

MACHINE LEARNING (911.285)

Summer Term 2023

03/07/23

Motivating example

	height (f)	color ($\in [0, 1]$)	Tasty (0) / Non-Tasty (1)
Papaya 1	100	0.1	yes (1)
- - 2	:	:	:
:	:	:	:
Papaya N	500	0.8	no (0)
	$\in \mathbb{R}$	$\in \mathbb{R}$	

TRAINING DATA

Our goal would be to find

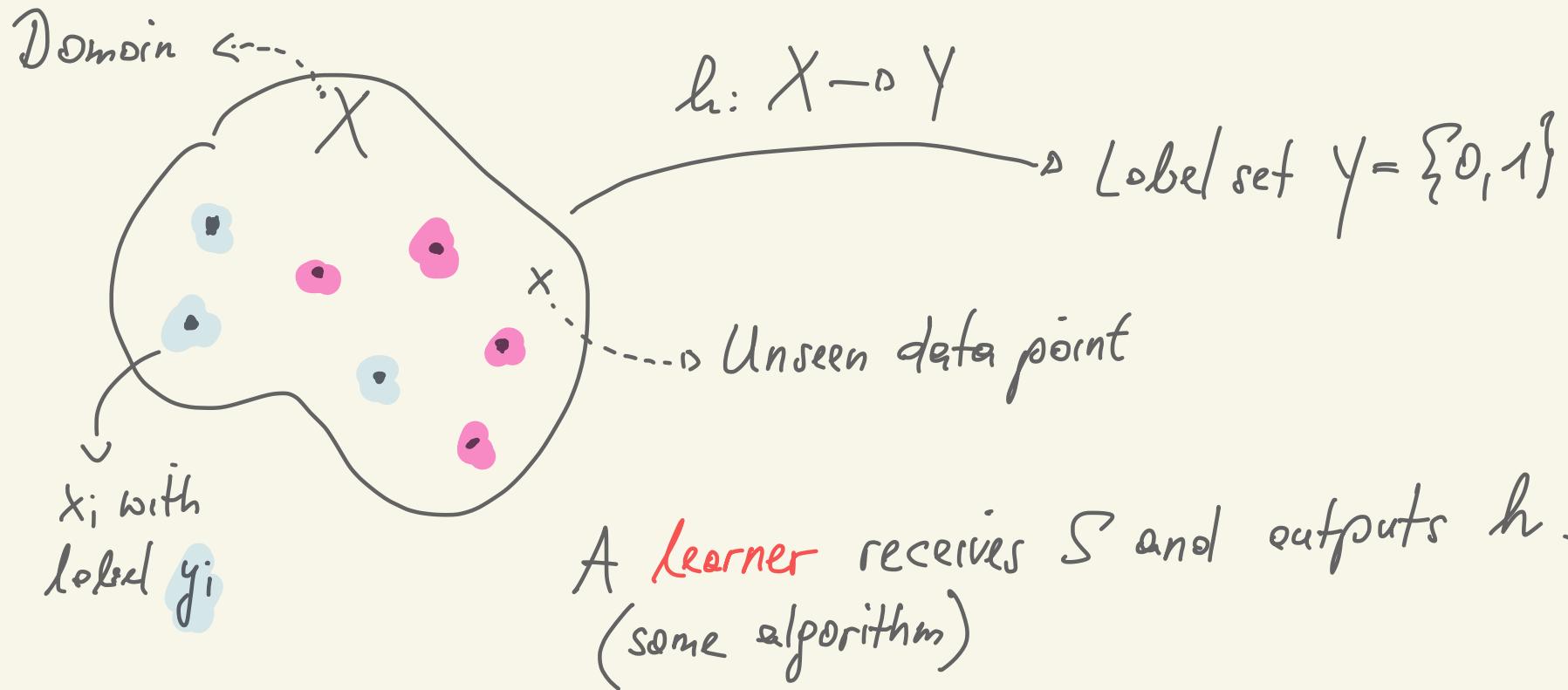
$$h: \mathbb{R}^2 \rightarrow \{0, 1\}$$

that will take, as input, a 2-dimensional vector and output whether a papaya is tasty or not. We call h a **hypothesis**.

What data is available to us?

$$S = \left((x_1, y_1), \dots, (x_N, y_N) \right), \text{ in our case } x_i \in \mathbb{R}^2 \\ y_i \in \{0, 1\}$$

What do we mean by learning?



Two assumptions that we will make initially is as follows:

- All the x_i 's are drawn independently and identically distributed from some (unknown) distribution (D) over the domain X .
 - The x_i 's are labeled by some (unknown) function $f: X \rightarrow Y$ (true labeling function). This means

what do we care about?

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

subset of X

all the points where the hypothesis h differs from the true labeling function.

MACHINE LEARNING (911.235)

Summer Term 2023

PRELIMINARIES

08/14/23

$\mathcal{P}(X)$... power set of X

Remember that we care about $A = \{x \in X : h(x) \neq f(x)\}$. If $X = \mathbb{R}^n$ and $A \in \mathcal{P}(X)$, we already have a problem.

Measure problem : find $\mu: \mathcal{P}(\mathbb{R}^n) \rightarrow [0, \infty]$ with

1. $A_i \in \mathcal{P}(\mathbb{R}^n), i \in \mathbb{N}$, pairwise disjoint, we want

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

2. if $A, B \in \mathcal{P}(\mathbb{R}^n)$ are congruent, we want

$$\mu(A) = \mu(B)$$

3. we want $\mu([0,1]^n) = 1$

! The measure problem is unsolvable for all $n \in \mathbb{N}$.

We will constrain ourselves to sets that are an element of some σ -Algebra over the domain X .

Def. (σ -Algebra): Let S be a non-empty set. A family of sets $\mathcal{F} \subset \mathcal{P}(S)$ is called a σ -Algebra over S , if

1. $S \in \mathcal{F}$

2. From $A \in \mathcal{F}$, it follows that $A^c = S \setminus A \in \mathcal{F}$

3. From $A_i \in \mathcal{F}, i \in \mathbb{N}$, it follows that

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

Example (smallest σ -Algebra that contains set A)

$$\{\emptyset, A, A^c, S\}$$

Def. (generator): Given $\Sigma \subset \mathcal{P}(S)$ (a family of sets) and Σ denoting the set of all σ -Algebras that contain Σ , we call

$$\sigma(\Sigma) = \bigcap_{\mathcal{F} \in \Sigma} \mathcal{F}$$

the σ -Algebra generated by Σ . Further, if it holds that for a σ -Algebra

A

$$\sigma(\Sigma) = A$$

we call Σ the generator of A .

Example: $\Sigma = \{\{\{1\}\}\}, S = \{1, 2, 3\}$

$$\sigma(\Sigma) = \sigma(\{\{\{1\}\}\}) = \{\emptyset, \{\{1\}\}, \{\{1\}, \{2, 3\}\}, \{\{1, 2, 3\}\}\}$$

Def. (Topological space): A topological space is a tuple (X, τ) with X a set and τ a collection of subsets of X with

1. $\emptyset \in \tau$ and $X \in \tau$

2. closed under union, i.e., $\{U_i\}_{i \in I} \subseteq \tau \Rightarrow \bigcup_{i \in I} U_i \in \tau$

3. closed under finite intersection, i.e.,

$$\{U_i\}_{i=1}^n \subseteq \tau \Rightarrow \bigcap_{i=1}^n U_i \in \tau$$

Remark: the elements of τ are called **open sets!**

Example: $X = \{1, 2, 3, 4\}$

$$\tau = \{\emptyset, \{1, 2, 3, 4\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

Def. (Borel σ -Algebra): given a topological space (S, \mathcal{O}) with \mathcal{O} denoting the system of open sets, then we call

$$\mathcal{B}(S) := \sigma(\mathcal{O})$$

the Borel σ -Algebra over S . Its elements are called Borel sets. For $S = \mathbb{R}^n$, we write $\mathcal{B}^n := \mathcal{B}(\mathbb{R}^n)$.

