STATISTICAL LEARNING THEORY

Summer Term 2025

Book: Shai Shalev-Schwartz
Understanding Machine Learning

Online Notes: http://rkwitt.org → Teaching
(All my handwritten notes will be available there!)

OTHER COURSE NAMES:

1) Advanced machine learning (for AIS students)
2) Machine learning (for Data Science students)

MOTIVATION

Example:

	$\in \mathbb{R}$ weight (in g)	$\in \mathbb{R}$ color ($\in [0,1]$)	Tasty?
Papaya 1	800	0.1	0
⋮	⋮	⋮	⋮
Papaya N	1200	0.7	1

0 ... non-tasty
1 ... tasty

Lets call the information in this table our training data. Based on that, we try to find

$$h : \mathbb{R} \times \mathbb{R} \to \{0, 1\}$$

i.e., a function that will take a two-dimensional vector as input (weight & color) and output a prediction of whether a papaya is tasty (1) or not (0).

We call such a function a hypothesis.

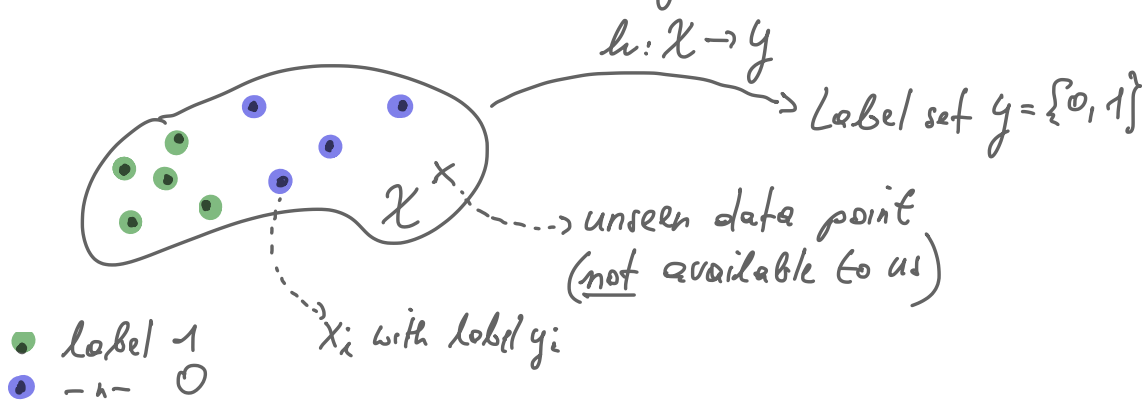
More formally, the data that is available to us comes in the form

$$\left((x_1, y_1), \ldots, (x_N, y_N)\right) = S$$

with $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$.

→ Domain $\mathcal{X} = \mathbb{R}^2$ → Label set $\mathcal{Y} = \{0, 1\}$

What do we mean by "learning"?



We say a learner (some algorithm) receives S and outputs h!

Two assumptions we will make initially:

(1) All the x_i's are drawn independently and identically (iid) from some (unknown) distribution D over the domain $\mathcal{X}$.

(2) The x_i's are labeled by some (unknown) function

$$f : \mathcal{X} \to \mathcal{Y},$$

called the true labeling function. This means

$$S = \left((x_1, \underbrace{f(x_1)}_{y_1}), \ldots\ldots, (x_N, \underbrace{f(x_N)}_{y_N}) \right)$$

What do we care about?

We care about

$$\{x \in \mathcal{X} : h(x) \neq f(x)\} = A$$

That is, all the points x in our domain $\mathcal{X}$, where the hypothesis h differs from the true labeling function f.

1. Domain set $\mathcal{X}$; we call $x \in \mathcal{X}$ an instance

2. Label set $\mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$

3. Training set $S = ((x_1, y_1), \dots, (x_m, y_m))$ with
$(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} =: \mathcal{Z}$

4. A learner receives S and outputs $h: \mathcal{X} \to \mathcal{Y}$ (i.e., a hypothesis)

Assumption: For now, we assume that the x_i's are drawn iid from some probability measure $\mathcal{D}$ over the domain $\mathcal{X}$ and labeled by some function $f: \mathcal{X} \to \mathcal{Y}$, so $y_i = f(x_i)$.

