Vapnik – Chervonenkis Dimension (VC)

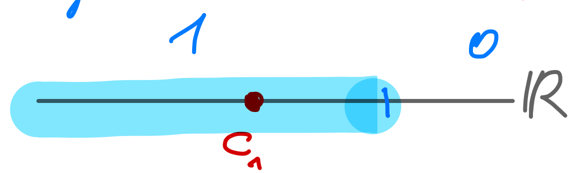
If we review the NFL proof, it seems intuitive to study how a hypothesis class H behaves on C.

<u>Def</u>: 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

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

as the restriction of H to C.

e.g. $C = \{c_1\}$

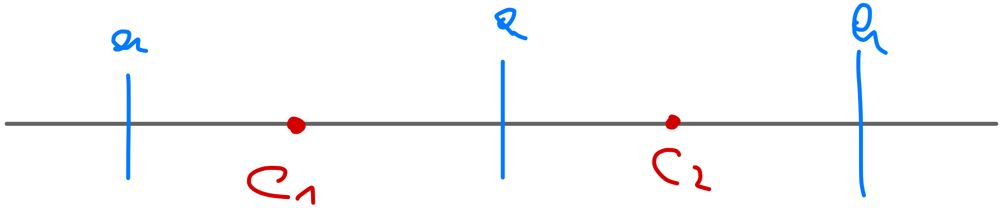


Class of thresholds on $\mathbb{R}$.

$$H_C = \{(0), (1)\}$$

$$|H_C| = 2^1 = 2$$

$C = \{c_1, c_2\}$



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

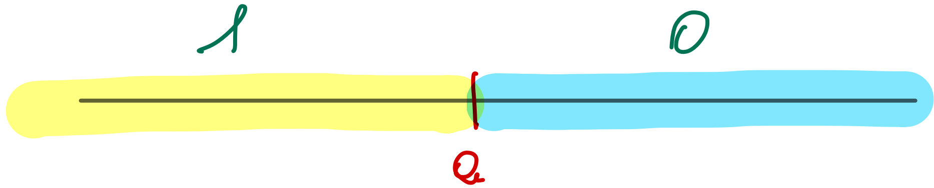
Threshold

$H_C = \{(0,0), (1,1), (1,0)\}, \quad |H_C| = 3$

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

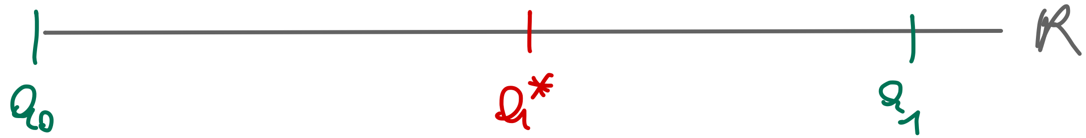
Example: H of infinite size but PAC learnable.

$$H^{thr} = \{h_a : a \in \mathbb{R}\}, \quad h_a : \mathbb{R} \to \{0, 1\}$$
$$x \mapsto h_a(x) = \mathbb{1}_{x < a}$$



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

Proof: We assume realizability $\Rightarrow \exists h^* \in H^{thr}$ s.t. $L_{D,f}(h^*) = 0$.
(Let a^* be the corresponding threshold).



Let a_0 be s.t. $D(\{x\in\mathbb{R}: x\in(a_0, a^*)\}) = \varepsilon$

Let a_1 be s.t. $D(\{x\in\mathbb{R}: x\in(a^*, a_1)\}) = \varepsilon$

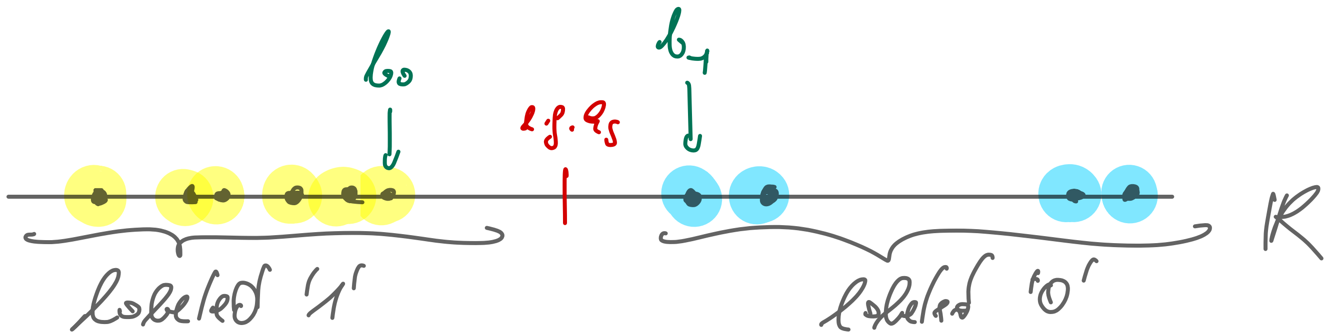
The special cases are:

- if $D(\{x\in\mathbb{R}: x\in(a_0, a^*)\}) < \varepsilon$, then set $a_0 = -\infty$
- if $D(\{x\in\mathbb{R}: x\in(a^*, a_1)\}) < \varepsilon$, then set $a_1 = +\infty$

We are given $S = ((x_1,y_1), \dots, (x_m,y_m))$. An ERM algorithm would be, e.g.:

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

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



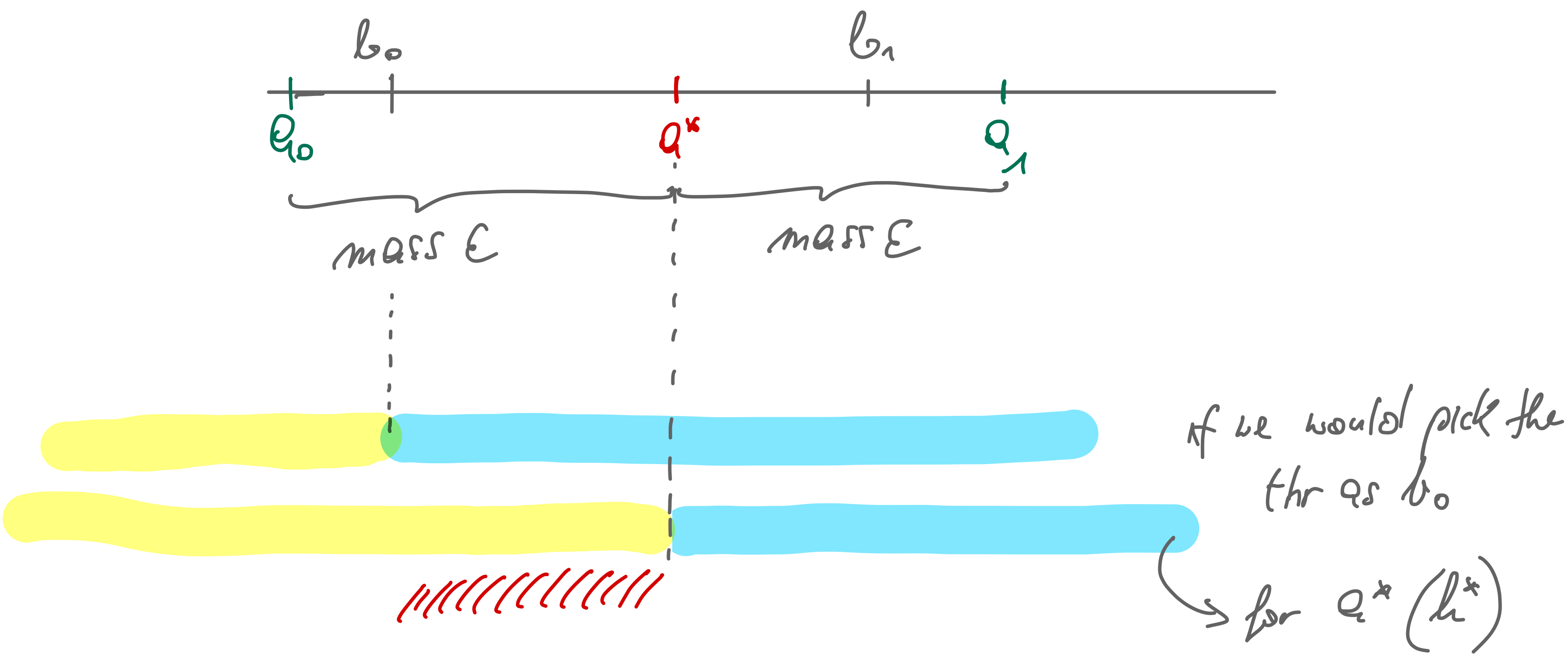
So, choose any threshold within (b_0, b_1). Lets denote that threshold by a_s (we let h_s be the corresponding ERM hypothesis).