Each of the following systems of sets are generators for $\mathcal{B}(\mathbb{R}^n)$:

$$- \left\{ U \subset \mathbb{R}^n : U \text{ open} \right\}$$

$$- \left\{ A \subset \mathbb{R}^n : A \text{ closed} \right\}$$

$$- \left\{]q, b] : q, b \in \mathbb{R}^n \text{ with } q \leq b \right\}$$

$$- \left\{]-\infty, c] : c \in \mathbb{R}^n \right\}$$

$$]q, b] =]q_1, b_1] \times \dots \times]q_n, b_n]$$

$$\begin{aligned} & \text{for } q = (q_1, \dots, q_n) \\ & b = (b_1, \dots, b_n) \end{aligned}$$

Convention: We extend \mathbb{R} by symbols " $-\infty$ " and " $+\infty$ " as

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$$

$$\overline{\mathcal{B}} = \sigma(\mathcal{B} \cup \{-\infty\} \cup \{+\infty\})$$

Def (Measurable space): if \mathcal{F} is a σ -Algebra over S , we call
 (S, \mathcal{F})

a measurable space.

Example: $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

Def. (measurable maps/functions): Given (S_1, \mathcal{F}_1) and (S_2, \mathcal{F}_2) measurable spaces, we call

$$f: S_1 \rightarrow S_2$$

a measurable function if $\forall E \in \mathcal{F}_2 : f^{-1}(E) \subset \mathcal{F}_1$

Example: $1_A : S \rightarrow \{0, 1\}$ indicator function of set $A \subset S$

$$\omega \mapsto 1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{else} \end{cases}$$

Consider 1_A as a function to \mathbb{R} .

- if $B \leq 0$, then $\{\omega : 1_A(\omega) < B\} = \emptyset$
- if $B > 1$, then $\{\omega : 1_A(\omega) < B\} = S$
- if $0 < B \leq 1$, then $\{\omega : 1_A(\omega) < B\} = S \setminus A$

$\Rightarrow 1_A$ is measurable if $A \in \mathcal{F}$ (S, \mathcal{F})

Def. (Measure) : Let (S, \mathcal{F}) be a measurable space. A function

$$\mu: \mathcal{F} \rightarrow \overline{\mathbb{R}}$$

is called a measure, if the following conditions hold:

1. $\mu(\emptyset) = 0$

2. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$

3. for every sequence $(A_n)_{n \in \mathbb{N}}$ of disjoint sets from \mathcal{F} , we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad // \sigma\text{-Additivity}$$

Example: $\mu_{\text{Count}}: \mathcal{F} \rightarrow \overline{\mathbb{R}}$

$$A \mapsto \begin{cases} |A|, & \text{if } A \text{ is finite} \\ \infty, & \text{else} \end{cases} \quad // \text{Counting measure}$$

Def. (Measure space): given a measurable space (S, \mathcal{F}) and a measure $\mu: \mathcal{F} \rightarrow \overline{\mathbb{R}}$, we call

$$(S, \mathcal{F}, \mu)$$

a measure space.

Some important properties: let $A, B, A_n \in \mathcal{F}, n \in \mathbb{N}$. Then

1. if A and B are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$

2. if $A \subset B$, then $\mu(A) \leq \mu(B)$

3.

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i) \quad // \sigma\text{-sub-additivity}$$

Def. (Probability space): given (S, \mathcal{F}, D) a measure space, with

$$D(S) = 1,$$

then we call (S, \mathcal{F}, D) a probability space.

Remark: A possibility to construct, based on a measure space $(S_1, \mathcal{F}_1, \mu)$ and a measurable function f from S_1 to S_2 (with (S_2, \mathcal{F}_2)),

another measure space $(S_2, \mathcal{F}_2, \mu_f)$:

$$\mu_f : \mathcal{F}_2 \rightarrow [0, \infty]$$

$$B \mapsto \mu_f(B) := \mu(f^{-1}(B)) \quad // \text{push-forward measure}$$

Def (Random variable): if $(S_1, \mathcal{F}_1, \mathcal{P})$ is a probability space and (S_2, \mathcal{F}_2) a measurable space, then a measurable function

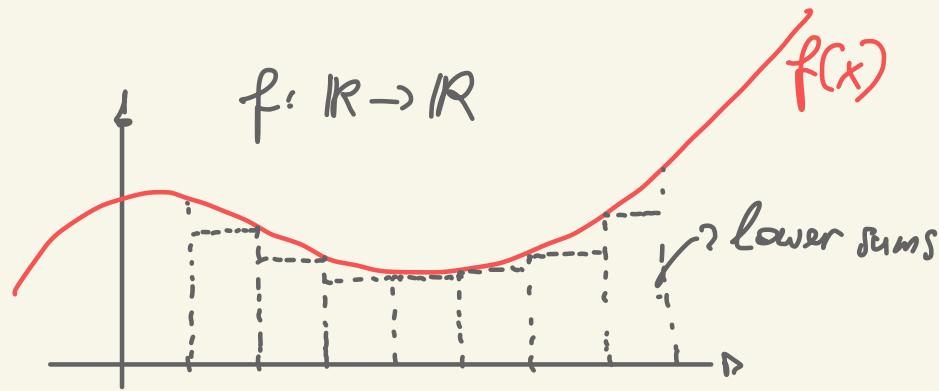
$$X: S_1 \rightarrow S_2$$

is called a random variable. For $S_2 = \mathbb{R}$, we call X a real random variable.

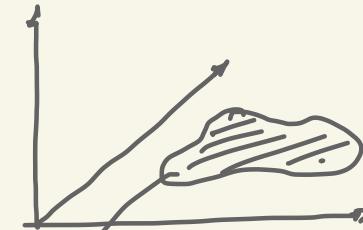
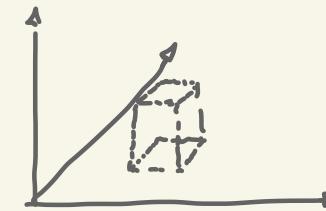
The push-forward measure \mathbb{P}_X on (S_2, \mathcal{F}_2) is also a prob.-measure.

$$\underbrace{\mathbb{P}_X(S_2)}_{\text{"distribution" of } X} = \mathcal{D}(X^{-1}(S_2)) = \mathcal{D}(S_1) = 1$$

Idea of Riemann integral

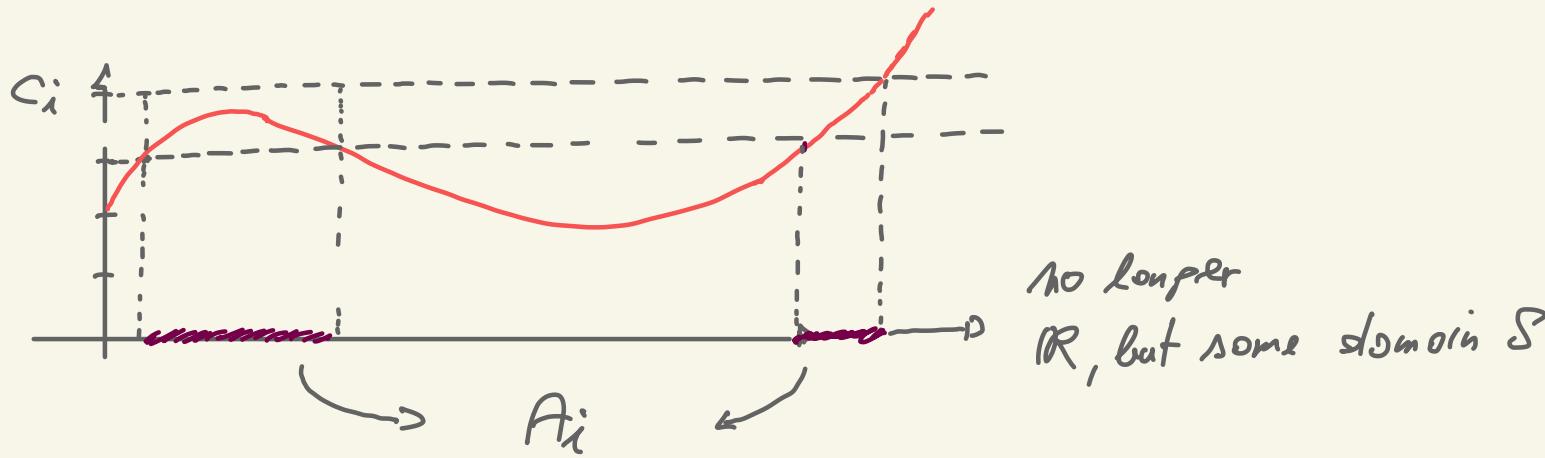


- Problems:
- Diff. to extend to higher dimensions
 - Reliance on continuity



so f is defined here
Q: How do partition the domain?

