CS 57800 – Statistical Machine Learning

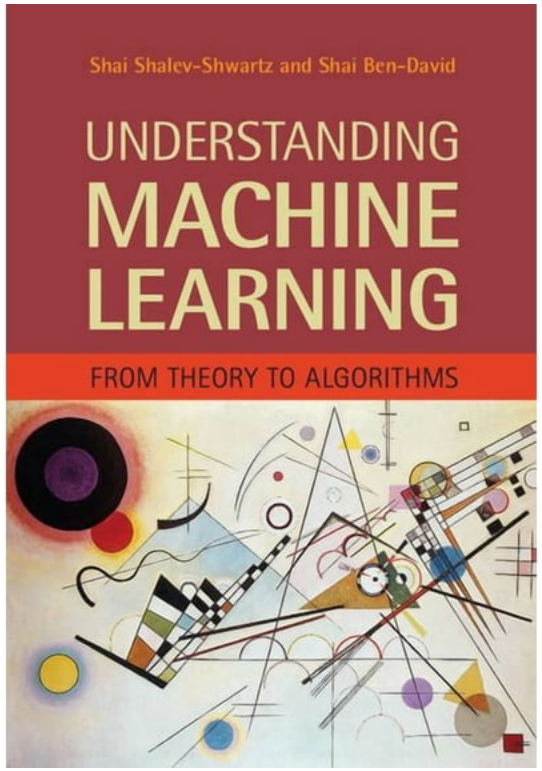
Statistical Learning Theory

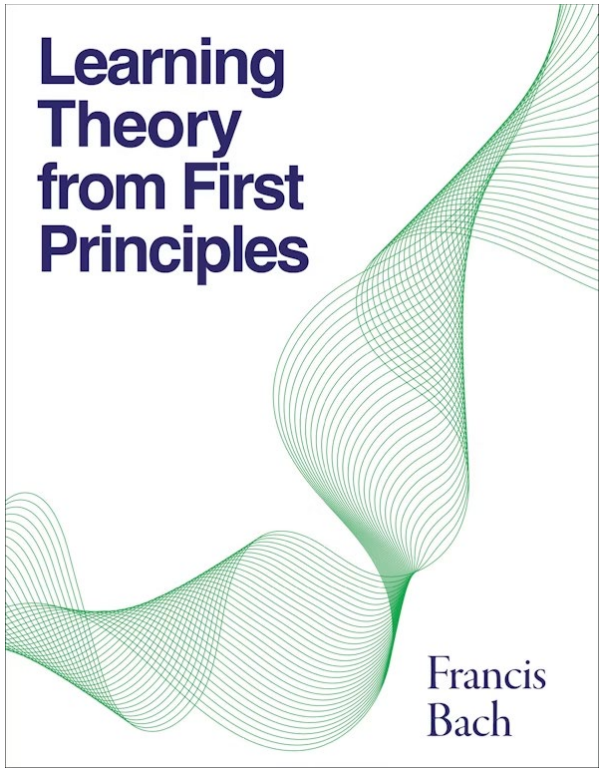
https://ruizhezhang.com/course_fall_2026.html

Outline

Intro to Learning Theory

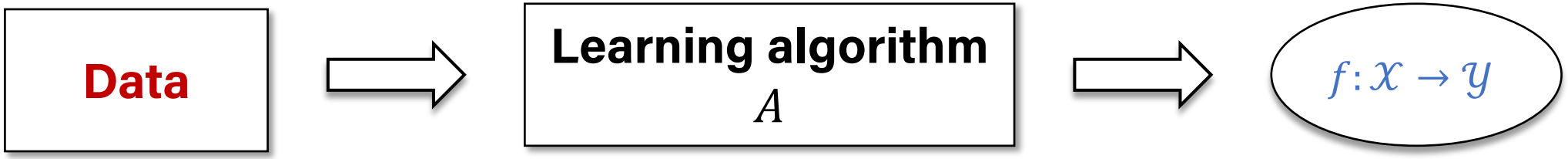
- Statistical Learning Framework
- PAC Learning
- The VC Dimension
- SVM





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:

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

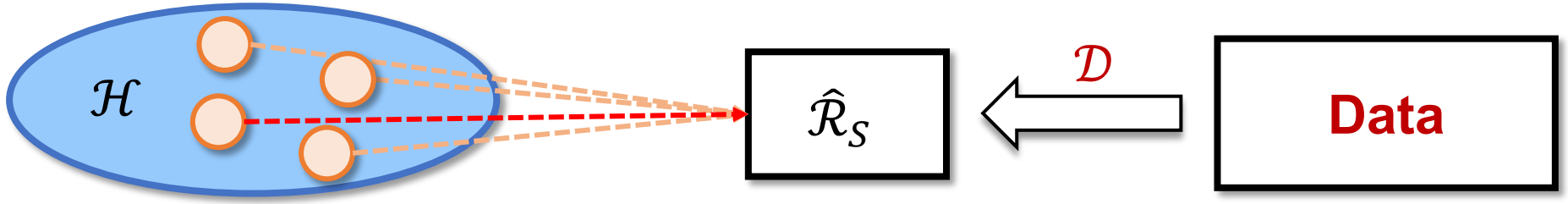
- Empirical risk / Empirical loss / Training error of f:

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

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

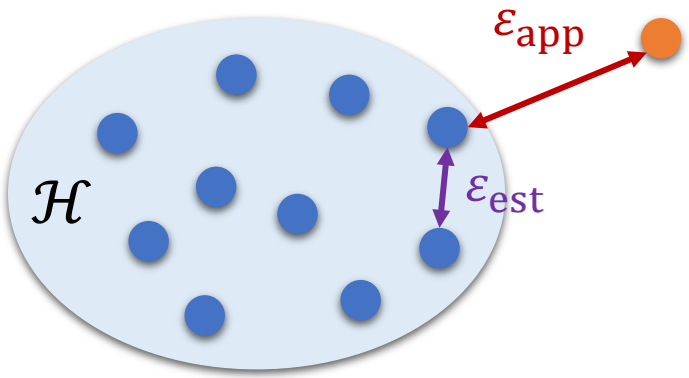
$$\mathcal{R}_\star \coloneqq \mathcal{R}(f_\star) = \int_{\mathcal{X}} \min_{z\in\mathcal{Y}} \mathbb{E}[\ell(z,y)|x=x']\, dp(x')$$

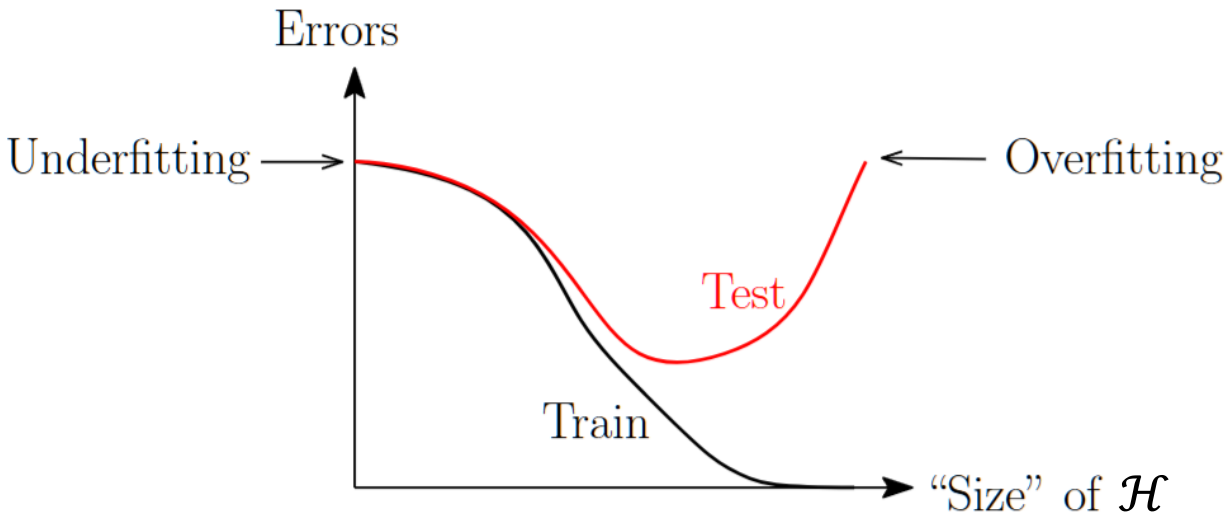
- **Exercise:** Show that the Bayes optimal predictor $f_\star$ is optimal, i.e., $\mathcal{R}_\star \le \mathcal{R}(f)$ for any $f\colon \mathcal{X}\to\mathcal{Y}$

