<u>Lemma 2.3</u> If $x_1, \dots, x_m$ are drawn i.i.d. according to D then for any h

$$\mathop{\mathbb{E}}_{S \sim D^m} [\hat{R}(h)] = R(h)$$

<u>Proof</u>

$$\mathop{\mathbb{E}}_{S \sim D^m} [\hat{R}(h)] = \mathop{\mathbb{E}}_{S \sim D^m} \frac{1}{m} \sum_{i=1}^{m} [\mathbb{1}_{h(x_i) \neq f(x_i)}]$$

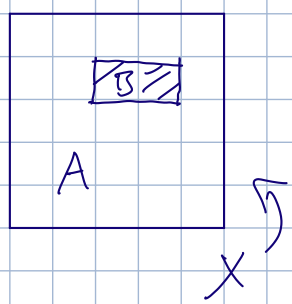
$$= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} [\mathbb{1}_{h(x_i) \neq f(x_i)}]$$

$$\underset{\text{i.i.d}}{=} \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{x \sim D} [\mathbb{1}_{h(x) \neq f(x)}] = \frac{1}{m} \cdot m R(h)$$

$$= R(h).$$

<u>Overfitting</u> The choice of a suitable hypothesis set $\mathcal{H}$ is important.

<u>Example</u> Let $X \subseteq \mathbb{R}^2$ be an axis aligned rectangle, e.g. $X = [0, 1]^2$, let $B \subsetneq X$ be another axis aligned rectangle, let $A = X \setminus B$.



Let $\mathbb{P}$ be a continuous probability distribution on X i.e. $\mathbb{P}(\{x\}) = 0 \;\; \forall x \in X$. with $\mathbb{P}(A) = \mathbb{P}(B) = 1/2$.

E.g. $X = [0, 1]^2$, $B = [0, \frac{1}{2}] \times [0, 1]$
$\mathbb{P}(M) = \mathrm{vol}(M)$

Let $f : X \to \{0, 1\}$ be given by

$$f(x) = \begin{cases} 0, & x \in A \\ 1, & x \in B \end{cases}$$

Given a sample $S = (x_1, \dots, x_m)$ with labels $y_i = f(x_i)$ choose hypothesis

$$h_S(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i \in [m] \\ 0 & \text{otherwise} \end{cases}$$

Then the empirical risk is

$$\hat{R}(h_S) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h_S(x_i) \neq f(x_i)} = 0,$$

since $h_S(x_i) = f(x_i)$ $\forall i \in [m] = \{1, \dots, m\}$

But the true risk is:

$$\mathbb{P}_{x \sim 1}\left(h_S(x) \neq f(x) \right) = \mathbb{P}\left(x \neq x_1, \dots, x_m \text{ and } f(x) = 1 \right\}$$

$$= \mathbb{P}\left(B \setminus \{x_1, \dots, x_m\} \right)$$

$$= \mathbb{P}(B) = \frac{1}{2}$$

$\mathbb{P}(\{x_1, \dots, x_m\}) = 0$ since $\mathbb{P}(\{x\}) = 0$ $\forall x \in X$

Hence $\hat{R}(h_S)$ is minimal $(= 0)$, but predictor is bad $R(h_S) = \mathbb{P}(h_S(x) \neq f(x)) = \frac{1}{2}$

not better than random guessing.

Thm 2.6

Let $\mathcal{H}$ be a finite set of functions $h: \mathcal{X} \to \{0,1\}$. Assume $f \in \mathcal{H}$ and let $\mathcal{A}$ be an algorithm that for each i.i.d sample $S = (x_1, \ldots, x_m)$ and labeled training data $(x_i, f(x_i))$, $i = 1, \ldots, m$ returns an $h_S \in \mathcal{H}$ with $\hat{R}(h_S) = 0$.