Idea of LEBESGUE integration: The key idea is to partition the range of a function $f: S \rightarrow \mathbb{R}$ in intervals, to get to an approximation by "elementary" functions.



We need a way to measure A_i . If we can do that, we can write

$$c_i \cdot \mu(A_i)$$

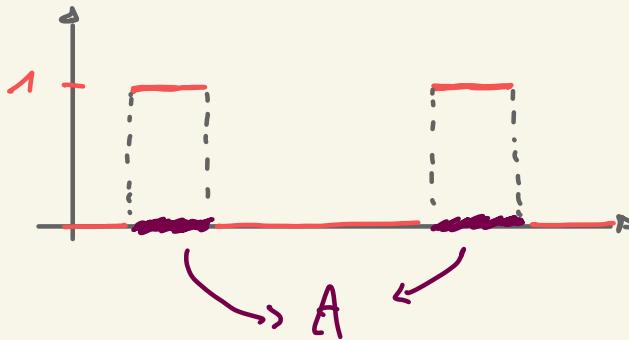
measure

.... This will allow us to eventually write

$$\sum_i c_i \mu(A_i) \text{ or } \int_S f d\mu$$

R

In a little bit more detail: In particular, let's look at case of "simple" functions. So, we have (S, F, μ) , $\mu: \bar{F} \rightarrow [0, \infty]$

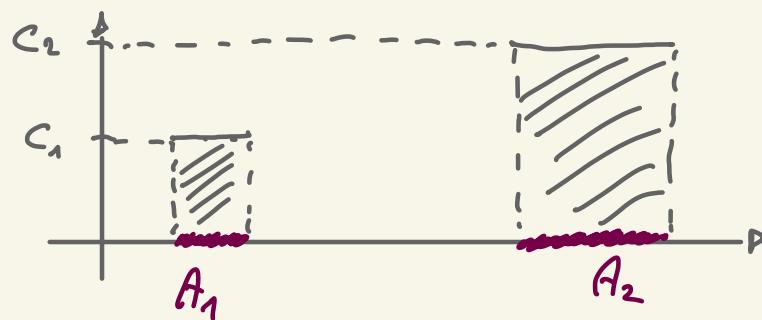


$\mathbb{1}_A : S \rightarrow \mathbb{R}, A \in \mathcal{F}$
 (Indicator function of set A)

Any reasonable interpretation of $\underline{1}_A$ should return $\mu(A)$.

$$\underbrace{I(\frac{1}{A})}_{\text{integral}} - \mu(A)$$

what is a "simple" function? $f(x) = \sum_{i=1}^n c_i \cdot 1_{A_i}(x)$ $A_i \in \mathcal{F}$
 $c_i \in \mathbb{R}$



we can define the integral of f via

$$I(f) = \sum_{i=1}^n c_i \cdot \mu(A_i)$$

Problem: $\mu: \mathcal{F} \rightarrow [0, \infty]$, $c_i \in \mathbb{R} \Rightarrow \text{so}$, we could get something like

$$10 \cdot \infty - 3 \cdot \infty$$

Solution: $\{f: S \rightarrow \mathbb{R} \mid f \text{ is 'simple' and } f \geq 0\} = T^+$

For $f \in T^+$, we have a representation of f as

$$f(x) = \sum_{i=1}^n c_i \cdot \mathbf{1}_{A_i}(x) \text{ with } c_i \geq 0$$

and we define the LEBESGUE integral as

$$I(f) = \sum_{i=1}^n c_i \cdot \mu(A_i) \text{ or } \int f \, d\mu, \text{ or } \int f(x) \, d\mu(x)$$

Properties: 1. For $f, g \in T^+$ and $\alpha, \beta \geq 0$

$$\int(\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu \quad \text{LINEARITY}$$

2. If $f, g \in T^+$ and $f \leq g$, it follows that

$$\int f d\mu \leq \int g d\mu \quad \text{MONOTONICITY}$$

! This construction extends to all measurable functions $f: S \rightarrow \mathbb{R}$.

Def (Expected value of a random variable): Given a probability space $(S, \mathcal{F}, \mathbb{P})$ and a (quasi-integrable) measurable random variable $X: S \rightarrow \mathbb{R}$, then

$$\mathbb{E}[X] = \int X d\mathbb{P}$$

is the expected value of X .

- Properties (some) :
1. if $X \geq 0$, then $\mathbb{E}[X] \geq 0$
 2. $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for random variables X and Y
 - ⋮

Our first inequality : MARKOV inequality

Def (Markov inequality): Given a prob. space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ and a non-negative (≥ 0) random variable, we have for $a > 0$

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a} \quad \left(\mathbb{P}[X \geq a] = \mathcal{P}\left(\{\omega \in \mathcal{S} : X(\omega) \geq a\}\right) \right)$$

Proof: We define $\varphi: S \rightarrow \mathbb{R}$ with

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

$\Rightarrow 0 \leq \varphi(x) \leq X(x)$. Hence, by monotonicity, we have

$$\begin{aligned} \int X dD &\geq \underbrace{\int \varphi dD}_{= a \cdot D(\{\omega \in S : X(\omega) \geq a\})} \end{aligned}$$

Divide by $a > 0$

$$\Rightarrow \frac{1}{a} \cdot \int X dD \geq D(\{\omega \in S : X(\omega) \geq a\}) = P[X \geq a]$$

Since $\int X dD$ is $E[X]$, we get $P[X \geq a] \leq \frac{E[X]}{a}$



Let's start with actual ML content

(ended here on 08/20)

Recap of our setup:

1. DOMAIN set X ; we call $x \in X$ an instance

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

3. TRAINING set $S = ((x_1, y_1), \dots, (x_m, y_m))$ with $x_i \in X, y_i \in Y$

4. A LEARNER that receives S and outputs

$$h: X \rightarrow Y$$

which we call a **hypothesis**.

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

We are interested in

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

"Generalization error"

The empirical version of this is

$$\frac{1}{m} \left| \left\{ i \in [m] : h(x_i) \neq f(x_i) \right\} \right| = \underbrace{L_S(h)}_{\text{Empirical error (or empirical risk)}} \quad ([m] = \{1, \dots, m\})$$

Convention: $S/x = (x_1, \dots, x_m)$

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

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

$$\mathbb{1}_{h(x_i) \neq f(x_i)} = \begin{cases} 1, & \text{if } h(x_i) \neq f(x_i) \\ 0, & \text{else} \end{cases}$$

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

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

$$= \frac{1}{m} \sum_{i=1}^m \underset{x \sim D}{\mathbb{P}} [h(x) \neq f(x)] \quad // \text{as } \mathbb{1} \dots \text{ only take on values } 0, 1$$

