3.2. Learnability via uniform convergence

- We now learn about an important concept to verify PAC-learnability for classes $\mathcal{H}$.

- Let us start with an observation about the ERM learning rule:

Proposition 3.6:

Given $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ we have for $A = \mathrm{ERM}_{\mathcal{H}}$ almost surely

$$\varepsilon_{\mathsf{est}}(\mathcal{H}, S) = \mathcal{R}_\mu(A(S)) - \inf_{h \in \mathcal{H}} \mathcal{R}_\mu(h) \ \leq \ 2 \sup_{h \in \mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mu(h)|.$$

Definition 3.7: Uniform convergence (UC)

A class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ satisfies the uniform convergence condition (w.r.t. a loss ℓ) if there exists

- a mapping $m_{\mathcal{H}}^{\mathsf{uc}} \colon (0,1)^2 \to \mathbb{N}$

such that

- for *any* data distribution μ on $\mathcal{X} \times \mathcal{Y}$
- *any* $\epsilon \in (0,1)$ and $\delta \in (0,1)$,

we have

$$\mathbb{P}_{\mu^m}\left(\sup_{h \in \mathcal{H}} |\mathcal{R}_\mu(h) - \mathcal{R}_S(h)| \leq \epsilon\right) \geq 1 - \delta \qquad \forall m \geq m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta).$$

Uniform convergence ensures that learning models generalize well by bounding the difference between training and test error for all hypotheses.

Corollary 3.8:

If a class $\mathcal{H}$ satisfies (UC) w.r.t. a loss ℓ, then $\mathcal{H}$ is also PAC-learnable w.r.t. ℓ with $A = \mathrm{ERM}_{\mathcal{H}}$ and

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon/2, \delta).$$

Tools to control $|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)|$

Theorem 3.9: (Law of Large Numbers, 1713)

Let Z_i, $i \in \mathbb{N}$, be i.i.d. with $\mathbb{E}[|Z_i|] < +\infty$. Then

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \xrightarrow[m\to\infty]{\mathbb{P}} \mathbb{E}[Z_1]$$

Jakob Bernoulli
(1655 – 1705)

- Yields with $Z_i := \ell(h, (X_i, Y_i))$, $(X_i, Y_i) \sim \mu$ i.i.d. the asymptotic result (i.e., relates to consistency)

$$|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| \quad \xrightarrow[|S|\to\infty]{\mathbb{P}} \quad 0$$

- How about non-asymptotic bounds for $|\mathcal{R}_\mu(h) - \mathcal{R}_S(h)|$ for finite sample sizes $m = |S|$?

Concentration inequalities

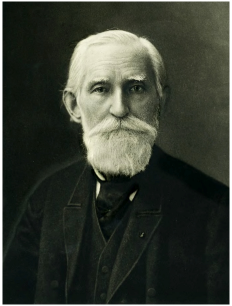
Proposition 3.10: Chebyschev inequality

Let $Z_1, \ldots, Z_m$ be i. i. d. with $\mathbb{V}[Z_i] < +\infty$. Then

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z_i]\right| > \epsilon\right) \leq \frac{\mathbb{V}[Z_i]}{m\epsilon^2}.$$

- Yields with $Z_i := \ell(h,(X_i,Y_i))$, $(X_i,Y_i) \sim \mu$ i.i.d.

$$\mathbb{P}\left(|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) \leq \frac{\mathbb{V}_\mu[\mu]}{m\epsilon^2}$$



Pafnuty L. Chebyshev
(1821 – 1894)

Lemma 3.11: Hoeffding's inequality

Let $Z_1, \ldots, Z_m$ be i. i. d. bounded random variables, i.e., $Z_i \in [a, b]$ almost surely for finite $a, b \in \mathbb{R}$. Then

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z_i]\right| > \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right).$$

Wasilly Hoeffding
(1914 – 1991)

- Yields sharper bounds than Chebyshev's inequality for bounded loss functions, e.g., for the 0-1 loss

$$\mathbb{P}\left(|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right).$$

