# Foundations for Machine Learning

L. Y. Stefanus

TU Dresden, June-July 2018

# Reference

- Shai Shalev-Shwartz and Shai Ben-David. UNDERSTANDING MACHINE LEARNING: From Theory to Algorithms. Cambridge University Press, 2014.

# The Bias-Complexity Tradeoff

## Error Decomposition

# How should we choose a good hypothesis class?

- To answer this question we decompose the error of an $\text{ERM}_\text{H}$ predictor into two components as follows.
- Let $h_S$ be an $\text{ERM}_\text{H}$ hypothesis. Then, we can write

$$L_D(h_S) = \varepsilon_{\text{app}} + \varepsilon_{\text{est}}$$

- where

$$\varepsilon_{\text{app}} = \min_{h \in H} L_D(h)$$

$$\varepsilon_{\text{est}} = L_D(h_S) - \varepsilon_{\text{app}}$$

# The Approximation Error $\varepsilon_{app}$

- The approximation error is the minimum risk achievable by a predictor in the hypothesis class. This term measures how much risk we have because we restrict ourselves to a specific class, namely, how much inductive bias we have.
- The approximation error does not depend on the sample size and is determined by the hypothesis class chosen.
- Enlarging the hypothesis class can decrease the approximation error.
- Under the realizability assumption, the approximation error is zero. In the agnostic case, however, the approximation error can be large.

# The Estimation Error $\varepsilon_{est}$

- The estimation error is the difference between the approximation error and the error achieved by the ERM predictor.
- The estimation error occurs because the empirical risk (i.e., training error) 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, or complexity, of the hypothesis class.

# The Estimation Error $\varepsilon_{est}$

- As we have studied, for a finite hypothesis class, $\varepsilon_{est}$ increases (logarithmically) with $|H|$ and decreases with m.
- We can think of the size of H as a measure of its complexity. Later we will study another complexity measure of hypothesis classes, called VC dimension.

# the bias-complexity tradeoff

- Since our goal is to minimize the total risk, we face a tradeoff, called the bias-complexity tradeoff.
- On one hand, choosing H to be a very rich class decreases the approximation error but at the same time might increase the estimation error, as a rich H might lead to overfitting.
- On the other hand, choosing 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.

# the bias-complexity tradeoff

- A great choice for H is the class that contains only one classifier -- the Bayes optimal classifier. But the Bayes optimal classifier depends on the underlying distribution D, which we do not know. Indeed, learning would have been unnecessary if we had known D.
- Learning theory studies how rich we can make H while still maintaining reasonable estimation error.
- In many cases, empirical research focuses on designing good hypothesis classes for a certain domain.

# The VC-Dimension

# Which classes H are PAC learnable?

- So far we have seen that finite classes are learnable, but that the class of all functions (over an infinite size domain) is not.
- What makes one class learnable and the other not learnable? Can infinite-size classes be learnable, and, if so, what determines their sample complexity?

# Infinite-Size Classes Can Be Learnable

- To show that the size of the hypothesis class is not the right characterization of its sample complexity, we first present a simple example of an infinite-size hypothesis class that is learnable.

# Infinite-Size Classes Can Be Learnable

## Example

- Let $H$ be the set of threshold functions over the real line, namely,

$$H = \{h_a : a \in \mathbb{R}\}$$

where $h_a : \mathbb{R} \to \{0,1\}$ is a function such that

$$h_a(x) = \mathbf{1}_{[x<a]} = \begin{cases} 1 & \text{if } x < a \\ 0 & \text{otherwise} \end{cases}$$

- This $H$ is of infinite size. Nevertheless, the following lemma shows that $H$ is learnable in the PAC model using the ERM algorithm.

# Lemma 6.1

Let $H$ be the class of threshold functions as defined earlier. Then, $H$ is PAC learnable, using the ERM rule, with sample complexity of

$$m_H(\epsilon,\delta) \leq \left\lceil \frac{\ln\left(\frac{2}{\delta}\right)}{\epsilon} \right\rceil.$$

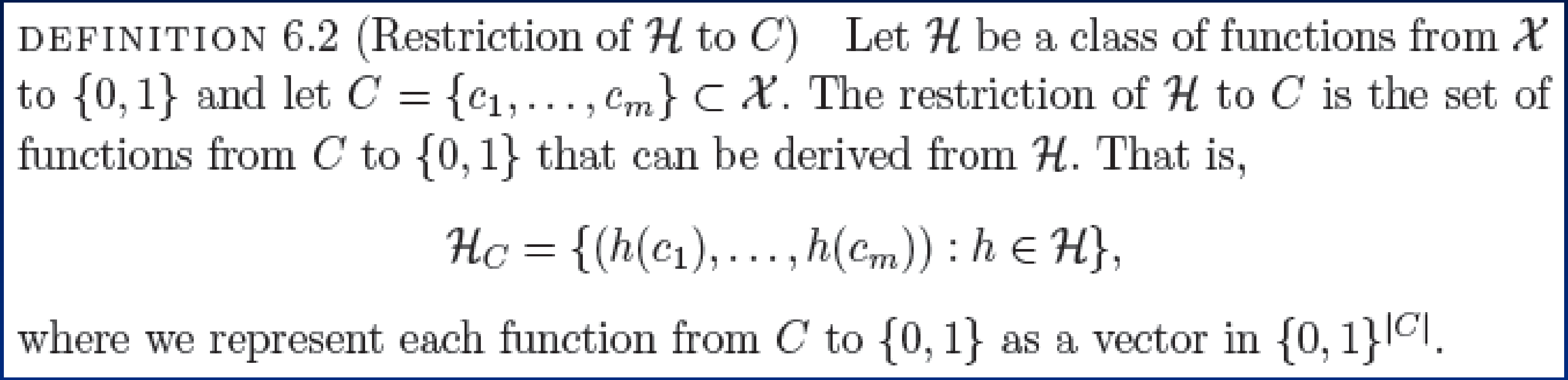
Proof: [Exercise, see the textbook]

