

# Machine Learning Theory 2022

## Lecture 2

**Tim van Erven**

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

- ▶ Review
- ▶ (Agnostic) PAC learning
- ▶ Agnostic PAC-learnability for finite classes
- ▶ Uniform convergence
- ▶ No-Free-Lunch Theorem (without proof)



homework roulette  
in the break

# Formal Setup Review

$$S = \begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_m \\ \mathbf{X}_m \end{pmatrix} \sim \mathcal{D}$$

**Risk:**  $L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h, \mathbf{X}, Y)]$  for  $(\mathbf{X}, Y) \sim \mathcal{D}$

**Empirical Risk:**  $L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, \mathbf{X}_i, Y_i)$  for  $(\mathbf{X}_i, Y_i)$  in  $S$

**Classification** (0/1-loss counts mistakes):

$$\ell(h, \mathbf{X}, Y) = \mathbf{1}\{h(\mathbf{X}) \neq Y\} = \begin{cases} 0 & \text{if } h(\mathbf{X}) = Y \\ 1 & \text{if } h(\mathbf{X}) \neq Y \end{cases}$$

**Regression** (Squared Error):

$$\ell(h, \mathbf{X}, Y) = (Y - h(\mathbf{X}))^2$$

# No Overfitting for (Multiclass) Classification

**Realizability assumption:** Exists perfect predictor  $h^* \in \mathcal{H}$ , i.e.  
 $\Pr(h^*(\mathbf{X}) = Y) = 1$ .

## Theorem (First Example of PAC-Learning)

Assume  $\mathcal{H}$  is **finite**, **realizability** holds. Choose any  $\delta \in (0, 1)$ ,  $\epsilon > 0$ .  
Then, for all  $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$ , ERM over  $\mathcal{H}$  guarantees

$$L_{\mathcal{D}}(h_S) \leq \epsilon \quad \text{with probability } \geq 1 - \delta.$$

NB Lower bound on  $m$  does not depend on  $\mathcal{D}$  or on  $h^*$ !

**PAC learning:** probably approximately correct

## (Agnostic) PAC Learning

- ▶ PAC learning (always for binary classification)
- ▶ Agnostic PAC learning for binary classification
- ▶ Agnostic PAC learning in general
- ▶ Improper Agnostic PAC learning in general

# Definition: PAC Learning (Binary Classification)

A hypothesis class  $\mathcal{H}$  is **PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$   
such that for all
- ▶ distributions  $\mathcal{D}$  for which **realizability** holds w.r.t.  $\mathcal{H}$
- ▶ and all  $\epsilon, \delta \in (0, 1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \quad \text{with probability } \geq 1 - \delta,$$
$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

# Definition: PAC Learning (Binary Classification)

A hypothesis class  $\mathcal{H}$  is **PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$

such that for all

- ▶ distributions  $\mathcal{D}$  for which **realizability** holds w.r.t.  $\mathcal{H}$
- ▶ and all  $\epsilon, \delta \in (0, 1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \quad \text{with probability } \geq 1 - \delta,$$
$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

## Sample complexity:

The function  $m_{\mathcal{H}}$  such that  $m_{\mathcal{H}}(\epsilon, \delta)$  is smallest possible for all  $\epsilon, \delta$

# No Overfitting for (Multiclass) Classification

## Theorem (First Example of PAC-Learning)

Assume  $\mathcal{H}$  is **finite**, **realizability** holds. Choose any  $\delta \in (0, 1)$ ,  $\epsilon > 0$ . Then, for all  $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$ , ERM over  $\mathcal{H}$  guarantees

$$L_{\mathcal{D}}(h_S) \leq \epsilon$$

with probability at least  $1 - \delta$ .

For binary classification this is equivalent to:

## Theorem

Every **finite** hypothesis class  $\mathcal{H}$  is PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

and learning algorithm ERM.