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

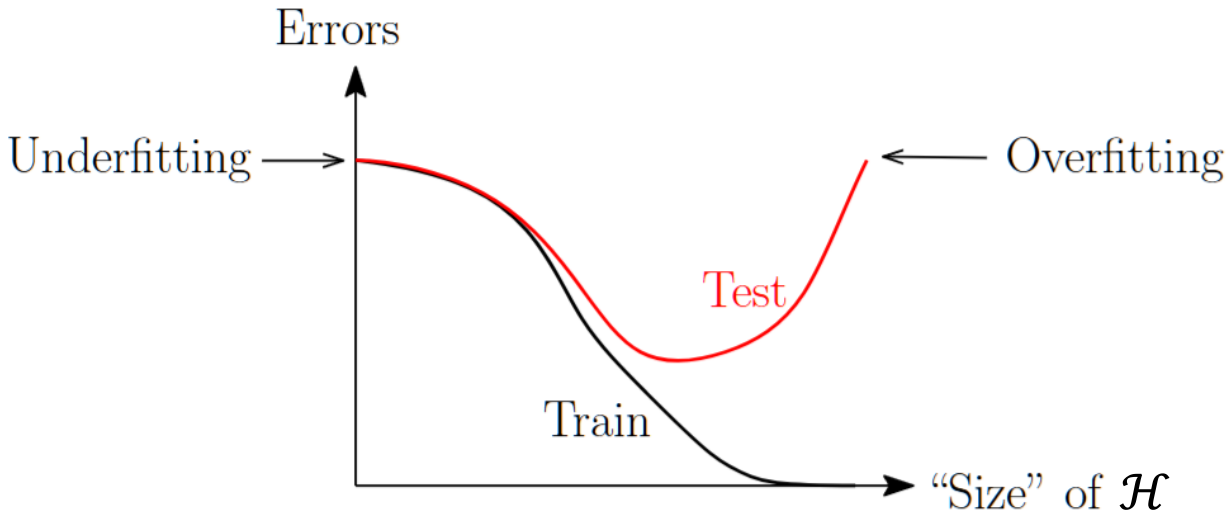
The goal is to find the best degree-P polynomial approximation of the function

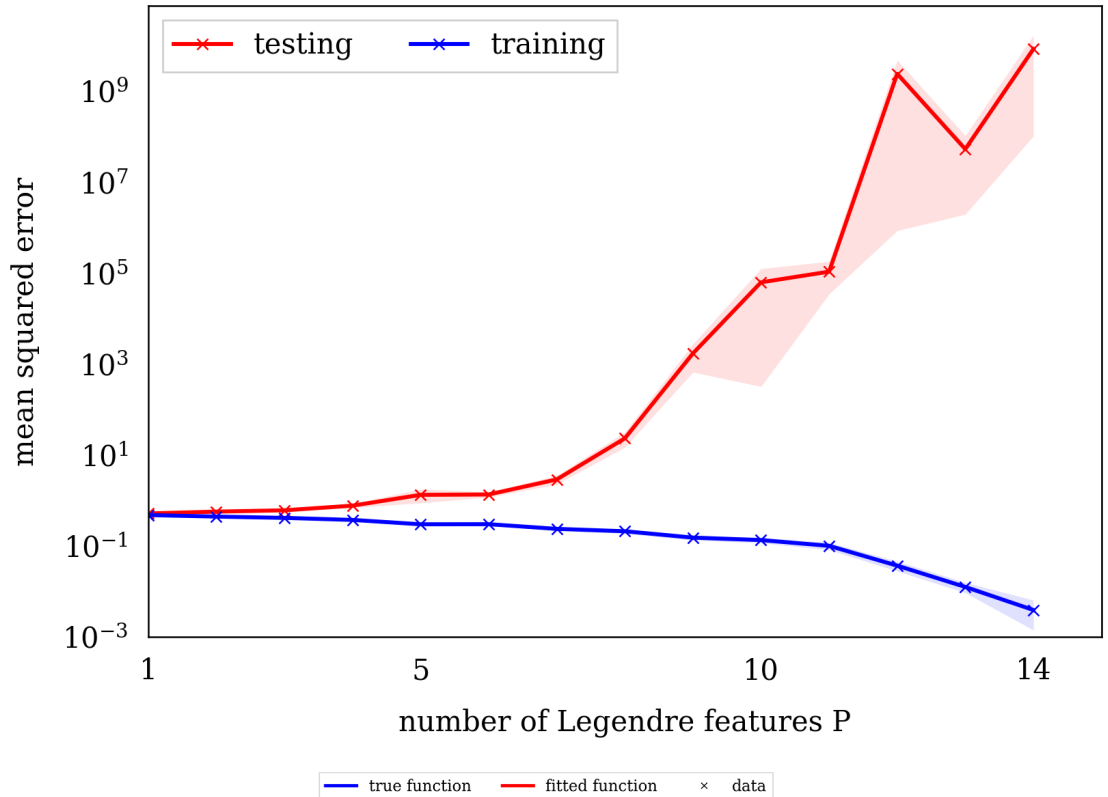
$$f(x) = 2x + \cos(25x)$$

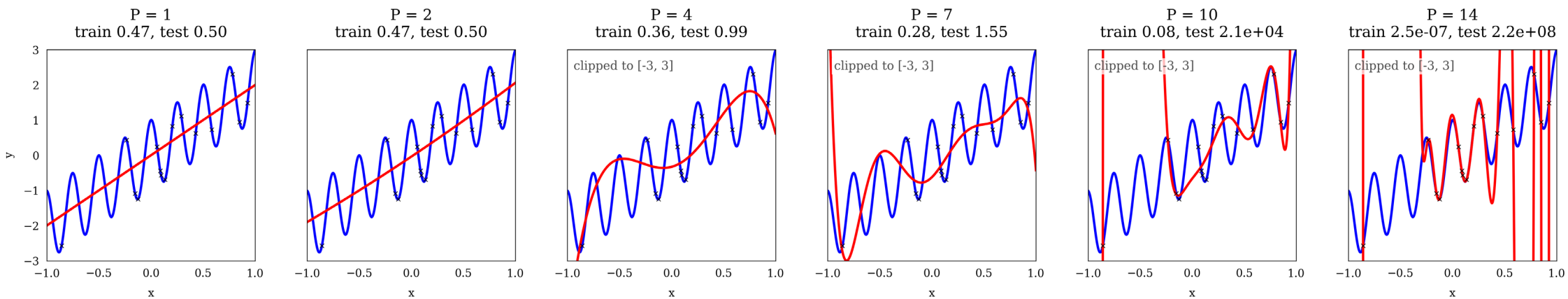
- Training data: $\{(x_i, f(x_i))\}_{i \le N}$, each $x_i \in [-1,1]$ is an i.i.d. uniform sample
- Hypothesis class:

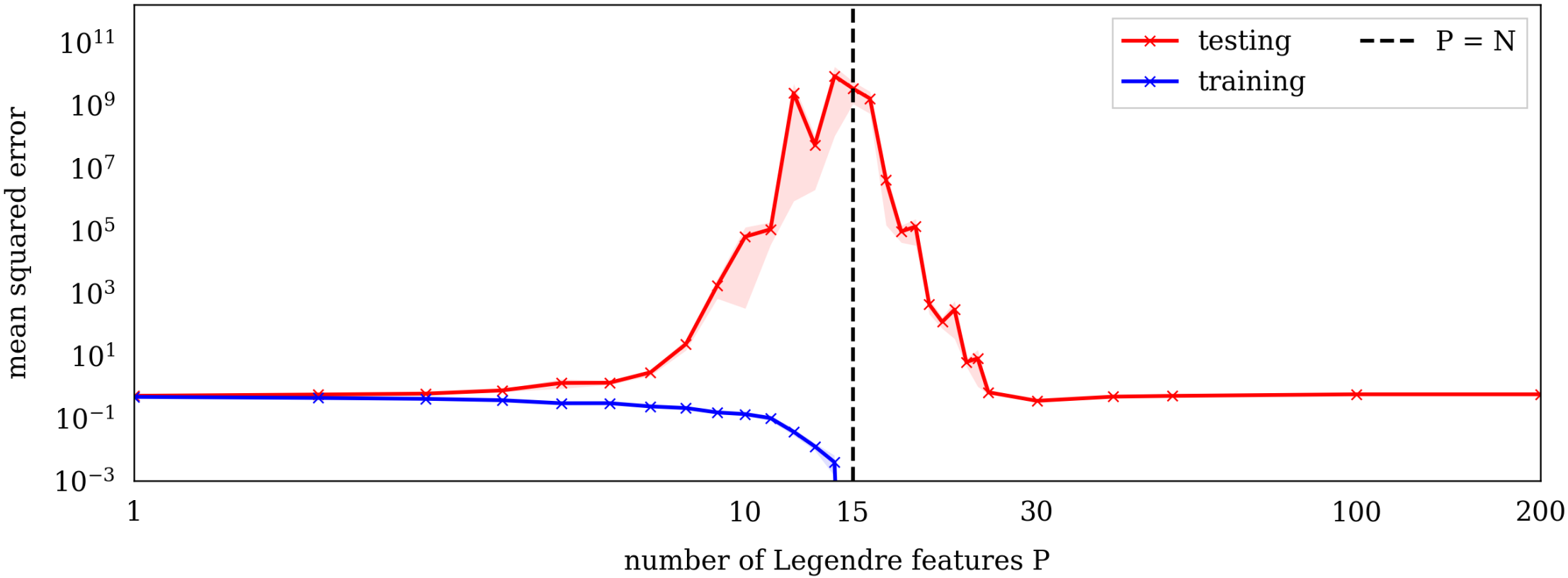
$$\mathcal{H} = \{\theta \in \mathbb{R}^P : \theta^\top \phi(x)\}, \qquad \phi(x) = \big(L_1(x), L_2(x), \dots, L_P(x)\big)$$

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,

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

- 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

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

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

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

(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.

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

ℓ is the 0-1 loss or bounded loss

■

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.

$$\Pr_{S\sim\mathcal{D}^m}[\mathrm{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B] \leq \sum_{f\in\mathcal{H}_B} \Pr_{S\sim\mathcal{D}^m}[\hat{\mathcal{R}}_S(f) = 0] = \sum_{f\in\mathcal{H}_B} \Pr_{S\sim\mathcal{D}^m}[\forall i \in [m], f(x_i) = y_i]$$

$$= \sum_{f\in\mathcal{H}_B} \left(\Pr_{(x,y)\sim\mathcal{D}}[f(x) = y]\right)^m$$

$$\leq \sum_{f\in\mathcal{H}_B} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[1 - \ell(f(x), y)]\right)^m$$

ℓ is the 0-1 loss or bounded loss

$$= \sum_{f\in\mathcal{H}_B} \left(1 - \mathcal{R}(f)\right)^m \leq \sum_{f\in\mathcal{H}_B} (1-\varepsilon)^m \leq |\mathcal{H}_B| e^{-\varepsilon m} \leq \delta$$

■

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 $x_1, \dots, x_k \in \mathcal{X}$
- 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 $S = \left(x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right)$

$$\Pr_{r,S,i}[A(S)(x_i) \neq r_i] = \mathbb{E}\left[\Pr_{r.i}\left[A(S)(x_i) \neq r_i \,\middle|\, x_{i_1}, r_{i_1}, \ldots, x_{i_m}, r_{i_m}\right]\right]$$

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

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

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:

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

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

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

Proof.

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

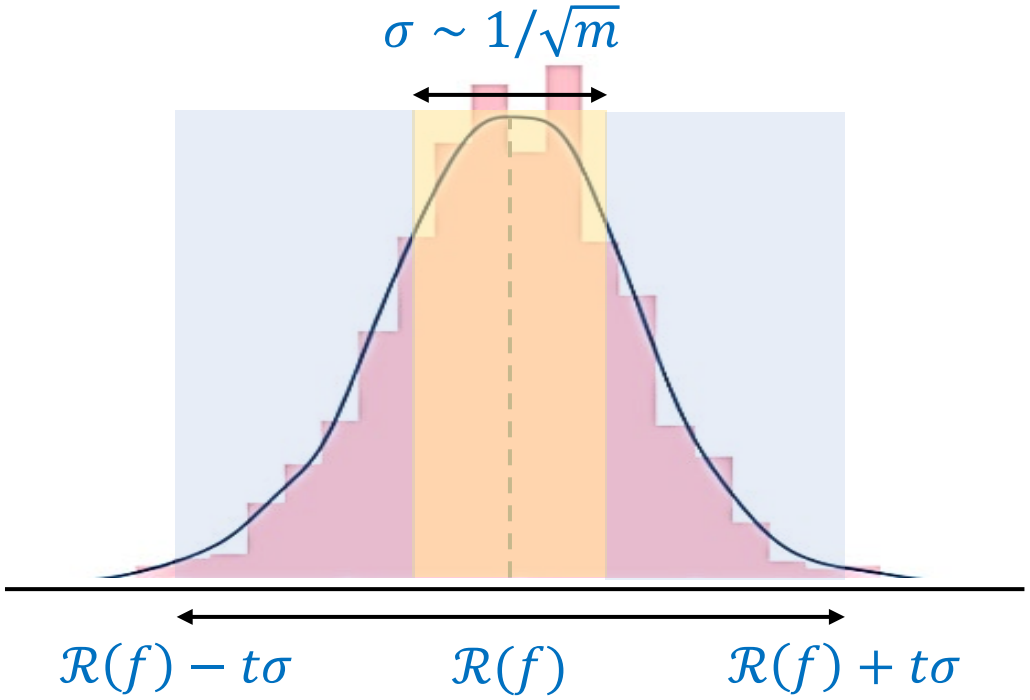
Random variable

- 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 / …



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

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

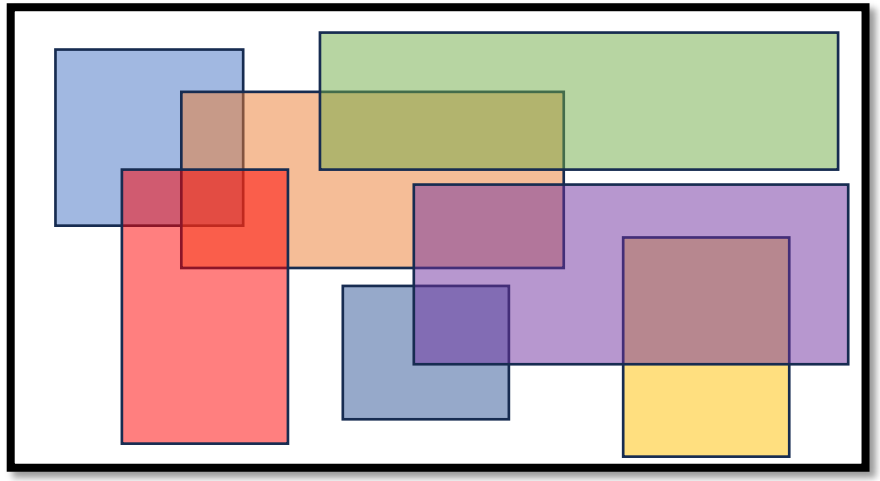
Alexei Chervonenkis and Vladimir Vapnik

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 $\mathcal{H}$: $2^{|C|} = |\mathcal{H}_C| \leq |\mathcal{H}| \quad \Longrightarrow \quad \mathrm{VCdim}(\mathcal{H}) \leq \log|\mathcal{H}|$
- 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" $|\mathcal{H}_C| = O\left(|C|^d\right)$ 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.

