

# Lecture 4: Non-Uniform PAC Learning

Tianpei Xie

Dec. 20th., 2022

## Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Non-Uniform PAC Learning</b>                     | <b>2</b> |
| 1.1      | Definitions . . . . .                               | 2        |
| 1.2      | Characterizing Non-Uniform Learnability . . . . .   | 3        |
| <b>2</b> | <b>Structrual Risk Minimization</b>                 | <b>3</b> |
| <b>3</b> | <b>Minimum Description Length and Occam's Razor</b> | <b>5</b> |
| 3.1      | Occam's Razor . . . . .                             | 5        |
| 3.2      | Minimum Description Length . . . . .                | 5        |

# 1 Non-Uniform PAC Learning

## 1.1 Definitions

- **Remark (*Relaxation of PAC Learning*)**

The notions of **PAC learnability** discussed so far allow *the sample sizes* to depend on the *accuracy* and *confidence parameters*, but they are **uniform** with respect to the **labeling rule** and **the underlying data distribution**. Consequently, classes that are learnable in that respect are *limited* (they must have a finite VC-dimension). In this chapter we consider more *relaxed, weaker notions of learnability*.

- Recall that the agnostic PAC-learning:

**Definition (*Agnostic PAC-Learning*)**

Let  $\mathcal{H}$  be a hypothesis set.  $\mathcal{A}$  is an **agnostic PAC-learning algorithm** if there exists a **polynomial function**  $\text{poly}(\cdot, \cdot, \cdot, \cdot)$  such that for any  $\epsilon > 0$  and  $\delta > 0$ , for all distributions  $\mathcal{P}$  over  $\mathcal{X} \times \mathcal{Y}$ , the following holds for any sample size  $m \geq \text{poly}(1/\epsilon, 1/\delta, d, \text{size}(c))$ :

$$\mathcal{P}^m \left\{ L(\mathcal{A}(\mathcal{D}_m)) - \inf_{h \in \mathcal{H}} L_{\mathcal{P}}(h) \leq \epsilon \right\} \geq 1 - \delta. \quad (1)$$

If  $\mathcal{A}$  further runs in  $\text{poly}(1/\epsilon, 1/\delta, d, \text{size}(c))$ , then  $\mathcal{H}$  is said to be **efficiently agnostic PAC-learnable**.

- We now introduce the concept of **competitiveness**:

**Definition (*Competitiveness*)**

A hypothesis  $g$  is  **$(\epsilon, \delta)$ -competitive** with another hypothesis  $g'$  if, *with probability higher than  $(1 - \delta)$* ,

$$L(g) \leq L(g') + \epsilon.$$

- **Definition (*Non-Uniform Learning*)**

Let  $\mathcal{H}$  be a hypothesis set.  $\mathcal{A}$  is an **non-uniform learning algorithm** if there exists a **polynomial function**  $\text{poly}(\cdot, \cdot, \cdot, \cdot, \cdot)$  such that for any  $\epsilon > 0$  and  $\delta > 0$ , for all distributions  $\mathcal{P}$  over  $\mathcal{X} \times \mathcal{Y}$ , the following holds for **any**  $h \in \mathcal{H}$ , any sample size  $m \geq \text{poly}(1/\epsilon, 1/\delta, \underline{h}, d)$ :

$$\mathcal{P}^m \{ L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}_m)) - L_{\mathcal{P}}(h) \leq \epsilon \} \geq 1 - \delta. \quad (2)$$

If  $\mathcal{A}$  further runs in  $\text{poly}(1/\epsilon, 1/\delta, h, d)$ , then  $\mathcal{H}$  is said to be **non-uniformly learnable**.

- **Remark (*Compare with Agnostic PAC Learning*)**

Both **agnostic PAC learning** and **non-uniform learning** require that the output hypothesis will be  **$(\epsilon, \delta)$ -competitive** with **every other hypothesis in the class**.

From the definition of **non-uniform learning**, we see that **the sample complexity**  $m \geq m(\epsilon, \delta, h)$ , which **depends on other hypothesis**  $h \in \mathcal{H}$ , while for agnostic PAC learning, the sample size  $m \geq m(\epsilon, \delta)$  is chosen *uniformly* over  $\mathcal{H}$ . It is easy to see that an **agnostic PAC learnable** class is also **non-uniformly learnable**.

