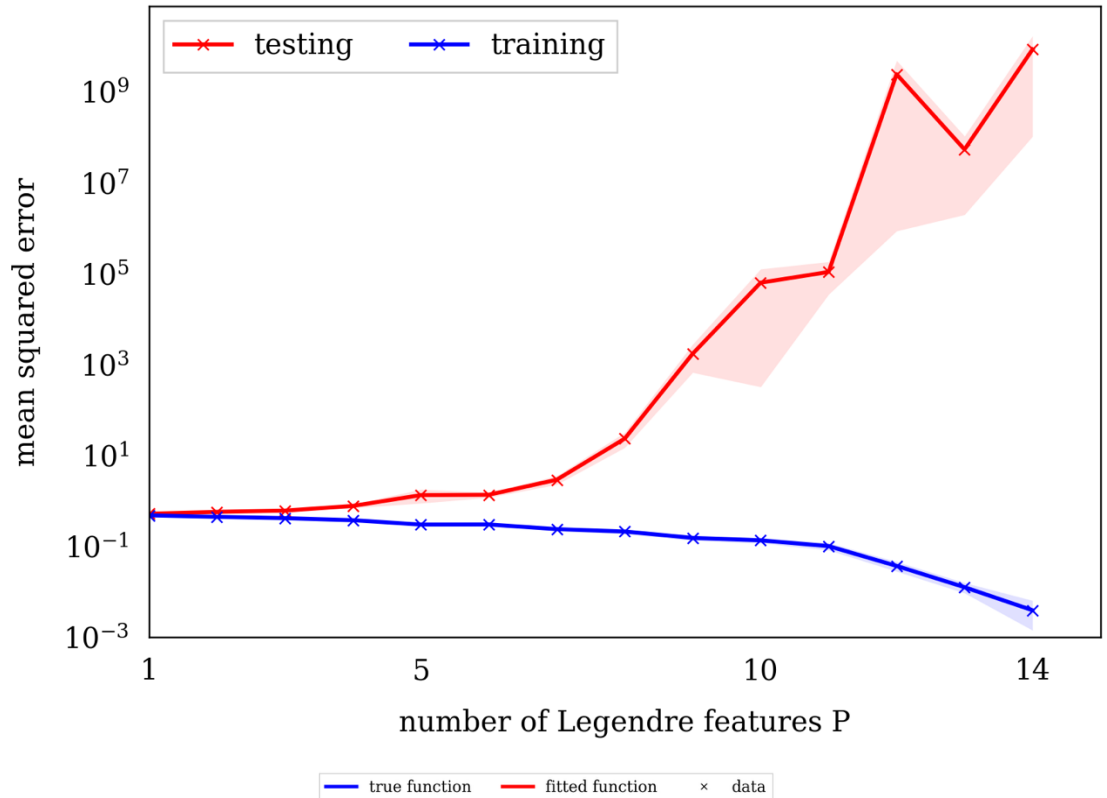
$$f(x) = 2x + \cos(25x)$$

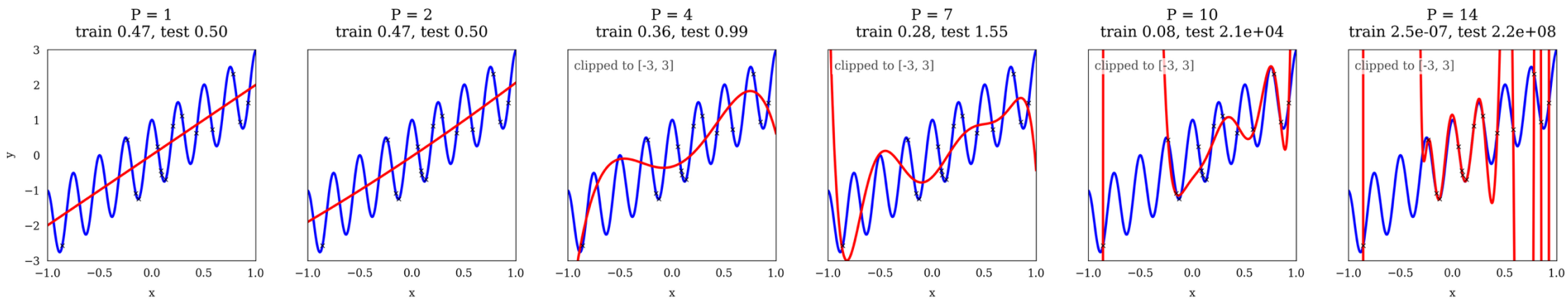
- Training data: $\{(x_i, f(x_i))\}_{i \le N}$, each $x_i \in [-1,1]$ is an i.i.d. uniform sample
- Hypothesis class:

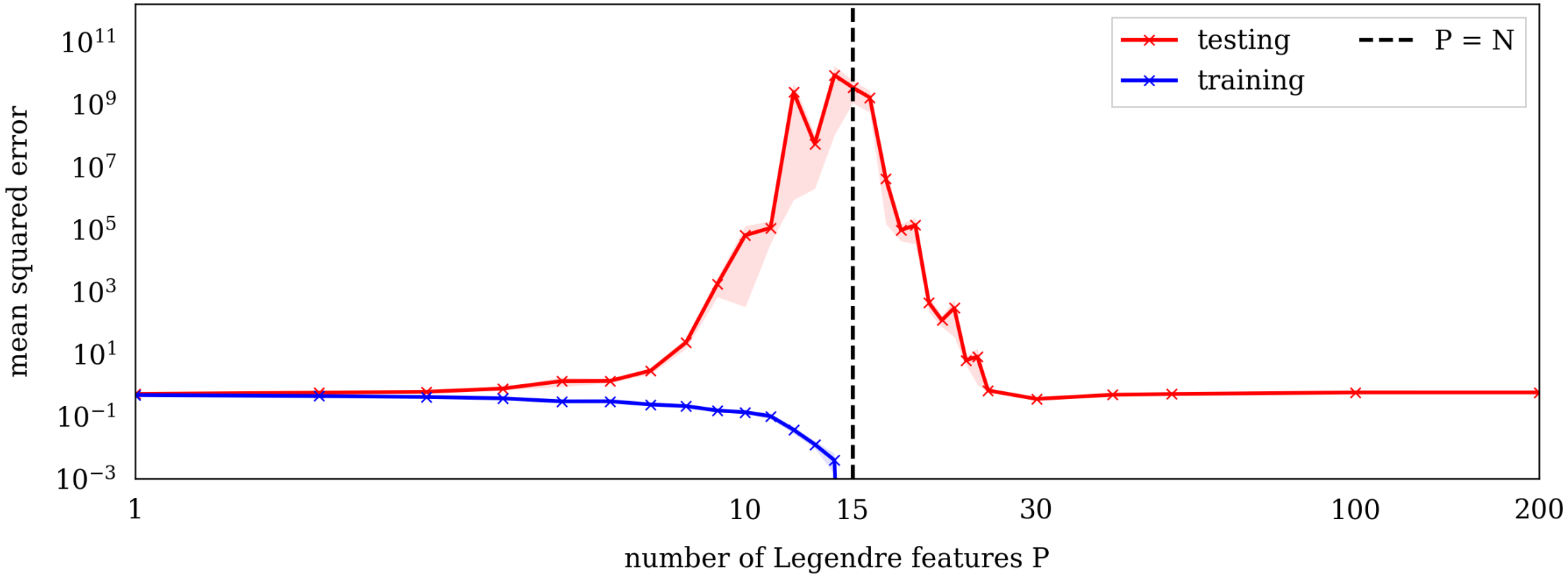
$$\mathcal{H} = \{\theta \in \mathbb{R}^P : \theta^\top \phi(x)\}, \qquad \phi(x) = \big(L_1(x), L_2(x), \ldots, L_P(x)\big)$$

Legendre polynomials




testing
training
mean squared error
10^{9}
10^{7}
10^{5}
10^{3}
10^{1}
10^{-1}
10^{-3}
1
5
10
14
number of Legendre features P


true function
fitted function
data
P = 1
train 0.47, test 0.50
P = 2
train 0.47, test 0.50
P = 4
train 0.36, test 0.99
P = 7
train 0.28, test 1.55
P = 10
train 0.08, test 2.1e+04
P = 14
train 2.5e-07, test 2.2e+08
clipped to [-3, 3]
x
y



(Schaeffer et al., arXiv:2303.14151)

PAC Learning

Probably approximately correct (PAC) learning

A hypothesis class $\mathcal{H}$ is (agnostic) PAC learnable if there exist a function $m_{\mathcal{H}}$ and a learning algorithm A such that for any $\delta \in (0,1)$ and $\varepsilon > 0$, for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. samples S generated by $\mathcal{D}$, the algorithm returns a hypothesis $A(S) : \mathcal{X} \to \mathcal{Y}$ satisfying

$$\Pr_{S \sim \mathcal{D}^m}\left[\mathcal{R}\big(A(S)\big) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq \varepsilon\right] \geq 1 - \delta$$

- Today's PAC learning definition is slightly different from the original one; see (Hanneke-Mehrotra-Velegkas-Zampetakis, arXiv:2605.13840)
- PAC learning has strong connections to many subfields of theoretical computer science apart from learning, including property testing, algorithmic game theory, computational complexity, statistical algorithms, and cryptography; see (Servedio, arXiv:2511.08791)

Leslie Valiant
Turing Award Laureate

PAC Learning

Given a hypothesis class $\mathcal{H}$, is it PAC learnable? How many samples do we need?