For h_s (with a_s as threshold) to have $L_{D,f}(h_s) \le \varepsilon$, it suffices that

1) $b_1 \le a_1$ AND

2) $b_0 \ge a_0$

Why?



Hence,

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

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

So, when does $b_0 < a_0$ happen? When there is no instance in S (labeled '1') s.t. $x \in (a_0, a^*)$.

We do know that $D(\{x \in \mathbb{R} : x \in (a_0, a^*)\}) = \varepsilon$. Hence, not seeing an instance in that interval has probability of $1-\varepsilon$, and not seeing an instance among m iid draws from D is $(1-\varepsilon)^m$.

$$\Rightarrow \mathbb{P}_{S \sim D^m}[b_0 < a_0] = (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

(same for $\mathbb{P}[b_1 > a_1]$)

We get $\mathbb{P}_{S \sim \mathcal{D}^m}[L_{D,f}(h_S) > \varepsilon] \leq 2e^{-\varepsilon m}$

$2e^{-\varepsilon m} < \delta$

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

We have seen that finiteness of H ($|H| < \infty$) is sufficient, but not necessary.

Def. (VC-Dimension): The VC-Dimension of H (i.e. a class of functions from $X \to \{0,1\}$) is the maximal size of a set $C \subset X$ that is shattered by H.

$VC(H)$

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.

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

$$\tau_H(m) = \max_{C \subset X, |C| = m} |H_C| \quad \left(\text{with } \tau_H(0) = 1\right)$$

We can also define the VC-dim as follows (using $\tau_H(m)$):

$$VC(H) = \max \{ m \in \mathbb{N}_0 , \tau_H(m) = 2^m \}$$

if max exists and ∞ otherwise.

<u>Lemma</u> (Sauer, Shelah, Perez "Sauer's lemma"): Let H be a hyp. class with $VC(H) \leq d$. Then, $\tau_H(m) = 2^m$ if $m \leq d$, but $\tau_H(m) \leq \left(\frac{em}{d}\right)^d$ if $m > d$.

(without proof).

VC dim. for finite H:

if we take any set C, then it's obvious that

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

Hence, if $|H| < 2^m$, then H cannot shatter C of size m. This implies that

$$\underbrace{VC(H)}_{\substack{\text{rem. that this is the} \\ \text{max. size of a set that is shattered.}}} \leq \log_2(|H|)$$

($VC(H) = d$ means that we have at least 2^d hyp.-s in H!)

Theorem: Let H be a hyp. 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 dist. D over $X \times Y$ and $\delta \in (0,1)$, we have that with prob. of at least $(1-\delta)$ over the choice of $S \sim D^m$

$$\forall h \in H: \quad |L_D(h) - L_S(h)| \leq c \cdot \sqrt{\frac{8 \log \left(\tau_H(2m) \cdot \frac{4}{\delta} \right)}{m}}$$

growth function!

Setting in Sauer's lemma, we get

Thm. says $2m$

$$\forall h \in H: |L_D(h) - L_S(h)| \leq c \cdot \sqrt{\frac{8 \cdot \log\left(\left(\frac{2em}{d}\right)^d \cdot \frac{4}{\delta}\right)}{m}}$$

The result from the theorem comes from

$$\mathbb{P}\left[\exists h \in H: |L_D(h) - L_S(h)| > \varepsilon\right] \leq 4 \cdot \tau_H(2m) \cdot e^{-m\varepsilon^2/8c^2},$$

setting the RHS $\leq \delta$ and solving for ε, i.e.

