CSCI699: Theory of Machine Learning Fall 2021

Lecture 3: Agnostic Learning and Uniform Convergence

Instructor: Vatsal Sharan *Scribe: Ta-Yang Wang & Yingxiao Ye*

Today

- Define Agnostic PAC
- Define Uniform Convergence (UC)
- UC $\implies$ Agnostic PAC Learning
- UC results
- Concentration of measure

1 Agnostic PAC Learning

Definition 1 (Agnostic PAC Learnability)**.** *A hypothesis $\mathcal{H}$ is agnostic PAC learnable if for every $\epsilon, \delta \in (0,1)$, there exists a function $n_{\mathcal{H}}(\epsilon, \delta)$ and a learning algorithm such that 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 i.i.d. 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 δ.*

Remark.

1. *PAC learning assumes realizbility while agnostic PAC learning does not. The realizability assumption requires that there exists h^* (Bayes optimal predictor) in the hypothsis class $\in \mathcal{H}$ such that $R(h^*) = 0$.*
2. *PAC learnable implies agnostic PAC learnable. Therefore Agnostic PAC learning is more general than PAC learning.*
3. *As an example, consider the hypothesis class*

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

   Learning $\mathcal{H}$ agnostically means finding a linear classifier which is at most ϵ-suboptimal compared to the best linear classifier on the data.

2 Uniform Convergence

- Ideally, we hope ERM outputs a predictor $\hat{h} \in \mathcal{H}$ which has minimum risk over training set
- We want to show that $\hat{h}$ is close to the minimization of population risk $(\min_{h \in \mathcal{H}} R(h))$.

- It **suffices** to show that **all** empirical risks of **all** members of $\mathcal{H}$ are close to their population risk.

- Recall our definitions. The risk of the predictor
$$R(h) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[\ell(h(x_i), y_i)].$$
And the empirical risk
$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i).$$

- Let $h_{s,\text{ERM}} = \arg\min_{h\in\mathcal{H}} \hat{R}(h)$ and $\widetilde{h} = \arg\min_{h\in\mathcal{H}} R(h)$ (Note that $\widetilde{h} \neq h^*$ since we do not assume realizability).

We have
$$R(h_{s,\text{ERM}}) - R(\widetilde{h}) = \underbrace{R(h_{s,\text{ERM}}) - \hat{R}(h_{s,\text{ERM}})}_{\text{trickier}} + \underbrace{\hat{R}(h_{s,\text{ERM}}) - \hat{R}(\widetilde{h})}_{\leq 0} + \underbrace{\hat{R}(\widetilde{h}) - R(\widetilde{h})}_{\text{easy to bound}}. \tag{1}$$

In (1), $\hat{R}(h_{s,\text{ERM}}) - \hat{R}(\widetilde{h}) \leq 0$ due to the definition of $h_{s,\text{ERM}}$.

Note that by definition,
$$\hat{R}(\widetilde{h}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\widetilde{h}(x_i), y_i)}_{\text{average of } n \text{ i.i.d. random variable}},$$

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

Therefore, by concentration inequalities we will introduce soon, it is not difficult to get
$$\Pr[|\hat{R}(\widetilde{h}) - R(\widetilde{h})| \geq \epsilon] \leq \delta$$
for any fixed hypothesis $\widetilde{h}$. Uniform convergence asks for such a deviation bound for every hypothesis in the hypothesis class.

Definition 2 (Uniform Convergence)**.** *A hypothesis class $\mathcal{H}$ has the* ***uniform convergence property*** *if for every $\epsilon, \delta \in (0,1)$, there exists a function $n_{\mathcal{H}}^{UC}(\epsilon,\delta)$ such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, if S is a training set of $n \geq n_{\mathcal{H}}^{UC}(\epsilon,\delta)$ samples drawn i.i.d. from $\mathcal{D}$, then with probability $1-\delta$,*
$$\forall\, h \in \mathcal{H}, |\hat{R}(h) - R(h)| \leq \epsilon.$$

Proposition 3 (UC $\implies$ 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 $n_{\mathcal{H}}(\epsilon,\delta) \leq n_{\mathcal{H}}^{UC}(\epsilon,\delta)$. Moreover, ERM is an algorithm which achieves this sample complexity.*

