

# Machine Learning Theory 2026

## Lecture 3

**Tim van Erven**

Download these slides now from [elo.mastermath.nl](http://elo.mastermath.nl)!

Focus on binary classification:

- ▶ Review
- ▶ Shattering and VC-dimension
- ▶ The Fundamental Theorem of PAC-Learning
- ▶ VC-dimension of Linear Predictors

# (Agnostic) PAC Learning

$\mathcal{H}$  is **agnostically PAC-learnable**:

Exist learner (selecting  $h_S \in \mathcal{H}$ ) that achieves, for finite  $m_{\mathcal{H}}(\epsilon, \delta)$ ,

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \quad \text{with probability } \geq 1 - \delta,$$

whenever  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ ,

for all  $\mathcal{D}, \epsilon, \delta$ .

# (Agnostic) PAC Learning

$\mathcal{H}$  is **agnostically PAC-learnable**:

Exist learner (selecting  $h_S \in \mathcal{H}$ ) that achieves, for finite  $m_{\mathcal{H}}(\epsilon, \delta)$ ,

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \quad \text{with probability } \geq 1 - \delta,$$

whenever  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ ,

for all  $\mathcal{D}, \epsilon, \delta$ .

$\mathcal{H}$  is **PAC-learnable** (only for binary classification):

Same, except only for  $\mathcal{D}$  for which **realizability** holds w.r.t.  $\mathcal{H}$ .

- ▶ Realizability: exists classifier  $h^* \in \mathcal{H}$  that is perfect for  $\mathcal{D}$
- ▶ Implies that  $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$

# What We Know So Far About Learnability

## Theorem (Finite Hypothesis Classes)

*Suppose loss range is  $[0, 1]$ . Finite hypothesis classes  $\mathcal{H}$  are **agnostically PAC-learnable** with ERM.*

# What We Know So Far About Learnability

## Theorem (Finite Hypothesis Classes)

Suppose loss range is  $[0, 1]$ . Finite hypothesis classes  $\mathcal{H}$  are **agnostically PAC-learnable** with ERM.

- ▶ Does not cover e.g. linear predictors

$$\mathcal{H} = \{h_{\mathbf{w},b}(\mathbf{X}) = \text{sign}(b + \langle \mathbf{w}, \mathbf{X} \rangle) \mid \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

# What We Know So Far About Learnability

## Theorem (Finite Hypothesis Classes)

Suppose loss range is  $[0, 1]$ . Finite hypothesis classes  $\mathcal{H}$  are **agnostically PAC-learnable** with ERM.

- ▶ Does not cover e.g. linear predictors

$$\mathcal{H} = \{h_{\mathbf{w}, b}(\mathbf{X}) = \text{sign}(b + \langle \mathbf{w}, \mathbf{X} \rangle) \mid \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

Let  $\mathcal{H}_{\text{all}}$  = all (measurable) functions from  $\mathcal{X}$  to  $\{-1, +1\}$

## Theorem (No-Free-Lunch)

Consider binary classification. For any  $\epsilon < 1/8$ ,  $\delta < 1/7$ , sample size  $m \leq |\mathcal{X}|/2$  is not enough to **PAC-learn**  $\mathcal{H}_{\text{all}}$ :

$$m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > \frac{|\mathcal{X}|}{2}.$$

Rest of today's lecture: focus on **binary classification**!

## Shattering and VC-Dimension

- ▶ VC-dimension of  $\mathcal{H}$  characterizes if  $\mathcal{H}$  is (agnostic) PAC-learnable!

# Consequences of No-Free-Lunch

No-Free-Lunch Theorem has **consequences even if  $\mathcal{H} \neq \mathcal{H}_{\text{all}}$** :

## Definition (Restriction of $\mathcal{H}$ to $\mathcal{C}$ )

For finite  $\mathcal{C} = \{x_1, \dots, x_k\} \subset \mathcal{X}$ , let  $\mathcal{H}_{\mathcal{C}} = \{(h(x_1), \dots, h(x_k)) \mid h \in \mathcal{H}\}$ .

- ▶ Obtain  $\mathcal{H}_{\mathcal{C}}$  by evaluating hypotheses in  $\mathcal{H}$  only on inputs in  $\mathcal{C}$ .