## 1.2 Characterizing Non-Uniform Learnability

- **Lemma 1.1** [Shalev-Shwartz and Ben-David, 2014]

Let  $\mathcal{H}$  be a hypothesis class that can be written as a **countable union** of hypothesis classes,

$$\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n,$$

where each  $\mathcal{H}_n$  enjoys the **uniform convergence** property. Then,  $\mathcal{H}$  is **non-uniformly learnable**.

- **Proposition 1.2 (Characterization of Non-Uniform Learnable Class)** [Shalev-Shwartz and Ben-David, 2014]

A hypothesis class  $\mathcal{H}$  of binary classifiers is **non-uniformly learnable if and only if** it is a **countable union of agnostic PAC learnable hypothesis classes**.

- **Example (Non-Uniform Learnable But Not Agnostic PAC Learnable)**

The following example shows that **non-uniform learnability** is a **strict relaxation** of **agnostic PAC learnability**; namely, there are hypothesis classes that are non-uniform learnable but are not agnostic PAC learnable.

Consider a binary classification problem with the instance domain being  $\mathcal{X} = \mathbb{R}$ . For every  $n \in \mathbb{N}$  let  $\mathcal{H}_n$  be the class of *polynomial classifiers* of degree  $n$ ; namely,  $\mathcal{H}_n$  is the set of all classifiers of the form  $h(x) = \text{sgn}(p(x))$  where  $p : \mathbb{R} \rightarrow \mathbb{R}$  is a polynomial of degree  $n$ . Let  $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$ . Therefore,  $\mathcal{H}$  is *the class of all polynomial classifiers* over  $\mathbb{R}$ .

It is easy to verify that  $VCdim(\mathcal{H}) = \infty$  while  $VCdim(\mathcal{H}_n) = n + 1$ . Hence,  $\mathcal{H}$  is *not PAC learnable*, while on the basis of Proposition above,  $\mathcal{H}$  is *non-uniformly learnable*.

## 2 Structural Risk Minimization

- **Remark (Encoding Prior Knowledge)**

So far, we have *encoded our prior knowledge* by *specifying a hypothesis class*  $\mathcal{H}$ , which we believe includes a good predictor for the learning task at hand.

Yet another way to express our prior knowledge is by *specifying preferences over hypotheses within*  $\mathcal{H}$ . In the *Structural Risk Minimization (SRM)* paradigm, we do so by first assuming that  $\mathcal{H}$  can be written as  $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$  and then specifying a **weight function**,  $w : \mathbb{N} \rightarrow [0, 1]$ , which assigns a **weight** to *each hypothesis class*,  $\mathcal{H}_n$ , such that a *higher weight* reflects a *stronger preference* for the hypothesis class.

- **Definition (Structural Risk Minimization (SRM) paradigm)**

Let  $\mathcal{H}$  be a hypothesis class that can be written as

$$\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n,$$

Assume that for each  $n$ , the class  $\mathcal{H}_n$  enjoys **the uniform convergence property**, i.e. *PAC-learnable* regardless underlying distribution, with a *sample complexity function*  $m_{\mathcal{H}_n}(\epsilon, \delta)$ . Let

us also define the function  $\epsilon_n : \mathbb{N} \times (0, 1) \rightarrow (0, 1)$  by

$$\epsilon_n(m, \delta) = \min \{ \epsilon \in (0, 1) : m_{\mathcal{H}_n}(\epsilon, \delta) \leq m \}. \quad (3)$$

In other words, we have a **fixed sample size**  $m$ , and we are interested in **the lowest possible upper bound** on the *gap* between *empirical* and *true risks* achievable by using a sample of  $m$  examples.

Note that it follows that for every  $m$  and  $\delta$ , with probability of at least  $1 - \delta$  over the choice of  $\mathcal{D}_m \sim \mathcal{P}$  we have that

$$\left| L_{\mathcal{P}}(h) - \widehat{L}_m(h) \right| \leq \epsilon_n(m, \delta), \quad \forall h \in \mathcal{H}_n.$$

Let  $w : \mathbb{N} \rightarrow [0, 1]$  be a function such that  $\sum_{n=1}^{\infty} w(n) \leq 1$ . We refer to  $w$  as a **weight function** over the hypothesis classes  $\mathcal{H}_1, \mathcal{H}_2, \dots$ . Such a weight function can *reflect the importance* that the learner attributes to **each hypothesis class**, or some *measure of the complexity* of different hypothesis classes.