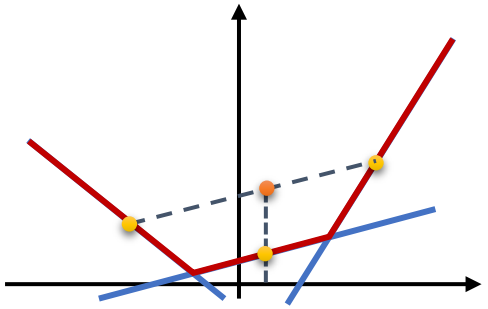
- By Markov's inequality, it suffices to prove: $\mathbb{E}_S \left[\max_{f \in \mathcal{H}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}_S(f) \right| \right] \leq O\left(\sqrt{\frac{1+\log \tau_{\mathcal{H}}(2m)}{m}} \right)$
- Since $\mathbb{E}_S\left[\hat{\mathcal{R}}_S(f)\right] = \mathcal{R}(f)$, we have

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

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 $C:=S\cup S'$ (only the $\mathcal{X}$ part). 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 $\sigma \in \{\pm 1\}^m$,

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

- 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,

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

$$\leq\tau_{\mathcal{H}}(2m)\exp\left(-\Omega(m\epsilon^2)\right)$$

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

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

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

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

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

■

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 $C = \{c_1, \dots, c_m\} \subset \mathcal{X}$
- Define two sets:
$$Y_0 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$
$$Y_1 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

The VC Dimension

- Define two sets:

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

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

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

- 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,

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

■

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

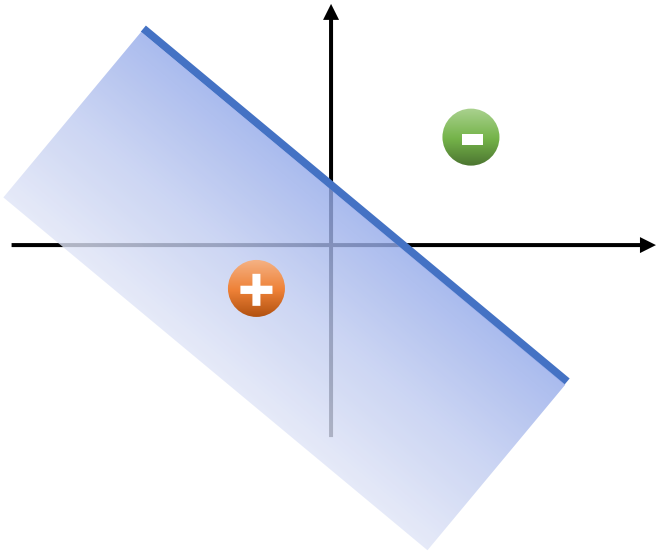
$$m(\varepsilon, \delta) = \Theta\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right)$$

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 $\Theta\left(\frac{d+\log(1/\delta)}{\varepsilon}\right)$

The VC Dimension



The Halfspaces: $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{-1,1\}$,

$$\mathcal{H} = \{\text{sgn}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

Theorem.

- The VC dimension of the class of homogeneous halfspaces (i.e., $b = 0$) in $\mathbb{R}^d$ is d
- The VC dimension of the class of the nonhomogeneous halfspaces is $d + 1$

The Fundamental Theorem of Statistical Learning gives that $\mathcal{H}$ is PAC learnable with sample complexity

$$\frac{d + \log(1/\delta)}{\varepsilon}$$

The VC Dimension

Theorem. The VC dimension of the class of homogeneous halfspaces (i.e., $b = 0$) in $\mathbb{R}^d$ is d

Proof.

- It is easy to see that $\{e_1, \dots, e_d\}$ is shattered by $\mathcal{H}$: for any labeling $(y_1, \dots, y_d) \in \{\pm 1\}^d$, let
$$w \coloneqq (y_1, \dots, y_d) \quad \Longrightarrow \quad \langle w, e_i \rangle = y_i$$
- For any $\{x_1, \dots, x_{d+1}\} \subset \mathbb{R}^d$, suppose it can be shattered by $\mathcal{H}$
- There exist $c_1, \dots, c_{d+1}$ that are not all zero such that $\sum_{i=1}^{d+1} c_i x_i = 0$
- Let $I \coloneqq \{i \in [d+1] : c_i > 0\}$ and $J \coloneqq \{i \in [d+1] : c_i < 0\}$
- There are two cases:
  I. Both I and J are non-empty
  II. One of them I is empty

The VC Dimension

- Let $I \coloneqq \{i \in [d+1] : c_i > 0\}$ and $J \coloneqq \{i \in [d+1] : c_i < 0\}$
- If $I \neq \emptyset$ and $J \neq \emptyset$, then consider the labeling:

$$y_i \coloneqq \operatorname{sgn}(c_i)$$

- By shattering, there exists a $w \in \mathbb{R}^m$ such that $\operatorname{sgn}(\langle w, x_i \rangle) = y_i$; that is,

$$\langle w, x_i \rangle \geq 0 \quad \text{if } i \in I, \qquad \langle w, x_j \rangle < 0 \quad \text{if } j \in J$$

- Note that

$$\sum_{i \in I} c_i x_i = \sum_{j \in J} |c_j| x_j$$

- Then, we have

$$0 \leq \sum_{i \in I} c_i \langle w, x_i \rangle = \sum_{j \in J} |c_j| \langle w, x_j \rangle < 0$$

Contradiction!

The VC Dimension

- It remains to consider the case when $I \neq \emptyset$ and $J = \emptyset$; that is,

$$\sum_{i \in I} c_i x_i = 0, \qquad c_i > 0 \quad \forall i \in I$$

- For any labeling with $y_i = -1$ for every $i \in I$, let w be the vector that realizes it, i.e.,

$$\langle w, x_i \rangle < 0 \qquad \forall i \in I$$

- Then, we have

$$0 = \sum_{i \in I} c_i \langle w, x_i \rangle < 0$$

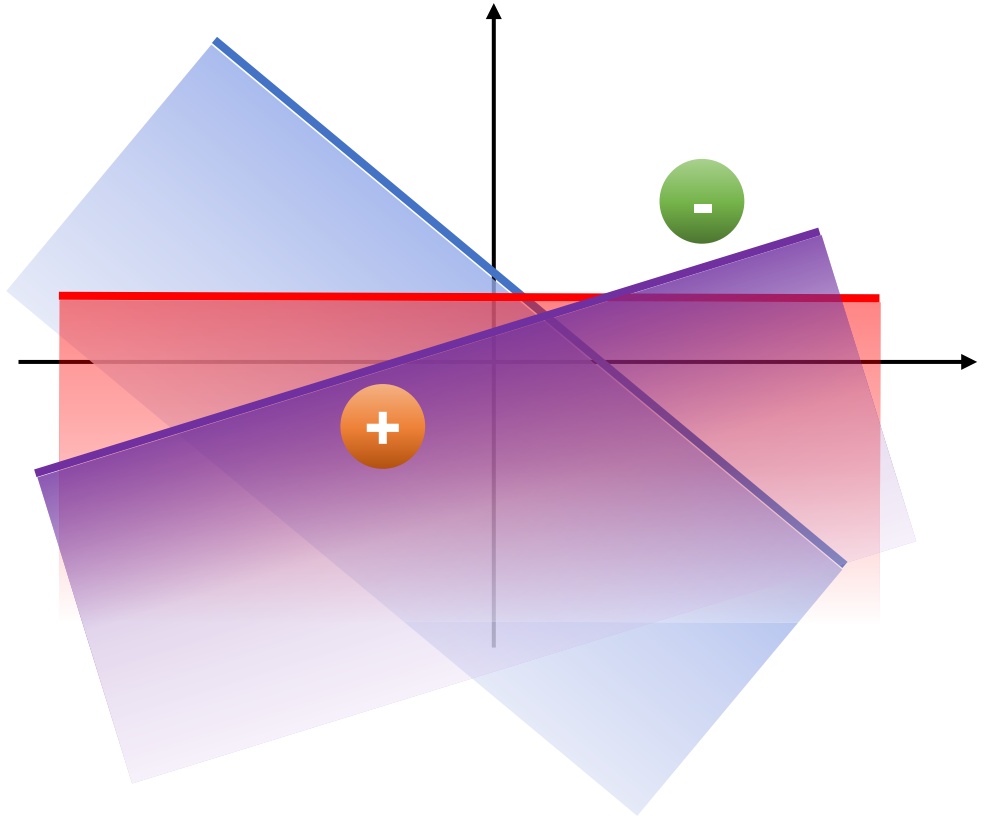
  Contradiction!