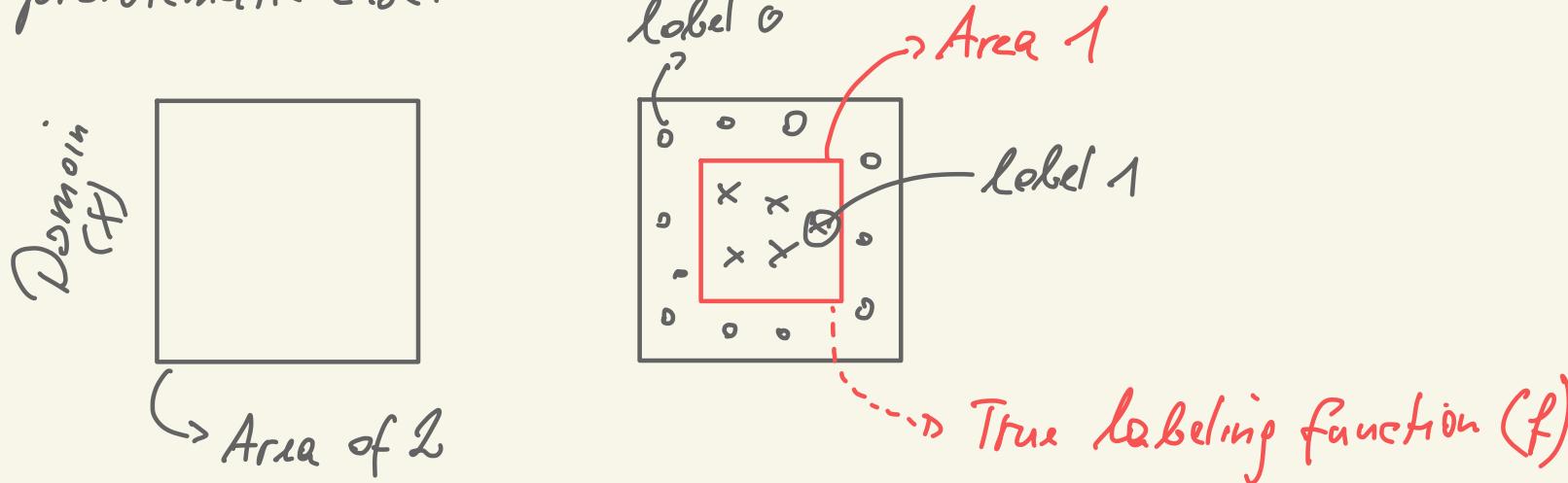
$$= \cancel{\frac{1}{m} \cdot m} \cdot \underset{x \sim D}{\mathbb{P}} [h(x) \neq f(x)]$$

$$= L_{D, f}(h) \quad // \text{this establishes the claim.}$$

Our first learning paradigm

Empirical risk minimization (ERM): As we only have access to the training data (S), it's natural to try to select \underline{h} such that the empirical risk is minimized. We call such an \underline{h} an empirical risk minimizer (h_S).

A problematic case:



- Say the distribution on X is uniform

- So if we have an ERM algorithm that returns h_S s.t.

$$h_S(x) = \begin{cases} y_i, & \text{if } \exists i \in [m] : x_i = x \\ 0, & \text{else} \end{cases} \quad // \text{lookup table}$$

Obviously h_S is correct on our training set $\Rightarrow L_S(h_S) = 0!$

But on unseen instances from $D (x \sim D)$, h_S is only correct 50% of the time (due to the ratio of areas \square and \square) $\Rightarrow L_{D,f}(h_S) = \frac{1}{2}!$

This is called overfitting!

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

$$\text{ERM}_H(S) \in \underset{h \in H}{\operatorname{argmin}} L_S(h)$$

Remark: In our previous example, we did not do this and allowed to memorize the training data!

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

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

Now, only ERM hypothesis h_S will attain 0 empirical error ($L_S(h_S) = 0$), as it competes with h^* (which obviously has 0 empirical error).

Hence, $L_{D,f}(h_S) > \varepsilon$ can only happen if we select a hypothesis with $L_S(h_r) = 0$ but $L_{D,f}(h_r) > \varepsilon$.

We can write

$$H_{\text{BAD}} = \left\{ h \in H : L_{D,f}(h) \geq \varepsilon \right\} \quad // \text{set of BAD hypotheses}$$

Also, we define

$$M = \left\{ S|_x : \exists h \in \underbrace{H}_{\text{BAD}}, L_S(h) = 0 \right\}$$

those are the ones with generalization error $\geq \varepsilon$

We observe

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

Emp. risk minimizer

$$M = \bigcup_{h \in H_{\text{BAD}}} \left\{ S|_x : L_S(h) = 0 \right\}$$

We get (upon measuring with D):

$$\begin{aligned}
 D^m \left(\{S|x : L_{D,f}(h_S) \geq \varepsilon\} \right) &\leq D^m \left(\bigcup_{h \in H_{BAD}} \{S|x : L_S(h) = 0\} \right) \\
 &\leq \sum_{h \in H_{BAD}} D^m \left(\{S|x : L_S(h) = 0\} \right)
 \end{aligned}$$

// due to
 "D-sub-additivity"
 "Union Bound"

Let's fix some $h \in H_{BAD}$:

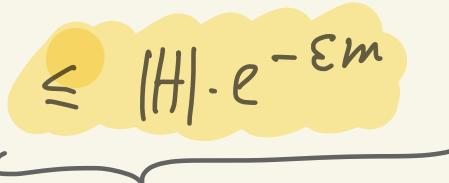
$$\begin{aligned}
 D^m \left(\{S|x : L_S(h) = 0\} \right) &= D^m \left(\{S|x : \forall i \in [m] : h(x_i) = f(x_i)\} \right) \\
 &= \prod_{i=1}^m D \left(\{x_i : h(x_i) = f(x_i)\} \right) \\
 &= \prod_{i=1}^m (1 - L_{D,f}(h)) \\
 &\leq \prod_{i=1}^m (1 - \varepsilon) \quad \text{as } h \in H_{BAD} \\
 &= (1 - \varepsilon)^m \leq e^{-\varepsilon m}
 \end{aligned}$$

// due to iid
 assumption

$$\Rightarrow \mathbb{D}^m \left(\{ \S_k : L_{D,f}(h_S) \geq \varepsilon \} \right) \leq \sum_{h \in H_{RAD}} e^{-\varepsilon m}$$

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

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



If we let this be $\leq \delta \in (0, 1)$ and solve for m , we get

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



error ↳ confidence

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 each labeling function $f: X \rightarrow Y$ and any distribution D over domain X (for which realizability holds), we have that with probability of at least $1 - \delta$ over the choice of S_{f_X} (of size m) it holds that every ERM hypothesis h_S satisfies

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

Interpretation: For sufficiently large m , ERM_H returns h_S (i.e., a hypothesis) that is PROBABLY APPROXIMATELY CORRECT (PAC).

This leads to:

Def. (PAC learnability): A hypothesis class H is **PAC learnable**, if there exists a function $m_H : (0, 1)^2 \rightarrow \mathbb{N}$ and a learning algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0, 1)$ and (II) every distribution D over domain X , and (III) every labeling function $f: X \rightarrow \{0, 1\}$, if (IV) realizability holds (with respect to D, H, f), then running A on $m \geq m_H(\varepsilon, \delta)$ iid instances drawn from D and labeled by f returns a hypothesis h such that with probability of at least $1 - \delta$ (over choice of S)

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

Def. (sample complexity): $m_H : (0, 1)^2 \rightarrow \mathbb{N}$ is called the sample complexity function. In particular, m_H returns the smallest integer such that the requirements for PAC learnability are satisfied.

We have already seen that finite hypothesis classes ($|H| < \infty$) are PAC learnable with

$$m_H(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon} \cdot \log \left(\frac{|H|}{\delta} \right) \right\rceil \quad \text{--- ceiling function}$$

We will now move to a more general setting.

[1] We will first release the realizability assumption.

(In this setting, the best we can hope for are guarantees relative to the "best" possible hypothesis in the class; $\min_{h \in \mathcal{H}} L_D(f(h))$)

Def. (Hoeffding inequality): Let X_1, \dots, X_m be iid random variables taking values in $[q_i, b_i]$ for $i \in [m]$. Then, it holds that

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

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

end

$$P[S_m - \mathbb{E}[S_m] < -\varepsilon] \leq e^{-\frac{2\varepsilon^2}{\sum_i (b_i - q_i)^2}}$$

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

(as we have $|q| \geq b \Leftrightarrow q \leq -b \text{ OR } q \geq b$)

Another useful form of this inequality is:

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

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

As a consequence of (*), we can say the following: fix $\varepsilon > 0$; then for any single $h: X \rightarrow Y$, we have

$$P\left[\left|L_S(h) - L_{D,f}(h)\right| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 m} \quad (a=0, b=1)$$

$S |_{X^n D^m}$

If we would set $2^{-2\varepsilon^2 m} = \delta$, and solve for ε , we would get

$$\varepsilon = \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

$$\Rightarrow L_{D,f}(h) \leq L_S(h) + \sqrt{\frac{1}{2m} \cdot \log\left(\frac{2}{\delta}\right)}$$

holds with probability of at least $1-\delta$ our choice of S

Remark: This result holds for a single h . However, we can easily get a bound that holds uniformly for all $h \in H, |H| < \infty$:

$$\underset{S \subseteq X^n \cup D^m}{\mathbb{P}} \left[\exists h \in H : |L_S(h) - L_{D, f}(h)| > \varepsilon \right] \leq \sum_{h \in H} 2 \cdot e^{-2\varepsilon^2 m} = 2|H| \cdot e^{-2\varepsilon^2 m}$$

(by the union bound)

We see that we did need realizability for that!