We are interested in:

$$\mathcal{D}\left(\{x \in \mathcal{X} : h(x) \neq f(x)\}\right) = \mathbb{P}_{x \sim \mathcal{D}}\left[h(x) \neq f(x)\right] = L_{\mathcal{D},f}(h)$$

"Generalization error"

The empirical version of that is

$$\frac{1}{m} \cdot \left|\{i \in \{1, \dots, m\} : h(x_i) \neq f(x_i)\}\right| = L_S(h)$$

"Empirical error"
(Empirical risk)

Notation:

$[m] = \{1, \dots, m\}$, $S|_x = (x_1, \dots, x_m)$

Claim: $\mathbb{E}_{S|x\sim D^m}[L_S(h)] = L_{D,f}(h)$

$S|x = (x_1, \dots, x_m)$

$L_S(h) = \frac{1}{m} \cdot \left|\{i \in [m] : h(x_i) \neq f(x_i)\}\right|$

$$\mathbb{E}_{S|x\sim D^m}[L_S(h)] = \mathbb{E}_{S|x\sim D^m}\left[\frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq f(x_i)}\right] \quad \text{// by def.}$$

$$= \frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{E}_{x\sim D}\left[\mathbb{1}_{h(x_i)\neq f(x_i)}\right] \quad \text{// by linearity of } \mathbb{E}[\cdot]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x\sim D}\left[\mathbb{1}_{h(x)\neq f(x)}\right] \quad \text{// as all } x_i\text{'s are drawn iid from } D$$

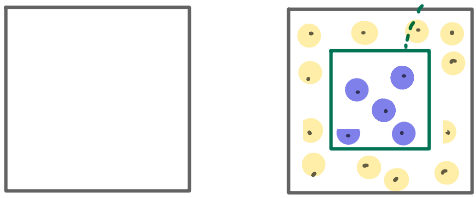
$$= \frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{P}\left[h(x)\neq f(x)\right]$$

$$= \frac{1}{\cancel{m}}\cdot\cancel{m}\cdot\mathbb{P}\left[h(x)\neq f(x)\right] = \underbrace{L_{D,f}(h)}_{\text{generalization error}}$$

Our first learning paradigm

Empirical Risk Minimization (ERM). As a learner only has access to the training data (S), it is "natural" to try to select h (our hypothesis) such that the empirical risk (emp. error) is minimized. We call such a h an **empirical risk minimizer** (an ERM hypothesis).

Example (of a problematic case):



Domain $\mathcal{X}$

Say the distribution on $\mathcal{X}$ is uniform.

- label 1
- label 0

Also, assume that the area of the domain (square) is 2 and the area of $\square$ is 1.

Now, say we have an ERM algorithm that returns h_S such that

$$h_S(x) = \begin{cases} y_i, \text{ if } \exists i \in \{1, \dots, m\} : x_i = x \\ 0, \text{ else} \end{cases}$$

(sort of a lookup table)

Obviously, h_S is correct on all instances in $S \rightarrow h_S$ is an emp. risk minimizer, meaning

$$L_S(h_S) = 0!$$

But, on unseen instances from D (which is uniform on $\mathcal{X}$), h_S is only correct 50% of the time $\rightarrow$

$$L_{D,f}(h_S) = \frac{1}{2}!$$

That's what we call overfitting!

Hypothesis class (H): We restrict searching for h to H, i.e., a class of functions from $\mathcal{X} \to \mathcal{Y}$ and we write

$$ERM_H(S) \in \underset{h \in H}{\arg\min}\, L_S(h)$$

ERM over finite hypothesis classes ($|H| < \infty$)

Assumption (realizability): $\exists\, h^* \in H$ with $L_{D,f}(h^*) = 0$.