- However, all these tools only hold for a **single, fixed hypothesis** h! We need a uniform bound

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) \leq \delta$$

PAC-learnability of finite classes

If the class $\mathcal{H}$ is finite, i.e., $\mathcal{H} = \{h_1, \ldots, h_n\}$, then we can apply the **union bound**

$$\begin{aligned}\mathbb{P}_{\mu^m}\left(\sup_{h\in\mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) &= \mathbb{P}_{\mu^m}\left(\exists h \in \mathcal{H} \colon |\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) \\ &= \mathbb{P}_{\mu^m}\left(\bigcup_{h\in\mathcal{H}} \{|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\}\right) \\ &\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{\mu^m}\left(|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right).\end{aligned}$$

Theorem 3.12:

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be finite and $\ell \colon \mathcal{H} \times \mathcal{D} \to \{0,1\}$ be the 0-1 loss. Then $\mathcal{H}$ satisfies (UC) w.r.t. ℓ with

$$m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil.$$

Hence, $\mathcal{H}$ is PAC-learnable with $A = \mathrm{ERM}_{\mathcal{H}}$ and $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon/2, \delta)$.

PAC-learnability of infinite classes

- We consider now a milestone of learning theory which establishes (UC) for arbitrary $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$.
- If the hypothesis class $\mathcal{H}$ is infinite, the union bound is not useful:

$$\mathbb{P}_{\mu^m}\left(\sup_{h\in\mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) \leq |\mathcal{H}| \sup_{h\in\mathcal{H}} \mathbb{P}_{\mu^m}\left(|\mathcal{R}_S(h) - \mathcal{R}_\mu(h)| > \epsilon\right) = \infty$$

- Luckily, a refined upper bound can be achieved by counting only those $h \in \mathcal{H}$ which yield different values on the training data $\{X_1, \ldots, X_m\}$, $m = |S|$

Definition 3.13:

Given a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and a finite set $M = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ we define the restriction of $\mathcal{H}$ to M by

$$\mathcal{H}_M := \{[h(x_1), \ldots, h(x_m)] \colon h \in \mathcal{H}\},$$

i.e., the set of all m-bits $\mathbf{b} \in \{0,1\}^m$ generated by an $h \in \mathcal{H}$ on M.

Example 3.14:

Heaviside hypotheses Let $\mathcal{X} = \mathbb{R}$ and consider the set of Heaviside classifiers

$$\mathcal{H} = \left\{\mathbb{1}_{[a,+\infty)} : a \in \mathbb{R}\right\} \qquad \text{where} \qquad \mathbb{1}_{[a,+\infty)}(x) = \begin{cases} 0, & x < a, \\ 1 & x \geq a. \end{cases}$$

How does $\mathcal{H}_M$ look like for various M?

- For $M = \{x_1\} \subset \mathbb{R}$ we have
$$\mathcal{H}_M = \{[0], [1]\}, \qquad |\mathcal{H}_M| = 2$$
- For $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 < x_2$, we have
$$\mathcal{H}_M = \{[0,0], [0,1], [1,1]\}, \qquad |\mathcal{H}_M| = 3$$
- For $M = \{x_1, x_2, x_3\} \subset \mathbb{R}$, $x_1 < x_2 < x_3$, we have
$$\mathcal{H}_M = \{[0,0,0], [0,0,1], [0,1,1], [1,1,1]\}, \qquad |\mathcal{H}_M| = 4$$
- For $M = \{x_1, \ldots, x_m\} \subset \mathbb{R}$, $x_1 < \ldots < x_m$, we have ... ?

Example 3.15: Interval hypotheses

Let $\mathcal{X} = \mathbb{R}$ and

$$\mathcal{H} = \{\mathbb{1}_{[a,b]} : a < b \in \mathcal{X}\}.$$

How does $\mathcal{H}_M$ look for various M?