[2] Next, we release our requirement of a "true" labeling function f .

We do this, by letting \mathcal{D} be a distribution over $X \times Y = \mathcal{Z}$. We need to adjust our definitions of empirical error and generalization error:

$$(1) \quad L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y) \sim \mathcal{D}} [h(x) \neq y] = \mathcal{D}(\{(x,y) \in X \times Y : h(x) \neq y\})$$

$$(2) \quad L_S(h) = \frac{1}{m} \cdot \left| \left\{ i \in [m] : h(x_i) \neq y_i \right\} \right|$$

Both $\boxed{1}$ and $\boxed{2}$ lead to:

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

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

(last slide on April 18)

General loss functions:

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

Example: 0-1 loss

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

Square loss

$$\ell^{\text{sq}}(h, (x, y)) = (y - h(x))^2 \quad // \text{e.g.: in regression problems}$$

↳ l.p. $\in \mathbb{R}$

In case of general loss functions, we adjust our def. of L_D and L_S as follows:

$$L_D(h) = \mathbb{E}_{\mathcal{Z}} [\ell(h, z)], \quad L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^m \ell(h, z_i)$$

Remark: $\mathbb{E}_{\tau} [e^{\ell(h, \tau)}] = 0 \cdot \mathbb{P}_{\tau} [h(\tau) = g] + 1 \cdot \mathbb{P}_{\tau} [h(\tau) \neq g]$

$$= \mathbb{P}_{\tau} [h(\tau) \neq g] \quad // \text{as before!}$$

Uniform convergence

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

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

Lemma: Assume that S is $\frac{\varepsilon}{2}$ -representative wrt. $\mathcal{Z}, \mathcal{H}, \ell$ and D .

Then, any hypothesis h_S returned by $\text{ERM}_H(S) \in \arg\min_{h \in H} L_S(h)$ satisfies

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

Proof: for any $h \in H$:

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

As this inequality chain holds for any $h \in H$, we conclude that

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



This leads us to the definition of uniform convergence.

Def. (Uniform convergence): A hypothesis class H has the uniform convergence (UC) property (with respect to ℓ , and loss function ℓ), if there exists $m_H^{uc}: (0,1)^2 \rightarrow \mathbb{N}$, such that for every $\varepsilon, \delta \in (0,1)$ and every distribution D over \mathcal{Z} , if S is an 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 has the UC property with m_H^{uc} , then H is agnostic PAC learnable with

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

Moreover, ERM is a successful APAC learner for H !

Claim: Finite H ($|H| < \infty$) are agnostic PAC learnable.

Given fixed $\varepsilon, \delta \in (0, 1)$, we want to show that

$$\mathcal{D}^m \left(\{S : \forall h \in H : |L_S(h) - L_D(h)| \leq \varepsilon\} \right) \geq 1 - \delta$$

Equivalently,

$$\mathcal{D}^m \left(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \varepsilon\} \right) < \delta$$

$\underbrace{\phantom{\mathcal{D}^m \left(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \varepsilon\} \right) < \delta}}$

$$\bigcup_{h \in H} \{S : |L_S(h) - L_D(h)| > \varepsilon\}$$

We bound

$$\mathcal{D}^m \left(\bigcup_{h \in H} \{S : |L_S(h) - L_D(h)| > \varepsilon\} \right) \leq \sum_{h \in H} \mathcal{D}^m \left(\{S : |L_S(h) - L_D(h)| > \varepsilon\} \right)$$

, by union bound

Let's fix $\ell = \ell^{0-1}$. As we know that $L_D(h) = \mathbb{E}_z [\ell(h, z)]$ and $L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i)$, we can use Hoeffding's inequality (with $b=1, a=0$):

$$\mathcal{D}^m \left(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \varepsilon\} \right) \leq \sum_{h \in H} 2 \cdot e^{-2\varepsilon^2 m} \\ = 2 \cdot |H| \cdot e^{-2\varepsilon^2 m}$$

Let $2 \cdot |H| e^{-2\varepsilon^2 m}$ be smaller than δ , i.e.,

$$2 \cdot |H| \cdot e^{-2\varepsilon^2 m} < \delta \iff m > \log \left(\frac{2 \cdot |H|}{\delta} \right) \cdot \frac{1}{2\varepsilon^2}$$

This gives us the sample complexity function m_H^{uc} for uniform convergence. In other words, for $\ell : H \times \mathcal{Z} \rightarrow [0, 1]$, we have shown

$$m_H^{uc}(\varepsilon, \delta) = \left\lceil \log \left(\frac{2 \cdot |H|}{\delta} \right) \cdot \frac{1}{2\varepsilon^2} \right\rceil$$

and by our earlier corollary we get for the sample complexity function m_H for agnostic PAC learnability

$$m_H(\varepsilon_{rd}) \leq m_H^{uc}\left(\frac{\varepsilon}{2}, d\right) \leq \left\lceil \frac{2 \cdot \log\left(\frac{2 \cdot |H|}{\delta}\right)}{\varepsilon^2} \right\rceil$$

This establishes the claim. \square

(ended dec. on April 25)

Question: Is there a "universal" learner?, i.e., a learner A without any prior knowledge (with respect to H or D, \dots) of a task, but can be challenged by any task (and achieves low $L_D(A(s))$). → specified by D

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

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

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

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

2 means that the learner fails on that task; 1 means that the task can be successfully learned by another learner (e.g. ERR_H for $H = \{f\}$).

Proof: Let $C \subset X$ be of size 2^m , $|C| = 2^m$. The number of possible labelings of C is $T = 2^{2^m}$. Let's denote the functions which realize these T labelings as f_1, \dots, f_T

For each f_i , we define

$$\underbrace{D_i(\{(x, y)\})}_{\text{defines a task}} = \begin{cases} \frac{1}{|C|}, & \text{if } y = f_i(x) \\ 0, & \text{else} \end{cases}$$

Hence, by construction, $L_{D_i}(f_i) = 0$.

With that in mind, we will show that for every learning algorithm A that receives m samples from $X \times \{0,1\}$, there exists a function $f: X \rightarrow \{0,1\}$ and a distribution D over $X \times \{0,1\}$ such that

$$1. L_D(f) = 0, \text{ and}$$

$$2. \mathbb{E}_{S \sim D^m} [L_D(A(S))] \geq \frac{1}{4} \quad (\times)$$

In particular, we show that for every learning alg. A receiving m samples from $C \times \{0,1\}$ and returning a function $A(S): X \rightarrow \{0,1\}$, we have

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m} [L_{D_i}(A(S))] \geq \frac{1}{4} \quad (\times \times)$$

$$[T] = \{1, \dots, T\}$$

Remark: (x) suffices for $L_D(A(\xi)) \geq \frac{1}{g}$ with prob. of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ (see PS exercise).

Let's show that (xx) holds:

We know that there are $(2m)^m - k$ possible training sequences of size m (remember that $|C|=2m$) from C . Let's call them

$$S_1, \dots, S_k$$

Also, let's call S_j^i the sequence S_j labeled by f_i , i.e.,

$$S_j^i = ((x_n, f_i(x_n)), \dots, (x_m, f_i(x_m)))$$

$$\text{with } S_j = (x_1, \dots, x_m)$$

Now, in case the distribution is D_i , then S_1^i, \dots, S_k^i are the possible training sequences that A can receive.