Now, any ERM hypothesis h_S will attain 0 empirical error ($L_S(h_S) = 0$) as h_S competes against h^* (which has $L_{D,f}(h^*) = 0$ and, obviously, $L_S(h^*) = 0$).

We know that $L_{D,f}(h_S) > \varepsilon$, $\varepsilon \in (0,1)$, can only happen if our learner selects a hypothesis h_S with $L_S(h_S) = 0$, BUT $L_{D,f}(h_S) > \varepsilon$.

We define $H_{BAD} = \{h \in H : L_{D,f}(h) > \varepsilon\}$ set of bad hypothesis!

Also, we define

$$M = \{S|_x : \exists h \in H_{BAD},\, L_S(h) = 0\}$$

Observation:

$$\{S|_x : L_{D,f}(h_S) > \varepsilon\} \subseteq \{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\} = M$$

(h_S: ERM hypothesis (Empirical risk minimizer))

Since $\{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\} = \bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}$

we have

$$D^m\left(\{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\}\right) = D^m\left(\bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}\right)$$

$$\leq \sum_{h \in H_{BAD}} D^m\left(\{S|_x : L_S(h) = 0\}\right)$$

(by σ-sub-additivity "union" bound)

Lets fix some $h \in H_{BAD}$:

$$D^m\left(\{S|_x : L_S(h) = 0\}\right) = D^m\left(\{S|_x : \forall i \in \{1,\dots,m\} : h(x_i) = f(x_i)\}\right)$$

(f: True labeling function)

$$\overset{\text{by iid assumption}}{=} \prod_{i=1}^{m} D\left(\{x_i : h(x_i) = f(x_i)\}\right)$$

$$= \prod_{i=1}^{m} D\left(\{x : h(x) = f(x)\}\right)$$

$$\overset{\text{(by definition of } L_{D,f})}{=} \prod_{i=1}^{m} \left(1 - \underbrace{L_{D,f}(h)}_{>\varepsilon}\right) \quad \left(\text{remember that } h \in H_{BAD}\right)$$

↳

$$\leq \prod_{i=1}^{m} (1-\varepsilon)$$

$$= (1-\varepsilon)^m$$

$$\leq e^{-\varepsilon m} \quad \text{(without proof)}$$

Overall, we have

$$D^m\left(\{S|_x : L_{D,f}(h_S) > \varepsilon\}\right) \leq \sum_{h \in H_{BAD}} D^m\left(\{S|_x : L_S(h) = 0\}\right)$$

$$\leq \sum_{h \in H_{BAD}} e^{-\varepsilon m}$$

$$= |H_{BAD}| \cdot e^{-\varepsilon m}$$

$$\leq |H| \cdot e^{-\varepsilon m} \quad \text{(because } H_{BAD} \subseteq H\text{)}$$

if we want $|H| \cdot e^{-\varepsilon m}$ to be less than some $\delta \in (0,1)$, we can solve for m and get:

$$|H| \cdot e^{-\varepsilon m} < \delta$$

$$\Rightarrow m > \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right)$$

ε: error

δ: we call $(1-\delta)$ confidence

In words, the probability of the generalisation error of an ERM hypothesis h_S being larger or equal to $\varepsilon \in (0,1)$ is upper bounded by $|\mathcal{H}| \cdot e^{-m\varepsilon}$ (to be understood as the probability over a random draw of a dataset S of size m, iid from $\mathcal{D}$).

Interlude (MARKOV inequality)

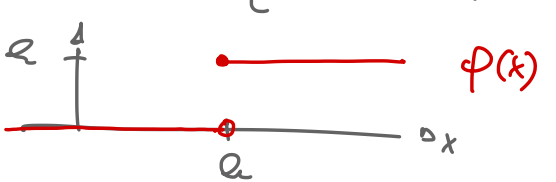
We have a probability space $(S, \mathcal{F}, \mathbb{P})$ and a non-negative random variable X $(X \geq 0)$.