- Therefore, no subset of size $d + 1$ can be shattered by $\mathcal{H}$
- Hence, $\mathrm{VCdim}(\mathcal{H}) = d$

■

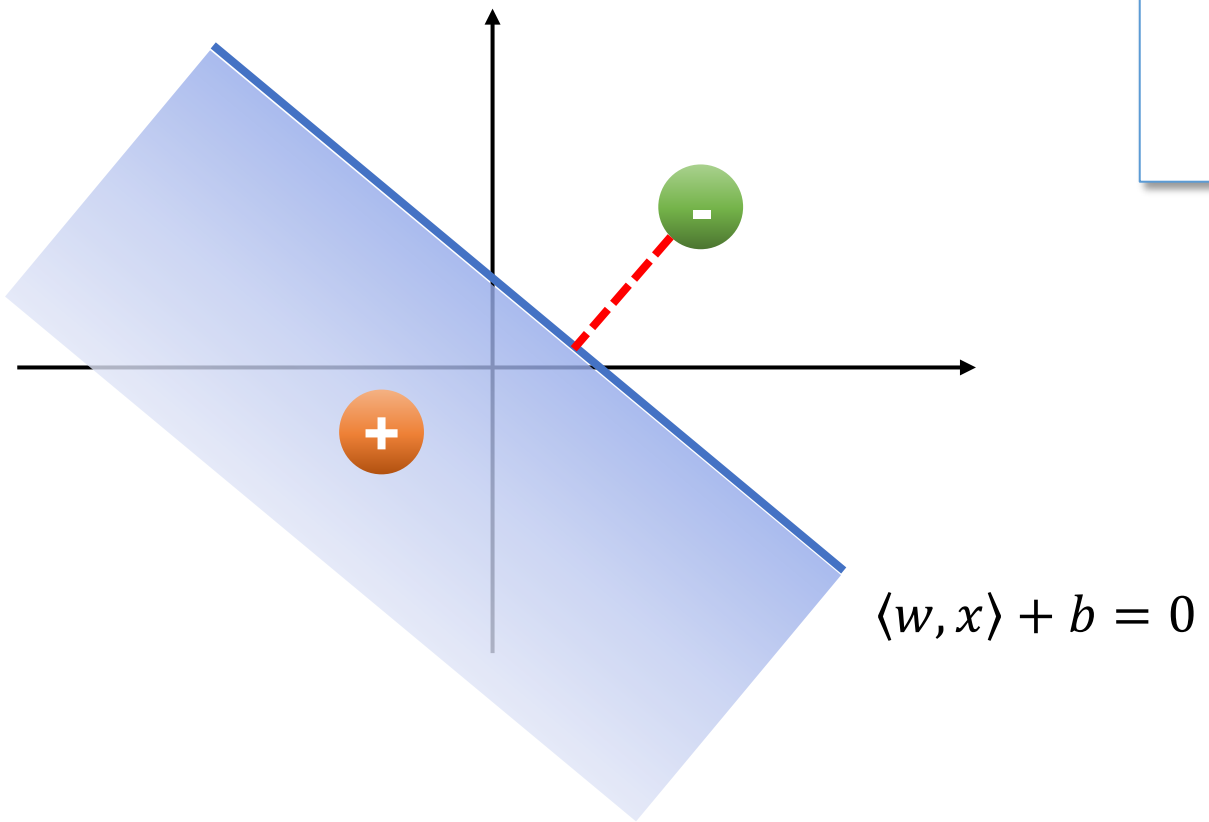
Support Vector Machine

Which halfspace is better?



Support Vector Machine

- **Distance to the hyperplane**



$\langle w, x \rangle + b = 0$

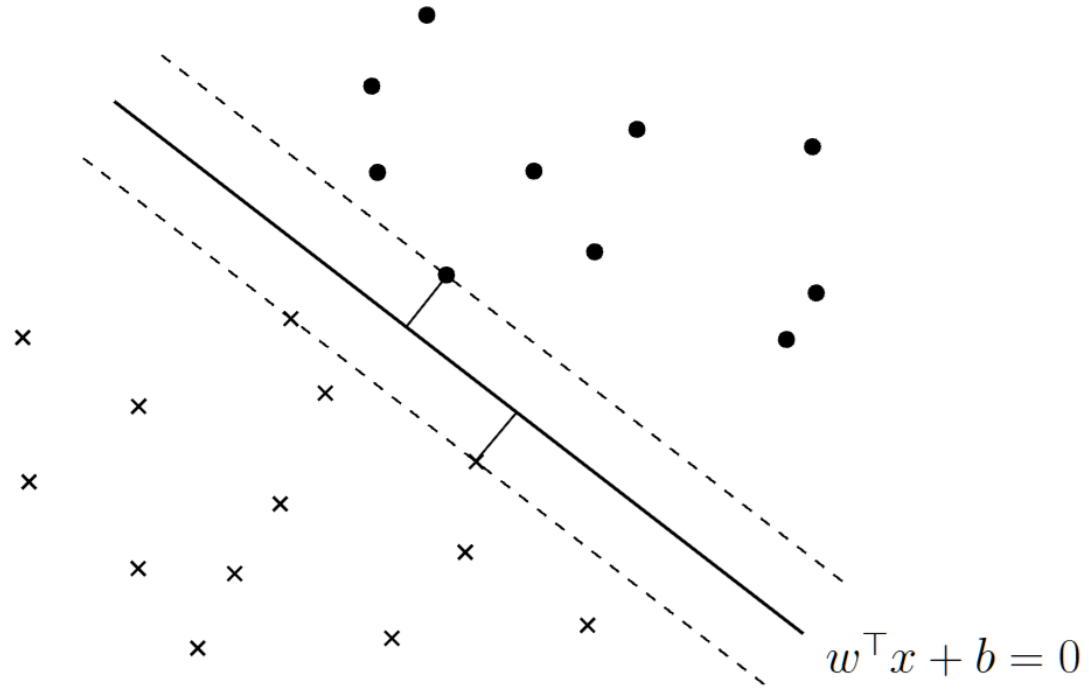
Geometric fact: The distance between a point x to the hyperplane (w, b) is

$$\frac{|\langle w, x \rangle + b|}{\|w\|}$$

Support Vector Machine

- **Margin:** the minimal distance between a point in the training set and the hyperplane

$$\min_{i\in[m]} \frac{|\langle w, x_i\rangle + b|}{\|w\|}$$



Support Vector Machine

- **Hard-SVM:** return an ERM hyperplane that separates the training set with the largest possible margin

$$\operatorname*{argmax}_{(w,b):\|w\|=1} \min_{i\in[m]} |\langle w, x_i\rangle + b| \qquad \text{s.t.} \qquad y_i(\langle w, x_i\rangle + b) > 0 \quad \forall i \in [m] \tag{1}$$

- Whenever (1) is feasible (i.e., linearly separable), it is equivalent to the following optimization:

$$\operatorname*{argmax}_{(w,b):\|w\|=1} \min_{i\in[m]} y_i(\langle w, x_i\rangle + b) \tag{2}$$

- It is also equivalent to the following quadratic optimization:

$$(w_0, b_0) \leftarrow \operatorname*{argmin}_{(w,b)} \|w\|^2 \qquad \text{s.t.} \qquad y_i(\langle w, x_i\rangle + b) \geq 1 \quad \forall i \in [m]$$

$$(w_\star, b_\star) \leftarrow \left(\frac{w_0}{\|w_0\|}, \frac{b}{\|w_0\|}\right) \tag{3}$$

Support Vector Machine

Duality:

$$\operatorname*{argmin}_{w} \|w\|^2 \qquad \text{s.t.} \qquad y_i\langle w, x_i\rangle \geq 1 \quad \forall i \in [m]$$

- Define a penalty function

$$g(w) \coloneqq \max_{\alpha\in\mathbb{R}^m:\alpha\geq 0} \sum_{i\in[m]} \alpha_i\big(1 - y_i(\langle w, x_i\rangle + b)\big) = \begin{cases} 0 & \text{if } y_i\langle w, x_i\rangle \geq 1 \ \forall i \in [m] \\ \infty & \text{otherwise} \end{cases}$$

- Then, we have

$$\min_w \frac{1}{2}\|w\|^2 + g(w) = \min_w \max_{\alpha\in\mathbb{R}^m:\alpha\geq 0} \frac{1}{2}\|w\|^2 + \sum_{i\in[m]} \alpha_i\big(1 - y_i(\langle w, x_i\rangle + b)\big)$$

$$= \max_{\alpha\in\mathbb{R}^m:\alpha\geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i\in[m]} \alpha_i\big(1 - y_i(\langle w, x_i\rangle + b)\big)$$

Slater's condition

- Thus, $w_\star = \sum_{i\in[m]} \alpha_i y_i x_i$ in the span of $\{x_1, \ldots, x_m\}$

Support Vector Machine

Duality:

Primal: $$\min_{w} \frac{1}{2}\|w\|^2 \qquad \text{s.t.} \qquad y_i\langle w, x_i\rangle \geq 1 \quad \forall i \in [m]$$

Dual: $$\max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \sum_{i\in[m]} \alpha_i - \frac{1}{2}\left\|\sum_{i\in[m]} \alpha_i\, y_i x_i\right\|^2$$