# Consequences of No-Free-Lunch

No-Free-Lunch Theorem has **consequences even if  $\mathcal{H} \neq \mathcal{H}_{\text{all}}$** :

## Definition (Restriction of $\mathcal{H}$ to $\mathcal{C}$ )

For finite  $\mathcal{C} = \{x_1, \dots, x_k\} \subset \mathcal{X}$ , let  $\mathcal{H}_{\mathcal{C}} = \{(h(x_1), \dots, h(x_k)) \mid h \in \mathcal{H}\}$ .

- ▶ Obtain  $\mathcal{H}_{\mathcal{C}}$  by evaluating hypotheses in  $\mathcal{H}$  only on inputs in  $\mathcal{C}$ .

## Corollary (Difficult Subsets of $\mathcal{H}$ )

*If exists finite  $\mathcal{C} \subset \mathcal{X}$  s.t.  $\mathcal{H}_{\mathcal{C}}$  contains all functions from  $\mathcal{C}$  to  $\{-1, +1\}$ , then sample size  $m \leq |\mathcal{C}|/2$  is not enough to PAC-learn  $\mathcal{H}$ .*

**Proof:** Restrict attention to  $\mathcal{D}$  supported on  $\mathcal{C}$  and apply no-free-lunch.

# Shattering

$\mathcal{H}_{\mathcal{C}}$ : evaluate hypotheses in  $\mathcal{H}$  only on inputs in  $\mathcal{C}$

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}_{\mathcal{C}} =$  all functions from  $\mathcal{C}$  to  $\{-1, +1\}$ ,  
i.e.  $|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}$ .

# Shattering

$\mathcal{H}_C$ : evaluate hypotheses in  $\mathcal{H}$  only on inputs in  $C$

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $C \subset \mathcal{X}$  if  $\mathcal{H}_C =$  all functions from  $C$  to  $\{-1, +1\}$ , i.e.  $|\mathcal{H}_C| = 2^{|C|}$ .

## Example (Axis-aligned Rectangles)

$\mathcal{H}_{\text{rec}}^2 = \{h_{(a_1, b_1, a_2, b_2)} \mid a_1 \leq b_1, a_2 \leq b_2\}$ , where

$$h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) = \begin{cases} +1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ -1 & \text{otherwise} \end{cases}$$

Exists a  $C$  of size 4 that is shattered by  $\mathcal{H}_{\text{rec}}^2$ , but not of size 5.

# Proof (Handwritten)

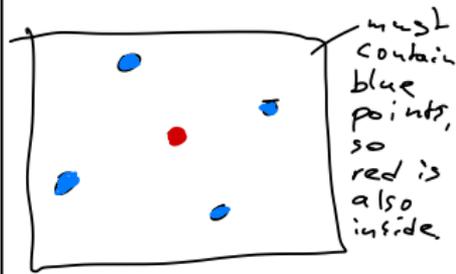
Need to show:

1. Exists  $\mathcal{C}$  of size 4 that is shattered
2. No  $\mathcal{C}$  of size 5 is shattered

can shatter:



Cannot Shatter:



Proof not size 5: if left-most, right-most, top-most and bottom-most point +1, then remaining point also +1

# VC-Dimension

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}_{\mathcal{C}} =$  all functions.

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

# VC-Dimension

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}_{\mathcal{C}} = \text{all functions}$ .

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) = \text{maximum size}$  of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

## Corollary (Difficult Subsets of $\mathcal{H}$ )

*If exists finite  $\mathcal{C} \subset \mathcal{X}$  such that  $\mathcal{H}$  shatters  $\mathcal{C}$ , then sample size  $m \leq |\mathcal{C}|/2$  is not enough to PAC-learn  $\mathcal{H}$ .*

# VC-Dimension

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}_{\mathcal{C}} =$  all functions.

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

## Corollary (Difficult Subsets of $\mathcal{H}$ )

*If exists finite  $\mathcal{C} \subset \mathcal{X}$  such that  $\mathcal{H}$  shatters  $\mathcal{C}$ , then sample size  $m \leq |\mathcal{C}|/2$  is not enough to PAC-learn  $\mathcal{H}$ .*

- ▶ Sample size  $m \leq \text{VCdim}(\mathcal{H})/2$  is not enough to PAC-learn  $\mathcal{H}$ .
- ▶ **If  $\text{VCdim}(\mathcal{H}) = \infty$ , then  $\mathcal{H}$  is not PAC-learnable.**