The Realizability Assumption: there exists an $h \in \mathcal{H}$ such that for any $(x, y) \sim \mathcal{D}$ satisfies $y = h(x)$. In other words, $\mathcal{R}(h) = 0$

Proposition. Every finite hypothesis class is PAC learnable with sampling complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}\right)$$

- There exists a learner A such that for any $m \geq \varepsilon^{-1} \log(|\mathcal{H}|/\delta)$,

$$\Pr_{S \sim \mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right) \leq \epsilon\right] \geq 1 - \delta$$

- With high probability over $S \sim \mathcal{D}^m$, the empirical risk minimization satisfies

Generalization bound $\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \underbrace{\hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)}_{=0} \leq O\left(\frac{\log(|\mathcal{H}|)}{m}\right)$

PAC Learning

Proposition. Every finite hypothesis class is PAC learnable with sampling complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}\right)$$

Proof.

- It suffices to prove that, for $m \geq \varepsilon^{-1} \log(|\mathcal{H}|/\delta)$, $\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) > \epsilon\right] \leq \delta$
- Let $\mathcal{H}_B \coloneqq \{f \in \mathcal{H} : \mathcal{R}(f) > \varepsilon\}$ be the set of "bad" hypothesis. That is,

$$\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) > \epsilon\right] = \Pr_{S\sim\mathcal{D}^m}[\mathrm{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B]$$

- Let $\mathcal{M} \coloneqq \left\{S : \exists f \in \mathcal{H}_B,\ \hat{\mathcal{R}}_S(f) = 0\right\}$ be the set of "misleading" datasets. Then, we have

$$\Pr_{S\sim\mathcal{D}^m}[\mathrm{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B] \leq \Pr_{S\sim\mathcal{D}^m}[S \in \mathcal{M}] = \Pr_{S\sim\mathcal{D}^m}\left[\exists f \in \mathcal{H}_B, \hat{\mathcal{R}}_S(f) = 0\right]$$

(Union bound)

PAC Learning

Proposition. Every finite hypothesis class is PAC learnable with sampling complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}\right)$$

Proof.

$$\begin{aligned}
\Pr_{S\sim\mathcal{D}^m}[\mathrm{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B] &\leq \sum_{f\in\mathcal{H}_B} \Pr_{S\sim\mathcal{D}^m}[\hat{\mathcal{R}}_S(f) = 0] = \sum_{f\in\mathcal{H}_B} \Pr_{S\sim\mathcal{D}^m}[\forall i \in [m], f(x_i) = y_i] \\
&= \sum_{f\in\mathcal{H}_B} \left(\Pr_{(x,y)\sim\mathcal{D}}[f(x) = y]\right)^m \\
&\leq \sum_{f\in\mathcal{H}_B} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[1 - \ell(f(x), y)]\right)^m \qquad \ell \text{ is the 0-1 loss or bounded loss} \\
&= \sum_{f\in\mathcal{H}_B} \left(1 - \mathcal{R}(f)\right)^m \leq \sum_{f\in\mathcal{H}_B} (1-\varepsilon)^m \leq |\mathcal{H}_B| e^{-\varepsilon m} \leq \delta
\end{aligned}$$

■

PAC Learning

"No Free Lunch" Theorem

Consider binary classification with 0–1 loss and $\mathcal{X}$ infinite. For any $m > 0$ and any learning algorithm A, there exists a probability distribution $\mathcal{D}$ such that

$$\mathcal{R}_\star = 0 \quad \text{and} \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right)\right] \geq 1/4$$

- No universal learner: for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner
- **Exercise:** Let $\mathcal{X}$ be infinite and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable

PAC Learning

"No Free Lunch" Theorem Consider binary classification with 0–1 loss and $\mathcal{X}$ infinite. For any $m > 0$ and any learning algorithm A, there exists a probability distribution $\mathcal{D}$ such that

$$\mathcal{R}_\star = 0 \quad \text{and} \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right)\right] \geq 1/4$$

Proof.

- Fix any $m > 0$. The idea is to construct a hard distribution $\mathcal{D}$ supported on $k \gg m$ data points such that the knowledge of the m training data point is not enough to predict the remaining $k - m$ data points
- Pick any k different elements $x_1, \ldots, x_k \in \mathcal{X}$
- For every $r \in \{0,1\}^k$, define a probability distribution $\mathcal{D}_r$:

$$\mathcal{D}_r\left((x_i, y)\right) := \begin{cases} 1/k & \text{if } y = r_i \\ 0 & \text{otherwise} \end{cases}, \qquad i \in [k], y \in \{0,1\}$$

- $\mathcal{R}_\star = 0$ for every $\mathcal{D}_r$

PAC Learning

"No Free Lunch" Theorem Consider binary classification with 0–1 loss and $\mathcal{X}$ infinite. For any $m > 0$ and any learning algorithm A, there exists a probability distribution $\mathcal{D}$ such that

$$\mathcal{R}_\star = 0 \quad \text{and} \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right)\right] \geq 1/4$$

Proof.