- For $M = \{x_1\} \subset \mathbb{R}$ we have again

$$\mathcal{H}_M = \{[0], [1]\}, \qquad |\mathcal{H}_M| = 2$$

- For $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 < x_2$, we have

$$\mathcal{H}_M = \{[0,0], [0,1], [1,0], [1,1]\}, \qquad |\mathcal{H}_M| = 4$$

- For $M = \{x_1, x_2, x_3\} \subset \mathbb{R}$, $x_1 < x_2 < x_3$, we have

$$\mathcal{H}_M = \{[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [0,1,1], [1,1,1]\}, \qquad |\mathcal{H}_M| = 7$$

- And for $M = \{x_1, \ldots, x_m\} \subset \mathbb{R}$, $x_1 < \ldots < x_m$?

The growth function

We are now interested in the maximal number of binary m-bits generated by $\mathcal{H}$ on arbitrary $x_1, \ldots, x_m \in \mathcal{X}$

Definition 3.16:

For a binary hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ its growth function $\tau_{\mathcal{H}} \colon \mathbb{N} \to \mathbb{N}$ is given by

$$\tau_{\mathcal{H}}(m) := \sup_{M \subset \mathcal{X} \colon |M| = m} |\mathcal{H}_M|.$$

Example 3.17:

Let $\mathcal{X} = \mathbb{R}$ and consider again the class of Heaviside classifiers

$$\mathcal{H} = \{\mathbb{1}_{[a,+\infty)} \colon a \in \mathbb{R}\}.$$

Then

$$\tau_{\mathcal{H}}(m) = m + 1 \qquad \forall m \in \mathbb{N}.$$

Theorem 3.18: Uniform Convergence Theorem (UCT)

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary hypothesis class and ℓ be the 0-1-loss. Then for any distribution μ on $\mathcal{D} = \mathcal{X} \times \{0,1\}$ and any $\epsilon \in (0,1)$ we have

$$\mathbb{P}_{\mu^m}\left(\sup_{h\in\mathcal{H}} |\mathcal{R}_\mu(h) - \mathcal{R}_S(h)| > \epsilon\right) \leq 4\,\tau_{\mathcal{H}}(2m)\,\exp\left(-\epsilon^2 m/8\right) \qquad \forall m \geq 2\ln(4)/\epsilon^2.$$

Remark: Why $\tau_{\mathcal{H}}(2m)$ and not $\tau_{\mathcal{H}}(m)$? Because the proof involves the step

$$\mathbb{P}_{S\sim\mu^m}\left(\sup_{h\in\mathcal{H}} |\mathcal{R}_\mu(h) - \mathcal{R}_S(h)| > \epsilon\right) \leq 2\,\mathbb{P}_{S,\tilde{S}\sim\mu^m}\left(\sup_{h\in\mathcal{H}} |\mathcal{R}_{\tilde{S}}(h) - \mathcal{R}_S(h)| > \epsilon/2\right)$$

Corollary 3.19:

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary hypothesis class and ℓ be the 0-1-loss. If $\tau_{\mathcal{H}}$ grows subexponentially, i.e., for any $\epsilon > 0$ exists a $c_\epsilon < \infty$ such that

$$\tau_{\mathcal{H}}(m) \leq c_\epsilon \exp(\epsilon m) \quad \forall m \in \mathbb{N},$$

then $\mathcal{H}$ satisfies the uniform convergence condition and is thus PAC-learnable by the ERM rule.

- Hence, the class of Heaviside hypotheses

$$\mathcal{H} = \{\mathbb{1}_{[a,+\infty)} : a \in \mathbb{R}\}$$

is an infinite PAC-learnable class on $\mathcal{X} = \mathbb{R}$, because $\tau_{\mathcal{H}}(m) = m + 1$.

- However, the class of sine hypotheses

$$\mathcal{H} = \{h = \operatorname{sgn}(\sin(w\,\cdot)) : w \in \mathbb{R}\}$$

is an infinite but not PAC-learnable class on $\mathcal{X} = \mathbb{R}$. In fact, it attains the upper bound

$$\tau_{\mathcal{H}}(m) = 2^m \qquad \forall m \in \mathbb{N}.$$

- So **which property of classes $\mathcal{H}$ determines the growth of $\tau_{\mathcal{H}}$** and, hence, their learnability?