Claim: for all $a > 0$, we have

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Proof: Define $\varphi : S \to \mathbb{R}$

$$\varphi(x) = \begin{cases} a, & \text{if } x \geq a \\ 0, & \text{else } (x < a) \end{cases}$$



We have $0 \leq \varphi(X(x)) \leq X(x)$ (for all $x \geq 0$)

11

$$\Rightarrow \int X \, dD \geq \underbrace{\int \varphi(X(x)) \, dD}_{a \cdot \underbrace{D(\{\omega \in S : X(\omega) \geq a\})}_{\mathbb{P}[X \geq a]}} \qquad \text{(by monotonicity)}$$

Since $a > 0$:

$$\frac{1}{a} \cdot \underbrace{\int X \, dD}_{\mathbb{E}[X] \text{ by def.}} \geq \mathbb{P}[X \geq a]$$

$$\Rightarrow \mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a} \qquad \square$$

Corollary: Let $|H| < \infty$ and $\varepsilon, \delta \in (0,1)$. Further, let m be an integer such that $m > \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right)$. Then, for any labeling function $f: X \to \{0,1\}$ and any distribution D (for which realizability holds), we have that with probability of at least $1-\delta$ over the choice of $S|_X$ of size m, every ERM hypothesis h_S satisfies

$$L_{D,f}(h_S) \leq \varepsilon.$$

Interpretation: For sufficiently large m, ERM_H returns a hypothesis h_S that is PROBABLY ($1-\delta$) APPROXIMATELY (ε) CORRECT (PAC).

Def. (PAC learnability): A hypothesis class H is **PAC learnable** if there exists a function $m_H : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0,1)$ and (II) every distribution D over the domain X, and (III) every labeling function $f: X \to \{0,1\}$, if (IV) realizability holds (with respect to H, D, f), then running A on

$$m \geq m_H(\varepsilon, \delta)$$

iid instances from D (labeled by f), returns a hypothesis h such that with probability of at least $1-\delta$, it holds that

$$\boxed{L_{D,f}(h) \leq \varepsilon .}$$

Terminology: $m_H : (0,1)^2 \to \mathbb{N}$ is called the **sample complexity (function)**.

We will now move to a more general setting.

(1) We will first remove the realizability assumption; this means, there is no longer a $h^* \in H$ with $L_{D,f}(h^*) = 0$. Now, the best possible thing that we can hope for is a guarantee relative to $\min_{h \in H} L_{D,f}(h)$.

↳ before, with realizability, this was 0!!!

Def. (Hoeffding inequality): Let $X_1, \dots, X_m$ iid random variables (RVs) taking values in $[a_i, b_i]$ for $i \in \{1, \dots, m\}$. Then, given that

$$S_m = \sum_{i=1}^{m} X_i$$

it holds that

(a) $\mathbb{P}\left[S_m - \mathbb{E}[S_m] > \varepsilon \right] \leq e^{\frac{-2\varepsilon^2}{\sum_i (b_i - a_i)^2}}$

(b) $\mathbb{P}\left[S_m - \mathbb{E}[S_m] < -\varepsilon \right] \leq$ — " —

and, upon combination (using union bound)

$$\mathbb{P}\left[\left|S_m - \mathbb{E}[S_m]\right| > \varepsilon\right] \leq 2 \cdot e^{\frac{-2\varepsilon^2}{\sum_i (b_i - a_i)^2}}$$

An alternative (quite useful) form is:

$$(\star) \quad \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > \varepsilon\right] \leq 2 \cdot e^{\frac{-2\varepsilon^2 m}{(b-a)^2}}$$

with $\mu = \mathbb{E}[X_i]$ and $\mathbb{P}[a \leq X_i \leq b] = 1$ for all $i \in \{1, \ldots, m\}$.

As a consequence of $(\star)$, we can already say the following: fix $\varepsilon > 0$; then, for a simple $h: X \to Y$, we have

$$\mathbb{P}_{S|_X \sim D^m}\left[\left|L_S(h) - L_{D,f}(h)\right| > \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 m}$$

since $a = 0$ and $b = 1$.