# Definition: PAC Learning (Binary Classification)

A hypothesis class  $\mathcal{H}$  is **PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$   
such that for all
- ▶ distributions  $\mathcal{D}$  for which **realizability** holds w.r.t.  $\mathcal{H}$
- ▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Definition: Agnostic PAC Learning (Binary Classification)

A hypothesis class  $\mathcal{H}$  is **Agnostic PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$

such that for all

- ▶ distributions  $\mathcal{D}$  ~~for which realizability holds w.r.t.  $\mathcal{H}$~~
- ▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Definition: Agnostic PAC Learning (~~Binary Classification~~) (In General)

A hypothesis class  $\mathcal{H}$  is **Agnostic PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$

such that for all

- ▶ distributions  $\mathcal{D}$  ~~for which realizability holds w.r.t.  $\mathcal{H}$~~
- ▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Definition: Agnostic PAC Learning (In General)

A hypothesis class  $\mathcal{H}$  is **Agnostic PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$

such that for all

- ▶ distributions  $\mathcal{D}$
- ▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Definition: Improper Agnostic PAC Learning (In General)

A hypothesis class  $\mathcal{H}$  is **Improperly Agnostic PAC-learnable** if there exist

- ▶ a **function**  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$
- ▶ and **learning algorithm** that outputs  $h_S \in \mathcal{H}$

such that for all

- ▶ distributions  $\mathcal{D}$
- ▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Agnostic PAC-Learnability for Finite Classes via Uniform Convergence

# Agnostic PAC-Learnability for Finite Classes

## Theorem (Bounded Loss, Finite Class)

Suppose  $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Then every **finite** hypothesis class  $\mathcal{H}$  is **agnostically** PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

# Agnostic PAC-Learnability for Finite Classes

## Theorem (Bounded Loss, Finite Class)

Suppose  $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Then every **finite** hypothesis class  $\mathcal{H}$  is **agnostically** PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

- ▶ Worse dependence on  $\epsilon$  compared to  $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$  for PAC-learnability

# Agnostic PAC-Learnability for Finite Classes

## Theorem (Bounded Loss, Finite Class)

Suppose  $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Then every **finite** hypothesis class  $\mathcal{H}$  is **agnostically** PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

- ▶ Worse dependence on  $\epsilon$  compared to  $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$  for PAC-learnability
- ▶ Losses with different range  $[a, b]$  can be reduced to  $[0, 1]$  range by subtracting  $a$  and dividing by  $(b - a)$ .

# Technical Tool: Uniform Convergence

A hypothesis class  $\mathcal{H}$  has the **uniform convergence** property if there exists

▶ a **function**  $m_{\mathcal{H}}^{\text{UC}} : (0, 1)^2 \rightarrow \mathbb{N}$

such that for all

▶ distributions  $\mathcal{D}$

▶ and all  $\epsilon, \delta \in (0, 1)$

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

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

# Uniform Convergence $\rightarrow$ Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

## Lemma

If  $\mathcal{H}$  has the **uniform convergence property**, then it is **agnostic PAC-learnable** with

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

# Uniform Convergence $\rightarrow$ Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

## Lemma

If  $\mathcal{H}$  has the **uniform convergence property**, then it is **agnostic PAC-learnable** with

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

- ▶ We will prove uniform convergence for finite  $\mathcal{H}$  and loss range  $[0, 1]$
- ▶ Then the desired agnostic PAC-learnability follows

## Proof (Handwritten)

To show, for  $h_S$  ERM hypothesis:

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \quad \text{with probability } \geq 1 - \delta,$$
$$\text{whenever } m \geq m_{\mathcal{H}}^{\text{UC}}\left(\frac{\epsilon}{2}, \delta\right).$$

Assuming uniform convergence, applied for  $\epsilon/2$ :

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2} \quad \text{with probability } \geq 1 - \delta,$$
$$\text{whenever } m \geq m_{\mathcal{H}}^{\text{UC}}\left(\frac{\epsilon}{2}, \delta\right).$$

Proof: On the event that  $|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2}$  for all  $h \in \mathcal{H}$ , we have for all  $h' \in \mathcal{H}$

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h') + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h') + \epsilon.$$

Then take the infimum over  $h'$ .