- Probabilistic method:

$$\max_{r\in\{0,1\}^k}\left\{\mathbb{E}_{S\sim\mathcal{D}_r^m}\left[\mathcal{R}\left(A(S)\right)\right]\right\} \geq \mathbb{E}_{r\sim\{0,1\}^k}\left[\mathbb{E}_{S\sim\mathcal{D}_r^m}\left[\mathcal{R}\left(A(S)\right)\right]\right]$$
$$= \mathbb{E}_{r,S}\left[\Pr_{i\sim[k]}[A(S)(x_i) \neq r_i]\right] = \Pr_{r,S,i}[A(S)(x_i) \neq r_i]$$

- *To prove the existence of some object with certain property, we devise a random construction and show that it works with positive probability. (Here, we use $Pr\left[X \geq \mathbb{E}[X]\right] > 0$)*

PAC Learning

"No Free Lunch" Theorem Consider binary classification with 0–1 loss and $\mathcal{X}$ infinite. For any $m > 0$ and any learning algorithm A, there exists a probability distribution $\mathcal{D}$ such that

$$\mathcal{R}_\star = 0 \quad \text{and} \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right)\right] \geq 1/4$$

Proof.

- Let $S = \left(x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right)$

$$\begin{aligned}
\Pr_{r,S,i}\left[A(S)(x_i) \neq r_i\right] &= \mathbb{E}\left[\Pr_{r,i}\left[A(S)(x_i) \neq r_i \,\middle|\, x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right]\right] \\
&\geq \mathbb{E}\left[\Pr_{r,i}\left[A(S)(x_i) \neq r_i \;\&\; x_i \notin \left\{x_{i_1}, \ldots, x_{i_m}\right\} \middle| x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right]\right] \\
&\geq \mathbb{E}\left[\frac{1}{2} \cdot \Pr_{i}\left[x_i \notin \left\{x_{i_1}, \ldots, x_{i_m}\right\} \middle| x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right]\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[1 - \frac{\left|\left\{x_{i_1}, \ldots, x_{i_m}\right\}\right|}{k}\right] \geq \frac{1}{2}\left(1 - \frac{m}{k}\right)
\end{aligned}$$

PAC Learning

"No Free Lunch" Theorem Consider binary classification with 0–1 loss and $\mathcal{X}$ infinite. For any $m > 0$ and any learning algorithm A, there exists a probability distribution $\mathcal{D}$ such that

$$\mathcal{R}_\star = 0 \quad \text{and} \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(A(S)\right)\right] \geq 1/4$$

Proof.

- Thus, we get that

$$\max_{r\in\{0,1\}^k}\left\{\mathbb{E}_{S\sim\mathcal{D}_r^m}\left[\mathcal{R}\left(A(S)\right)\right]\right\} \geq \frac{1}{2}\left(1-\frac{m}{k}\right) \geq \frac{1}{4}$$

  when $k \geq 2m$

■

PAC Learning

- We have shown that under the realizability assumption ($\exists h \in \mathcal{H}, \mathcal{R}(h) = 0$), every finite $\mathcal{H}$ is PAC learnable
- We can remove this assumption:

Theorem. Let $\mathcal{H}$ be a finite class, and ℓ be the loss function bounded in $[0,1]$. Then, $\mathcal{H}$ is agnostic PAC learnable using the ERM algorithm with sample complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

- A byproduct is the Uniform Convergence property: for any $\mathcal{D}$, with high probability over $S \sim \mathcal{D}^m$,

$$\max_{f \in \mathcal{H}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right| \leq O\left(\sqrt{\frac{\log(|\mathcal{H}|)}{m}}\right)$$

PAC Learning

Theorem. Let $\mathcal{H}$ be a finite class, and ℓ be the loss function bounded in $[0,1]$. Then, $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

Proof.

- First, we want to prove that for $m \geq m(\varepsilon, \delta)$,

$$\Pr_{S \sim \mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \geq \varepsilon\right] \leq \delta$$

- Let $h \coloneqq \arg\min_{f \in \mathcal{H}} \mathcal{R}(f)$. We have the following error decomposition:

$$\begin{aligned}
&\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \mathcal{R}(h) \\
&= \left(\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right) + \overbrace{\left(\hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S(h)\right)}^{\leq 0} + \left(\hat{\mathcal{R}}_S(h) - \mathcal{R}(h)\right) \\
&\geq \left(\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right) + \left(\hat{\mathcal{R}}_S(h) - \mathcal{R}(h)\right)
\end{aligned}$$

PAC Learning

Theorem. Let $\mathcal{H}$ be a finite class, and ℓ be the loss function bounded in $[0,1]$. Then, $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

Proof.

