# Lecture 3

Recap:

→ PAC learnability for finite hypothesis classes

→ No-free lunch theorem

→ Agnostic PAC learning

→ Define Bayes-optimal predictor for 0/1 loss.

$$h^*(x) = \mathbb{1}\left(\eta(x) \geq 1/2\right)$$

$$\eta(x) = Pr_D\left(y = 1 \mid x = x\right)$$

$$R(h^*) = \mathop{\mathbb{E}}_{x \sim D(x)}\left[\min\{\eta(x), 1 - \eta(x)\}\right]$$

Today

→ Define Agnostic PAC

→ Define Uniform Convergence (UC)

→ UC => Agnostic PAC Learning

→ VC results

→ Concentration of measure

# Agnostic PAC

**Definition (Agnostic PAC learnability).**

A hypothesis $\mathcal{H}$ is agnostic PAC learnable if for every $\epsilon, \delta \in (0,1)$ $\exists$ a function $n_{\mathcal{H}}(\epsilon, \delta)$ & a learning algorithm s.t. for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if the algorithm is run on $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ samples drawn iid from $\mathcal{D}$, then the algorithm returns a hypothesis $\hat{h}$ with $R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \epsilon$, except with probability $\delta$.

PAC learning assumes realizability.

Realizability : $h^*$ (Bayes optimal predictor) $\in \mathcal{H}$.

$\therefore$ Agnostic PAC more general than PAC learning

e.g. $\mathcal{H} = \{ h_w(x) : \mathbb{1}(w^T x > 0),\ w \in \mathbb{R}^d \}$

(linear classifiers)

Agnostic PAC for this $\mathcal{H}$ asks we do as well as best possible linear classifier.

# Uniform Convergence

Idea: ERM outputs $\hat{h} \in \mathcal{H}$, which has minimum risk over training set

Want: $\hat{h}$ is close to minimizer of population risk ($\min_{h \in \mathcal{H}} R(h)$).

Suffices: All empirical risks of all members of $\mathcal{H}$ are close to their population risk.

$$\tilde{h} = \operatorname*{argmin}_{h \in \mathcal{H}} R(h) \qquad (\text{Note that } \tilde{h} \neq h^*)$$

$$R(h_{S,ERM}) - R(\tilde{h}) = R(h_{S,ERM}) - \hat{R}(h_{S,ERM}) + \hat{R}(h_{S,ERM}) - \hat{R}(\tilde{h}) + \hat{R}(\tilde{h}) - R(\tilde{h})$$

trickier ↙ $\leq 0$ ↙ lossy to bound ↙

$$\hat{R}(\tilde{h}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\tilde{h}(x_i), y_i)$$

avg. of $n$ i.i.d. random variable

$$\forall i \in [n], \quad \mathbb{E}_{(x_i, y_i) \sim D} [\ell(\tilde{h}(x_i), y_i)] = R(\tilde{h})$$

$$\Pr\left[\left|\hat{R}(\tilde{h}) - R(\tilde{h})\right| \geq \varepsilon\right] \leq \delta.$$

Holds for any fixed hypothesis

**Definition (Uniform convergence)**

A hypothesis class $\mathcal{H}$ has the uniform convergence property if for every $\varepsilon, \delta \in (0,1)$ there exists a function $n_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ such that for every dist. $D$ over $X \times Y$, if $S$ is a training set of $n \geq n_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ samples drawn iid from $D$, then w.p. $1-\delta$,

$$\forall h \in \mathcal{H}, \quad \left|\hat{R}(h) - R(h)\right| \leq \varepsilon.$$

**Proposition (UC => Agnostic PAC learning)**

If $\mathcal{H}$ has the UC property with $n_{\mathcal{H}}^{UC}(\epsilon, \delta)$, then $\mathcal{H}$ is Agnostic PAC learnable with sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq n_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Moreover, ERM is an algorithm which achieves this sample complexity.

**Proof**

Let $S$ be a sample of size $n \geq n_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.

By definition,

$$\forall h \in \mathcal{H}, \ |\hat{R}(h) - R(h)| \leq \epsilon/2.$$

Consider ERM $h_{S,ERM}$ & let $\tilde{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} R(h)$.

$$R(h_{S,ERM}) - R(\tilde{h}) = \underbrace{R(h_{S,ERM}) - \hat{R}(h_{S,ERM})}_{\leq \epsilon/2} + \underbrace{\hat{R}(h_{S,ERM}) - \hat{R}(\tilde{h})}_{\leq 0} + \underbrace{\hat{R}(\tilde{h}) - R(\tilde{h})}_{\leq \epsilon/2}$$

