

# Machine Learning Theory 2024

## Lecture 4

**Tim van Erven**

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

Focus on binary classification:

- ▶ Review
- ▶ Fundamental theorem: quantitative version
- ▶ VC-dimension controls growth function

# 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**.

VC-dimension **characterizes** (agnostic) PAC-learnability  
and uniform convergence!

► Still to prove: 6  $\rightarrow$  1

# Uniform Convergence

$\mathcal{H}$  has the **uniform convergence** property:

For finite  $m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta)$ ,

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \quad \text{with probability } \geq 1 - \delta,$$

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

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

# Shattering and VC-Dimension

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

For finite  $\mathcal{C} = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset \mathcal{X}$ , let  $\mathcal{H}_{\mathcal{C}} = \{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_k)) \mid h \in \mathcal{H}\}$ .

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

# Shattering and VC-Dimension

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

For finite  $\mathcal{C} = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset \mathcal{X}$ , let  $\mathcal{H}_{\mathcal{C}} = \{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_k)) \mid h \in \mathcal{H}\}$ .

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

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}$  can classify the elements of  $\mathcal{C}$  in all possible ways, i.e.  $|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}$ .

# Shattering and VC-Dimension

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

For finite  $\mathcal{C} = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset \mathcal{X}$ , let  $\mathcal{H}_{\mathcal{C}} = \{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_k)) \mid h \in \mathcal{H}\}$ .

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

## Definition (Shattering)

$\mathcal{H}$  **shatters** a finite set  $\mathcal{C} \subset \mathcal{X}$  if  $\mathcal{H}$  can classify the elements of  $\mathcal{C}$  in all possible ways, i.e.  $|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}$ .

## 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

# Fundamental Theorem: Quantitative Version

# Fundamental Theorem: Quantitative Version

Does the VC-dimension also characterize the **sample complexity** of PAC-learning? Yes!



# Fundamental Theorem: Quantitative Version

Does the VC-dimension also characterize the **sample complexity** of PAC-learning? Yes!

## Theorem

Consider binary classification. Suppose  $\text{VCdim}(\mathcal{H}) = v < \infty$ . Then there exist absolute constants  $C_1, C_2 > 0$  such that

1. Uniform convergence:

$$C_1 \frac{v + \ln(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{v + \ln(1/\delta)}{\epsilon^2}$$

2. Agnostic PAC-learning:

$$C_1 \frac{v + \ln(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{v + \ln(1/\delta)}{\epsilon^2}$$

3. PAC-learning:

$$C_1 \frac{v + \ln(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{v \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}.$$

# Uniform Convergence Upper Bound

Upper bound from previous slide that we want to prove:

## Theorem

Consider binary classification. Suppose  $\text{VCdim}(\mathcal{H}) \leq v < \infty$ . Then there exists an absolute constant  $C > 0$  such that

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \quad \text{with probability} \geq 1 - \delta,$$

whenever

$$m \geq C \frac{v + \ln(1/\delta)}{\epsilon^2}.$$

# Uniform Convergence Upper Bound

Upper bound from previous slide that we want to prove:

## Theorem

Consider binary classification. Suppose  $\text{VCdim}(\mathcal{H}) \leq v < \infty$ . Then there exists an absolute constant  $C > 0$  such that

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \quad \text{with probability} \geq 1 - \delta,$$

whenever

$$m \geq C \frac{v \ln(1/\epsilon) + \ln(1/\delta) + 1}{\epsilon^2}.$$

- ▶ Extra factor  $\ln(1/\epsilon)$  is only logarithmic
- ▶ It could be avoided with a more involved argument (using a technique called chaining)

# Uniform Convergence Upper Bound

Upper bound from previous slide that we want to prove:

## Theorem

Consider binary classification. Suppose  $\text{VCdim}(\mathcal{H}) \leq v < \infty$ . Then there exists an absolute constant  $C > 0$  such that

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \quad \text{with probability} \geq 1 - \delta,$$

whenever

$$m \geq C \frac{v \ln(1/\epsilon) + \ln(1/\delta) + 1}{\epsilon^2}.$$

- ▶ Extra factor  $\ln(1/\epsilon)$  is only logarithmic
- ▶ It could be avoided with a more involved argument (using a technique called chaining)
- ▶  $v = 0 \Rightarrow |\mathcal{H}| = 1$  is trivial, so can assume  $v > 0$  w.l.o.g.