# VC-Dimension: Examples

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

## Example (Axis-Aligned Rectangles)

$$\text{VCdim}(\mathcal{H}_{\text{rect}}^2) = 4$$

# VC-Dimension: Examples

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

### Example (Axis-Aligned Rectangles)

$$\text{VCdim}(\mathcal{H}_{\text{rect}}^2) = 4$$

### Example (Finite Hypothesis Classes)

$$\text{VCdim}(\mathcal{H}) \leq \dots?$$

# VC-Dimension: Examples

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

### Example (Axis-Aligned Rectangles)

$$\text{VCdim}(\mathcal{H}_{\text{rect}}^2) = 4$$

### Example (Finite Hypothesis Classes)

$$\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

# VC-Dimension: Examples

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

### Example (Axis-Aligned Rectangles)

$$\text{VCdim}(\mathcal{H}_{\text{rect}}^2) = 4$$

### Example (Finite Hypothesis Classes)

$$\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

### Example (Step Functions)

$$\mathcal{H} = \{h_a \mid a \in \mathbb{R}\} \text{ where } h_a(x) = \begin{cases} -1 & \text{if } x \leq a \\ +1 & \text{if } x > a \end{cases}$$

$$\text{VCdim}(\mathcal{H}) = \dots?$$

# VC-Dimension: Examples

## Definition (Vapnik-Chervonenkis (VC) Dimension)

- ▶  $\text{VCdim}(\mathcal{H}) =$  **maximum size** of finite set  $\mathcal{C} \subset \mathcal{X}$  **shattered** by  $\mathcal{H}$
- ▶  $\text{VCdim}(\mathcal{H}) = \infty$  if there is no maximum

### Example (Axis-Aligned Rectangles)

$$\text{VCdim}(\mathcal{H}_{\text{rect}}^2) = 4$$

### Example (Finite Hypothesis Classes)

$$\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

### Example (Step Functions)

$$\mathcal{H} = \{h_a \mid a \in \mathbb{R}\} \text{ where } h_a(x) = \begin{cases} -1 & \text{if } x \leq a \\ +1 & \text{if } x > a \end{cases}$$

$$\text{VCdim}(\mathcal{H}) = 1$$

# The Fundamental Theorem of PAC-Learning

## Theorem

*For binary classification, the following are equivalent:*

1.  $\mathcal{H}$  has the **uniform convergence** property.
2. Any **ERM** rule is a successful agnostic PAC-learner for  $\mathcal{H}$ .
3.  $\mathcal{H}$  is **agnostic PAC-learnable**.
4.  $\mathcal{H}$  is **PAC-learnable**.
5. Any **ERM** rule is a successful PAC-learner for  $\mathcal{H}$ .
6.  $\mathcal{H}$  has **finite VC-dimension**.

# The Fundamental Theorem of PAC-Learning

## Theorem

*For binary classification, the following are equivalent:*

1.  $\mathcal{H}$  has the **uniform convergence** property.
2. Any **ERM** rule is a successful agnostic PAC-learner for  $\mathcal{H}$ .
3.  $\mathcal{H}$  is **agnostic PAC-learnable**.
4.  $\mathcal{H}$  is **PAC-learnable**.
5. Any **ERM** rule is a successful PAC-learner for  $\mathcal{H}$ .
6.  $\mathcal{H}$  has **finite VC-dimension**.

## Main Points:

- ▶ PAC-learnability and agnostic PAC-learnability are equivalent
- ▶ VC-dimension characterizes both!

# The Fundamental Theorem of PAC-Learning

## Theorem

*For binary classification, the following are equivalent:*

1.  $\mathcal{H}$  has the **uniform convergence** property.
2. Any **ERM** rule is a successful agnostic PAC-learner for  $\mathcal{H}$ .
3.  $\mathcal{H}$  is **agnostic PAC-learnable**.
4.  $\mathcal{H}$  is **PAC-learnable**.
5. Any **ERM** rule is a successful PAC-learner for  $\mathcal{H}$ .
6.  $\mathcal{H}$  has **finite VC-dimension**.

## Main Points:

- ▶ PAC-learnability and agnostic PAC-learnability are equivalent
- ▶ VC-dimension characterizes both!