Note that, by construction (of the D_i 's), all these training sequences have equal probability of being drawn / sampled.

$$\Rightarrow \mathbb{E}_{S \sim D_i^m} [L_{D_i}(A(S))] = \frac{1}{k} \cdot \sum_{j=1}^k L_{D_i}(A(S_j^i)) \quad (\text{xxx})$$

\hookrightarrow # possible training seq. of size m

(lets remember what we want to show)

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m} [L_{D_i}(A(S))] \geq \frac{1}{q}$$

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m} [L_{D_i}(A(S))] = \max_{i \in [T]} \frac{1}{k} \cdot \sum_{j=1}^k L_{D_i}(A(S_j^i)) \quad \text{by (xxx)}$$

$$\geq \frac{1}{T} \cdot \sum_{i=1}^T \frac{1}{k} \cdot \sum_{j=1}^k L_{D_i}(A(S_j^i)) \quad // \text{Qs max} \geq \text{avg.}$$



$$= \frac{1}{k} \cdot \sum_{j=1}^k \frac{1}{T} \cdot \sum_{i=1}^T \mathbb{L}_{D_i}(A(\xi_j^i))$$

$$\geq \underset{j \in [k]}{\min} \left[\frac{1}{T} \cdot \sum_{i=1}^T \mathbb{L}_{D_i}(A(\xi_j^i)) \right]$$

lets fix some $j \in [k]$: As ξ_j is of size m , but C is of size 2^n , there are instances from C which we have not seen. Lets call them

$$V_1, \dots, V_p$$

We also know that $p \geq m$.

For every $h: C \rightarrow \{0,1\}$ and every i , it holds that

$$L_{D_i}(h) = \frac{1}{2m} \cdot \sum_{x \in C} \mathbb{1}_{h(x) \neq f_i(x)}$$

$$\geq \frac{1}{2m} \sum_{r=1}^p \mathbb{1}_{h(v_r) \neq f_i(v_r)}$$

$$\geq \frac{1}{2\rho} \cdot \sum_{r=1}^p \mathbb{1}_{h(v_r) \neq f_i(v_r)} \quad (\text{xxx})$$

(as $\rho \geq m$)

Combining results gives

$$\frac{1}{T} \sum_{i=1}^T L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^T \frac{1}{2\rho} \cdot \sum_{r=1}^p \mathbb{1}_{\underbrace{A(S_j^i)(v_r)}_{\downarrow} + f_i(v_r)}$$

A ran on S_j^i gives some $h: C \rightarrow \{0,1\}$ and we know that (xxx) holds for every $h: C \rightarrow \{0,1\}$

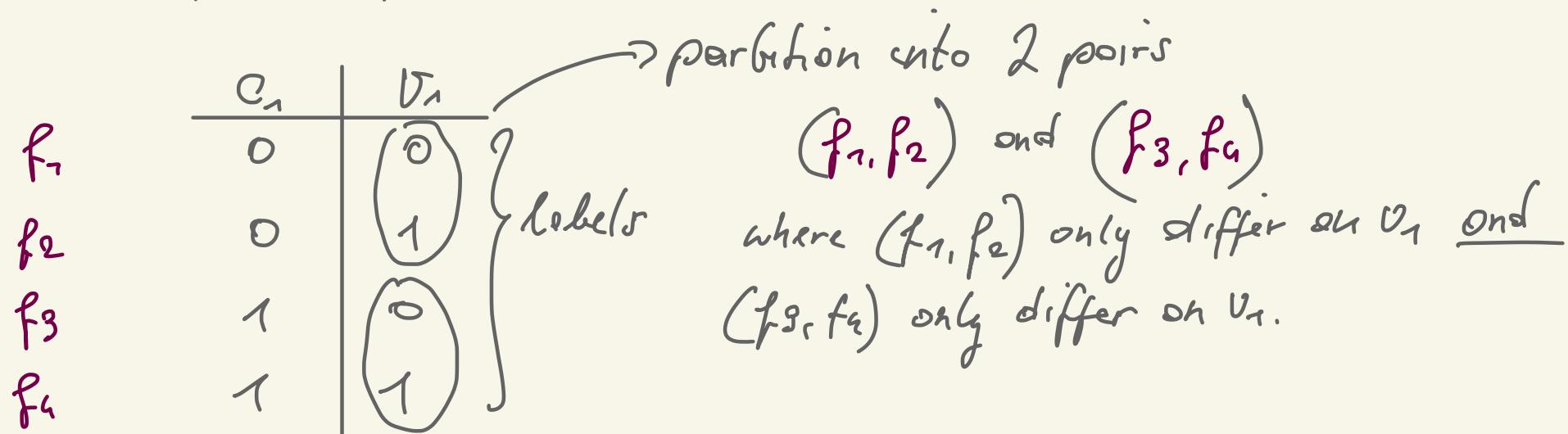
$$= \frac{1}{2} \cdot \frac{1}{P} \sum_{h=1}^P \frac{1}{T} \cdot \sum_{i=1}^T \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

$$\geq \frac{1}{2} \cdot \min_{h \in [P]} \frac{1}{T} \cdot \sum_{i=1}^T \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_h)} \quad // \text{as } \min \geq \text{avg}$$

Let's fix some $r \in [P]$:

Remember that we have f_1, \dots, f_T (for $T = 2^{2m}$ labelings)

Example: say $m=1$, $|C|=2$



In general, we can partition into $\frac{T}{2}$ disjoint pairs of functions (f_i, f'_i) . For every pair (f_i, f'_i) we have that for every $c \in C$ $f_i(c) + f'_i(c)$ if and only if $c = v_r$. For such a pair, we have $S_j^{(i)} = S_j^{(i')}$.

Consequently, we have

$$\frac{1}{T} \sum_{j=1}^T A(S_j^{(i)})(v_r) + f_i(v_r) + \frac{1}{T} \sum_{j=1}^T A(S_j^{(i')})(v_r) + f'_i(v_r) = 1$$

Remembering that we had $\frac{1}{T} \sum_{i=1}^T \frac{1}{T} \sum_{j=1}^T A(S_j^{(i)})(v_r) + f_i(v_r)$ and now re-writing the sum as a sum over all disjoint pairs, we get in combination with ~~$\times \times \times \times$~~

$$\cancel{\frac{1}{T} \cdot \frac{T}{2}} = \frac{1}{2}$$

Overall,

$$\max_{i \in [T]} \mathbb{E}_{S_i \sim \mathcal{D}_i^m} \left[L_{\mathcal{D}_i}(A(s)) \right] \geq \frac{1}{2} \cdot \min_{r \in [p]} \underbrace{\frac{1}{T} \cdot \sum_{j=1}^T \prod_{i=1}^n A(s_j^i)(v_r)}_{= \frac{1}{2}} = f_i(v_r)$$

$$= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Our earlier steps of fixing choices in the min-operations are irrelevant as we get a constant lower bound!

Corollary: let X be an infinitely large domain and let H be the set of all functions from $X \rightarrow \{0, 1\}$. Then, H is not PAC learnable.

□

Vapnik - Chervonenkis Dimension (VC dimension)

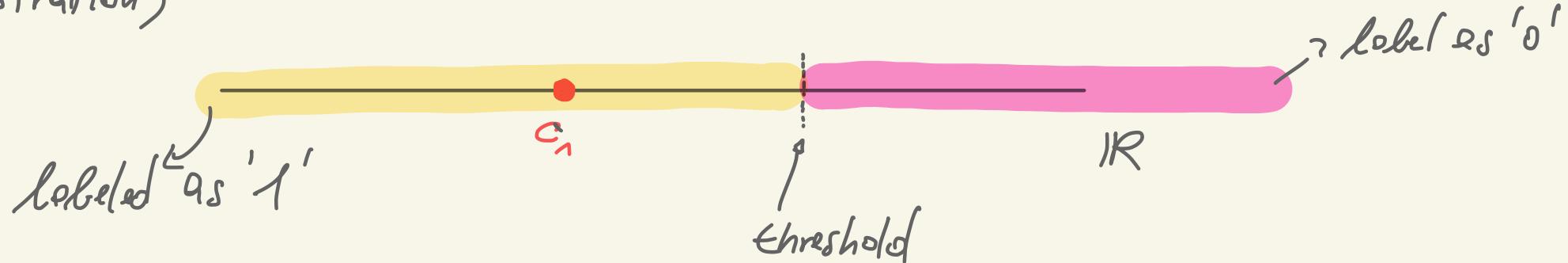
if we review the proof of the NFL theorem, it seems intuitive to study how a hyp. class H behaves on C (i.e., the set from NFL thm.).