- Suppose $w^\star$ is the optimal primal solution and $\alpha^\star$ is the optimal dual solution
- Strong duality implies that

$$\frac{1}{2}\|w^\star\|^2 = \sum_{i\in[m]} \alpha_i^\star - \frac{1}{2}\left\|\sum_{i\in[m]} \alpha_i^\star\, y_i x_i\right\|^2$$

Support Vector Machine

Duality:

Primal: $$\min_{w} \frac{1}{2}\|w\|^2 \qquad \text{s.t.} \qquad y_i\langle w, x_i\rangle \geq 1 \quad \forall i \in [m]$$

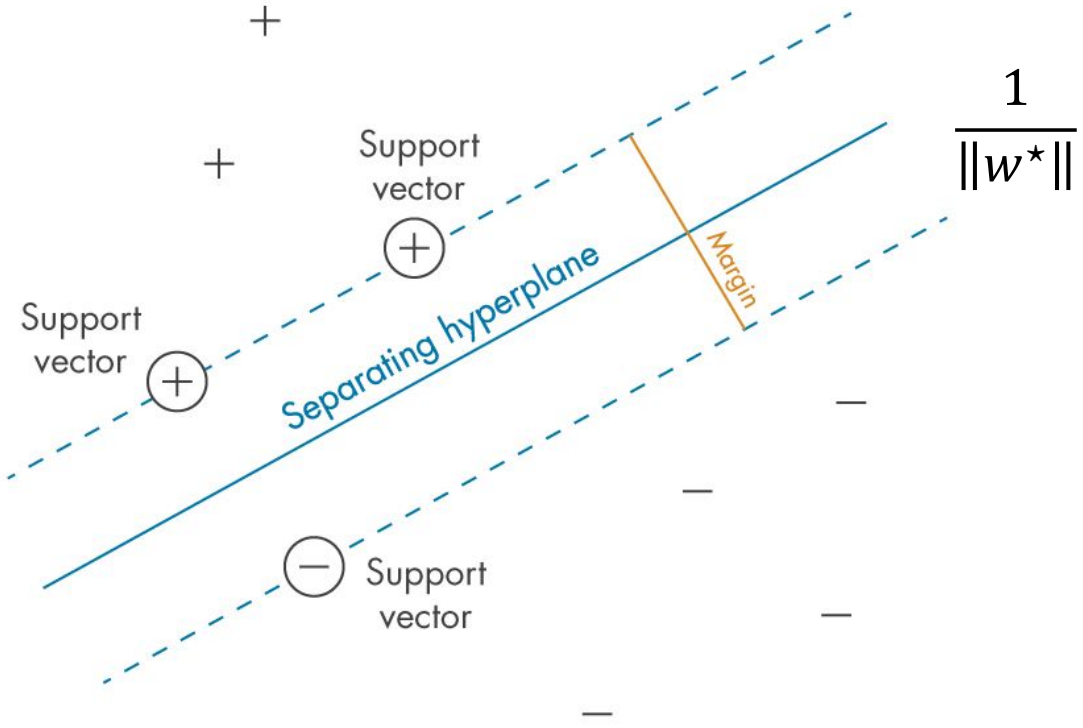
Dual: $$\max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \sum_{i\in[m]} \alpha_i - \frac{1}{2}\left\|\sum_{i\in[m]} \alpha_i\, y_i x_i\right\|^2$$

- Suppose $w^\star$ is the optimal primal solution and $\alpha^\star$ is the optimal dual solution
- Strong duality implies that

$$0 = \frac{1}{2}\left\|w^\star - \sum_{i\in[m]} \alpha_i^\star\, y_i x_i\right\|^2 + \sum_{i\in[m]} \alpha_i^\star (y_i\langle w^\star, x_i\rangle - 1)$$

$$w^\star = \sum_{i\in[m]} \alpha_i^\star\, y_i x_i, \qquad \alpha_i^\star (y_i\langle w^\star, x_i\rangle - 1) = 0 \quad \forall i \in [m]$$

Support Vector Machine



$$\frac{1}{\|w^\star\|}$$

What is the sample complexity of hard-SVM?

The Rademacher Complexity

- Let $\mathcal{F} \coloneqq \ell \circ \mathcal{H} = \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$. Then, for any $f \in \mathcal{F}$, we have

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

- **Representativeness** of S w.r.t. $\mathcal{F}$: the largest gap between the true error of a function f and its empirical error

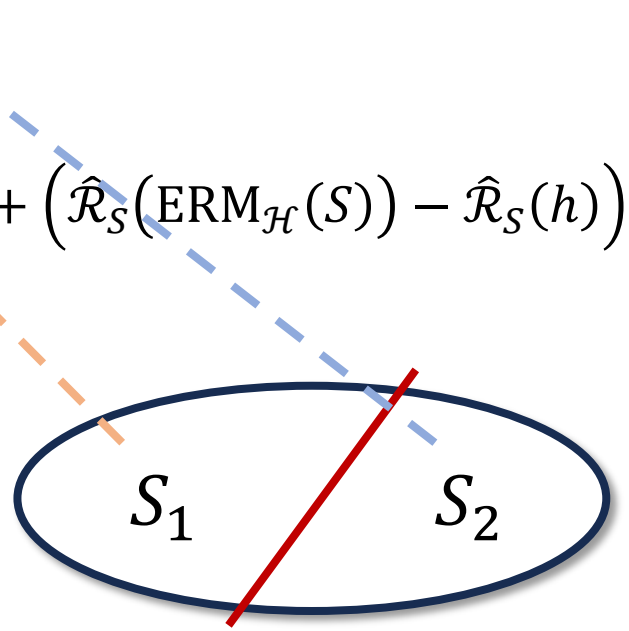
$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \coloneqq \sup_{f\in\mathcal{F}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}_S(f)\right|$$

- If $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \le \varepsilon/2$, then

$$\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \inf_{h\in\mathcal{H}} \mathcal{R}(h) = \left(\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right) + \left(\hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S(h)\right) + \left(\hat{\mathcal{R}}_S(h) - \mathcal{R}(h)\right) \le \varepsilon$$

- How to estimate $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)$ using samples in S only?

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{|S_1|}\sum_{i\in S_1} f(x_i, y_i) - \frac{1}{|S_2|}\sum_{j\in S_2} f\left(x_j, y_j\right)\right|$$



The Rademacher Complexity

$$\widehat{\mathrm{Rep}}_{\mathcal{D}}(\mathcal{F}, S) \coloneqq \sup_{f \in \mathcal{F}} \left| \frac{1}{|S_1|} \sum_{i \in S_1} f(x_i, y_i) - \frac{1}{|S_2|} \sum_{j \in S_2} f(x_j, y_j) \right|$$

- Suppose $|S_1| = |S_2| = m/2$. And let $\sigma \in \{-1,1\}^m$ be such that

$$\sigma_i = 1 \quad \text{if } i \in S_1, \qquad \sigma_i = -1 \quad \text{if } i \in S_2$$

$$\widehat{\mathrm{Rep}}_{\mathcal{D}}(\mathcal{F}, S) = \frac{m}{2} \cdot \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{m} \sigma_i f(x_i, y_i) \right|$$

- The Rademacher complexity generalizes this idea by considering the expectation with respect to a random $\sigma \sim \{-1,1\}^m$

The Rademacher Complexity

- Let $\mathcal{F} \circ S \coloneqq \{(f(x_1, y_1), \ldots, f(x_m, y_m)) : f \in \mathcal{F}\} \subset \mathbb{R}^m$
- Let $\sigma \in \{-1,1\}^m$ be a random vector such that each coordinate is an independent Rademacher random variable:

$$\Pr[\sigma_i = \pm 1] = 1/2$$

- The **Rademacher complexity** of $\mathcal{F}$ with respect to S is defined as:

$$\mathrm{R}(\mathcal{F} \circ S) \coloneqq \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i, y_i) \right] = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{v \in \mathcal{F} \circ S} \langle \sigma, v \rangle \right]$$

Lemma.

$$\mathbb{E}_{S \sim \mathcal{D}^m}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)] \leq 2\mathbb{E}_{S \sim \mathcal{D}^m}[\mathrm{R}(\mathcal{F} \circ S)]$$

- The proof uses the symmetrization technique
- See (Shalev-Shwartz-Ben-David, Sec. 26.1) for more details

The Rademacher Complexity

Lemma.