# Motivation for DC-dimension

- Lemma 6.1 shows that while finiteness of H is a sufficient condition for learnability, it is not a necessary condition.
- We demonstrate that a property called the VC-dimension of a hypothesis class gives the correct characterization of its learnability.
- In the proof of the No-Free-Lunch theorem, we have shown that without restricting the hypothesis class, for any learning algorithm, an adversary can construct a distribution for which the learning algorithm will perform poorly, while there is another learning algorithm that can succeed on the same distribution. To do so …

# Motivation for DC-dimension

- To do so, the adversary used a finite set $C \subset X$ and considered a family of distributions that are concentrated on elements of $C$. Each distribution was derived from a “true” target function from $C$ to $\{0,1\}$.
- To make any algorithm fail, the adversary used the power of choosing a target function from the set of all possible functions from $C$ to $\{0,1\}$.
- When considering PAC learnability of a hypothesis class $H$, the adversary is restricted to constructing distributions for which some hypothesis $h \in H$ achieves a zero risk. Since we are considering distributions that are concentrated on elements of $C$, we should study how $H$ behaves on $C$, which leads to the following definition.

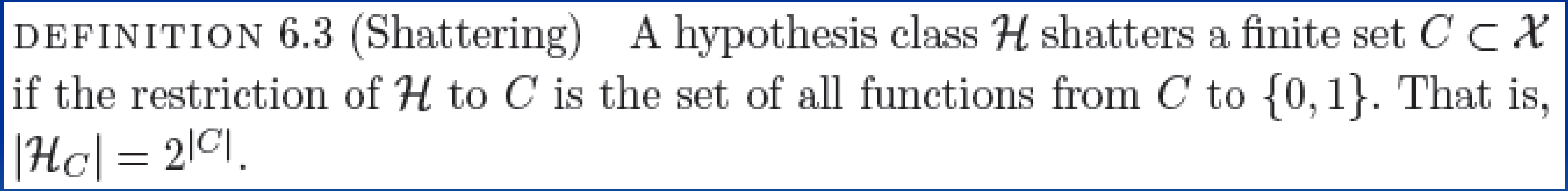


DEFINITION 6.2 (Restriction of $\mathcal{H}$ to $C$) Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to $C$ is the set of functions from $C$ to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\},$$

where we represent each function from $C$ to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

If the restriction of H to C is the set of all functions from C to {0,1}, then we say that H shatters the set C.



DEFINITION 6.3 (Shattering) A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to $C$ is the set of all functions from $C$ to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.

# Examples of shattering (1)

- Let $H$ be the class of threshold functions over $\mathbb{R}$.
- Take a set $C = \{c_1\}$. Now, if we take $a = c_1 + 1$, then we have $h_a(c_1) = 1$, and if we take $a = c_1 - 1$, then we have $h_a(c_1) = 0$. Therefore, $H_C$ is the set of all functions from $C$ to {0,1}, and $H$ shatters $C$.
- Now take a set $C = \{c_1, c_2\}$, where $c_1 \leq c_2$. No $h \in H$ can account for the labeling (0, 1), because any threshold that assigns the label 0 to $c_1$ must assign the label 0 to $c_2$ as well. Therefore not all functions from $C$ to {0,1} are included in $H_C$; hence $C$ is not shattered by $H$.

# Examples of shattering (2)

- Let $H$ be the class of intervals over $\mathbb{R}$, namely,

$$H = \{h_{a,b}: a, b \in \mathbb{R}, a < b\},$$

  where $h_{a,b}: \mathbb{R} \to \{0,1\}$ is a function such that

$$h_{a,b}(x) = \mathbf{1}_{[x \in (a,b)]}.$$

- Take the set $C = \{c_1, c_2\}$, where $c_1 \leq c_2$. Then $H$ shatters $C$.
- Now take a set $C = \{c_1, c_2, c_3\}$ where $c_1 \leq c_2 \leq c_3$. Then the labeling (1,0,1) cannot be obtained by an interval and therefore $H$ does not shatter $C$ .