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

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

as the restriction of H to C .

Example : $C = \{c_1\}$, $c_1 \in \mathbb{R}$; lets look at the class H of
(Illustration)



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

$$C = \{c_1, c_2\}$$



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

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

Example: The class of thresholds on \mathbb{R} is PAC learnable.

$$H^{\text{thr}} = \{h_\alpha : \alpha \in \mathbb{R}\}, \quad h_\alpha : \mathbb{R} \rightarrow \{0, 1\}$$

$$x \mapsto h_\alpha(x) = \begin{cases} 1 & x < \alpha \\ 0 & \text{else} \end{cases}$$

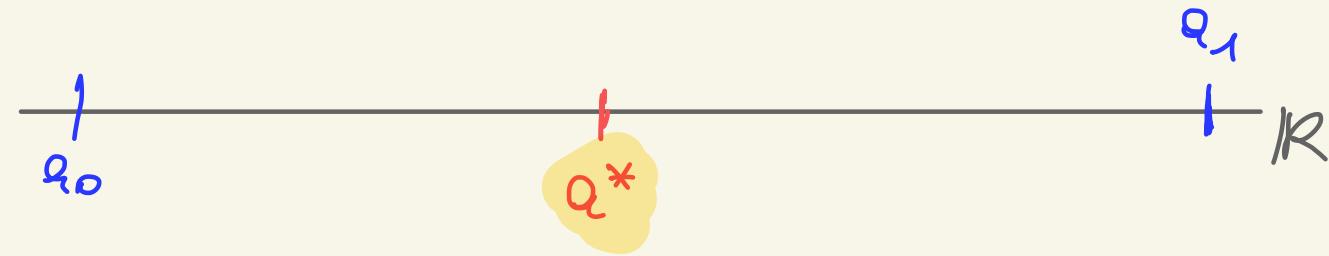
$$= \begin{cases} 1, & \text{if } x < \alpha \\ 0, & \text{else} \end{cases}$$

Claim: H^{thr} is PAC learnable with

$$m_{H^{\text{thr}}}(\epsilon, \delta) \leq \left\lceil \log\left(\frac{2}{\delta}\right) \cdot \frac{1}{\epsilon} \right\rceil$$

Proof: we assume realizability $\Rightarrow \exists h^* \in H^{\text{thr}}$ such that $L_{D,f}(h^*) = 0$.

We let α^* be the corresponding threshold.



Let q_0 be such that $D(\{x \in \mathbb{R} : x \in (q_0, q^*)\}) = \varepsilon$
 Let q_1 be such that $D(\{x \in \mathbb{R} : x \in (q^*, q_1)\}) = \varepsilon$

The special cases are:

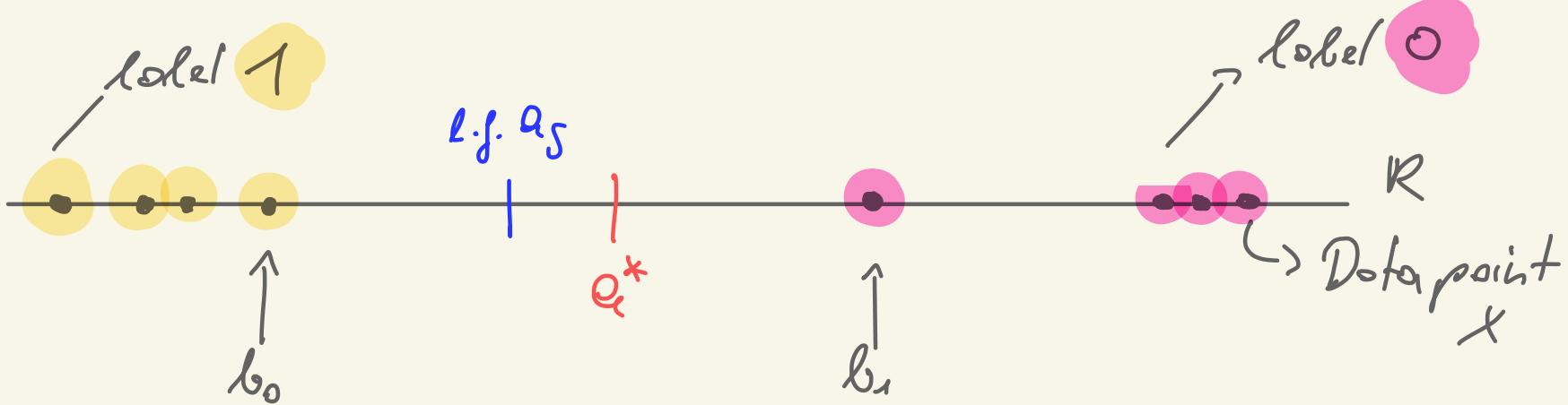
- if $D(\{x \in \mathbb{R} : x \in (q_0, q^*)\}) < \varepsilon$, then set q_0 to $-\infty$
- if $D(\{x \in \mathbb{R} : x \in (q^*, q_1)\}) < \varepsilon$, then set q_1 to $+\infty$

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

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

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

Visually,

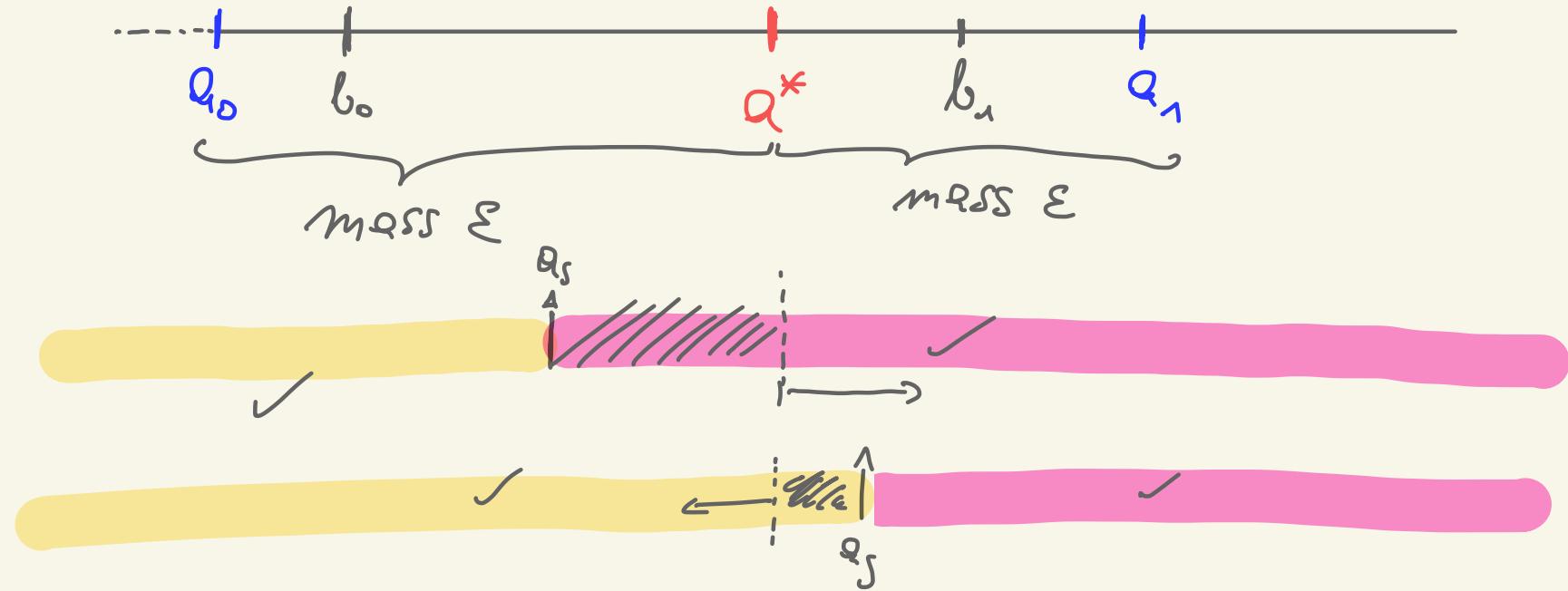


So, ERK can choose any threshold within (b_0, b_1) . Let's denote the threshold of an ERK hyp. by α_S and let h_S be the corresponding hypothesis.

