

# Învățare Automată (Machine Learning)



Bogdan Alexe,

[bogdan.alexe@fmi.unibuc.ro](mailto:bogdan.alexe@fmi.unibuc.ro)

Master Informatică, anul I, 2018-2019, cursul 4

# Recap - PAC learnability of a class $\mathcal{H}$

A hypothesis class  $\mathcal{H}$  is called **PAC learnable** if there exists a function  $m_{\mathcal{H}}: (0,1)^2 \rightarrow \mathbb{N}$  and a learning algorithm  $A$  with the following property:

- for every  $\varepsilon > 0$  (*accuracy*  $\rightarrow$  “approximately correct”)
- for every  $\delta > 0$  (*confidence*  $\rightarrow$  “probably”)
- for every labeling  $f \in \mathcal{H}$  (*realizability case*)
- for every distribution  $\mathcal{D}$  over  $\mathcal{X}$

when we run the learning algorithm  $A$  on a training set  $S$ , consisting of  $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$  examples sampled i.i.d. from  $\mathcal{D}$  and labeled by  $f$  the algorithm  $A$  returns a hypothesis  $h_S \in \mathcal{H}$  such that, with probability at least  $1-\delta$  (over the choice of examples),  $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ .

$$P_{S \sim D^m}(L_{f,D}(h_S) \leq \varepsilon) \geq 1 - \delta \Leftrightarrow P_{S \sim D^m}(L_{f,D}(h_S) > \varepsilon) < \delta$$

# Recap – agnostic PAC learnability of a class $\mathcal{H}$

A hypothesis class  $\mathcal{H}$  is called *agnostic PAC learnable* if there exists a function  $m_{\mathcal{H}}: (0,1)^2 \rightarrow \mathbb{N}$  and a learning algorithm  $A$  with the following property:

- for every  $\varepsilon > 0$  (*accuracy*  $\rightarrow$  “approximately correct”)
- for every  $\delta > 0$  (*confidence*  $\rightarrow$  “probably”)
- for every distribution  $\mathcal{D}$  over  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

when we run the learning algorithm  $A$  on a training set  $S$ , consisting of  $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$  examples sampled i.i.d. from  $\mathcal{D}$  the algorithm  $A$  returns a hypothesis  $A(S)$  from  $\mathcal{H}$  such that, with probability at least  $1-\delta$  (over the choice of examples) it holds that:

$$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

- if the realizability assumption holds, agnostic PAC = PAC
- in agnostic PAC learning, a learner can still declare success if its error is not much larger than the best error achievable by a predictor from the class  $\mathcal{H}$ .

# PAC vs. Agnostic PAC learning

	PAC	Agnostic PAC
Distribution	$\mathcal{D}$ over $\mathcal{X}$	$\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
Truth	$f \in \mathcal{H}$	not in class or doesn't exist
Risk	$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x : h(x) \neq f(x)\})$	$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x, y) : h(x) \neq y\})$
Training set	$(x_1, \dots, x_m) \sim \mathcal{D}^m$ $\forall i, y_i = f(x_i)$	$((x_1, y_1), \dots, (x_m, y_m)) \sim \mathcal{D}^m$
Goal	$L_{\mathcal{D},f}(A(S)) \leq \epsilon$	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

# A note of Caution

The fact that  $\mathcal{H}$  is agnostically PAC learnable using the ERM paradigm doesn't mean that the result is any good.

It only means that you can be reasonable sure the ERM paradigm gives you a result that is close to the optimal result.

If the optimal result is bad (because, for example, the hypothesis class  $\mathcal{H}$  fits the data really badly) the ERM paradigm will also give you a bad result.

PAC doesn't tell you that your hypothesis class  $\mathcal{H}$  fits the data well, it only tells you that, if it fits well, the ERM paradigm will probably give you a reasonable good hypothesis.