# Uniform Convergence for Finite Classes

## Lemma (Bounded Loss, Finite Class)

Suppose  $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Then every **finite** hypothesis class  $\mathcal{H}$  has the **uniform convergence property** with

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil.$$

To show:

$$\Pr \left( \sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leq \epsilon \right) \geq 1 - \delta$$

$$\text{whenever} \quad m \geq \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

## Proof (Handwritten)

$$\Pr \left( \sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leq \epsilon \right) \stackrel{?}{\geq} 1 - \delta$$

$$\Pr \left( \sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| > \epsilon \right) \stackrel{?}{\leq} \delta$$

$$\Pr \left( \text{exists } h \in \mathcal{H} : |L_S(h) - L_D(h)| > \epsilon \right) \stackrel{?}{\leq} \delta$$

Part I (union bound):

$$\Pr \left( \text{exists } h \in \mathcal{H} : |L_S(h) - L_D(h)| > \epsilon \right) \leq \sum_{h \in \mathcal{H}} \Pr \left( |L_S(h) - L_D(h)| > \epsilon \right)$$

Part II (Hoeffding's inequality): Let  $Z_i = \ell(h, \mathbf{X}_i, Y_i) \in [0, 1]$ .

$$\Pr \left( |L_S(h) - L_D(h)| > \epsilon \right) = \Pr \left( \left| \frac{1}{m} \sum_{i=1}^m Z_i - \mathbb{E}[Z] \right| > \epsilon \right) \stackrel{\text{Hoeffding}}{\leq} 2e^{-2m\epsilon^2}$$

## Proof Continued (Handwritten)

Part I+II:

$$\begin{aligned}\Pr(\text{exists } h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) &\leq \sum_{h \in \mathcal{H}} \Pr(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \\ &\leq |\mathcal{H}| 2e^{-2m\epsilon^2} \stackrel{?}{\leq} \delta\end{aligned}$$

$$\text{Yes, for } m \geq \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2}$$

# Putting Everything Together

## Theorem (Bounded Loss, Finite Class)

Suppose  $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Then every **finite** hypothesis class  $\mathcal{H}$  has the uniform convergence property with

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil,$$

and is therefore **agnostically** PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

# No-Free-Lunch Theorem

# No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

## Theorem (No-Free-Lunch)

Let  $A$  be **any learning algorithm** for binary classification. If  $m \leq |\mathcal{X}|/2$ , then there exists a distribution  $\mathcal{D}$  such that

1. There exists a perfect predictor  $f$  with  $L_{\mathcal{D}}(f) = 0$ .
2.  $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$  for  $S \sim \mathcal{D}^m$ .

# No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

## Theorem (No-Free-Lunch)

Let  $A$  be **any learning algorithm** for binary classification. If  $m \leq |\mathcal{X}|/2$ , then there exists a distribution  $\mathcal{D}$  such that

1. There exists a perfect predictor  $f$  with  $L_{\mathcal{D}}(f) = 0$ .
2.  $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$  for  $S \sim \mathcal{D}^m$ .

Interpretation:

- ▶  $\mathcal{H}_{\text{all}}$  = all functions from  $\mathcal{X}$  to  $\{-1, +1\}$
- ▶  $m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$  for any  $\epsilon < 1/8$ ,  $\delta < 1/7$

# No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

## Theorem (No-Free-Lunch)

Let  $A$  be **any learning algorithm** for binary classification. If  $m \leq |\mathcal{X}|/2$ , then there exists a distribution  $\mathcal{D}$  such that

1. There exists a perfect predictor  $f$  with  $L_{\mathcal{D}}(f) = 0$ .
2.  $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$  for  $S \sim \mathcal{D}^m$ .

Interpretation:

- ▶  $\mathcal{H}_{\text{all}}$  = all functions from  $\mathcal{X}$  to  $\{-1, +1\}$
- ▶  $m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$  for any  $\epsilon < 1/8$ ,  $\delta < 1/7$

## Corollary

Suppose  $|\mathcal{X}| = \infty$ . Then  $\mathcal{H}_{\text{all}}$  is not PAC-learnable.

# No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

## Theorem (No-Free-Lunch)

Let  $A$  be **any learning algorithm** for binary classification. If  $m \leq |\mathcal{X}|/2$ , then there exists a distribution  $\mathcal{D}$  such that

1. There exists a perfect predictor  $f$  with  $L_{\mathcal{D}}(f) = 0$ .
2.  $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$  for  $S \sim \mathcal{D}^m$ .

## Proof Intuition:

- ▶ Suppose  $\mathcal{D}$  is uniform on  $2m$  points in  $\mathcal{X}$ , and  $Y = f(X)$  for some unknown function  $f$ .
- ▶ From  $S$  we only know  $f(X)$  for  $m$  observed points.
- ▶ Without any assumptions about  $f$ , learner cannot do better than random guessing on  $m$  unobserved points.