Importantly, this only holds for a single h!!!

What we want is a bound that holds uniformly over all $h \in H$.

So, for $|H| < \infty$ (finite classes), we can get (by the union bound):

$$\mathbb{P}_{S|_x \sim D^m}\left[\exists h \in H : |L_S(h) - L_{D,f}(h)| > \varepsilon\right] \leq \sum_{h \in H} 2e^{-2\varepsilon^2 m}$$

$$\leq |H| \cdot 2e^{-2\varepsilon^2 m}$$

Remark: we did not need realizability for that!!! But, we pay the price of ε^2 vs. ε.

(2) Next, we will get rid of the true labeling function $f: X \to Y$ (which labeled our training instances x_i) by letting D be a distribution over $X \times Y$, i.e., over domain and label space (Y).

Let $Z := X \times Y$. We need to adjust our definitions of empirical error and generalization error.

$$L_D(h) = \mathbb{P}_{(x,y)\sim D}\left[h(x) \neq y\right] = D\left(\{(x,y) \in X \times Y : h(x) \neq y\}\right)$$

$$L_S(h) = \frac{1}{m} \cdot \left|\{i \in \{1,\dots,m\} : h(x_i) \neq y_i\}\right|$$

Def. (Agnostic PAC learnability): A hypothesis class H of functions $h: X \to Y$ is agnostic PAC learnable if $\exists\, m_H : (0,1)^2 \to \mathbb{N}$ and an algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0,1)$ and (II) every distribution D over $X \times Y = Z$, when running A on m iid instances from D with $m \geq m_H(\varepsilon, \delta)$, A returns a hypothesis h such that with probability of at least $1-\delta$ (over the choice of S)

$$L_D(h) \leq \min_{h' \in H} L_D(h') + \varepsilon$$

One more generalization:

Loss function:

$$\ell : H \times (X \times Y) \to \mathbb{R}_+$$

Example: "0-1 loss"

$$\ell^{0-1}(h, (x,y)) = \begin{cases} 1, & \text{if } h(x) \neq y \\ 0, & \text{else} \end{cases}$$

With this, we can write

$$L_D(h) = \mathop{\mathbb{E}}_{z \sim D} \left[\ell(h, z) \right] \quad \text{with} \quad z = (x, y)$$

- $$L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h, z_i)$$

If we set in $\ell \equiv \ell^{0-1}$, then

$$L_D(h) = \mathop{\mathbb{E}}_{z \sim D} \left[\ell(h, z) \right] = \mathop{\mathbb{E}}_{z \sim D} \left[\underbrace{\ell^{0-1}(h, z)}_{\text{only takes values 0 or 1}} \right]$$

$$= 0 \cdot \mathop{\mathbb{P}}_{(x,y) \sim D} \left[h(x) = y \right] + 1 \cdot \mathop{\mathbb{P}} \left[h(x) \neq y \right]$$

$$= \mathop{\mathbb{P}} \left[h(x) \neq y \right]$$

Uniform Convergence

Def. (ε-representative sample): A sample S is called ε-representative with respect to $Z = X \times Y$, hypothesis class H, loss function ℓ, and distribution D, if

$$\forall h \in H: \quad |L_S(h) - L_D(h)| \leq \varepsilon .$$

Lemma: Assume that S is $\varepsilon/2$-representative wrt. Z, H, ℓ and D. Then, any hypothesis h_S returned by $ERM_H(S) \in \operatorname{argmin}_{h' \in H} L_S(h')$ satisfies

$$\boxed{L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon}$$

Proof: for any $h \in H$:

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \varepsilon/2 && \text{// by def. of } \varepsilon/2\text{-rep.} \\ &\leq L_S(h) + \varepsilon/2 && \text{// since } h_S \text{ is an ERM hypothesis} \\ &\leq L_D(h) + \varepsilon/2 + \varepsilon/2 && \text{// by def. of } \varepsilon/2\text{-rep.} \\ &= L_D(h) + \varepsilon \end{aligned}$$

