Uniform Convergence.

The goal is to prove that finite $\mathcal{H}$ are agnostic PAC learnable by the ERM algo.

$$A(S) \in \arg\min L_S(h).$$

The general idea is to show that if $|\mathcal{H}| < +\infty$ and $|S|$ is large enough (how large is to be determined) then $L_S(h)$ is sufficiently close to $L_{\mathcal{D}}(h)$ uniformly in $h \in \mathcal{H}$.

This will allow to check the condition:

$$\Pr_S \left\{ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \geq 1 - \delta.$$

for $|S|$ large enough $|S| = m > m_{\mathcal{H}}(\epsilon, \delta)$ for some $m_{\mathcal{H}}(\epsilon, \delta)$.

Definition: a training set S is ϵ-representative if $\forall h \in \mathcal{H}$: $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

In other words $L_S(h)$ is uniformly close to $L_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$. ["closeness" does not depend on h].

Lemma: Let S be $\frac{\epsilon}{2}$-representative. Let

$$h_S = ERM_{\mathcal{H}}(S) \equiv \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$$ be an

empirical risk minimizer. Then

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon .$$

Proof: $L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2}$; by $\frac{\epsilon}{2}$-repr def $\forall h$.

$L_{\mathcal{D}}(h) - \epsilon \leq L_S(h) \leq L_{\mathcal{D}}(h) + \epsilon$

$\leq L_S(h) + \frac{\epsilon}{2}$; by h_S minimizer def

$\leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2}$; by $\frac{\epsilon}{2}$-repr def $\forall h$

$= L_{\mathcal{D}}(h) + \epsilon .$ ∎

Lemma: Let $|\mathcal{H}| < +\infty$ a finite hyp class and let $\ell : \mathcal{H} \times \mathcal{Z} \to \{0,1\}$ the 0-1 loss fct (for classification say). Then $\mathcal{H}$ has the uniform convergence property:

$$\begin{cases} \Pr_S \{ \forall h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \} \geq 1-\delta \\ \Pr_S \{ S \text{ is } \epsilon\text{-representative} \} \geq 1-\delta \end{cases}$$

for training sets S with size $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$

where,

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

Note: The uniform convergence property just means that the empirical risk $L_S(h)$ is uniformly close to the true risk $L_{\mathcal{D}}(h)$ with high probability w.r.t S if $|S|$ is large enough.

Proof of Lemma.

We show the statement for the converse event:

$$\Pr\{\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} \overset{?}{\leq} \delta$$

$$\Pr\{\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \Pr\Big\{\bigcup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\Big\}$$

$$\underset{\text{union bound}}{\leq} \sum_{h \in \mathcal{H}} \Pr\{|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$$

$$= \sum_{h \in \mathcal{H}} \Pr\Big\{\Big|\underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)}_{\text{empirical mean}} - \underbrace{\mathbb{E}_{\mathcal{D}}\, \ell(h, z)}_{\text{true mean}}\Big| > \epsilon\Big\}$$

$$\leq \sum_{h \in \mathcal{H}} 2 \exp(-2 m \epsilon^2)$$ by Hoeffding's inequ. for $z_i \overset{iid}{\sim} \mathcal{D}$

$$= |\mathcal{H}|\, 2 e^{-2 m \epsilon^2}.$$

Imposing $|\mathcal{H}|\, 2 e^{-2m\epsilon^2} \leq \delta$ implies that

$$m \geq \frac{1}{2\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

Hoeffding inequality: Reminder or see exercises.

Let $\vartheta_1 \dots \vartheta_m$ be iid with $\mathbb{E}(\vartheta_i) = \mu$

and $a \leq \vartheta_i \leq b$. (RV have bounded support)

$$\Pr\left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \vartheta_i - \mu \right| > \epsilon \right\} \leq 2 e^{-2m \frac{\epsilon^2}{(a-b)^2}}$$

Note: for us $|a-b| = 1$ for 0-1 loss.

since $\vartheta_i = \ell(h, z_i) \in \{0, 1\}$.

but proof works also for any bounded loss fct.

Now we arrive at the main result of this chapter:

<u>Corollary</u>: If $\mathcal{H}$ is finite then it is agnostic PAC learnable with the ERM rule with sample complexity $m^{UC}_{\mathcal{H}}(\epsilon, \delta) = \left\lceil \frac{1}{2\epsilon^2} \log \frac{|\mathcal{H}|}{2\delta} \right\rceil$.

<u>Proof</u>: For $m \geq m^{UC}_{\mathcal{H}}(\epsilon, \delta)$ we have that

$$\Pr\{S \text{ is } \tfrac{\epsilon}{2}\text{-repr}\} \geq 1-\delta$$

We also proved that:

$$S \text{ is } \tfrac{\epsilon}{2}\text{-repr} \implies L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

so $$\Pr\{S \text{ is } \tfrac{\epsilon}{2}\text{-repr}\} \leq \Pr\{L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\}$$

Thus we obtain

$$\Pr\{L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\} \geq 1-\delta$$

$\blacksquare$