Recap.

Probably Approximately Correct (PAC) learning

Learner
- does not know distribution D or true hypothesis f
- chooses parameters for accuracy ε and confidence δ.
- training data $S = \{(x_i, y_i = f(x_i))\}$, of size $m(\varepsilon, \delta)$
- outputs a hypothesis h s.t.

$$\mathbb{P}(L_{D,f}(h) \leq \varepsilon) \geq 1 - \delta$$

then, this learner is Probably (w.p. $\geq 1-\delta$) Approximately (error $\leq \varepsilon$) Correct.

Empirical Risk Minimization (ERM)

- train error: $L_S(h) \triangleq \frac{1}{m} |\{i \in [m] : h(x_i) \neq y_i\}|$

- ERM: $h_S \in \underset{h \in H}{\arg\min}\ L_S(h)$

  ↑ hypothesis class of some complexity, like $|H|$

- claim: If $|H| < \infty$, f, D are realizable, and sample complexity: $m \geq \frac{\log(\frac{|H|}{\delta})}{\varepsilon}$, then for any (ε, δ)

$$\text{PAC learnable}: \quad \mathbb{P}_S(L_{D,f}(h_S) > \varepsilon) \leq \delta$$

↑ ERM

Lecture 11. Keywords: Uniform Convergence, Agnostically PAC learnable, Hoeffding's Inequality.

Relaxing the realizability assumption: $y_i = f(x_i)$, $f \in \mathcal{H}$.
↑ deterministic

Data generating distribution $\mathcal{D}$ is over $X \times Y$: $z = (x, y) \sim \mathcal{D}$
↑ stochastic = random

True error: $L_{\mathcal{D}}(h) \triangleq \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$

Empirical error: $L_S(h) \triangleq \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$

Goal: find h minimizing $L_{\mathcal{D}}(h)$

if given oracle access to $\mathcal{D}$, then Bayes Optimal Predictor achieves lower error than any other predictor [Theory HW 1, Q2] due 5/14

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}_{\mathcal{D}}(y=1|x) \geq 1/2 \\ 0 & \text{else} \end{cases}$$

* Discuss.

Agnostic PAC learnability.

A class of hypotheses, $\mathcal{H}$, is **Agnostic PAC Learnable** if $\exists m_{\mathcal{H}}$ and a learning algorithm s.t. for ε, δ & $\mathcal{D}$,

running the Algorithm with $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ samples returns

h s.t. $L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$

Relaxed accuracy requirement.

Relaxing Binary Classification: what needs to change in our framework?

- multi-class classification $\mathcal{Y} = [K]$

  loss: $\ell_h(x_i, y_i) = \mathbb{I}(h(x_i) \neq y_i)$

- Regression $\mathcal{Y} = \mathbb{R}$

  loss: $\ell_h(x_i, y_i) = (h(x_i) - y_i)^2$ (same)

Risk function $L_D(h) \triangleq \mathbb{E}_{(x,y)\sim D}[\ell(h, (x,y))]$

Empirical Risk function $L_S(h) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(h, (x_i, y_i))$

[Q 3.3 from book.] Problem Definition:

$\mathcal{X} = \mathbb{R}^2$

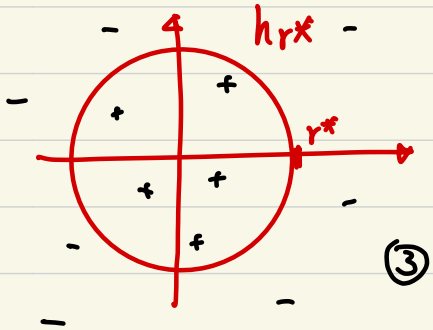
$\mathcal{Y} = \{0, 1\}$

$\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ ($|\mathcal{H}| = \infty$), $h_r(x) = \mathbb{I}\{\|x\| \leq r\}$ ($\|\cdot\|$: L_2 norm)

Prove that $\mathcal{H}$ is PAC learnable (assuming realizability) with. $m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil \frac{\log 1/\delta}{\varepsilon} \rceil$

solution: ① Under realizability assumption ⟶ fix $x \sim D$, and f_{r^*}

② Define : Algorithm : return smallest circle that includes all +.



$H_{BAD} = \{h_r : r \leq \tilde{r} < r^*, \text{where } P_{x\sim D}(h_{\tilde{r}}(x) \neq h_{r^*}(x)) = \varepsilon\}$

$= \{h_r : r \leq \tilde{r} < r^*, \text{where } P_{x\sim D}(\tilde{r} \leq \|x\| \leq r^*) = \varepsilon\}$

③ Analysis : claim: $A_{lgo} \leq r^*$

$P_S(A_{lgo} < \tilde{r}) = P(\|x_i\| < \tilde{r})^m$

$\leq (1-\varepsilon)^m \leq e^{-\varepsilon m} \leq e^{-\log 1/\delta} \leq \delta$ ($m \geq \frac{1}{\varepsilon}\log\frac{1}{\delta}$)

Chapter 4

Uniform Convergence is sufficient for learnability.