# Today's lecture: Overview

- The No-Free-Lunch theorem
- The Bias-Complexity tradeoff
- Uniform Convergence for Agnostic PAC learnability

# The No-Free-Lunch theorem

# Prior knowledge

*Empirical Risk Minimization* (ERM) = learning paradigm that returns a predictor  $h$  that minimizes  $L_S(h)$ ,  $S$  –training sequence of examples sampled i.i.d. from an unknown distribution  $\mathcal{D}$  over a domain  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

- ERM might overfit if we are not careful

To guard against overfitting we introduced some prior knowledge (inductive bias)

- hypothesis class =  $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$
- revised ERM rule: apply the ERM learning paradigm over  $\mathcal{H}$
- for the training sample  $S$ , the  $\text{ERM}_{\mathcal{H}}$  learner chooses a predictor  $h \in \mathcal{H}$  with the lowest possible error over  $S$ :

$$\text{ERM}_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\text{argmin}} L_S(h);$$



# Universal learner?

Do we need prior knowledge ( $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ ) for the success of learning?

Consider  $\mathcal{H}$  the set of all functions from  $\mathcal{X}$  to  $\mathcal{Y}$ ,  $\mathcal{H} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$ . This class represents lack of prior knowledge: every member of it is a good candidate.

Maybe there exists some kind of universal learner = a learner who has no prior knowledge about a certain task and is ready to be challenged by any task.

Does there exist a learning algorithm  $A$  and a training set size  $m$  such that for every distribution  $\mathcal{D}$ , if  $A$  receives  $m$  i.i.d. samples from  $\mathcal{D}$  is there a big chance it outputs a predictor  $h$  that has a low risk?

The No-Free-Lunch theorem states that no such universal learner exists. For binary classification prediction tasks ( $\mathcal{Y} = \{0,1\}$ ) for every learner there exists a distribution on which it fails (the learner will output a hypothesis with large generalization error)

# The No-Free-Lunch theorem

## Theorem (No-Free-Lunch)

Let  $A$  be any learning algorithm for the task of binary classification with respect to the 0–1 loss over a domain  $\mathcal{X}$ . Let  $m$  be any number smaller than  $|\mathcal{X}|/2$ , representing a training set size.

Then, there exists a distribution  $\mathcal{D}$  over  $\mathcal{X} \times \{0,1\}$  such that:

1. there exists a function  $f: \mathcal{X} \rightarrow \{0, 1\}$  with  $L_{\mathcal{D}}(f) = 0$ .
2. with probability of at least  $1/7$  over the choice of  $S \sim \mathcal{D}^m$  we have that  $L_{\mathcal{D}}(A(S)) \geq 1/8$ .

- *In other words, for every learning algorithm  $A$  there are cases for which this algorithm will fail whereas there is another learner (e.g. a trivial successful learner in this case would be an ERM learner with the hypothesis class  $\mathcal{H} = \{f\}$ , or more generally, ERM with respect to any finite hypothesis class that contains  $f$  and whose size satisfies the equation  $m \geq 8 \log(7|H|/6)$  that solves the task. It simply means that an adversary can use the fact that  $A$  has no clue what happens on the other half of the domain. We cannot learn perfectly without the proper background knowledge.*

# Intuition of the proof

Let  $C \subseteq \mathcal{X}$ , such that  $|C| = 2m$ ,  $C = \{c_1, c_2, \dots, c_{2m}\}$ .

Any learning algorithm that observes only  $m$  training examples = half of the instances in  $C = \{c_1, c_2, \dots, c_m\}$  has no information on what should be the labels of the rest of the instances in  $C = \{c_{m+1}, c_{m+2}, \dots, c_{2m}\}$ .

We expect that on the other half it will predict correctly half of them. Over all it will do on average 25% mistakes (50% mistakes on the unseen 50% data). There exists a distribution  $\mathcal{D}_i$  over  $C \times \{0, 1\}$  on which the learning algorithm will make  $\geq 25\%$  mistakes.

