COMP 585 **Spring 2024**

Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*

March 19, 2024 *Scriber: Jefferson Hernandez*

Lecture 21

1 Fundamental Theorem of PAC Learning

In this lecture, we aim to prove one of the implications of the fundamental theorem of PAC learning. We aim to show that,

Theorem 1. *A Finite VCdim implies Uniform Convergence.*

Recall that the definition of uniform convergence is:

$$\forall \varepsilon, \delta; \quad \varepsilon \in [0,1], \exists m(\varepsilon, \delta, \mathcal{C}) \quad \text{s.t.} \quad \Pr\left[\sup_{c \in \mathcal{C}} |\text{err}(c) - \hat{\text{err}}(c)| < \varepsilon\right] \geq 1 - \delta$$

With a VCdim $= d$, it has been shown that the best bound on the value of $m(\varepsilon, \delta, \mathcal{C})$ are $m = O\left(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}\right)$ for the realizable case and $m = O\left(\frac{d+\log(1/\delta)}{\varepsilon^2}\right)$ for the agnostic case.

In this lecture, we will show a weaker version of this, specifically that $m = O\left(\frac{d}{(\delta\varepsilon)^2}\log\left(\frac{d}{\delta\varepsilon}\right)\right)$

We start with the following lemma.

Lemma 1.1. *For a concept class $\mathcal{C}$, with growth function $\tau_{\mathcal{C}}(m)$, we have that for all distributions $\mathcal{D}$ and parameters ε, δ, and sample set S of size m:*

$$\forall \varepsilon, \delta; \quad \varepsilon \in [0,1], \exists m(\varepsilon, \delta, \mathcal{C}) \quad s.t. \quad Pr\left[\sup_{c \in \mathcal{C}} |err(c) - \hat{err}(c)| < \varepsilon\right] \geq 1 - \delta,$$

implies that

$$\varepsilon = \frac{4 + \sqrt{\log(\tau(2m))}}{\delta\sqrt{2m}} = \frac{4 + \sqrt{\log(\frac{m}{d})}}{\sqrt{2m}},$$

using Sauer lemma to bound the growth function as $\tau(2m) \leq \left(\frac{2me}{d}\right)^d \approx m^d$. This also implies the number of samples needed is of the order $m = O\left(\frac{d}{(\delta\varepsilon)^2}\log\left(\frac{d}{\delta\varepsilon}\right)\right)$

To prove Lemma 1.1, we need to recall the following:

1. Jensen's inequality. For all convex functions $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$.

2. If $\Pr[x,y] = \Pr[y,x]$ then $\mathbb{E}_{x,y}[f(x,y)] = \mathbb{E}_{y,x}[f(y,x)]$.
3. Given $\sigma = \{\sigma_1, \cdots, \sigma_n\} \quad \forall a A \leq f(\sigma_i) \implies A \leq \mathbb{E}_\sigma[f(\sigma)]$

And the following lemma:

Lemma 1.2. *Let X be a random variable and $x' \in \mathbb{R}$ be a scalar and assume that there exists $a > 0$ and $b \geq e$ such that for all $t \geq 0$ we have $Pr[|X - x'| > t] \leq 2be^{-t^2/a^2}$. Then,* $\mathbb{E}[|X - x'|] \leq a(2 + \sqrt{\log(b)})$.

Proof. We will show

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} |\text{err}(c) - \hat{\text{err}}(c)|\right] \leq \frac{4 + \sqrt{\log(\frac{m}{d})}}{\sqrt{2m}}$$

which implies Lemma 1.1 by markvov's inequality.

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} |\text{err}(c) - \hat{\text{err}}(c)|\right] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} |\mathbb{E}_{S'\sim\mathcal{D}^m}[\hat{\text{err}}_{S'}(c)] - \hat{\text{err}}_S(c)|\right] \\
&\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} |\hat{\text{err}}_{S'}(c) - \hat{\text{err}}_S(c)|\right] \\
&\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} \frac{1}{m}\left|\sum_{i=1}^m \mathbb{1}_{c(x_i')\neq y_i'} - \mathbb{1}_{c(x_i)\neq y_i}\right|\right] \\
&\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{c\in\mathcal{C}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\mathbb{1}_{c(x_i')\neq y_i'} - \mathbb{1}_{c(x_i)\neq y_i}\right|\right] \quad \forall \sigma_i \in \{-1,+1\}^m \\
&\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\mathbb{E}_{\sigma_i\in\{-1,+1\}^m}\left[\sup_{c\in\mathcal{C}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\mathbb{1}_{c(x_i')\neq y_i'} - \mathbb{1}_{c(x_i)\neq y_i}\right|\right]
\end{aligned}$$

Let us restrict the class $\mathcal{C}$ to $\mathcal{C}_{S\cup S'}$ defined as

$$\mathcal{C}_{S\cup S'} = \{z = (z_1, \cdots z_m, z_1' \cdots z_m')|\exists c \in \mathcal{C} \quad \text{s.t.} c(x_i) = z_i \wedge c(x_i') = z_i', \}$$

then we have

$$\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\mathbb{E}_{\sigma_i\in\{-1,+1\}^m}\left[\sup_{c\in\mathcal{C}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\mathbb{1}_{z_i'\neq y_i'} - \mathbb{1}_{z_i\neq y_i}\right|\right]$$

Next, fix S and S', and since we restrict ourselves to the class $\mathcal{C}_{S\cup S'}$. Then, we replace the supremum with a maximum over the restricted class. Therefore,

$$\mathbb{E}_{S,S'\sim\mathcal{D}^m}\mathbb{E}_{\sigma_i\in\{-1,+1\}^m}\left[\sup_{c\in\mathcal{C}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\mathbb{1}_{z_i'\neq y_i'} - \mathbb{1}_{z_i\neq y_i}\right|\right] = \mathbb{E}_{\sigma_i\in\{-1,+1\}^m}\left[\max_{c\in\mathcal{C}_{S\cup S'}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\mathbb{1}_{z_i'\neq y_i'} - \mathbb{1}_{z_i\neq y_i}\right|\right]$$

Fix some $c \in \mathcal{C}_{S\cup S'}$ and denote $\theta_c = \sum_{i=1}^m \sigma_i\mathbb{1}_{z_i'\neq y_i'} - \mathbb{1}_{z_i\neq y_i}$. Since $\mathbb{E}[\theta_c] = 0$ and θ_c is an average of independent variables, each of which takes values in $[-1,1]$, we have by Hoeffding's inequality that for every $\varepsilon > 0$,

$$\Pr[|\theta_c| > \varepsilon] \leq 2\exp(-2m\varepsilon^2).$$

Applying the union bound over $c \in \mathcal{C}_{S \cup S'}$, we obtain that for any $\varepsilon > 0$,

$$\Pr\left[\max_{c \in \mathcal{C}_{S \cup S'}} |\theta_c| > \varepsilon\right] \leq 2|\mathcal{C}|\exp(-2m\varepsilon^2).$$

Finally, using Lemma 1.2 we get

$$\mathbb{E}\left[\max_{c \in \mathcal{C}_{S \cup S'}} |\theta_c|\right] \leq \frac{4 + \sqrt{\log(|\mathcal{C}|)}}{\sqrt{2m}}.$$

Combining all with the definition of the growth function τ_C from previous lecture, we have shown that

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{c \in \mathcal{C}} |\mathrm{err}(c) - \hat{\mathrm{err}}(c)|\right] \leq \frac{4 + \sqrt{\log(\tau_C(2m))}}{\sqrt{2m}}.$$

□