$$4\tau_H(2m)\, e^{-m\varepsilon^2/8c^2} \leq \delta$$

$$e^{-m\varepsilon^2/8c^2} \leq \frac{\delta}{4 \cdot \tau_H(2m)} \quad \Big| \log \,\&\, \cdot -1$$

$$m\varepsilon^2/8c \geq \log\left(\tau_H(2m) \cdot \frac{4}{\delta}\right)$$

$$\varepsilon \geq c \cdot \sqrt{\frac{8 \log(\dots)}{m}}$$

We could look at the bound in the theorem diff., i.e., trying to get a sample complexity (with $c = 1$, as in 0-1 loss):

$$\left|L_D(h) - L_S(h)\right| \leq \sqrt{\frac{8 \log\left(\tau_H(2m)\frac{4}{\delta}\right)}{m}} \leq \sqrt{\frac{8 \log\left(\left(\frac{2em}{d}\right)^d \cdot \frac{4}{\delta}\right)}{m}} \quad (*)$$

If we want to be at most ε, i.e.,

$$\sqrt{\frac{8 \log\left(\left(\frac{2em}{d}\right)^d \cdot \frac{4}{\delta}\right)}{m}} \leq \varepsilon \qquad \Big|^2$$

$$8 \log\left(\left(\frac{2em}{d}\right)^d \frac{4}{\delta}\right) \leq \varepsilon^2 m$$

$$m \geq \frac{8}{\varepsilon^2}\left[\log\left(\left(\frac{2em}{d}\right)^d\right) + \log\left(\frac{4}{\delta}\right)\right]$$

$$m \geq \frac{8}{\varepsilon^2} \cdot \left[d \cdot \log\left(\frac{2em}{d}\right) + \log\left(\frac{4}{\delta}\right) \right]$$

$$= \frac{8}{\varepsilon^2} \cdot \left[d \cdot \log(m) + d \cdot \log\left(\frac{2e}{d}\right) + \log\left(\frac{4}{\delta}\right) \right]$$

$$= \frac{8}{\varepsilon^2} \cdot d \log(m) + \frac{8}{\varepsilon^2} \left[d \log\left(\frac{2e}{d}\right) + \log\left(\frac{4}{\delta}\right) \right]$$

still contains m!

<u>Lemma</u>: let $a \geq 1$ and $b > 0$. Then, $x \geq 4a \log(2a) + 2b \Rightarrow x \geq a \log(x) + b$.

(A.2 in book).

In our case: $a = \frac{8}{\varepsilon^2} d$, $b = \frac{8}{\varepsilon^2} \cdot [\ldots\ldots]$ meaning we get the <u>sufficient condition</u>

$$m \geq \frac{32 d}{\varepsilon^2} \cdot \log\left(\frac{16 d}{\varepsilon^2}\right) + \frac{16}{\varepsilon^2} \cdot \left[d \log\left(\frac{2e}{d}\right) + \log\left(\frac{4}{\delta}\right) \right]$$

Assume (which is reasonable)

$\boxed{d \geq 2e}$

$$\Rightarrow \quad m \geq \frac{32d}{\varepsilon^2} \cdot \log\left(\frac{16d}{\varepsilon^2}\right) + \frac{16}{\varepsilon^2} \cdot \log\left(\frac{4}{\delta}\right)$$

This gives us our sample complexity function $m_{\mathcal{H}}^{UC}$ for <u>unif. convergence!</u>

Hence, finite VC $\Rightarrow$ UC !!

Fundamental theorem of stat. learning

Let H be a hyp. class of functions from $X \to \{0,1\}$ and let ℓ be the 0-1 loss. Then, the following statements are equivalent:

$\boxed{1}$ H has the UC property

$\boxed{2}$ Any ERM alg. is a successful agnostic PAC learner for H

$\boxed{3}$ H is agnostic PAC learnable

$\boxed{4}$ Any ERM alg. is a successful PAC learner for H

$\boxed{5}$ H is PAC learnable

$\boxed{6}$ H has finite VC dimension

$\boxed{1} \longrightarrow \boxed{2}$ we have shown this already (see $\frac{\varepsilon}{2}$-rep. samples etc.)

$\boxed{2} \longrightarrow \boxed{3}$ trivial as ERM is a succ. APAC learner $\to$ this is the alg.

$\boxed{3} \longrightarrow \boxed{4}$ trivial as assuming realizability gives us PAC learnability via ERM

$\boxed{4} \longrightarrow \boxed{5}$ same arg. as in $\boxed{2} \to \boxed{3}$

$\boxed{5} \longrightarrow \boxed{6}$ follows from NFL, so does $\boxed{4} \to \boxed{6}$

$\boxed{6} \longrightarrow \boxed{1}$ we have just seen!