$$\begin{aligned}
\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \mathcal{R}(h) \geq \varepsilon\right] &\leq \Pr_{S\sim\mathcal{D}^m}\left[\left(\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right) + \left(\hat{\mathcal{R}}_S(h) - \mathcal{R}(h)\right) \geq \varepsilon\right] \\
&\leq \Pr_{S\sim\mathcal{D}^m}\left[\exists f \in \mathcal{H}, \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right| \geq \varepsilon/2\right] \\
&\leq \sum_{f\in\mathcal{H}} \Pr_{S\sim\mathcal{D}^m}\left[\left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right| \geq \varepsilon/2\right]
\end{aligned}$$

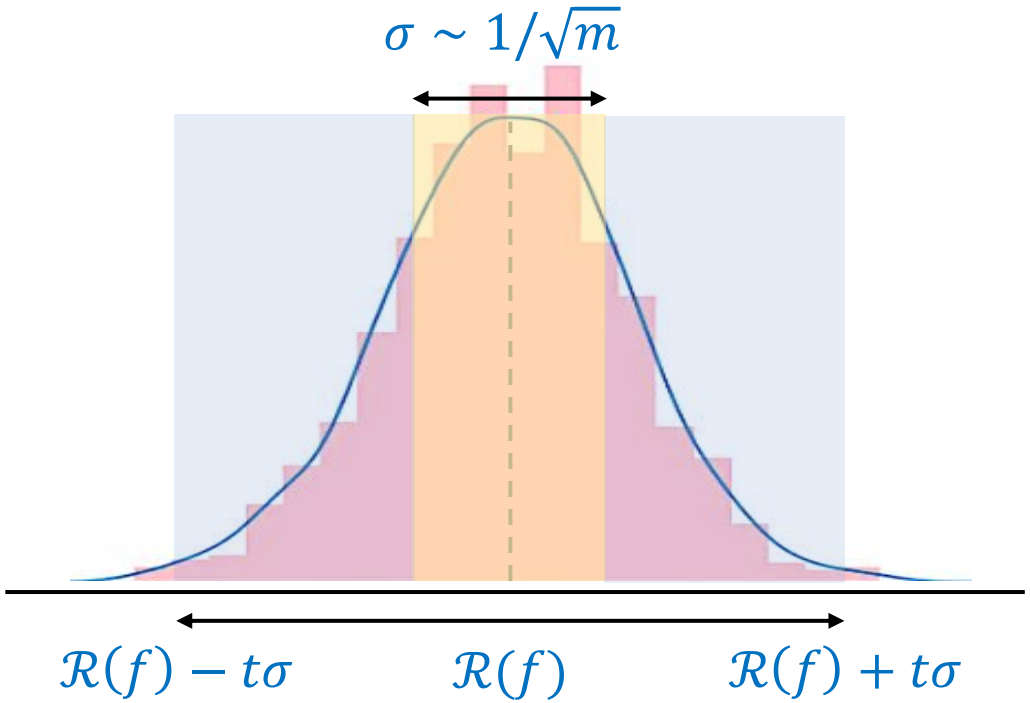
Random variable

- Note that

$$\hat{\mathcal{R}}_S(f) = \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i), \qquad \mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)], \qquad \mathbb{E}\left[\hat{\mathcal{R}}(f)\right] = \mathcal{R}(f)$$

Detour: Concentration

Central limit theorem / Hoeffding / Chernoff / …



$$\Pr[|X - \mathbb{E}[X]| \geq t\sigma] \leq \exp\left(-\Omega(t^2)\right)$$

Hoeffding's inequality

Let $X_1, \ldots, X_m$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$. Then,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

https://ruizhezhang.com/teaching_2026/Lecture%203.pdf

PAC Learning

Theorem. Let $\mathcal{H}$ be a finite class, and ℓ be the loss function bounded in $[0,1]$. Then, $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

Proof.

- By Hoeffding's inequality,

$$\Pr_{S\sim\mathcal{D}^m}\left[\left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right| \geq \varepsilon/2\right] \leq \exp\left(-\Omega(m\varepsilon^2)\right)$$

- Thus,

$$\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \inf_{f\in\mathcal{H}} \mathcal{R}(f) \geq \varepsilon\right] \leq |\mathcal{H}| \cdot \exp\left(-\Omega(m\varepsilon^2)\right) \leq \delta$$

  as long as

$$m \geq \Omega\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

■

PAC Learning

Theorem. Let $\mathcal{H}$ be a finite class, and ℓ be the loss function bounded in $[0,1]$. Then, $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m(\varepsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$$

What about infinite class $|\mathcal{H}| = \infty$?