ERM hyp. ←

Overall $L_D(h_S) \leq L_D(h) + \varepsilon$ (✗)

Finally, as (*) holds for <u>any</u> $h \in H$, we conclude that

$$L_D(h_S) \leq \min_{h' \in H} L_D(h) + \varepsilon.$$

∎

Def. (uniform convergence (UC)): A hyp. class H has the UC property with respect to $Z = X \times Y$ and loss function ℓ, if there exists $m_H^{UC} : (0,1)^2 \to \mathbb{N}$, such that for any $\varepsilon, \delta \in (0,1)$ and every distribution D over $Z = X \times Y$, if S is and iid sample of size $m \geq m_H^{UC}(\varepsilon, \delta)$ from D, then with probability of at least $1-\delta$ (over the choice of S), S is ε-representative.

Corollary: If H satisfies the UC property with m_H^{UC}, then H is <u>agnostic PAC learnable</u> with

$$m_H(\varepsilon, \delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$$

by ERM.

Question: Is there a universal learner? By universal we mean a learner without any prior knowledge of the learning task (given by D), but can be challenged by any task and still returns $A(S)$ with low $L_D(A(S))$.

This question is answered by the following theorem:

Thm (No-Free Lunch): Let A be a learning algorithm for the task of binary classification, with respect to the 0-1 loss, over domain $\mathcal{X}$. Also, let m be any number ($\in \mathbb{N}$) smaller than $|\mathcal{X}|/2$ (representing the size of S). Then, there exists a distribution D over $\mathcal{X} \times \{0,1\}$ such that

1. $\exists f : \mathcal{X} \to \{0,1\}$ with $L_D(f) = 0$!

and

2. with probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$, we have

$$L_D(A(S)) \geq \frac{1}{8}.$$

Interpretation: 1 means that the task can be learned successfully by another learner (e.g. using $H = \{f\}$) and 2 means that A fails on that task.

Corollary: Let X be infinitely large and H be the class of all functions from $X \to \{0,1\}$, then H is not PAC learnable.

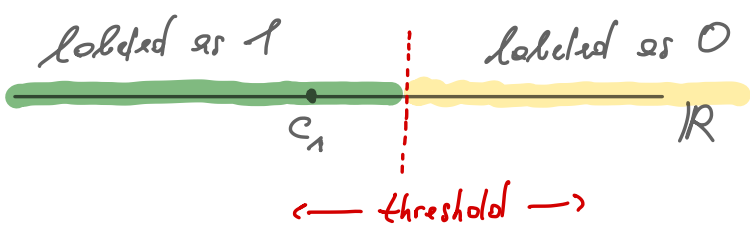
VAPNIK – CHERVONENKIS Dimension (VC-Dimension)

Def.: Let H be a class of functions from $X \to \{0,1\}$ and let $C \subset X$, $C = \{c_1, \dots, c_m\}$. We define

$$H_C = \{(h(c_1), h(c_2), \dots, h(c_m)) : h \in H\}$$

as the restriction of H to the set C.

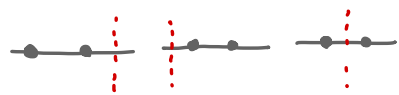
Example: $C = \{c_1\}$, $c_1 \in \mathbb{R}$ and let H^{thr} be the class of thresholds on the real line.



$$H_C^{thr} = \{(1), (0)\} \Rightarrow |H_C^{thr}| = 2 = 2^1$$

Lets take $C = \{c_1, c_2\}$, wlog $c_1 \leq c_2$

We get $H_C^{thr} = \{(1,1), (0,0), (1,0)\} \Rightarrow |H_C^{thr}| = 3$



But, we would have $2^2 = 4$ possible labelings of $\{c_1, c_2\} = C$.

Def. (Shattering) H shatters a finite set C of size m if $|H_C| = 2^m$.