Apply Markov inequality to show the final result:

$$P_{S \sim D^m} (L_D(A(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$$

# Idea of the proof

To prove the theorem, we will construct a set of distributions such that there exist a distribution on which  $A$  will fail. The basic ingredients are:

- let  $C \subseteq X$ , such that  $|C| = 2m$ ,  $C = \{c_1, c_2, \dots, c_{2m}\}$
- there exist  $T = 2^{2m}$  functions from  $C$  to  $\{0, 1\}$
- denote these functions by  $f_1, \dots, f_T$ .
  - $f_1$  assigns to all elements of  $C$  the value 0:  $f_1(c_i) = 0$ ,  $i = 1, 2, \dots, 2m$
  - $f_2$  assigns to all elements of  $C$  value 0 except  $c_1$ :  $f_2(c_1) = 1, f_2(c_i) = 0$ ,  $i = 2, 3, \dots, 2m$
  - $f_3$  assigns to all elements of  $C$  value 0 except  $c_2$ :  $f_3(c_2) = 1, f_3(c_i) = 0$ ,  $i = 1, 3, \dots, 2m$
  - ....
  - $f_T$  assigns to all elements of  $C$  value 1:  $f_T(c_i) = 1$ ,  $i = 1, 2, \dots, 2m$
- for each function  $f_i$  define the distribution  $\mathcal{D}_i$  over  $C \times \{0, 1\}$  such that:

$$\mathcal{D}_i(\{(x, y)\}) = \begin{cases} 1/|C| & \text{if } y = f_i(x) \\ 0 & \text{otherwise.} \end{cases}$$

- $\mathcal{D}_i$  is perfect for  $f_i$ . It will only generate samples  $(x, y)$  on which  $f_i$  is correct ( $y = f_i(x)$ ). Hence  $L_{\mathcal{D}_i}(f_i) = 0$ .

# Idea of the proof

For every learning algorithm  $A$  that receives a sample  $S$  of  $m$  elements of  $C \times \{0, 1\}$  and returns a function  $A(S) : C \rightarrow \{0, 1\}$  we have:

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m} [L_{\mathcal{D}_i}(A(S))] \geq 1/4.$$

This part is the main core of the proof. It is very technical, we skip it, see the details in the book. The intuition tells us that on average we make errors on 25% of examples. But now we are interested in the worse case scenario (the maximum error).

Thus, we have (by choosing a maximizing  $i$  for the above inequality) that for every learning algorithm  $A'$  that receives  $m$  examples from  $\mathcal{X} \times \{0, 1\}$  there exist a function  $f : \mathcal{X} \rightarrow \{0, 1\}$  and a distribution  $\mathcal{D}$  over  $\mathcal{X} \times \{0, 1\}$ , such that  $L_{\mathcal{D}}(f) = 0$  and:

$$\mathbb{E}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A'(S))] \geq 1/4.$$

# Idea of the proof

$$\mathbb{E}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A'(S))] \geq 1/4.$$

Use the Markov's inequality to derive the following lemma:

**Lemma B.1.** *Let  $Z$  be a random variable that takes values in  $[0, 1]$ . Assume that  $\mathbb{E}[Z] = \mu$ . Then, for any  $a \in (0, 1)$ ,*

$$\mathbb{P}[Z > 1 - a] \geq \frac{\mu - (1 - a)}{a}.$$

Apply the lemma for  $Z = L_{\mathcal{D}}(A'(S))$ ,  $a = 7/8$ ,  $\mu = \mathbb{E}[Z] \geq 1/4$ . We obtain:

$$P(Z \geq 1 - \frac{7}{8}) \geq \frac{\frac{1}{4} - (1 - \frac{7}{8})}{\frac{7}{8}} \Leftrightarrow P(Z \geq \frac{1}{8}) \geq \frac{\frac{1}{4} - \frac{1}{8}}{\frac{7}{8}} = \frac{\frac{1}{8}}{\frac{7}{8}} = \frac{1}{7}$$

$$\mathbb{P}[L_{\mathcal{D}}(A'(S)) \geq 1/8] \geq 1/7$$

# Last time: Universal concept class $\mathcal{U}_n$

- $B^n$  = set of boolean  $n$ -tuples,  $|B| = 2^n$
- want to learn arbitrary subsets of  $B^n$
- $\mathcal{U}_n = \{h: B^n \rightarrow \{0,1\}\}$  - the concept class formed by all subsets of  $B^n$
- $\mathcal{U}_n$  – universal class
- $|\mathcal{U}_n| = 2^{2^n}$  – finite, so is PAC learnable with  $m_{\mathcal{H}}(\varepsilon, \delta)$  in the order of  $m$ :

$$m \geq \left\lceil \frac{1}{\varepsilon} \left( 2^n \log(2) + \log\left(\frac{1}{\delta}\right) \right) \right\rceil$$