Definition: (ε-representative sample)

Training set S is ε-representative if $\forall h \in \mathcal{H}$

$$|L_S(h) - L_D(h)| \le \varepsilon$$

Intuitively, S is large enough and good representation of D.

$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$: Empirical Risk $\longleftrightarrow$ population Risk : $L_D(h) = \mathbb{E}_D[\ell(h,z)]$

$\arg\min_{h \in \mathcal{H}} L_S(h)$ ERM $\longleftrightarrow$ best in $\mathcal{H}$. $\arg\min_{h \in \mathcal{H}} L_D(h)$.

Lemma. Assume we are given a training set S that is $\frac{\varepsilon}{2}$-representative, then any output from ERM satisfies

$$L_D(h_S) \le \min_{h \in \mathcal{H}} L_D(h) + \varepsilon$$

How do we apply this lemma? Assume $m_{\mathcal{H}} \ge 1$ $\square$ $\Rightarrow$ (Uniform Convergence) S is $\varepsilon/2$-representative for $\mathcal{H}$ $\Rightarrow$ (Lemma) ERM is agnostic PAC

Proof of Lemma: by triangular inequality: $a \le b + |a-b|$

$$L_D(h_S) \overset{\Delta \text{ ineq.}}{\le} |L_D(h_S) - L_S(h_S)| + L_S(h_S)$$

$$\le \varepsilon/2 + L_S(h) \quad , \forall h \in \mathcal{H}$$

($\varepsilon/2$: $\uparrow$ representative S; $L_S(h)$: $\uparrow$ def ERM)

$$\overset{\Delta \text{ ineq.}}{\le} \varepsilon/2 + |L_S(h) - L_D(h)| + L_D(h)$$

$$\le \varepsilon/2 + \varepsilon/2 + L_D(h).$$

To show ERM is Agnostic PAC learner, we are left to show that, with probability $1-\delta$, S is $\frac{\epsilon}{2}$-representative.

Definition (Uniform Convergence)

A hypothesis class $\mathcal{H}$ has the uniform convergence property, if $\exists\, m_{\mathcal{H}}^{UC}$ s.t. for every ϵ, δ and every D, the following holds:

if $m \geq m_{\mathcal{H}}^{U.C}(\epsilon, \delta)$ then S is ϵ-representative w.p $\geq 1-\delta$.

Corollary 1. $\mathcal{H}$ is agnostically PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$

and ERM is the Agnostic PAC learner.

Next, we want to show that finite $\mathcal{H}$, i.e. $|\mathcal{H}| < \infty$, is agnostic PAC learnable ← by showing uniform convergence.

Claim. Any finite $\mathcal{H}$, i.e., $|\mathcal{H}| < \infty$, is uniformly convergent with

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{\epsilon} \right\rceil$$

Corollary 2. ERM is agnostic PAC learner with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \underset{\text{Corollary 1}}{\leq} m_{\mathcal{H}}^{U.C}\left(\frac{\epsilon}{2}, \delta\right) \underset{\text{Claim.}}{\leq} \left\lceil \frac{2 \log \frac{2|\mathcal{H}|}{\delta}}{\epsilon} \right\rceil$$

we are left to prove the claim.

$\leftrightarrow$ Q. assuming $0 \le \ell(h, z_i) \le 1$, what is the sample size we need to guarantee $|L_S(h) - L_D(h)| \le \epsilon$ w.p $1-\delta$?

n.t.s (need to show) $\quad \mathbb{P}_S(\{S : \forall h \in H, |L_S(h) - L_D(h)| \le \epsilon\}) \ge 1-\delta$

n.t.s $\quad \leftrightarrow \mathbb{P}_S(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$

$$\text{LHS} \overset{\text{union bound}}{\le} \sum_{h \in H} \underbrace{\mathbb{P}_S(|L_S(h) - L_D(h)| > \epsilon)}_{\mathbb{P}(|\frac{1}{m}\sum_{i=1}^m \ell(h, z_i) - \mathbb{E}[\ell(h,z)]| > \epsilon) \quad (*)}$$

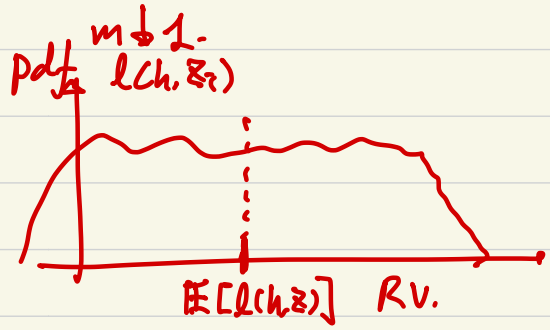
Recall: $L_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h, z_i)$ ← Average of iid r.v.s, $z_i = (x_i, y_i)$