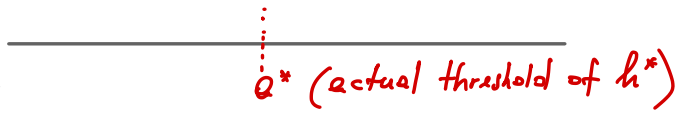
Claim: The class H^{thr} on $X = \mathbb{R}$ is PAC learnable with $m_{H^{thr}}(\varepsilon, \delta) \leq \left\lceil \log\left(\frac{2}{\delta}\right) \cdot \frac{1}{\varepsilon} \right\rceil$.

$$H^{thr} = \{h_a : a \in \mathbb{R}\}$$

$$h_a : \mathbb{R} \to \{0,1\}, \quad h_a(x) = \begin{cases} 1, & \text{if } x < a \\ 0, & \text{else} \end{cases}$$

(or $h_a(x) = \mathbb{1}_{x<a}$)

We assume realizability, i.e., $\exists h^* \in H^{thr}$ such that $L_{D,f}(h^*) = 0$.



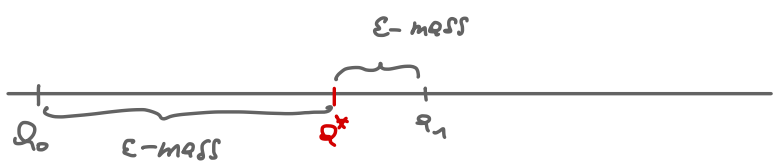
We let $a_0 \in \mathbb{R}$ be such that

$$\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a_0, a^*)\}\right) = \varepsilon$$

and let $a_1 \in \mathbb{R}$ be such that

$$\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a^*, a_1)\}\right) = \varepsilon$$

Special cases: In case $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a_0, a^*)\}\right) < \varepsilon$, set $a_0 = -\infty$
and in case $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a^*, a_1)\}\right) < \varepsilon$, set $a_1 = +\infty$.



Next, we are going to define an ERM algorithm (needed for PAC)

Pick $b_0 \in \mathbb{R}$ and $b_1 \in \mathbb{R}$ from $S = ((x_1, y_1), \ldots, (x_m, y_m))$ as

$$b_0 = \max\{x : (x, 1) \in S\}$$
$$b_1 = \min\{x : (x, 0) \in S\}$$

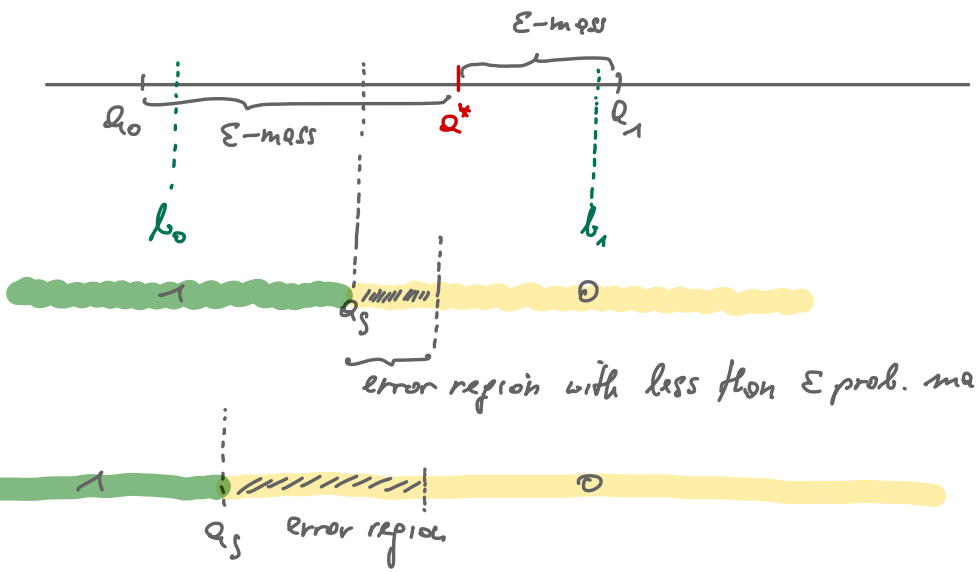
Then, pick <u>any</u> threshold within (b_0, b_1) and call the corresponding hypothesis h_S. We see, h_S always has 0 empirical error on S.