- $\mathcal{U}_n$  – finite class, from the No-Free-Lunch theorem we need to see  $m > 2^{n-1}$  training examples otherwise there exists a task  $(f, \mathcal{D})$  on which the learner will fail

$$P_{S \sim D^m} (L_D(A(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$$

- for  $\varepsilon = 1/8$ ,  $\delta = 6/7$ ,  $n = 10$ ,  $m \geq 5679$  (all examples are  $2^{10} = 1024$ ), no matter how  $\mathcal{D}$  looks like we have:

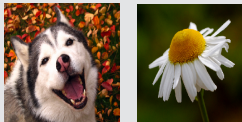
$$P_{S \sim D^m} (L_D(A(S)) \geq \frac{1}{8}) \leq \frac{1}{7}$$

# From theory to practice

- in practice we are interested in learning to solve a task for specific distributions: e.g. distribution of natural images
- we are not interested to generalize over all distributions (unnatural images: images with wrong labels, with shuffled pixels in an image, etc)

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

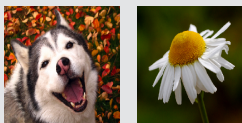
Data with original labels



dog

flower

Data with **random** labels



bus

dog

Deep Neural Networks  
easily fit random labels.



# No-Free-Lunch and prior knowledge

Consider  $\mathcal{H}$  the set of all functions from  $\mathcal{X}$  to  $\mathcal{Y}$ ,  $\mathcal{H} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$ . This class represents lack of prior knowledge: every member of it is a good candidate.

According to the No-Free-Lunch theorem, any learning algorithm (in particular the ERM predictor) that chooses its output from hypothesis in  $\mathcal{H}$  will fail on some learning task. Therefore,  $\mathcal{H}$  is not PAC learnable.

## Corollary

Let  $\mathcal{X}$  be an infinite domain set and let  $\mathcal{H}$  be the set of all functions from  $\mathcal{X}$  to  $\{0, 1\}$ . The  $\mathcal{H}$  is not PAC learnable.

*Proof:* Choose  $\varepsilon < 1/8$ ,  $\delta < 6/7$  and then apply the No-Free-Lunch theorem. There exists an  $f$  in  $\mathcal{H}$  and a distribution  $\mathcal{D}$  such that  $L_{\mathcal{D}}(f) = 0$ , but no matter how many examples a learning algorithm  $A$  will receive it will fail on the task as  $P(L_{\mathcal{D}}(A(S)) \geq 1/8) \geq 1/7$  (contradiction with PAC learnability).

# No-Free-Lunch and prior knowledge

We can escape the hazards foreseen by the No-Free-Lunch theorem by using our prior knowledge about a specific learning task, to avoid the distributions that will cause us to fail when learning that task.

Such prior knowledge can be expressed by restricting our hypothesis class. But how should we choose a good hypothesis class?

On the one hand, we want to believe that this class includes the hypothesis that has no error at all (in the PAC setting), or at least that the smallest error achievable by a hypothesis from this class is indeed rather small (in the agnostic setting).

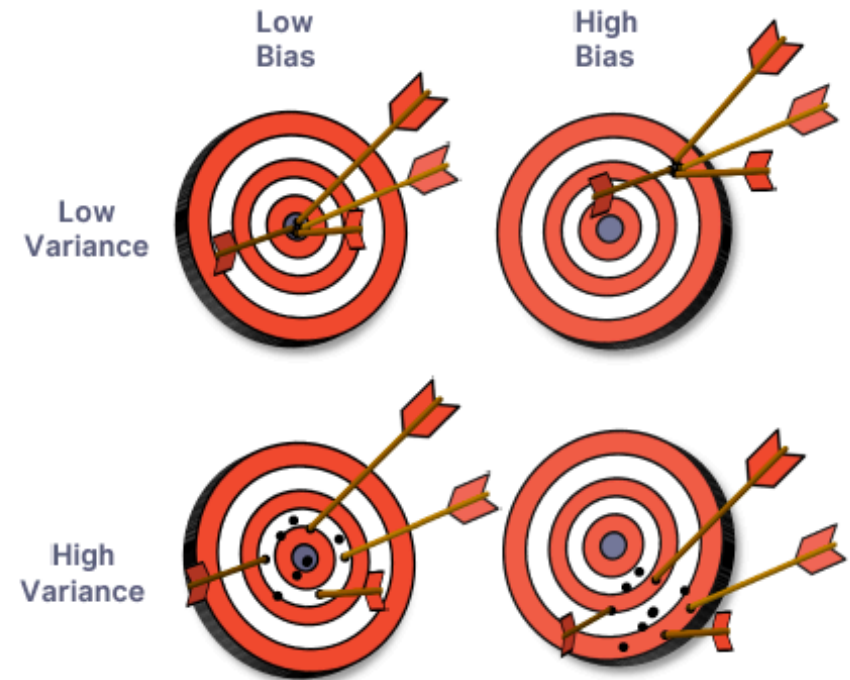
We cannot simply choose the richest class – the class of all functions over the given domain. Discuss this tradeoff next.