The goal of a **Structural Risk Minimization (SRM) rule** is to find a hypothesis  $h \in \mathcal{H}$  that minimizes a certain upper bound on the true risk by **choosing weight** in a “**bound minimization**” manner. In particular, the SRM solves the following problem:

$$\min_{h \in \mathcal{H}} \left\{ \widehat{L}_m(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta) \right\} \quad (4)$$

where

$$n(h) := \min \{ n : h \in \mathcal{H}_n \}. \quad (5)$$

- We have the following proposition:

**Proposition 2.1** [Shalev-Shwartz and Ben-David, 2014]

Let  $w : \mathbb{N} \rightarrow [0, 1]$  be a function such that  $\sum_{n=1}^{\infty} w(n) \leq 1$ . Let  $\mathcal{H}$  be a hypothesis class that can be written as  $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$ , where for each  $n$ ,  $\mathcal{H}_n$  satisfies **the uniform convergence property** with a sample complexity function  $m_{\mathcal{H}_n}(\epsilon, \delta)$ . Let  $\epsilon_n$  be as defined in Equation 3.

Then, for every  $\delta \in (0, 1)$  and distribution  $\mathcal{P}$ , with probability of at least  $1 - \delta$  over the choice of  $\mathcal{D}_m \sim \mathcal{P}^m$ , the following bound **holds (simultaneously)** for every  $n \in \mathbb{N}$  and  $h \in \mathcal{H}_n$ .

$$\left| L_{\mathcal{P}}(h) - \widehat{L}_m(h) \right| \leq \epsilon_n(m, w(n) \cdot \delta).$$

Therefore, for every  $\delta \in (0, 1)$  and distribution  $\mathcal{P}$ , with probability of at least  $1 - \delta$  it holds that

$$L_{\mathcal{P}}(h) \leq \widehat{L}_m(h) + \min_{n: h \in \mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta). \quad (6)$$

- **Remark (Bias for Lower Risk vs Bias for Smaller Estimation Error Tradoff)**  
Unlike the *ERM* paradigm discussed in previous chapters, we *no longer just care about the empirical risk*,  $\widehat{L}_m(h)$ , but we are willing to **trade** some of our **bias** toward **low empirical risk** with a **bias** toward **classes** for which  $\epsilon_{n(h)}(m, w(n(h)) \cdot \delta)$  is **smaller**, for the sake of a smaller estimation error.

- **Remark** By *Hoeffding's inequality*, each singleton class has the uniform convergence property with rate  $m_{\mathcal{H}_n}(\epsilon, \delta) = \frac{\log(2/\delta)}{2\epsilon^2}$  so SRM rule (4) becomes

$$\begin{aligned} & \min_{h \in \mathcal{H}} \left\{ \hat{L}_m(h) + \sqrt{\frac{-\log(w(n)) + \log(2/\delta)}{2m}} \right\} \\ \Rightarrow & \min_{h \in \mathcal{H}} \left\{ \hat{L}_m(h) + \sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}} \right\} \end{aligned} \quad (7)$$

- **Proposition 2.2 (SRM for Non-Uniform Learning)** [Shalev-Shwartz and Ben-David, 2014]

Let  $\mathcal{H}$  be a hypothesis class such that  $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$ , where each  $\mathcal{H}_n$  has the uniform convergence property with sample complexity  $m_{\mathcal{H}_n}$ . Let  $w : \mathbb{N} \rightarrow [0, 1]$  be such that

$$w(n) = \frac{6}{\pi^2 n^2}.$$

Then,  $\mathcal{H}$  is **non-uniformly learnable** using the **SRM rule** with rate

$$m_{\mathcal{H}}(\epsilon, \delta, h) \leq m_{\mathcal{H}_n} \left( \frac{\epsilon}{2}, \frac{6\delta}{(\pi n(h))^2} \right).$$

- **Remark (SRM as Resource Allocation)**

Consider SRM as **simultaneously run**  $n$  PAC learning algorithm on different hypothesis classes with **shared sample size**  $m$  so that it need to **allocate resources** to the hypothesis class  $\mathcal{H}_n$  in some optimal way to minimize the overall gap between true risk and empirical risk.

### 3 Minimum Description Length and Occam's Razor

#### 3.1 Occam's Razor

- **Remark (Occam's Razor)**

A **short explanation** (that is, a hypothesis that has a short length) tends to be **more valid** than a **long explanation**.