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F},S)] \le 2\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{R}(\mathcal{F}\circ S)]$$

Corollary. For any $h \in \mathcal{H}$,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \mathcal{R}(h)\right] \le 2\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{R}(\ell\circ\mathcal{H}\circ S)]$$

Proof.

- By the definition of $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F},S)$, we have

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}\left(\mathrm{ERM}_{\mathcal{H}}(S)\right) - \hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right] \le 2\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{R}(\ell\circ\mathcal{H}\circ S)]$$

- For any fixed $h \in \mathcal{H}$,

$$\mathcal{R}(h) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{R}}_S(h)\right] \ge \mathbb{E}_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{R}}_S\left(\mathrm{ERM}_{\mathcal{H}}(S)\right)\right]$$

■

How to convert this expectation bound to a high-probability bound?

Detour: Concentration (II)

$$\mathbb{E}[X] \leq ?? \quad \Longrightarrow \quad \Pr[X \geq ???] \leq \delta$$

General approach: Markov inequality

For any nonnegative random variable X,

$$\Pr\left[X \geq \frac{\mathbb{E}[X]}{\delta}\right] \leq \delta$$

- We have with probability $1 - \delta$ over $S \sim \mathcal{D}^m$,

$$\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \inf_{h \in \mathcal{H}} \mathcal{R}(h) \leq \frac{2\mathbb{E}_{S' \sim \mathcal{D}^m}[\mathrm{R}(\ell \circ \mathcal{H} \circ S')]}{\delta}$$

Detour: Concentration (II)

$$\mathbb{E}[X] \leq ?? \quad \Longrightarrow \quad \Pr[X \geq ???] \leq \delta$$

Specialized approach: concentration inequalities

- If the random variable X has a specific tail condition (e.g., subgaussian tail), then we have
$$\Pr[X \geq \mathbb{E}[X] + c] \leq \Pr[|X - \mathbb{E}[X]| \geq c] \leq e^{-\mathcal{I}(c)}$$
- That is,
$$\Pr[X \geq \mathbb{E}[X] + \mathcal{I}^{-1}(\log 1/\delta)] \leq \delta$$

Example:

- Hoeffding's inequality: for any fixed $h \in \mathcal{H}$, with probability $1 - \delta$,
$$\left|\mathcal{R}(h) - \hat{\mathcal{R}}_S(h)\right| \lesssim \sqrt{\frac{\log(1/\delta)}{m}}$$

Detour: Concentration (II)

$$\mathbb{E}[X] \leq ?? \quad \Longrightarrow \quad \Pr[X \geq ???] \leq \delta$$

Specialized approach: concentration inequalities

McDiarmid's Inequality

Let $\mathbb{V}$ be some set and let $f : \mathbb{V}^m \to \mathbb{R}$ be a function that satisfies the bounded differences condition: for all $i \in [m]$ and for all $x_1, \ldots, x_m, x_i' \in \mathbb{V}$, we have

$$|f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \leq c$$

Let $X_1, \ldots, X_m$ be independent random variables supported on $\mathbb{V}$. Then, with probability $1 - \delta$,

$$|f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)]| \leq O\left(c\sqrt{m \log(1/\delta)}\right)$$

The Rademacher Complexity

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \coloneqq \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}_S(f) \right| = \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \frac{1}{m} \sum_{i=1}^{m} f(x_i, y_i) \right|$$

- Suppose that the loss function is bounded, i.e., $|\ell(y, z)| \leq 1$
- Then, $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)$ satisfies the bounded differences condition with parameter $\frac{2}{m}$
- Recall that $\mathbb{E}_{S \sim \mathcal{D}^m}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)] \leq 2\mathbb{E}_{S \sim \mathcal{D}^m}[\mathrm{R}(\mathcal{F} \circ S)]$
- McDiarmid's inequality gives that, with probability $1 - \delta$,

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \leq \mathbb{E}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)] + O\left(\frac{2}{m}\sqrt{m \log(1/\delta)}\right)$$

$$\leq O\left(\mathbb{E}_{S' \sim \mathcal{D}^m}[\mathrm{R}(\mathcal{F} \circ S')] + \sqrt{\frac{\log(1/\delta)}{m}}\right)$$

The Rademacher Complexity

Theorem. Assume that the loss function ℓ is bounded by 1. Then, with probability $1-\delta$,

$$\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \inf_{h\in\mathcal{H}} \mathcal{R}(h) \lesssim \mathrm{R}(\ell \circ \mathcal{H} \circ S) + \sqrt{\frac{\log(1/\delta)}{m}}$$

Proof.

- We already show that with probability at least $1-\delta$,

$$\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \hat{\mathcal{R}}_S\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) \le \mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \lesssim \mathbb{E}_{S'\sim\mathcal{D}^m}[\mathrm{R}(\mathcal{F}\circ S')] + \sqrt{\frac{\log(1/\delta)}{m}}$$

- $\mathrm{R}(\mathcal{F}\circ S') = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i, y_i)\right]$ also satisfies the bounded differences condition with parameter $\frac{2}{m}$

The Rademacher Complexity

Proof.

- We already show that with probability at least $1 - \delta/2$,

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \lesssim \mathbb{E}_{S' \sim \mathcal{D}^m}[\mathrm{R}(\mathcal{F} \circ S')] + \sqrt{\frac{\log(1/\delta)}{m}}$$

- By McDiarmid's inequality again, with probability $1 - \delta/2$,

$$|\mathrm{R}(\mathcal{F} \circ S) - \mathbb{E}_{S' \sim \mathcal{D}^m}[\mathrm{R}(\mathcal{F} \circ S')]| \lesssim \sqrt{\frac{\log(1/\delta)}{m}}$$

- Thus, by union bound, with probability $1 - \delta$,

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) \lesssim \mathrm{R}(\ell \circ \mathcal{H} \circ S) + \sqrt{\frac{\log(1/\delta)}{m}}$$

The Rademacher Complexity

Proof.

- By the error decomposition:

$$\begin{aligned}\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \inf_{h\in\mathcal{H}} \mathcal{R}(h) &\leq \Big(\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \hat{\mathcal{R}}_S\big(\mathrm{ERM}_{\mathcal{H}}(S)\big)\Big) + \Big(\hat{\mathcal{R}}_S(h) - \mathcal{R}(h)\Big) \\ &\leq 2\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)\end{aligned}$$

- Thus, with probability at least $1-\delta$,

$$\mathcal{R}\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - \inf_{h\in\mathcal{H}} \mathcal{R}(h) \lesssim \mathrm{R}(\ell \circ \mathcal{H} \circ S) + \sqrt{\frac{\log(1/\delta)}{m}}$$

■

The Rademacher Complexity

Properties of the Rademacher complexity

- If $A \subseteq A' \subseteq \mathbb{R}^m$, then $\mathrm{R}(A) \leq \mathrm{R}(A')$
- For any $A \subseteq \mathbb{R}^m$, $c \in \mathbb{R}$, and $a_0 \in \mathbb{R}^m$,
$$\mathrm{R}(\{ca + a_0 : a \in A\}) = |c|\mathrm{R}(A)$$
- Let $A \subseteq \mathbb{R}^m$ and let $A' = \left\{\sum_{i=1}^N c_i a_i : N \in \mathbb{N}, a_i \in A, c_i \geq 0, \|c\|_1 = 1\right\}$. Then, $\mathrm{R}(A') = \mathrm{R}(A)$

Massart lemma. Let $A = \{a_1, \ldots, a_N\} \subset \mathbb{R}^m$. Define $\overline{a} \coloneqq \frac{1}{N}\sum_{i=1}^N a_i$. Then,

$$\mathrm{R}(A) \leq \max_{i \in [N]} \|a_i - \overline{a}\| \frac{\sqrt{2 \log N}}{m}$$

Contraction lemma. For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz: for all $x, y \in \mathbb{R}$, we have $|\phi_i(x) - \phi_i(y)| \leq \rho|x - y|$. For $A \subset \mathbb{R}^m$, let $\phi \circ A \coloneqq \left\{\left(\phi_1(a_1), \ldots, \phi_m(a_m)\right) : a \in A\right\} \subset \mathbb{R}^m$. Then,

$$\mathrm{R}(\phi \circ A) \leq \rho \cdot \mathrm{R}(A)$$

The Rademacher Complexity

Lemma. Let $\mathcal{H}_2 \coloneqq \{x \mapsto \langle w, x\rangle : \|w\| \le 1\}$. For any $S = (x_1, \dots, x_m)$,

$$\mathrm{R}(\mathcal{H}_2 \circ S) = \mathrm{R}(\{(\langle w, x_1\rangle, \dots, \langle w, x_m\rangle) : \|w\| \le 1\}) \le \frac{\max_i \|x_i\|}{\sqrt{m}}$$