# The Bias-Complexity tradeoff

# Practical ML (sem. 1): Bias vs. Variance trade-off

## Bias vs. Variance

- There are two sources of error in ML models.
- **Bias** (or **systematic error**) comes from the inability of the algorithm to model the true relationship between features and label (*underfitting*).
  - Bias error cannot be corrected by adding more training samples, but by increasing the complexity of the model.
- **Variance** (or **random error**) comes from sensitivity to small fluctuations in the training data, causing the algorithm to model random noise. (*overfitting*)
  - Variance error can be corrected by adding more training samples or by decreasing the complexity of the model.
- There is usually a *tradeoff* between bias and variance.



# The error decomposition

Decompose the error of an  $\text{ERM}_{\mathcal{H}}$  predictor that chooses  $h_S$  from a restricted class  $\mathcal{H}$  into two components:

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}$$

where :  $\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ ,  $\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}$

## *The approximation error ( $\epsilon_{\text{app}}$ )*

- the minimum risk achievable by a predictor in the hypothesis class  $\mathcal{H}$
- it measures how well our hypothesis class  $\mathcal{H}$  fits the distribution
- it is independent of the particular sample
- it is determined only by  $\mathcal{H}$ , enlarging it decreases the approximation error
- under the realizability assumption, the approximation error is zero.
- it is the **bias** term

# The error decomposition

Decompose the error of an  $\text{ERM}_{\mathcal{H}}$  predictor that chooses  $h_S$  from a restricted class  $\mathcal{H}$  into two components:

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}$$

where :  $\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ ,  $\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}$

## *The estimation error ( $\epsilon_{\text{est}}$ )*

- it measures how well our particular sample let us estimate the best classifier
- the empirical risk is only an estimate of the true risk, and so the predictor minimizing the empirical risk is only an estimate of the predictor minimizing the true risk.
- the quality of this estimation depends on the training set size and on the size (complexity) of  $\mathcal{H}$
- it varies with samples
- it is the *variance/complexity* term

# The bias – complexity tradeoff

On one hand, choosing  $\mathcal{H}$  to be a very rich class decreases the approximation error but at the same time might increase the estimation error, as a rich  $\mathcal{H}$  might lead to overfitting .

On the other hand, choosing  $\mathcal{H}$  to be a very small set reduces the estimation error but might increase the approximation error or, in other words, might lead to underfitting .

A great choice for  $\mathcal{H}$  is the class that contains only one classifier – the Bayes optimal classifier. But the Bayes optimal classifier depends on the underlying distribution  $\mathcal{D}$ , which we do not know (indeed, learning would have been unnecessary had we known  $\mathcal{D}$ ).

# Uniform Convergence



# Sufficient learning condition for agnostic PAC learnability

- given  $\mathcal{H}$ , the  $\text{ERM}_{\mathcal{H}}$  learning paradigm works as follows:
  - based on a received training sample  $S$  of examples drawn i.i.d from an unknown distribution  $\mathcal{D}$  over a domain  $\mathcal{Z}$ ,  $\text{ERM}_{\mathcal{H}}$  evaluates the risk (error) of each  $h$  in  $\mathcal{H}$  on  $S$  and outputs a member  $h_S = \text{ERM}_{\mathcal{H}}(S)$  that minimizes the empirical error  $L_S(h_S)$ ;
  - we want that  $h_S$  will generalize wrt true data probability distribution  $\mathcal{D}$ , i.e  $L_{\mathcal{D}}(h_S)$  is small;
  - it suffices to ensure that the empirical risks of all  $h$  in  $\mathcal{H}$  are good approximations of their true risk
- we need that *uniformly* over all hypothesis  $h$  in the hypothesis class  $\mathcal{H}$ , the empirical risk based on  $S$  will be close to true risk for all possible probability distributions  $\mathcal{D}$  over the domain  $\mathcal{Z}$