- No Free Lunch theorem tells that $\mathcal{H}$ cannot be all functions from an infinite domain $\mathcal{X}$ to $\{0,1\}$
- **Exercise:** Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H}$ be the set of indicator functions of closed intervals:

  $$\mathbf{1}\{\mathrm{a} \leq x \leq b\}, \qquad \forall a \leq b$$

  Then $\mathcal{H}$ can be PAC-learned using ERM with sample complexity $O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$

The VC Dimension

Shattering: A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$, i.e., for $C = \{c_1, \dots, c_m\}$,

$$\mathcal{H}_C := \{(h(c_1), \dots, h(c_m)) : h \in \mathcal{H}\} = \{(y_1, \dots, y_m) : y_i \in \{0,1\}\}$$

That is, $|\mathcal{H}_C| = 2^{|C|}$

	c_1	c_2	c_3	c_4	c_5
h_1	0	0	0	0	0
h_2	1	1	1	1	1
h_3	1	1	1	0	0
h_4	0	0	0	1	1

The VC Dimension

Shattering: A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$, i.e., for $C = \{c_1, \dots, c_m\}$,

$$\mathcal{H}_C \coloneqq \{(h(c_1), \dots, h(c_m)) : h \in \mathcal{H}\} = \{(y_1, \dots, y_m) : y_i \in \{0,1\}\}$$

That is, $|\mathcal{H}_C| = 2^{|C|}$

VC-dimension

The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size, we say $\mathrm{VCdim}(\mathcal{H}) = \infty$

- If $\mathrm{VCdim}(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC learnable

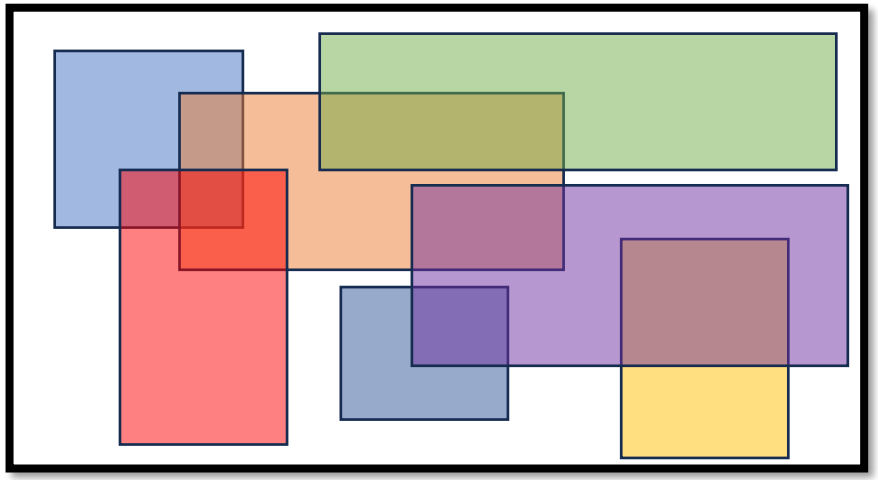
Alexei Chervonenkis and Vladimir Vapnik

The VC Dimension

The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

Examples:

- Finite $\mathcal{H}$: $2^{|C|} = |\mathcal{H}_C| \leq |\mathcal{H}| \implies \mathrm{VCdim}(\mathcal{H}) \leq \log|\mathcal{H}|$
- Axis aligned rectangles: $\mathrm{VCdim}(\mathcal{H}) = 4$



The VC Dimension

The Fundamental Theorem of Statistical Learning.

Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the 0−1 loss. Then, the following are equivalent:

1) $\mathcal{H}$ is agnostic PAC learnable
2) $\mathcal{H}$ is PAC learnable
3) $\mathcal{H}$ has finite VC dimension

- We already have *1) ⟹ 2)* and *2) ⟹ 3)*. It remains to show that *3) ⟹ 1)*
- It is very similar to the agnostic PAC learning theorem for finite $\mathcal{H}$:

  Let $\mathcal{H}$ be a finite class. Then, $\mathcal{H}$ is agnostic PAC learnable with sample complexity $O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$
- The VC dimension bounds the "effective size" of $\mathcal{H}$

The VC Dimension

Lemma (Sauer-Shelah-Perles). Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all m,

$$\tau_{\mathcal{H}}(m) \coloneqq \max_{C \subset \mathcal{X} : |C| \leq m} |\mathcal{H}_C| \leq \sum_{i=0}^{d} \binom{m}{i}$$

In particular, if $m \geq d + 1$, then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

- For any subset C, the "effective size" $|\mathcal{H}_C| = O\big(|C|^d\big)$ instead of $2^{|C|}$

- We only prove the uniform convergence property, which will give agnostic PAC learnability

Proposition (Uniform Convergence). Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ we have

$$\max_{f \in \mathcal{H}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}_S(f) \right| \leq \frac{\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta \sqrt{m}}$$

The VC Dimension

Proposition. Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every $\mathcal{D}$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ we have

$$\max_{f\in\mathcal{H}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right| \leq O\left(\frac{\sqrt{1+\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{m}}\right)$$

Proof.