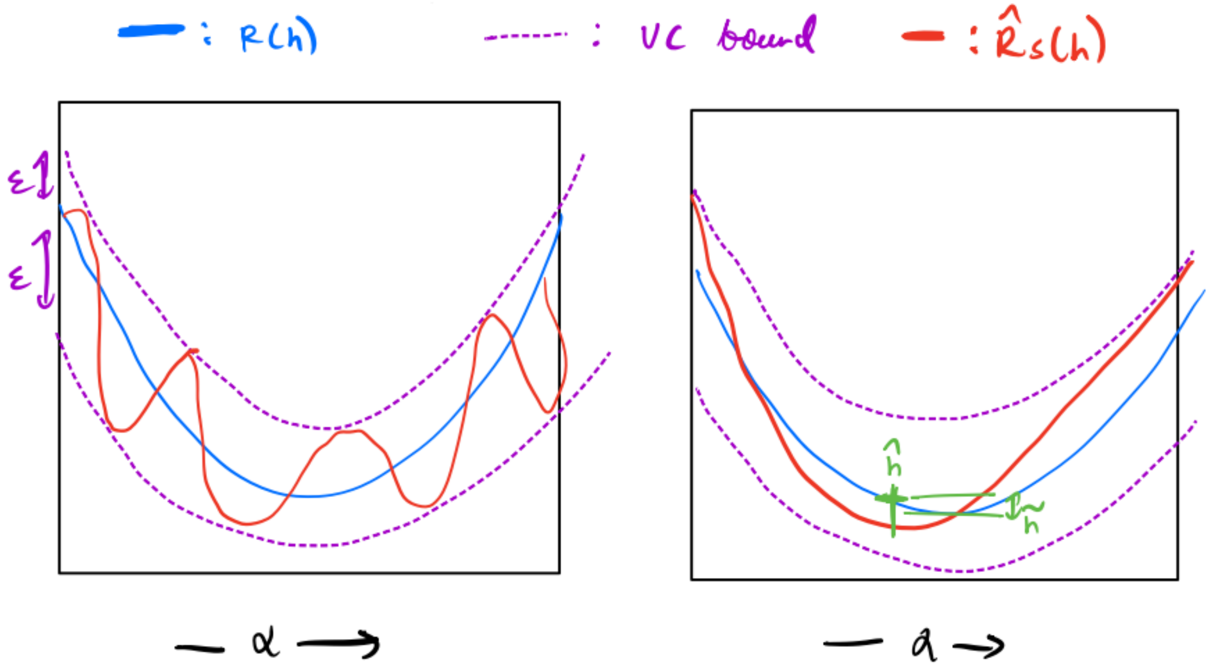

Figure 1: Consider the hypothesis class $\mathcal{H}$ is parameterized by single parameter ($\alpha \in \mathbb{R}$) and has the UC property. Then the deviation of the empirical risk from the true risk is bounded by ϵ. The figure on the right consider a stronger case: $\hat{R}(h)$ is a convex function with respect to the parameter α. In this case, there exists a unique local minimizer (and thus it is the global minimizer) so optimization method e.g. stochastic gradient descent algorithm can find the optimal solution. In contrast, in the figure on the left, there are multiple local minimizers. SGD can converge to anyone of them, which may lead to a suboptimal solution. Uniform convergence does not distinguish between these landscapes, but there is recent work on doing this.

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

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

Consider ERM $h_{s,\mathrm{ERM}}$ and let $\widetilde{h} = \underset{h \in \mathcal{H}}{\arg\min}\, R(h)$.

$$R(h_{s,\mathrm{ERM}}) - R(\widetilde{h}) = \underbrace{R(h_{s,\mathrm{ERM}}) - \hat{R}(h_{s,\mathrm{ERM}})}_{\leq \epsilon/2} + \underbrace{\hat{R}(\hat{h}_{s,\mathrm{ERM}}) - \hat{R}(\widetilde{h})}_{\leq 0} + \underbrace{\hat{R}(\widetilde{h}) - R(\widetilde{h})}_{\leq \epsilon/2}.$$

□

Theorem 4 (Agnostic PAC for finite classes)**.** *Let $\mathcal{H}$ be a class $|\mathcal{H}| < \infty$. Then $\mathcal{H}$ is agnostic-PAC learnable with*

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

Proof. We will show:

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

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

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

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

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

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

(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 5 (Hoeffdings inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq x_i \leq b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,*