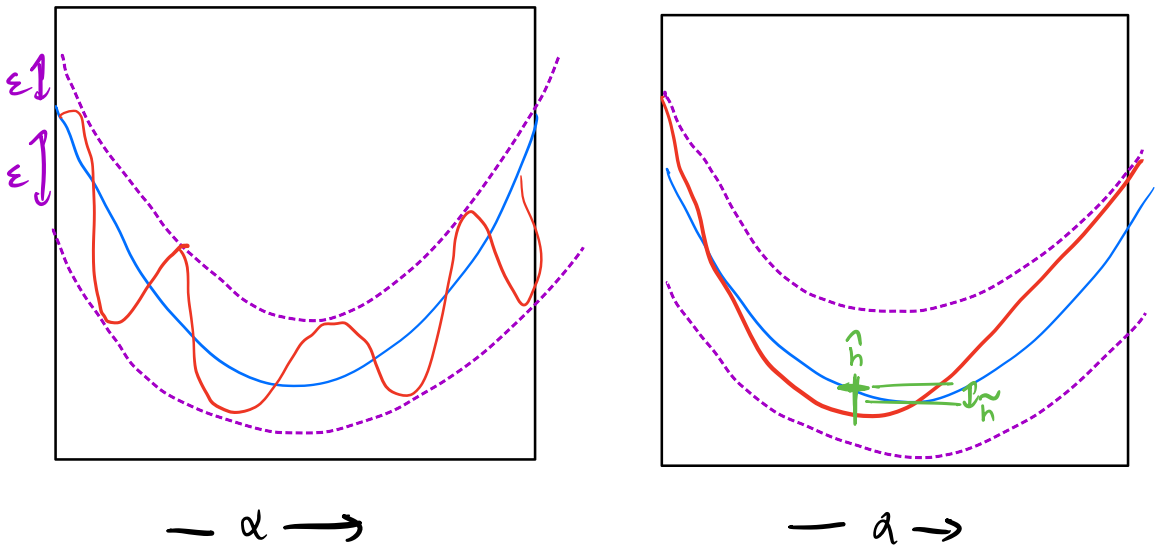
$$R(h_{S,ERM}) - R(\tilde{h}) \leq \epsilon.$$

■

Consider : $\mathcal{H}$ is parameterized by single parameter

($\alpha \in \mathbb{R}$)

— : $R(h)$ ------ : VC bound — : $\hat{R}_S(h)$



**Theorem (Agnostic PAC for finite classes)**

Let $\mathcal{H}$ be a class $|\mathcal{H}| < \infty$. Then $\mathcal{H}$ is agnostic-PAC learnable with

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

**Proof** (via Uniform convergence)

We will show:

(1) For any fixed $h \in \mathcal{H}$ and $\epsilon > 0$,

$$\Pr\left[ |\hat{R}(h) - R(h)| \leq \epsilon \right] \geq 1 - 2e^{-2n\epsilon^2}.$$

(2) For any $\epsilon > 0$,

$$\Pr\left[ \forall h \in \mathcal{H}, |\hat{R}(h) - R(h)| \leq \epsilon \right] \geq 1 - 2|\mathcal{H}| e^{-2n\epsilon^2}.$$

(3) For $n \geq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$, with probability $1-\delta$,

$$|\hat{R}(h) - R(h)| \leq \epsilon \quad \forall h \in \mathcal{H}.$$

(4) By UC, $\mathcal{H}$ is agnostic PAC learnable with $n \geq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$ samples.

## Proof of (1)

**Lemma (Hoeffding's inequality):**

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \le X_i \le b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \le \epsilon\right] \ge 1 - 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Given Hoeffding's, we prove (1).

Take each $X_i = \ell(h(x_i), y_i)$

Since $\ell(h(x_i), y_i) = \mathbb{1}(h(x_i) \neq y_i)$

$X_i \in \{0,1\} \Rightarrow a_i = 0, \; b_i = 1.$

$$\Pr\left[\left|\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\hat{R}(h)} - \mathbb{E}\left[\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{R(h)}\right]\right| \le \epsilon\right] \ge 1 - 2\exp(-2n\epsilon^2)$$

Proof of (2).

$$\Pr[\forall h \in \mathcal{H}, |\hat{R}(h) - R(h)| \le \varepsilon]$$

$$= 1 - \Pr\left[\bigcup_{i=1}^{|\mathcal{H}|} |\hat{R}(h) - R(h)| > \varepsilon\right]$$

$$\ge 1 - \sum_{i=1}^{|\mathcal{H}|} \Pr[|\hat{R}(h_i) - R(h_i)| > \varepsilon]$$

$$\ge 1 - |\mathcal{H}| (2 \exp(-2n\varepsilon^2))$$

Proof of (3).

Set $n \ge \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$

failure prob. $\le \delta$.