#### 3.2 Minimum Description Length

- **Remark (Efficient Prior Knowledge Encoding)**

See that the SRM optimize the following objective:

$$\min_{h \in \mathcal{H}} \left\{ \hat{L}_m(h) + \sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}} \right\}$$

It follows that in this case, **the prior knowledge** is solely determined by the **weight** we assign to each hypothesis. We assign **higher weights** to hypotheses that we **believe are more likely to be the correct one**, and in the learning algorithm we prefer hypotheses that have higher weights.

- **Definition (*Description Language of Hypothesis Class*)**

Let  $\mathcal{H}$  be the hypothesis class we wish to describe. Fix some *finite set*  $\Sigma$  of **symbols** (or “characters”), which we call the **alphabet**. For concreteness, we let  $\Sigma = \{0, 1\}$ . A **string** is a *finite sequence of symbols* from  $\Sigma$ ; for example,  $\sigma = (0, 1, 1, 1, 0)$  is a string of *length* 5. We denote by  $|\sigma|$  **the length of a string**. The set of all finite length strings is denoted  $\Sigma^*$ .

A **description language for  $\mathcal{H}$**  is a function  $d : \mathcal{H} \rightarrow \Sigma^*$  mapping each member  $h$  of  $\mathcal{H}$  to a string  $d(h)$ .  $d(h)$  is called **the description of  $h$** , and its *length* is denoted by  $|h|$ .

- **Remark (*Restriction of  $\mathcal{H}$  on  $\mathcal{D}_m$* )**

The **restriction of  $\mathcal{H}$  to  $\mathcal{D}$**  is the set of functions from  $\mathcal{D}$  to  $\{0, 1\}$  that can be derived from  $\mathcal{H}$ . That is,

$$\mathcal{H}_{\mathcal{D}} := \{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\}.$$

For each  $h \in \mathcal{H}$ ,  $h_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}} \subset \{0, 1\}^*$  is a **description** of  $h$  which is a **binary string of fixed length  $m$** . It is *not* a *prefix-free string* and is **data-dependent** which is not preferred in MDL.

- **Definition (*Prefix-Free String*)**

For every **distinct**  $h, h'$ ,  $d(h)$  is **not a prefix** of  $d(h')$ .

That is, we do not allow that any string  $d(h)$  is *exactly the first  $|h|$  symbols of any longer string  $d(h')$* .

- **Lemma 3.1 (*Kraft's Inequality*).**

If  $S \subseteq \{0, 1\}^*$  is a **prefix-free set of strings**, then

$$\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1$$

- **Remark** In light of Kraft's inequality, any prefix-free description language of a hypothesis class,  $\mathcal{H}$ , gives rise to a **weighting function**  $w$  over that hypothesis class where

$$w(h) = \frac{1}{2^{|h|}}.$$

- **Proposition 3.2 (*Generalization Bound by Description Length*)** [Shalev-Shwartz and Ben-David, 2014]

Let  $\mathcal{H}$  be a hypothesis class and let  $d : \mathcal{H} \rightarrow \{0, 1\}^*$  be a **prefix-free description language** for  $\mathcal{H}$ . Then, for every sample size,  $m$ , every confidence parameter,  $\delta > 0$ , and every probability distribution,  $\mathcal{P}$ , with probability greater than  $1 - \delta$  over the choice of  $\mathcal{D}_m \sim \mathcal{P}^m$  we have that,

$$L_{\mathcal{P}}(h) \leq \widehat{L}_m(h) + \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \quad (8)$$

where  $|h|$  is the **length of description  $d(h)$**  of  $h$ .

- **Definition (*Minimum Description Length (MDL) learning paradigm*)**

With the definition of description language of hypothesis, the goal of a **Minimum Description Length (MDL) learning** is to find a hypothesis  $h \in \mathcal{H}$  such that

$$\min_{h \in \mathcal{H}} \left\{ \widehat{L}_m(h) + \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \right\} \quad (9)$$

In particular, it suggests *trading off* empirical risk for *saving description length*.

- **Remark** (*Choose Description Language Independent From the Data*)

As we know from the basic Hoeffding's bound, if we **commit** to any hypothesis **before** seeing the data, then we are guaranteed a rather small estimation error term.

Choosing a description language (or, equivalently, some weighting of hypotheses) is a **weak form of committing** to a hypothesis. Rather than committing to a *single* hypothesis, we *spread* out our commitment among many. As long as it is done **independently** of the training sample, our generalization bound holds. Just as the choice of a *single hypothesis* to be evaluated by a sample can be arbitrary, so is *the choice of description language*.

## References

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.