ERM hyp. ↓ h_S — threshold a_S

By our construction of a_0 and a_1, the claim is that for h_S to have $L_{D,f}(h_S) \le \varepsilon$, it suffices that

(a)	$b_1 \le a_1$	AND
(b)	$b_0 \ge a_0$	

(*)



1st exam date: June 24 (10am) - last lecture

If we write down (*) formally, we have

$$\mathbb{P}\left[L_{D,f}(h) > \varepsilon\right] \leq \mathbb{P}\left[(b_0 < a_0) \text{ OR } (b_1 > a_1)\right]$$

$$\leq \mathbb{P}\left[b_0 < a_0\right] + \mathbb{P}\left[b_1 > a_1\right] \quad \text{// union bound}$$

To upper-bound $\mathbb{P}\left[L_{D,f}(h) > \varepsilon\right]$, we need to check when $b_0 < a_0$ (or $b_1 > a_1$). Answer: when there is no data point x in S that is labeled '1' such that $x \in (a_0, a^*)$.

We know that $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a_0, a^*)\}\right) = \varepsilon$ (by construction of a_0); hence, not seeing a data point in (a_0, a^*) has probability $1-\varepsilon$. Consequently, not seeing a data point in m iid samples (inside (a_0, a^*)) has probability $(1-\varepsilon)^m$.

$\Rightarrow$ Overall, we have

$$\mathbb{P}\left[L_{D,f}(h) > \varepsilon\right] \leq 2 \cdot \underbrace{(1-\varepsilon)^m}_{\leq e^{-\varepsilon m}} \leq \underbrace{2e^{-\varepsilon m}}_{< \delta}$$

$$\Rightarrow \boxed{m > \frac{1}{\varepsilon} \log\left(\frac{2}{\delta}\right)}$$

sample complexity

We have seen that "finiteness" of $\mathcal{H}$ is sufficient but not necessary !!! for PAC learnability.

Def. (VC-Dimension): The VC-dimension of H (i.e., a class of functions from $X \to \{0,1\}$), written as

$$VC(H),$$

is the maximal size of a set $C \subset X$ that is shattered by H.

Theorem: Let H be a class of functions from $X \to \{0,1\}$. If, H has infinite VC-dimension, then H is not PAC learnable. (follows immediately from NFL theorem).

Def. (growth function): Let H be a class of functions from $X \to \{0,1\}$. The growth function of H $\tau_H : \mathbb{N}_0 \to \mathbb{N}_0$ is defined as

$$\tau_H(m) = \max_{\substack{C \subset X \\ |C| = m}} |H_C|$$

Lemma (Sauer, Shelah, Perez, "Sauer's lemma"): Let H be a class of functions from $X \to \{0,1\}$ with $VC(H) = d < \infty$. Then,

$$\tau_H(m) = 2^m \text{ if } m \leq d, \text{ but}$$

$$\tau_H(m) = \left(\frac{em}{d}\right)^d \text{ if } m > d$$

(d = VC-dimension)

Lets take any set C (of size m) and $|H| < \infty$ (finite).

We know

$$|H_C| \leq |H|$$

So, if $|H| < 2^m$, then H can obviously not shatter C of size m. This implies

$$VC(H) \leq \log_2(|H|)$$

Theorem: Let H be a class of functions from $X \to \{0,1\}$ and $\ell : H \times X \times Y \to [0, c]$, $c > 0$, a loss function. For any distribution D over $X \times Y$ and $\delta \in (0,1)$, we have with probability of at least $1-\delta$ (over the choice of $S \sim D^m$)

$$\forall h \in H : \left| L_D(h) - L_S(h) \right| \leq c \cdot \sqrt{\frac{\delta \cdot \log\left(\tau_H(2m)\frac{4}{\delta}\right)}{m}}$$

(without proof).