Proof of (4)

UC $\Rightarrow$ Agnostic PAC

■

## Concentration inequalities

Let $\{x_1, x_2, \ldots\}$ be sequence of random variables $\mathbb{E}[x_i] = \mu$, $\mathrm{Var}(x_i) = \sigma^2 < \infty$. Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

How close is $\bar{x}_n$ to $\mu$?

## Central Limit Theorem

**Thm (Central Limit Thm):** As $n \to \infty$

$$(\bar{x}_n - \mu) \xrightarrow{\text{dist.}} \mathcal{N}(0, \sigma^2/n).$$

$\bar{x}_n$ converges to $\mu$ at rate of $O\left(\frac{\sigma}{\sqrt{n}}\right)$.

**Markov's bound.**

Let $x$ be a non-negative random variable.

Then for any $t > 0$, we have

$$\Pr[x \geq t] \leq \frac{\mathbb{E}[x]}{t}.$$

**Proof.**

$$\mathbb{E}[x] = \sum_{i=0}^{\infty} i \Pr[x = i]$$

$$\geq \sum_{i=t}^{\infty} i \Pr[x = i] \quad \text{(non-negativity)}$$

$$\geq t \cdot \sum_{i=t}^{\infty} \Pr[x = i]$$

$$\Rightarrow \Pr[x \geq t] \leq \mathbb{E}[x]/t. \quad \blacksquare$$

**Exercise:**
Show Markov's bound is tight.

Only applies to non-negative.

Can work with variance.

**Chebyshev's inequality:**

Let $X$ be a random variable. Then for any $t > 0$.

$$\Pr[|X - \mathbb{E}(X)| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}.$$

**Proof.**

$$\Pr[|X - \mathbb{E}(X)| \geq t] = \Pr[|X - \mathbb{E}(X)|^2 \geq t^2]$$

Apply Markov's inequality to the non-negative r.v. $(X - \mathbb{E}(X))^2$,

$$\Pr[(X - \mathbb{E}(X))^2 \geq t^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{t^2}$$

$$= \frac{\mathrm{Var}[X]}{t^2}.$$

■

Is this enough to get Hoeffding's?

**Lemma (Hoeffding's inequality):**

Let $x_1, x_2, \ldots, x_n$ be independent random variables such that $a_i \le x_i \le b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right| \le \epsilon\right] \ge 1 - 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$a_i = 0$

$b_i = 1$

$\mathbb{E}[x_i] = \mu \quad \forall i.$

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathrm{Var}[\bar{x}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[x_i] \le \frac{1}{n}$$

$$\mathbb{E}[\bar{x}_n] = \mu$$

For any r.v. $x_1, x_2$ (independent),

$\mathrm{Var}[x_1 + x_2] = \mathrm{Var}[x_1] + \mathrm{Var}[x_2]$

$\mathrm{Var}[x_i] = \mathbb{E}[x_i^2] - \mathbb{E}[x]^2 \le 1$

By Chebyshev,

$$\Pr\left[\left|\bar{f}_n - \mu\right| \geq \varepsilon\right] \leq \frac{\mathrm{Var}[\bar{f}_n]}{\varepsilon^2}$$

$$\leq \frac{1}{n\varepsilon^2}.$$

But Hoeffding's inequality says,

$$\Pr\left[\left|\bar{f}_n - \mu\right| \geq \varepsilon\right] \leq 2\exp(-2n\varepsilon^2).$$

Idea: No need to stop at second moment.

for any positive integer $k$,

$$\Pr\left[\left|f - \mathbb{E}(f)\right| \geq t\right] = \Pr\left[\left|f - \mathbb{E}(f)\right|^k \geq t^k\right]$$

$$\leq \frac{\mathbb{E}\left(\left|f - \mathbb{E}(f)\right|^k\right)}{t^k}.$$

Consider functions other than polynomial.

For $\lambda \geq 0$,

$$\Pr[X - \mathbb{E}(X) \geq t] = \Pr[e^{\lambda(X - \mathbb{E}(X))} \geq e^{\lambda t}]$$

$$\leq \frac{\mathbb{E}\left(e^{\lambda(X - \mathbb{E}(X))}\right)}{e^{\lambda t}}$$

$$\leq \inf_{\lambda \geq 0} \frac{\mathbb{E}\left(e^{\lambda(X - \mathbb{E}(X))}\right)}{e^{\lambda t}}$$

**Chernoff-style bounds.**

$$\log\left(\Pr[X - \mathbb{E}(X) \geq t]\right) \leq \inf_{\lambda \geq 0} \left(\log\left(\mathbb{E}\left(e^{\lambda(X - \mathbb{E}(X))}\right)\right) - \lambda t\right).$$

**Moment-generating function.**

For any r.v. $X$, the moment generating function

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Exercise: $\left.\frac{\partial^k M_X(t)}{\partial t^k}\right|_{t=0} = \mathbb{E}[X^k]$