For h_S (with α_S as threshold) to have $L_{D,f}(h_S) \leq \varepsilon$, it suffices that

- 1) $b_1 \leq \alpha_S$ AND
- 2) $b_0 \geq \alpha_S$

Why?



In other words,

$$\begin{aligned}
 P[L_{D,f}(h_s) > \epsilon] &\leq P[(b_0 < q_0) \vee (b_1 > q_1)] \\
 &\leq P[(b_0 < q_0)] + P[(b_1 > q_1)] \quad \text{Union bound}
 \end{aligned}$$

OR

So, when does $b_0 < q_0$ happen? When there is no data point x in S that is labeled as 1 s.t. $x \in (q_0, q^*)$.

We do know that $D(\{x \in R : x \in (q_0, q^*)\}) = \varepsilon$, hence not seeing an instance in that interval (q_0, q^*) has probability of $1 - \varepsilon$. Not seeing a data point in m iid samples thus has probability of $\underline{(1-\varepsilon)^m} \leq e^{-\varepsilon m}$.

$$\Rightarrow \underset{S \sim D^m}{P}[b_0 < q_0] = (1 - \varepsilon)^m \leq e^{-\varepsilon m}; \text{ same for } P[b_1 > q_1]$$

We get, overall

$$P[L_{\text{Diff}}(h_S) > \varepsilon] \leq 2 \cdot e^{-\varepsilon m} \quad \begin{matrix} \rightarrow \text{we want this to be } < \delta \\ \uparrow \\ \text{ERH} \\ \text{hyp.} \end{matrix} \Rightarrow m > \log\left(\frac{2}{\delta}\right) \cdot \frac{1}{\varepsilon}$$



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

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

Theorem : Let H be a class of functions from $X \rightarrow \{0, 1\}$. If H has infinite VC-dimension, then H is not PAC learnable.

without proof (see NFL).

Def. (growth function): Let H be a class of functions from $X \rightarrow \{0, 1\}$.

The **growth function** of H , $\gamma_H : \mathbb{N}_0 \rightarrow \mathbb{N}_0$, is defined as

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

Remark: Alternatively, we could define the VC-dim. via γ_H :

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

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

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

$$\tilde{\epsilon}_H(m) \leq \left(\frac{e \cdot m}{d}\right)^d \text{ if } m > d.$$

without proof!

VC-dimension of 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

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

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

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

growth function (evaluated on $2m$)

without proof!

Fundamental theorem of statistical learning

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

- [1] H has the uniform convergence property
- [2] Any ERM algorithm is a successful agnostic PAC learner for H
- [3] H is agnostic PAC learnable
- [4] Any ERM algorithm is a successful PAC learner for H
- [5] H is PAC learnable
- [6] H has finite VC dimension

LINEAR PREDICTORS

June 6, 23

We define

$$\mathcal{L}_d = \left\{ h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R} \right\} \quad h_{w,b}(x) = \langle w, x \rangle + b$$

as the class of affine functions. From \mathcal{L}_d we get different predictors via composition with some

$$\phi : \mathbb{R} \rightarrow \mathcal{Y}$$

Further, we can use

$$w' = (b, w_1, \dots, w_d)^T \in \mathbb{R}^{d+1}$$

$$x' = (1, x_1, \dots, x_d)^T \in \mathbb{R}^{d+1}$$

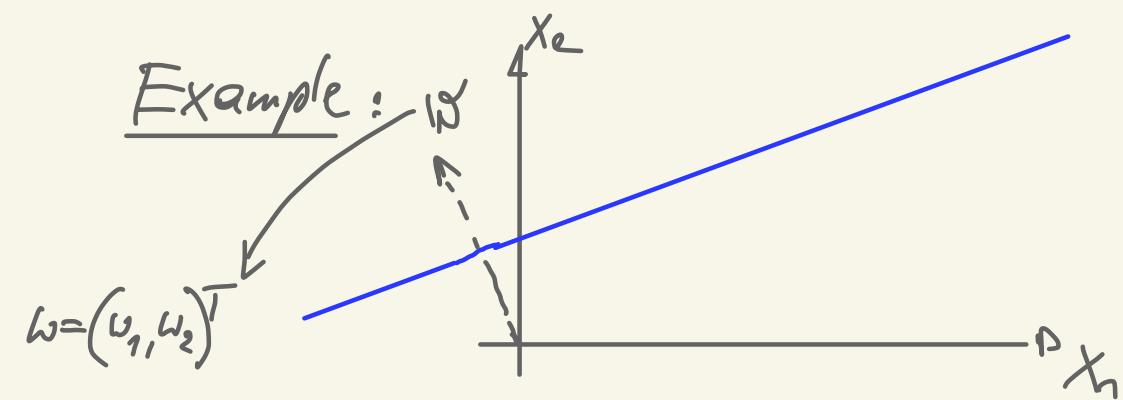
to write

$$h_{w,b}(x) = \langle w', x' \rangle.$$

If $\phi: \mathbb{R} \rightarrow \mathcal{Y}$ is $\phi = \text{sign}$, we get so called *halfspace predictors*:

$$HS_d = \phi \circ L_d = \left\{ x \mapsto \text{sign}(h_{w,b}(x)) : h_{w,b} \in L_d \right\}$$

Example:



Question: What is the VC-dimension of

$$\mathcal{F} = \left\{ x \mapsto \text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d \right\}$$

VC(\mathcal{F}) = ?

Claim: $VC(\mathcal{F}) = d$

[1] we show that $VC(\mathcal{F}) \geq d$ (lower bound). Let

$$X = \{c_1, \dots, c_d\} \text{ with } c_i \in \mathbb{R}^d \text{ for all } i \in \{1, \dots, d\}$$

$$c_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad c_d = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad \left(\text{Basis vectors in } \mathbb{R}^d \right)$$

Let $w = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}$ {
 label of 1st point
 label of d-th point}

} for any given labeling of c_1, \dots, c_d
 $\{\pm 1\}$

Notation c_{ij} j-th coordinate of c_i

$$\langle w_i, c_i \rangle - \sum_{j=1}^d w_j \cdot c_{ij} = 1 \cdot w_i = y_i \quad \checkmark$$

$\Rightarrow VC(\mathcal{F}) \geq d$ (as we have found a set of size d that is shattered).

[2] We will show the upper-bound $VC(\mathcal{F}) < d+1$ via contradiction.

Assume that $d+1$ points,

$$X = \{c_1, \dots, c_{d+1}\},$$

are shattered by \mathcal{F} .

This means that $\exists w_1, \dots, w_{2^{d+1}}$ weight vectors that realize all 2^{d+1} labelings of the $d+1$ points.

Getting all "possible labelings" would look like this:

$$M = \begin{pmatrix} w_1^T c_1 & w_1^T c_1 & \dots & & w_1^T c_{d+1} \\ w_1^T c_2 & w_2^T c_2 & & & \vdots \\ w_1^T c_3 & w_2^T c_3 & & & \vdots \\ \vdots & \vdots & & & \vdots \\ w_1^T c_{d+1} & w_2^T c_{d+1} & & & w_{2^{d+1}}^T c_{d+1} \end{pmatrix}$$

2^{d+1} columns

d+1 rows

we can also write this as XW

$$X = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{d+1} \end{pmatrix}, \quad W = \begin{pmatrix} | & | & & & | \\ w_1 & w_2 & \dots & w_{2^{d+1}} \end{pmatrix}$$

Taking $\text{sign}(XW)$ componentwise gives all possible labelings.

Let $M = XW$.

We know that $\text{rank}(M) \leq \min\{\text{rank}(X), \text{rank}(W)\} = 5$
(by virtue of $\text{rank}(\cdot)$).