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

Given Hoeffding's, we prove (1): take each $X_i = \ell(h(x_i), y_i)$. Since $\ell(h(x_i), y_i) = 1(h(x_i) \neq y_i)$, $X_i \in \{0, 1\}$, and thus $a_i = 0, b_i = 1$. Therefore, we have

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

Proof of (2)

$$\begin{aligned}
\Pr[\forall\, h \in \mathcal{H}, |\hat{R}(h) - R(h)| \leq \epsilon] &= 1 - \Pr\left[\bigcup_{i=1}^{|\mathcal{H}|} |\hat{R}(h) - R(h)| > \epsilon\right] \\
&\geq 1 - \sum_{i=1}^{|\mathcal{H}|} \Pr[|\hat{R}(h_i) - R(h_i)| > \epsilon] \\
&\geq 1 - |\mathcal{H}|(2\exp(-2n\epsilon^2))
\end{aligned}$$

Proof of (3)

Set $n \geq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$ in the previous step, then we have the failure probability is bounded bt δ for any $h \in \mathcal{H}$ i.e. $\Pr[\forall\, h \in \mathcal{H}, |\hat{R}(h) - R(h)| \leq \epsilon] \geq 1 - \delta$.

Proof of (4)

In (3), we have shown that $\mathcal{H}$ has the uniform convergence property, and thus by Proposition 3, we have $\mathcal{H}$ is agnostic PAC learnable with $n \geq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$ samples. □

3 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 $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. How close is $\overline{X_n}$ to $\mathcal{H}$?

Central Limit Theorem

The Central Limit Theorem gives an asymptotic answer to this question.

Theorem 6 (Central Limit Theorem)**.** *As $n \to \infty$*

$$\left(\overline{X_n} - \mu\right) \xrightarrow{\textit{in distribution}} \mathcal{N}(0, \sigma^2/n).$$

$\overline{X_n}$ converges to $\mathcal{H}$ at rate of $O(\frac{\sigma}{\sqrt{n}})$.

However, we are interested in bounds which hold non-asymptotically for finite n. Concentration inequalities provide this, the simplest of which is Markov's bound.

Markov's bound

Proposition 7. *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.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \Pr[X = i] \\
&\geq \sum_{i=t}^{\infty} i \Pr[X = i] \quad \text{(non-negativity)} \\
&\geq t \cdot \sum_{i=1}^{t} \Pr[X = i] \\
\implies & \Pr[X \geq t] \leq \mathbb{E}[X]/t.
\end{aligned}$$

□

Exercise: Show Markov's bound is tight.

Markov's inequality only applies to non-negative random variables. But we can work with the variance of a random variables to apply it to general random variables.

Chebyshev's inequality

Proposition 8. *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 random variable $(X - \mathbb{E}(X))^2$,

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

□

We can ask if this is already enough to get Hoeffding's inequality. Recall Hoeffding's inequality.

Lemma 9 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq x_i \leq b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,*

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

- Here, we let $a_i = 0, b_i = 0$, and $\mathbb{E}[X_i] = \mu, \ \forall\, i$.
- By definition, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and thus $\mathbb{E}[\overline{X}_n] = \mu$.

- Also, note that $\mathrm{Var}[\overline{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \leq \frac{1}{n}$. $(\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \leq 1)$

By Chebyshev,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq \frac{\mathrm{Var}[\overline{X}_n]}{\epsilon^2} \leq \frac{1}{n\epsilon^2}.$$

But Hoeffding's inequality says,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq 2\exp(-2n\epsilon^2).$$

Idea: No need to stop at second moment. In fact for any positive integer k,

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

We can even consider functions other than polynomials. For any $\lambda > 0$,

$$\begin{aligned}\Pr[X - \mathbb{E}(X) \geq t] &= \Pr\left[e^{\lambda(X - \mathbb{E}(X))} \geq e^{\lambda t}\right] \\ &\leq \frac{\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]}{e^{\lambda t}}.\end{aligned}$$

Since the inequality holds for any $\lambda > 0$, to get the best possible bound, we have

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

Chernoff-style bounds

This brings us to Chernoff-style bounds which take the following form.

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

Moment-generating function

The quantity that appears inside the logarithm of the Chernoff bound above is known as the *moment-generating function*. Formally, for any random variable X, the moment generating function $M_X(t)$ is defined as $M_X(t) = \mathbb{E}(e^{tx})$. An Exercise is to verify that

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

which is where the name moment-generating function comes from.