# $\epsilon$ - Representative

- how well you can learn a hypothesis depends on the quality of that sample:
  - you can't learn anything from a bad sample
  - a bad sample will make a bad hypothesis to look good and a good one to look bad
- when is a sample good?
  - a sample is good if the estimated quality (the loss) of a hypothesis on that sample is very close to its true error

**Definition** ( $\epsilon$  – representative sample)

A sample  $S$  is called  $\epsilon$  – representative wrt domain  $Z$ , hypothesis class  $\mathcal{H}$ , loss function  $\ell$  and distribution  $\mathcal{D}$  if:

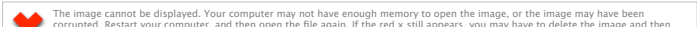
$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)] \quad L_S(h) = \frac{1}{m} \sum_{z \in S} \ell(h, z)$$

# $\epsilon$ – Representative Samples are Good

## Lemma

Let  $S$  be a sample that is  $\epsilon/2$  – representative wrt domain  $\mathcal{Z}$ , hypothesis class  $\mathcal{H}$ , loss function  $\ell$  and distribution  $\mathcal{D}$ . Then any output of  $\text{ERM}_{\mathcal{H}}(S)$  i.e any  $h_S \in \text{argmin}_h L_S(h)$  satisfies:


$$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

## Proof

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \epsilon/2 \leq \min_h L_S(h) + \epsilon/2 \leq \min_h L_{\mathcal{D}}(h) + \epsilon/2 + \epsilon/2$$



$S$  is  $\epsilon/2$  – representative sample

# Uniform convergence

If  $\epsilon$ -representative samples allows us to learn as good as possible, we can agnostically PAC learn if we can guarantee that we will almost always get (with probability  $1 - \delta$ )  $\epsilon$ -representative sample.

## **Definition** (*uniform convergence*)

A hypothesis class  $\mathcal{H}$  has the *uniform convergence property* wrt a domain  $Z$ , loss function  $\ell$  if:

- there exists a function  $m_H^{UC} : (0,1)^2 \rightarrow \mathbb{N}$
- such that for all  $(\epsilon, \delta) \in (0,1)^2$
- and for any probability distribution  $\mathcal{D}$  over  $Z$

if  $S$  is a sample of  $m \geq m_H^{UC}(\epsilon, \delta)$  examples drawn i.i.d. according to  $\mathcal{D}$ , then, with probability of at least  $1 - \delta$ ,  $S$  is  $\epsilon$ -representative.

The term *uniform* refers to having a fixed sample size that works for all members of  $\mathcal{H}$  and over all possible probability distributions  $\mathcal{D}$  over the domain  $Z$

# A tool to prove PAC learnability

- uniform convergence serves as a tool to prove that we can PAC learn a hypothesis class  $\mathcal{H}$

## Corollary

If hypothesis class  $\mathcal{H}$  has the uniform convergence property with function  $m_H^{UC}$  then  $\mathcal{H}$  is agnostically PAC learnable with the sample complexity:

$$m_H(\varepsilon, \delta) \leq m_H^{UC}(\varepsilon / 2, \delta)$$

Moreover, the  $\text{ERM}_{\mathcal{H}}$  paradigm is a successful agnostic PAC learner for  $\mathcal{H}$ .