# Proof Approach

Will define **growth function**  $\tau_{\mathcal{H}}(m)$ . Then

**Part I: Growth function controls uniform convergence:**

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leq c \sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c \sqrt{\frac{\ln(2/\delta)}{m}} \quad \text{with probability } \geq 1 - \delta,$$

**Part II: VC-dimension controls growth function:**

$$\ln \tau_{\mathcal{H}}(m) \leq v \ln \left( \frac{em}{v} \right) \quad \text{for } m > v.$$

- ▶ Finish: combine Parts I and II, and find lower bound on  $m$  s.t.  
 $\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leq \epsilon.$

**Proof Part II:  
VC-dimension Controls Growth Function**

# Growth Function

- ▶ **Finite**  $\mathcal{H}$  have the uniform convergence property.
- ▶ How do we measure the **size of infinite**  $\mathcal{H}$ ?

**Growth function:** effective size of  $\mathcal{H}$  at sample size  $m$ :

$$\tau_{\mathcal{H}}(m) = \max_{\mathcal{C} \subset \mathcal{X}: |\mathcal{C}|=m} |\mathcal{H}_{\mathcal{C}}|$$

- ▶ Interpretation: How many truly different hypotheses are there when we only observe  $m$  inputs  $\mathcal{C} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ ?
- ▶ If  $\mathcal{H}$  is finite, then  $\tau_{\mathcal{H}}(m) \leq |\mathcal{H}|$

# Sauer's Lemma

Growth function:  $\tau_{\mathcal{H}}(m) = \max_{|C|=m} |\mathcal{H}_C|$

## Lemma (Sauer-Shelah-Perles)

Suppose  $\text{VCdim}(\mathcal{H}) \leq v < \infty$ . Then the growth function is bounded by

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^v \binom{m}{i} \leq \begin{cases} 2^m & \text{if } m \leq v \\ \left(\frac{em}{v}\right)^v & \text{if } m > v. \end{cases}$$

- ▶ VC-dimension  $v$  determines switch from exponential to polynomial growth in  $m$ .
- ▶ Case  $m > v$  is what we need to show for Part II.



Sauer's Lemma For all  $\mathcal{H}$  and all  $m$

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^V \binom{m}{i},$$

$$\text{where } \tau_{\mathcal{H}}(m) = \max_{|C|=m} |\mathcal{H}_C|$$

Proof:

Will show: For any  $\mathcal{H}$  and  $C$  of size  $|C|=m$

$$\begin{aligned} |\mathcal{H}_C| &\stackrel{(1)}{\leq} \left| \{ B \subseteq C : \mathcal{H} \text{ shatters } B \} \right| \\ &\stackrel{(2)}{\leq} \sum_{i=0}^V \binom{m}{i} \end{aligned}$$

(2):  $\mathcal{H}$  shatters  $B \Rightarrow |B| \leq V$

nr of sets  $B \subseteq C$  with  $|B|=i$  is  $\binom{m}{i}$

summing over  $i=0, \dots, V$  implies (2).

(1)  $|H_C| \leq |\{B \subseteq C : H \text{ shatters } B\}|$  for any  $|C| = m$  and any  $H$

By induction in  $m$ :

$m = 1$ :

$|H_C| = 1 \Rightarrow C$  is not shattered by  $H$   
so only  $B = \emptyset$  is shattered by  $H$

$\Rightarrow$  r.h.s is 1

$|H_C| = 2 \Rightarrow C$  is shattered and  $B = \emptyset$  is shattered

$\Rightarrow$  r.h.s. = 2.

$m \geq 2$ : Suppose (1) holds for all  $m < k$ .

To show: (1) holds for  $m = k$ .

