# Chapter 7

### Motivation

So far, the notions of PAC learnability are uniform with respect to the labeling rule and the underlying data distribution.

Recall that:

> The term *uniform* refers to having a fixed sample size that works for all members of $H$ and over all possible probability distributions over the domain.

We want to consider weaker notions of learnability in this chapter.

## Nonuniform Learnability

*Nonuniform learnability* allows the sample size to be nonuniform w.r.t. the different hypotheses with which the learner is competing.

A hypothesis $h$ is $(\epsilon, \delta)$-competitive with another hypothesis $\eta$ if, with probability higher than $(1-\delta)$, $L_D(h) \leq L_D(\eta) + \epsilon$.

In nonuniform learnability, we allow the sample size to be of the form $m_H(\epsilon, \delta, h)$, so it now depends on the hypothesis with which we are competing.

### Definition 7.1

> A hypothesis class $H$ is nonuniformly learnable if there exists a learning algorithm, $A$, and a function $m_H^{NUL} : (0,1)^2 \times H \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$ and for every $h \in H$, if $m \geq m_H^{NUL}(\epsilon, \delta, h)$, then for every distribution $D$, with probability of at least $1-\delta$ over the choice of $S \sim D^m$, it holds that
>
> $$L_D(A(S)) \leq L_D(h) + \epsilon$$

In both agnostic PAC learning and nonuniform learnability, we require that the output hypothesis will be $(\epsilon, \delta)$-competitive with every other hypothesis in the class. But the difference is that, in nonuniform learnability, the sample size $m$ may depend on the hypothesis to which the error of $A(S)$ is compared.

Nonuniform learnability is a relaxation of agnostic PAC learnability.

### Theorem 7.2

> A hypothesis class of binary classifiers is nonuniformly learnable iff it is a countable union of agnostic PAC learnable hypothesis classes.

### Theorem 7.3

> Let H be a hypothesis class that can be written as a countable union of hypothesis classes, $H = \cup_{n \in \mathbb{N}} H_n$, where each $H_n$ enjoys the uniform convergence property. Then, $H$ is nonuniformly learnable.

# Structural Risk Minimisation

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

Another way to express our prior knowledge is by specifying preferences over hypothesis *within* $H$.

In the Structural Risk Minimisation (SRM) paradigm, we do so by first assuming that $H$ can be written as $H = \cup_{n \in \mathbb{N}} H_n$ and then specifying a weight function, $w : \mathbb{N} \to [0, 1]$, which assigns a weight to each hypothesis class, $H_n$, such that a higher weight reflects a stronger preference for the hypothesis class.

Let $H = \cup_{n \in \mathbb{N}} H_n$. assume that for each $n$, the class $H_n$ enjoys the uniform convergence property with a sample complexity function $m^{UC}_{H_n}(\epsilon, \delta)$.

Define the function $\epsilon_n : \mathbb{N} \times (0, 1) \to (0, 1)$:

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0, 1) : m^{UC}_{H_n}(\epsilon, \delta) \leq m\}$$

So, for a given sample size $m$, we are interested in the lowest $(\min)$ possible upper bound $(\epsilon)$ on the gap between the empirical and true risk achievable by using some sample of size $m$.

From the definitions of uniform convergence and $\epsilon_n$, it follows that for every $m$ and $\delta$, with probability of at least $1 - \delta$ over the choice of $S \sim D^m$ we have that

$$\forall h \in H_n, |L_D(h) - L_S(h)| \leq \epsilon_n(m, \delta)$$

#### Weighting Function

Let $w : \mathbb{N} \to [0, 1]$ be a function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. It is a *weight function* over the hypothesis classes $H_1, H_2, \ldots$.

Such a weight function can reflect the importance that the learner attributes to each hypothesis class, or some measure of the compliexty of different hypothesis classes. If $H$ is a finite union of $N$ hypothesis classes, one can simply assign the same weight of $1/N$ to all hypothesis classes. This equal weighting

corresponds to not including any prior knowledge/preference of any class. If you believe that some class is more likely to contain the correct target function, then it should be assigned a larger weight. When $H$ is a countably infinite union of hypothesis classes, a uniform weighting is not possible but other weighting schemes may work.

The SRM rule follows a *bound minimisation* approach. This means that the goal of the paradigm is to find a hypothesis which minimises a certain upper bound on the true risk.

## Theorem 7.4

> Assume we have a weighting function as defined above. Let $H$ be a hypothesis class that can be written as a countable union of hypothesis classes, where each class satisfies the uniform convergence property with a sample complexity function $m_{H_n}^{UC}$. Assume we have a function $\epsilon_n$ as defined above. Then, for every $\delta \in (0,1)$ and distribution $D$, with probability of at least $1-\delta$ over the choice of $S \sim D^m$, the following bound holds (simulatenously) for every $n \in \mathbb{N}$ and $h \in H_n$.
>
> $$|L_D(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$
>
> Therefore, for every $\delta \in (0,1)$ and distribution $D$, with probability of at least $1-\delta$ it holds that $\forall h \in H, L_D(h) \leq L_S(h) + \min_{n:h \in H_n} \epsilon_n(m, w(n) \cdot \delta)$.