Then for $\varepsilon, \delta > 0$ if

$$m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log(1/\delta)\right)$$

then with prob. $\geq 1-\delta$

$$R(h_S) \leq \varepsilon .$$

Proof: $R(h_S)$ depends on training sample S and is difficult to evaluate directly. Instead, we bound it as follows:

Let $0 < \varepsilon < 1$, and let $\mathcal{H} = \{h_1, \ldots, h_n\}$

$$n = |\mathcal{H}| = \# \mathcal{H}$$

$$\mathbb{P}(R(h_S) > \varepsilon) = \mathbb{P}(\hat{R}(h_S) = 0 \text{ and } R(h_S) > \varepsilon)$$

$$\leq \mathbb{P}(\hat{R}(h) = 0 \text{ and } R(h) > \varepsilon \text{ for some } h \in \mathcal{H})$$

$$= \mathbb{P}\Big(\{\hat{R}(h_1) = 0 \ \& \ R(h_1) > \varepsilon\} \cup \{\hat{R}(h_2) = 0 \ \& \ R(h_2) > \varepsilon\}$$

$$\cup \ldots \cup \{\hat{R}(h_n) = 0 \ \& \ R(h_n) > \varepsilon\}\Big)$$

$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}\Big(\{\hat{R}(h) = 0\} \cap \underbrace{\{R(h) > \varepsilon\}}_{\text{deterministic event (indep. of } S\text{)}} \Big)$$

$$= \sum_{\substack{h \in \mathcal{H} \\ R(h) > \varepsilon}} \mathbb{P}(\hat{R}(h) = 0)$$

Fix $h \in \mathcal{H}$ with $R(h) > \varepsilon$ and recall

$$\hat{R}(h) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{h(x_j) \neq f(x_j)}$$

$$\Rightarrow \mathbb{P}(\hat{R}(h) = 0) = \mathbb{P}(h(x_j) = f(x_j) \text{ for } j = 1, \ldots, m)$$

$$\overset{\text{i.i.d.}}{=} \prod_{j=1}^{m} \mathbb{P}(h(x_j) = f(x_j))$$

$$= \prod_{j=1}^{m} \Big(1 - \underbrace{\mathbb{P}(h(x_j) \neq f(x_j))}_{= R(h) > \varepsilon}\Big)$$

$$< \prod_{j=1}^{m} (1 - \varepsilon) = (1 - \varepsilon)^m$$

$$\Rightarrow \quad \mathbb{P}(R(h_S) > \varepsilon) \leq \sum_{\substack{h \in \mathcal{H} \\ R(h) > \varepsilon}} (1-\varepsilon)^m \leq |\mathcal{H}| (1-\varepsilon)^m$$

$$\leq |\mathcal{H}| e^{-\varepsilon m}$$

$1 - x \leq e^{-x} \quad \forall x \in \mathbb{R}$

$$\rightsquigarrow \quad \mathbb{P}(R(h_S) \leq \varepsilon) \geq 1 - |\mathcal{H}| e^{-\varepsilon m}$$

With $\delta := |\mathcal{H}| e^{-\varepsilon m}$ we obtain

$$\mathbb{P}(R(h_S) \leq \varepsilon) \geq 1 - \delta$$

$\Rightarrow$

For $\delta \in (0,1)$ and $m \geq \frac{1}{\varepsilon} \log\left(|\mathcal{H}|/\delta\right) = \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log(1/\delta)\right)$

$$\mathbb{P}(R(h_S) \leq \varepsilon) \geq 1 - \delta$$

$\ulcorner$ solve $\delta = |\mathcal{H}| e^{-\varepsilon m}$ for m

$$\Rightarrow \quad e^{\varepsilon m} = |\mathcal{H}|/\delta$$

$$\Rightarrow \quad m = \frac{1}{\varepsilon} \log\left(|\mathcal{H}|/\delta\right) \quad \lrcorner$$

<u>Conclusion</u> For a finite hypothesis set $\mathcal{H}$, a consistent learning algorithm $\mathcal{A}$ is a PAC-learning algorithm with sample complexity polynomial (even linear) in $1/\varepsilon$ and logarithmic in $|\mathcal{H}|$ and $1/\delta$.

$\log|\mathcal{H}|$ may be interpreted as the number of bits required to represent $\mathcal{H}$ (up to a constant factor).

Note that $f \in \mathcal{H}$ guarantees that ERM always returns an h_S with $\hat{R}(h_S) = 0$.