Let  $C = \{x_1, \dots, x_k\}$  be arbitrary.

Want to apply inductive assumption, so  
define

$$C' = \{x_2, \dots, x_k\}$$

Let  $Y_0 = H_{C'} = \{ (y_2, \dots, y_k) \mid \exists y_1 \text{ s.t.}$

$$(y_1, y_2, \dots, y_k) \in H_C \}$$

Then  $|Y_0| \leq |H_C|$  under counts  $|H_C|$ , because

$y_1 = -1$  and  $y_1 = +1$  may both satisfy

So let's count how often this happens:

$$Y_1 = \{ (y_2, \dots, y_k) \mid \forall y_1 \text{ s.t. } (y_1, y_2, \dots, y_k) \in H_C \}$$

Thus

$$|H_C| = |Y_0| + |Y_1|$$

Will show:

$$i) |Y_0| \leq |\{B \subseteq C : x_1 \notin B, \mathcal{H} \text{ shatters } B\}|$$

$$ii) |Y_1| \leq |\{B \subseteq C : x_1 \in B, \mathcal{H} \text{ shatters } B\}|$$

So together:

$$|H_C| = |Y_0| + |Y_1| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|,$$

which is to be shown.

i) Recall that

$$C' = \{x_2, \dots, x_k\}, \quad Y_0 = H_{C'}$$

(induction)

$$|Y_0| = |H_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}|$$

$$= |\{B \subseteq C : x_1 \notin B, \mathcal{H} \text{ shatters } B\}|$$

ii)  $|Y_1| \leq |\{B \subseteq C : x_1 \in B, H \text{ shatters } B\}|$

Define  $H' = \{h \in H \mid \exists h' \in H \text{ s.t. } h \text{ and } h' \text{ agree on } C'$

$h'(x_i) = h(x_i) \text{ for } i = 2, \dots, k$

but  $h'(x_1) \neq h(x_1)\}$

Observe:

\*  $H'$  shatters  $B \subseteq C' \iff H'$  shatters  $B \cup \{x_1\}$

\*  $|Y_1| = |H'_C|$  (induction)

$|Y_1| = |H'_C| \leq |\{B \subseteq C' : H' \text{ shatters } B\}|$

$= |\{B \subseteq C' : H' \text{ shatters } B \cup \{x_1\}\}|$

$= |\{B \subseteq C : x_1 \in B, H' \text{ shatters } B\}|$

$\leq |\{B \subseteq C : x_1 \in B, H \text{ shatters } B\}|$

□

# The Final Inequality (Handwritten)

## Lemma

$$\sum_{i=0}^v \binom{m}{i} \leq \begin{cases} 2^m & \text{if } m \leq v \\ \left(\frac{em}{v}\right)^v & \text{if } m > v \end{cases}$$

**Proof:** Will use binomial theorem:  $(x + y)^m = \sum_{i=0}^m \binom{m}{i} x^i y^{m-i}$ .

$m \leq v$ :  $\binom{m}{i} = 0$  for  $i > m$ , so  $\sum_{i=0}^v \binom{m}{i} = \sum_{i=0}^m \binom{m}{i}$ . Then apply binomial theorem with  $x = y = 1$ .

$m > v$ : [Simpler proof from Anthony and Bartlett, *Neural Network Learning: Theoretical Foundations*, 1999]

$$\begin{aligned} \sum_{i=0}^v \binom{m}{i} &\leq \left(\frac{m}{v}\right)^v \sum_{i=0}^v \binom{m}{i} \left(\frac{v}{m}\right)^i \leq \left(\frac{m}{v}\right)^v \sum_{i=0}^m \binom{m}{i} \left(\frac{v}{m}\right)^i \\ &= \left(\frac{m}{v}\right)^v \left(1 + \frac{v}{m}\right)^m \leq \left(\frac{m}{v}\right)^v (e^{v/m})^m = \left(\frac{em}{v}\right)^v \end{aligned}$$

(First equality follows from binomial theorem with  $x = 1, y = \frac{v}{m}$ .)