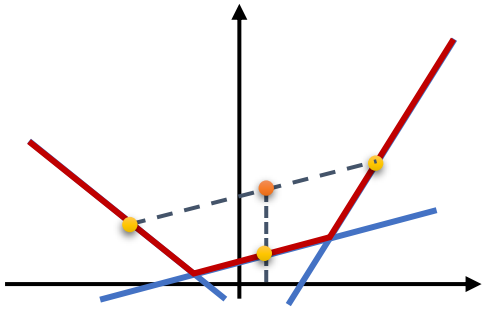
- It suffices to prove that $\mathbb{E}_S\left[\max_{f\in\mathcal{H}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right|\right] \leq O\left(\sqrt{\frac{1+\log \tau_{\mathcal{H}}(2m)}{m}}\right)$
- Recall that $\mathbb{E}_S\left[\hat{\mathcal{R}}_S(f)\right] = \mathcal{R}(f)$. We can use the reverse direction:

$$\begin{aligned}\mathbb{E}_S\left[\max_{f\in\mathcal{H}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right|\right] &= \mathbb{E}_S\left[\max_{f\in\mathcal{H}} \left|\mathbb{E}_{S'}\left[\hat{\mathcal{R}}_{S'}(f)\right] - \hat{\mathcal{R}}_S(f)\right|\right] \\ &\leq \mathbb{E}_S\left[\max_{f\in\mathcal{H}} \left\{\mathbb{E}_{S'}\left[\left|\hat{\mathcal{R}}_{S'}(f) - \hat{\mathcal{R}}_S(f)\right|\right]\right\}\right]\end{aligned}$$

The VC Dimension

- By Jensen's inequality,

$$\max_{f\in\mathcal{H}}\left\{\mathbb{E}_{S'}\left[\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|\right]\right\}\leq\mathbb{E}_{S'}\left[\max_{f\in\mathcal{H}}\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|\right]$$



- Thus,

$$\mathbb{E}_S\left[\max_{f\in\mathcal{H}}\left|\mathcal{R}(f)-\hat{\mathcal{R}}_S(f)\right|\right]\leq\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}}\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|\right]$$

- Fix $S,S'\sim\mathcal{D}^m$, and let $C\coloneqq S\cup S'$. Then,

$$\max_{f\in\mathcal{H}}\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|=\max_{f\in\mathcal{H}_C}\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|$$

$$\mathbb{E}_S\left[\max_{f\in\mathcal{H}}\left|\mathcal{R}(f)-\hat{\mathcal{R}}_S(f)\right|\right]\leq\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\hat{\mathcal{R}}_{S'}(f)-\hat{\mathcal{R}}_S(f)\right|\right]$$

$$=\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right|\right]$$

The VC Dimension

- Note the symmetry between S and S': for any $\sigma \in \{\pm 1\}^m$,

$$\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right|\right]=\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right)\right|\right]$$

- Thus, we can take expectation over a uniformly random $\sigma \sim \{\pm 1\}^n$:

$$\text{LHS}=\mathbb{E}_{S,S'}\mathbb{E}_{\sigma}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right)\right|\right]$$

- $\sigma_i\left(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right)$ is a mean-zero random variable in $\{-1,0,1\}$. By Hoeffding's inequality,

$$\Pr_{\sigma}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right)\right|\geq\epsilon\right]\leq\exp\left(-\Omega(m\epsilon^2)\right)$$

The VC Dimension

- Note the symmetry between S and S': for any $\sigma \in \{\pm 1\}^m$,

$$\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\right|\right]=\mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\big)\right|\right]$$

- Thus, we can take expectation over a uniformly random $\sigma \sim \{\pm 1\}^n$:

$$\text{LHS}=\mathbb{E}_{S,S'}\mathbb{E}_{\sigma}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\big)\right|\right]$$

- $\sigma_i\big(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\big)$ is a mean-zero random variable in $\{-1,0,1\}$. By Hoeffding's inequality,

$$\begin{aligned}\Pr_{\sigma}\left[\max_{f\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(f(x_i'),y_i')-\ell(f(x_i),y_i)\big)\right|\geq\epsilon\right]&\leq|\mathcal{H}_C|\exp\big(-\Omega(m\epsilon^2)\big)\\&\leq\tau_{\mathcal{H}}(2m)\exp\big(-\Omega(m\epsilon^2)\big)\end{aligned}$$