Proof.

$$m\mathrm{R}(\mathcal{H}_2 \circ S) = \mathbb{E}_\sigma\left[\sup_{v \in \mathcal{H}_2 \circ S} \langle \sigma, v\rangle\right] = \mathbb{E}_\sigma\left[\sup_{\|w\| \le 1} \sum_{i=1}^m \sigma_i \langle w, x_i\rangle\right]$$

$$= \mathbb{E}_\sigma\left[\sup_{\|w\| \le 1} \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle\right] = \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|\right]$$

Jensen's inequality

$$\le \left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|^2\right]\right)^{1/2} = \left(\sum_{i=1}^m \|x_i\|^2\right)^{1/2}$$

$$\le \sqrt{m}\max\|x_i\|$$

■

Sample Complexity of Hard-SVM

Theorem. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X} \times \{\pm 1\}$ such that there exists $w^\star$ with

$$\Pr_{(x,y)\sim\mathcal{D}}[y_i\langle w^\star, x_i\rangle \geq 1] = 1, \qquad \text{and} \qquad \|x\| \leq R$$

Let w_S be the solution of Hard-SVM:

$$\operatorname*{argmin}_w \|w\|^2 \qquad \text{s.t.} \qquad y_i\langle w, x_i\rangle \geq 1 \quad \forall i \in [m]$$

Then, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$,

$$\Pr_{(x,y)\sim\mathcal{D}}[y \neq \operatorname{sgn}(\langle w_S, x\rangle)] \leq O\left(\frac{R\|w^\star\| + \log(1/\delta)}{\sqrt{m}}\right)$$

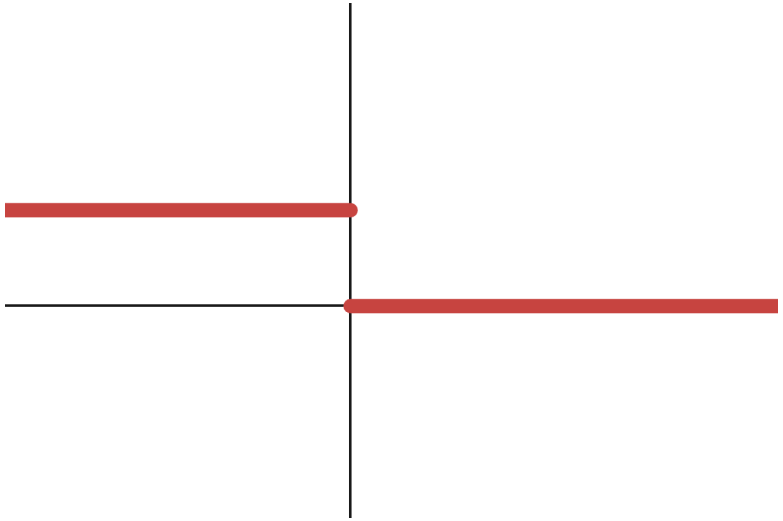
- The sample complexity is $O\left(\frac{R^2\|w^\star\|^2}{\epsilon^2}\right)$, **without** any explicit dependence on the dimension
- The VC dimension of the class of halfspaces is d

Sample Complexity of Hard-SVM

- We know that with probability $1-\delta$,

$$\mathcal{R}(h) - \hat{\mathcal{R}}_S(h) \lesssim \mathrm{R}(\ell \circ \mathcal{H} \circ S) + \sqrt{\frac{\log(1/\delta)}{m}}$$

- $\Pr_{(x,y)\sim\mathcal{D}}[y \neq \mathrm{sgn}(\langle w_S, x\rangle)] = \mathbb{E}[\ell_{0-1}(\mathrm{sgn}(\langle w_S, x\rangle), y)]$
- However, the function $a \mapsto \ell_{0-1}(\mathrm{sgn}(a), y)$ is not Lipschitz



Sample Complexity of Hard-SVM

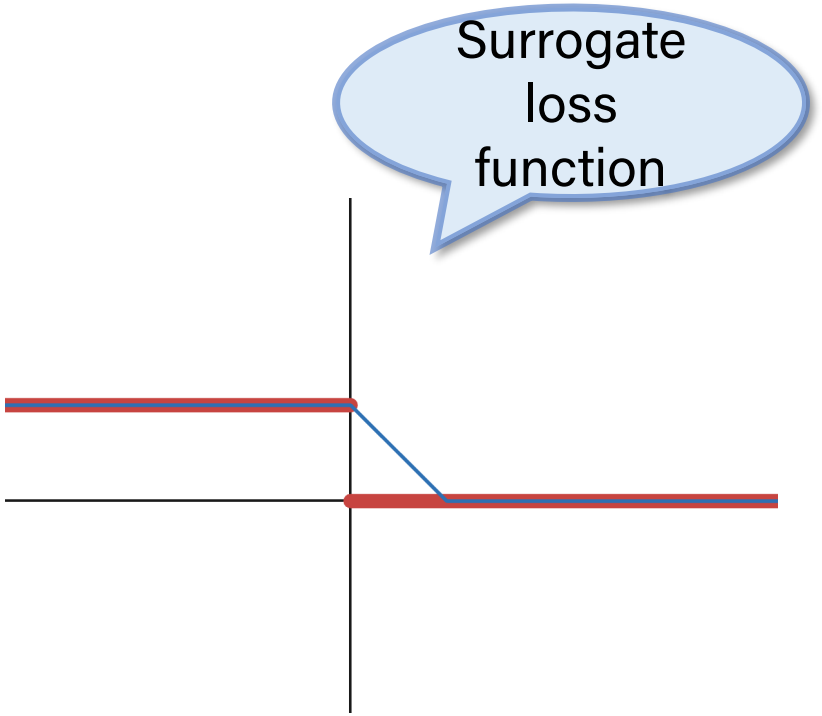
- We know that with probability $1-\delta$,

$$\mathcal{R}(h) - \hat{\mathcal{R}}_S(h) \lesssim \mathrm{R}(\ell \circ \mathcal{H} \circ S) + \sqrt{\frac{\log(1/\delta)}{m}}$$

- $\Pr_{(x,y)\sim\mathcal{D}}[y \neq \mathrm{sgn}(\langle w_S, x\rangle)] = \mathbb{E}[\ell_{0-1}(\mathrm{sgn}(\langle w_S, x\rangle), y)]$
- However, the function $a \mapsto \ell_{0-1}(\mathrm{sgn}(a), y)$ is not Lipschitz
- Define the ramp loss

$$a \mapsto \ell_{\mathrm{ramp}}(a, y) \coloneqq \min\{1, \max\{0, 1 - ya\}\}$$

- For $y \in \{\pm 1\}$, $\ell_{\mathrm{ramp}}(\cdot, y)$ is 1-Lipschitz and bounded by 1
- $\Pr_{(x,y)\sim\mathcal{D}}[y \neq \mathrm{sgn}(\langle w, x\rangle)] \leq \mathbb{E}\left[\ell_{\mathrm{ramp}}(\langle w, x\rangle, y)\right] = \mathcal{R}(h_w)$



Sample Complexity of Hard-SVM

- Let $\mathcal{H} := \{x \mapsto \langle w, x\rangle : \|w\| \le \|w^\star\|\}$
- Then, with probability 1, w_S (the Hard-SVM solution) is in $\mathcal{H}$ and $\hat{\mathcal{R}}_S(w_S) = 0$
- By Contraction lemma,

$$\mathrm{R}\left(\ell_{\mathrm{ramp}} \circ \mathcal{H} \circ S\right) \le \mathrm{R}(\mathcal{H} \circ S) = \|w^\star\| \cdot \mathrm{R}(\mathcal{H}_2 \circ S) \le \|w^\star\| \frac{\max_i \|x_i\|}{\sqrt{m}} \le \frac{\|w^\star\| R}{\sqrt{m}}$$

- Thus, with probability $1 - \delta$,

$$\Pr_{(x,y)\sim\mathcal{D}}[y \neq \mathrm{sgn}(\langle w_S, x\rangle)] \le O\left(\frac{\|w^\star\| R + \sqrt{\log(1/\delta)}}{\sqrt{m}}\right)$$

■