# Finite classes are agnostic PAC learnable

## Theorem

Let  $\mathcal{H}$  be a finite hypothesis class, let  $\mathcal{Z}$  be a domain and let  $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow [0,1]$  be a loss function. Then  $\mathcal{H}$  has the uniform convergence property with sample complexity:

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

Moreover, the class  $\mathcal{H}$  is agnostically PAC learnable using the ERM paradigm with sample complexity:

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

# Proof - Finite classes are agnostic PAC learnable

- uniform convergence serves as a tool to prove that we can PAC learn a hypothesis class  $\mathcal{H}$
- to prove that finite hypothesis classes have the uniform convergence property, we need to:
  - for fixed  $\epsilon$  and  $\delta$
  - find a sample size  $m$
  - such that for any distribution  $\mathcal{D}$  over  $\mathcal{Z}$
  - and a sample  $S = (z_1, z_2, \dots, z_m)$  of examples i.i.d from  $\mathcal{D}$
  - with probability at least  $1 - \delta$
  - it holds that for all  $h \in \mathcal{H}$   $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$ .

That is:  $\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta$ .



$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

# Proof - union bound

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\},$$

Use the union bound to obtain:

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).$$

For a sufficiently large  $m$ , each summand of the right-hand side of this inequality is small enough.

Show that for any fixed hypothesis  $h$  (which is chosen in advance prior to the sampling of the training set), the gap between the true and empirical risks,  $|L_S(h) - L_{\mathcal{D}}(h)|$ , is likely to be small.



# Proof - Hoeffding's inequality

**Lemma** (Hoeffding's Inequality). *Let  $\theta_1, \dots, \theta_m$  be a sequence of i.i.d. random variables and assume that for all  $i$ ,  $\mathbb{E}[\theta_i] = \mu$  and  $\mathbb{P}[a \leq \theta_i \leq b] = 1$ . Then, for any  $\epsilon > 0$*

$$\mathbb{P} \left[ \left| \frac{1}{m} \sum_{i=1}^m \theta_i - \mu \right| > \epsilon \right] \leq 2 \exp \left( -2m \epsilon^2 / (b - a)^2 \right).$$

Apply in our case by setting:

$$\theta_i = l(h, z_i) \quad L_S(h) = \frac{1}{m} \sum_{z \in S} l(h, z) = \frac{1}{m} \sum_i \theta_i \quad L_D(h) = \mu \quad a = 0, b = 1$$

Then, we have:

$$\mathcal{D}^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) = \mathbb{P} \left[ \left| \frac{1}{m} \sum_{i=1}^m \theta_i - \mu \right| > \epsilon \right] \leq 2 \exp \left( -2m \epsilon^2 \right).$$

# Proof - final step

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} 2 \exp(-2m\epsilon^2) \\ &= 2|\mathcal{H}| \exp(-2m\epsilon^2)\end{aligned}$$

Choose  $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$

Then, we have:

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta.$$

# Beyond the result

By going from realizability to agnostic, we go:

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

The denominator goes from  $\epsilon$  to  $\epsilon^2$ , which means that for the same of accuracy the minimal sample size grows by a factor of  $1/\epsilon$ .

# Uniform convergence for agnostic PAC learning?

## **Definition** (*uniform convergence*)

A hypothesis class  $\mathcal{H}$  has the *uniform convergence property* wrt a domain  $\mathcal{Z}$ , loss function  $\ell$  if:

- there exists a function  $m_H^{UC} : (0,1)^2 \rightarrow \mathbb{N}$
- such that for all  $(\epsilon, \delta) \in (0,1)^2$
- and for any probability distribution  $\mathcal{D}$  over  $\mathcal{Z}$

if  $S$  is a sample of  $m \geq m_H^{UC}(\epsilon, \delta)$  examples drawn i.i.d. according to  $\mathcal{D}$ , then, with probability of at least  $1 - \delta$ ,  $S$  is  $\epsilon$ -representative.

## **Definition** ( $\epsilon$ – representative sample)

A sample  $S$  is called  $\epsilon$  – representative wrt domain  $\mathcal{Z}$ , hypothesis class  $\mathcal{H}$ , loss function  $\ell$  and distribution  $\mathcal{D}$  if:

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

# Next time

<b>6</b>	<b>The VC-Dimension</b>	43
6.1	Infinite-Size Classes Can Be Learnable	43
6.2	The VC-Dimension	44
6.3	Examples	46
6.4	The Fundamental Theorem of PAC learning	48
6.5	Proof of Theorem 6.7	49
6.6	Summary	53
6.7	Bibliographic remarks	53
6.8	Exercises	54