The VC Dimension

- Using the fact that for any nonnegative random variable X,

$$\mathbb{E}[X] = \int_0^\infty \Pr[X \geq t]\, dt\,,$$

we have

$$\mathbb{E}_\sigma \left[\max_{f \in \mathcal{H}_C} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i) \big) \right| \right]$$

$$= \int_0^\infty \Pr_\sigma \left[\max_{f \in \mathcal{H}_C} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i) \big) \right| \geq \epsilon \right] d\epsilon$$

$$\leq \int_0^\infty \min\left\{ 1, \tau_{\mathcal{H}}(2m) \exp\left(-\Omega(m\epsilon^2) \right) \right\} d\epsilon$$

$$= O\left(\frac{\sqrt{1 + \log \tau_{\mathcal{H}}(2m)}}{\sqrt{m}} \right)$$

The VC Dimension

∫24_0^∞▒ ⟦min{1, τ_H (2m) exp(-Ω(mϵ^2)) } dϵ⟧

Worked for 1m 55s

Let

$$A = \tau_{\mathcal{H}}(2m), \qquad \exp\bigl(-\Omega(m\epsilon^2)\bigr) \le e^{-cm\epsilon^2}$$

for some constant $c > 0$. Consider

$$I = \int_0^\infty \min\left\{1, Ae^{-cm\epsilon^2}\right\} d\epsilon.$$

The two terms inside the minimum are equal when

$$Ae^{-cm\epsilon_*^2} = 1 \quad \Longrightarrow \quad \epsilon_* = \sqrt{\frac{\log A}{cm}}.$$

Thus, for $A \ge 1$,

$$\begin{aligned} I &= \int_0^{\epsilon_*} 1\, d\epsilon + A\int_{\epsilon_*}^\infty e^{-cm\epsilon^2}\, d\epsilon \\ &= \sqrt{\frac{\log A}{cm}} + \frac{A\sqrt{\pi}}{2\sqrt{cm}}\,\operatorname{erfc}\left(\sqrt{\log A}\right). \end{aligned}$$

Using the Gaussian-tail bound

$$\int_x^\infty e^{-cmt^2}\, dt \le \frac{e^{-cmx^2}}{2cmx},$$

we obtain

$$I \le \sqrt{\frac{\log A}{cm}} + \frac{1}{2\sqrt{cm\log A}}.$$

Therefore, in a form valid even when A is close to 1,

$$\int_0^\infty \min\left\{1, \tau_{\mathcal{H}}(2m)e^{-\Omega(m\epsilon^2)}\right\} d\epsilon = O\left(\sqrt{\frac{1+\log \tau_{\mathcal{H}}(2m)}{m}}\right).$$

The VC Dimension

Lemma (Sauer-Shelah-Perles). Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all m,

$$\tau_{\mathcal{H}}(m) \coloneqq \max_{C \subset \mathcal{X} : |C| \leq m} |\mathcal{H}_C| \leq \sum_{i=0}^{d} \binom{m}{i}$$

The number of subsets in C of size $\leq d$

Proof.

- We claim that for any $\mathcal{H}$ and $C \subset \mathcal{X}$,
$$|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$$
- We prove this claim by induction. Suppose it holds for $k < m$
- Fix $\mathcal{H}$ and $C = \{c_1, \ldots, c_m\}$
- Define two sets:
$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$
$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

The VC Dimension

- Define two sets:

$$Y_0 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$
$$Y_1 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

- $|Y_0| + |Y_1| = |\mathcal{H}_C|$
- Since $Y_0 = \mathcal{H}_{C'}$ with $C' \coloneqq \{c_2, \dots, c_m\}$, we have

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subseteq C : c_1 \notin C \wedge \mathcal{H} \text{ shatters } B\}|$$

- Define $\mathcal{H}' \subset \mathcal{H}$ containing pairs of functions that only differ in c_1:

$$\mathcal{H}' \coloneqq \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } h(c_i) = h'(c_i)\ \forall 2 \leq i \leq m \wedge 1 - h(c_1) = h'(c_1)\}$$

- Since $Y_1 = \mathcal{H}'_{C'}$, we have

$$|Y_1| = \left|\mathcal{H}'_{C'}\right| \leq |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}| = \left|\left\{B \subseteq C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\right\}\right|$$
$$= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}|$$
$$\leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}|$$

The VC Dimension

- Thus,

$$
\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\leq |\{B \subseteq C : c_1 \notin C \wedge \mathcal{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}| \\
&= |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|
\end{aligned}
$$

■

The VC Dimension

The Fundamental Theorem of Statistical Learning (quantitate version).

Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the 0−1 loss. Suppose $\mathrm{VCdim}(\mathcal{H}) = d$. Then

1) $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$m(\varepsilon, \delta) = \Theta\left(\frac{1 + \log(1/\delta)}{\varepsilon^2}\right)$$

2) $\mathcal{H}$ is PAC learnable with sample complexity

$$\Omega\left(\frac{d + \log(1/\delta)}{\epsilon}\right) \leq m(\varepsilon, \delta) \leq O\left(\frac{d\log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$$

Steve Hanneke (2016): $\mathcal{H}$ is improperly PAC learnable with sample complexity $\Theta\left(\frac{d+\log(1/\delta)}{\varepsilon}\right)$