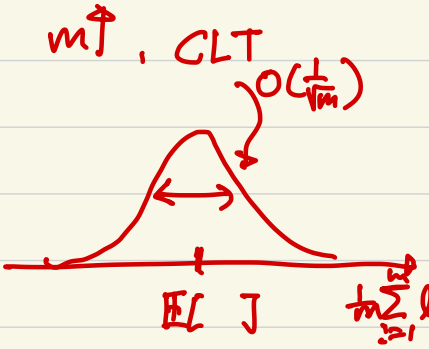
$L_D(h) = \mathbb{E}_{z \sim D}[\ell(h,z)] = \mathbb{E}[L_S(h)]$ ← expectation of each term

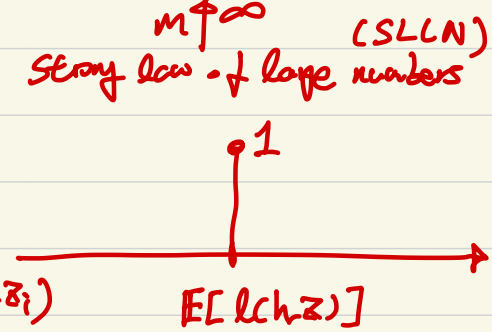
* we need to show that the measure (distribution) of R.V. $L_S(h)$ concentrates around its mean.

* the most important tool in statistical analysis for learning.
  - concentration of measure
  - tail bound

Q. How fast (in terms of # of samples) does the measure of empirical mean concentrate around its expected value







End of lecture

Tail bound: $\mathbb{P}(\frac{1}{m}\sum_{i=1}^m \ell(h, z_i) - \mathbb{E}[\ ] > \epsilon) \le F(m, \epsilon)$.

* Hoeffding's Inequality

Let $\theta_1, \theta_2, \dots, \theta_m$ be a set of i.i.d. random variables with $\mu \triangleq \mathbb{E}[\theta_i]$, satisfying

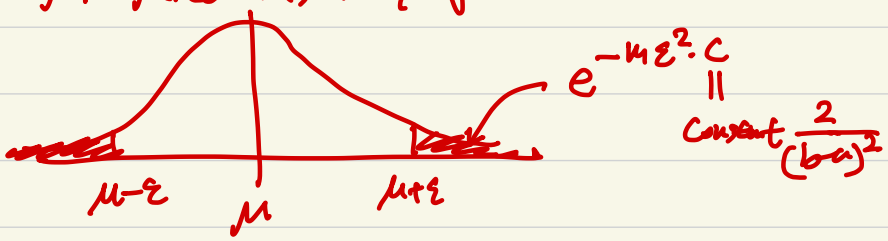
$$\mathbb{P}(a \le \theta \le b) = 1, \qquad \text{[Boundedness]}$$

then for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right) \le 2 \cdot e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

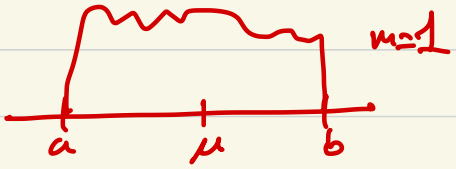
* first example of concentration of measure.

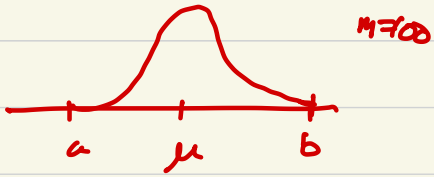
for fixed m, varying ε



$\text{constant} = \frac{2}{(b-a)^2}$

for fixed ε, varying m.





---

letting $\theta_i = \ell(h, z_i)$, $\underset{a}{0} \le \ell(h, z_i) \le \underset{b}{1}$, $\mu = L_D(h)$

Hoeffding's $\Rightarrow$ $\mathbb{P}_S\left(|L_S(h) - L_D(h)| > \varepsilon\right) = \mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right)$

$$\le 2 \cdot e^{-2m\varepsilon^2}$$

$$(*) \le \sum_{h \in H} 2e^{-2m\varepsilon^2} \le 2|H| \cdot e^{-2m\varepsilon^2} \le \delta$$

let $m(\varepsilon, \delta) \ge \frac{\log\left(\frac{2|H|}{\delta}\right)}{2\varepsilon^2}$

* Summary.

if $m \geq \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2}$

then $\mathcal{H}$ has uniform convergence

Lemma: $(\frac{\epsilon}{2}, \delta)$-uniform convergence

↓

(ϵ, δ) PAC learnable with ERM

if $m \geq \left\lceil \frac{2 \cdot \log(\frac{2|\mathcal{H}|}{\delta})}{\epsilon} \right\rceil \geq m_{\mathcal{H}}^{U.C}(\epsilon/2, \delta) \geq m_{\mathcal{H}}(\epsilon, \delta)$.