## Other Observations:

- ▶ Finite VC-dimension is equivalent to uniform convergence
- ▶ ERM always works for (agnostic) PAC-learning

# VC-Dimension of Linear Predictors (Halfspaces)

$$\mathcal{H}_{\text{lin}}^d = \{h_{\mathbf{w},b} \mid \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\},$$

where

$$h_{\mathbf{w},b}(\mathbf{X}) = \begin{cases} +1 & \text{if } b + \langle \mathbf{w}, \mathbf{X} \rangle \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

for  $\mathbf{X} \in \mathbb{R}^d$

## Theorem

$$\text{VCdim}(\mathcal{H}_{\text{lin}}^d) = d + 1$$

- ▶ For many (but not all!) hypothesis classes VC-dimension equals number of parameters

$x \in \mathbb{R}^d$

VC-dim for  
halfspaces

$$h_{w,b}(x) = \begin{cases} +1 & \text{if } b + \langle w, x \rangle \geq 0 \\ -1 & \text{o.w.} \end{cases}$$

$$\mathcal{H} = \{ h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

I. VC-dim  $\geq d+1$

To show: exists  $C \subset \mathbb{R}^d$  of size  $|C| = d+1$   
that is shattered by  $\mathcal{H}$ .

Take  $C = \{0, e_1, \dots, e_d\}$ .

Let  $y_0, y_1, \dots, y_d \in \{-1, +1\}$   
be arbitrary.

Now take  $b = \frac{y_0}{2}$ ,  $w = (y_1, \dots, y_d)$

$$\begin{aligned} \text{Then } b + \langle w, 0 \rangle &= \frac{y_0}{2} \\ b + \langle w, e_i \rangle &= \frac{y_0}{2} + y_i \end{aligned} \left. \vphantom{\begin{aligned} \text{Then } b + \langle w, 0 \rangle &= \frac{y_0}{2} \\ b + \langle w, e_i \rangle &= \frac{y_0}{2} + y_i \end{aligned}} \right\} \begin{array}{l} \text{correct} \\ \text{sign.} \end{array}$$

II. VC-dim  $< d+2$ :

To show: If  $C \subset \mathbb{R}^d$  of size  $|C| = d+2$ ,  
then  $C$  is not shattered by  $\mathcal{H}$ .

exists labels  $y_1, \dots, y_{d+2}$  that  
cannot be realized by any  
 $hw, b$ .

Let  $C = \{x_1, \dots, x_{d+2}\}$  be arbitrary.  
 To choose:  $y_1, \dots, y_{d+2}$

$$C_{-1} = \{x_i \in C : y_i = -1\}$$

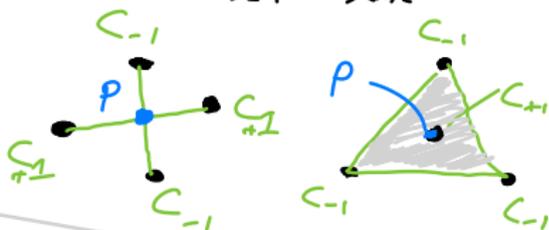
$$C_{+1} = \{x_i \in C : y_i = +1\}$$

$C_j$  classified correctly  
 $\Downarrow$  (linearity)

all points in convex hull of  $C_j$   
 assigned class  $j$

$\Downarrow$   
 contradiction for  $p$   
 in intersection of  
 convex hulls of  $C_{-1}$  and  $C_{+1}$ .

Can we always find  $C_{-1}$  and  $C_{+1}$  for which  
 convex hulls intersect? YES!



Radon's Theorem: Any  $C = \{x_1, \dots, x_{d+2}\} \subset \mathbb{R}^d$   
 can be partitioned into two (disjoint)  
 subsets  $C_{-1}$  and  $C_{+1}$  whose convex hulls  
 intersect.

Proof: Let  $a_1, \dots, a_{d+2}$  (not all zero) be a solution to

$$\sum_{i=1}^{d+2} a_i x_i = 0, \quad \sum_{i=1}^{d+2} a_i = 0$$

$\uparrow$   $d$  constraints       $\leftarrow$  1 constraint

$$\text{Let } C_{-1} = \{x_i : a_i < 0\}$$

$$C_{+1} = \{x_i : a_i \geq 0\}$$

Then both convex hulls contain

$$p = \sum_{x_i \in C_{+1}} \frac{a_i}{A} x_i = \sum_{x_j \in C_{-1}} \frac{-a_j}{A} x_j$$

where

$$A = \sum_{x_i \in C_{+1}} a_i = \sum_{x_j \in C_{-1}} -a_j \quad \square$$