



# Computational Learning Theory

## Table of Contents

12.1 Basic Knowledge	288
12.2 PAC Learning	289
12.3 Finite Hypothesis Space	292
12.4 VC Dimension	295
12.5 Rademacher Complexity	300
12.6 Stability	306
12.7 Further Reading	309
References	313

## 12.1 Basic Knowledge

As the name suggests, *computational learning theory* is about “learning” by “computation” and is the theoretical foundation of machine learning. It aims to analyze the difficulties of learning problems, provides theoretical guarantees for learning algorithms, and guides the algorithm design based on theoretical analysis.

Given a data set  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ , where  $\mathbf{x}_i \in \mathcal{X}$ . In this chapter, we focus on binary classification problems (i.e.,  $y_i \in \mathcal{Y} = \{-1, +1\}$ ) unless otherwise stated. Suppose there is an underlying unknown distribution  $\mathcal{D}$  over all samples in  $\mathcal{X}$ , and all samples in  $D$  are drawn independently from the distribution  $\mathcal{D}$ , that is, *i.i.d.* samples.

Let  $h$  be a mapping from  $\mathcal{X}$  to  $\mathcal{Y}$ , and its generalization error is

$$E(h; \mathcal{D}) = P_{\mathbf{x} \sim \mathcal{D}}(h(\mathbf{x}) \neq y). \quad (12.1)$$

The empirical error of  $h$  over  $D$  is

$$\widehat{E}(h; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(h(\mathbf{x}_i) \neq y_i). \quad (12.2)$$

Since  $D$  contains *i.i.d.* samples drawn from  $\mathcal{D}$ , the expectation of the empirical error of  $h$  equals to the generalization error. When it is clear from the context, we abbreviate  $E(h; \mathcal{D})$  and  $\widehat{E}(h; D)$  as  $E(h)$  and  $\widehat{E}(h)$ , respectively. The maximum error we can tolerate for a learned model, also known as the *error parameter*, is an upper bound of  $E(h)$ , denoted by  $\epsilon$ , where  $E(h) \leq \epsilon$ .

The rest of this chapter studies the gap between the empirical error and the generalization error. A mapping  $h$  is said to be consistent with  $D$  if the empirical error of  $h$  on the data set  $D$  is 0. For any two mappings  $h_1, h_2 \in \mathcal{X} \rightarrow \mathcal{Y}$ , their difference can be measured by the *disagreement*

$$d(h_1, h_2) = P_{\mathbf{x} \sim \mathcal{D}}(h_1(\mathbf{x}) \neq h_2(\mathbf{x})). \quad (12.3)$$

For ease of reference, we list a few frequently used inequalities below

- **Jensen’s inequality:** for every convex function  $f(x)$ , we have

$$f(\mathbb{E}(x)) \leq \mathbb{E}(f(x)). \quad (12.4)$$

## 12.1 Basic Knowledge

- **Hoeffding's inequality** (Hoeffding 1963): if  $x_1, x_2, \dots, x_m$  are  $m$  independent random variables with  $0 \leq x_i \leq 1$ , then, for any  $\epsilon > 0$ , we have

$$P\left(\frac{1}{m} \sum_{i=1}^m x_i - \frac{1}{m} \sum_{i=1}^m \mathbb{E}(x_i) \geq \epsilon\right) \leq \exp(-2m\epsilon^2), \quad (12.5)$$

$$P\left(\left|\frac{1}{m} \sum_{i=1}^m x_i - \frac{1}{m} \sum_{i=1}^m \mathbb{E}(x_i)\right| \geq \epsilon\right) \leq 2 \exp(-2m\epsilon^2). \quad (12.6)$$

- **McDiarmid's inequality** (McDiarmid 1989): if  $x_1, x_2, \dots, x_m$  are  $m$  independent random variables, and for any  $1 \leq i \leq m$ , the function  $f$  satisfies

$$\sup_{x_1, \dots, x_m, x'_i} |f(x_1, \dots, x_m) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c_i,$$

then, for any  $\epsilon > 0$ , we have

$$P(f(x_1, \dots, x_m) - \mathbb{E}(f(x_1, \dots, x_m)) \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right), \quad (12.7)$$

$$P(|f(x_1, \dots, x_m) - \mathbb{E}(f(x_1, \dots, x_m))| \geq \epsilon) \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right). \quad (12.8)$$

## 12.2 PAC Learning

Probably Approximately Correct (PAC) learning theory (Valiant 1984) is one of the most fundamental components of computational learning theory.

Let  $c$  denote a *concept*, which provides a mapping from the sample space  $\mathcal{X}$  to the label space  $\mathcal{Y}$ , and  $c$  determines the ground-truth label  $y$  of the sample  $x$ . A concept  $c$  is said to be a target concept if  $c(x) = y$  holds for every sample  $(x, y)$ . The set of all target concepts that we wish to learn is called a *concept class*, denoted by  $\mathcal{C}$ .

The set of all possible concepts for a given learning algorithm  $\mathcal{L}$  is called a *hypothesis space*, denoted by  $\mathcal{H}$ . Since the ground-truth concept class is unknown to learning algorithms,  $\mathcal{H}$  and  $\mathcal{C}$  are usually different. A learning algorithm constructs  $\mathcal{H}$  by collecting all concepts that are believed to be the target concepts. Since it is unknown whether the collected concepts are ground-truth target concepts,  $h \in \mathcal{H}$  is referred to as a *hypothesis*.

The hypothesis space of a learning algorithm  $\mathcal{L}$  is different from the hypothesis space of the learning problem as discussed in Sect. 1.3.

*esis*, which provides a mapping from the sample space  $\mathcal{X}$  to the label space  $\mathcal{Y}$ .

If  $c \in \mathcal{H}$ , then  $\mathcal{H}$  contains a hypothesis that can correctly classify all instances, and such a learning problem is said to be *separable* or *consistent* with respect to the learning algorithm  $\mathcal{L}$ . If  $c \notin \mathcal{H}$ , then  $\mathcal{H}$  does not contain any hypothesis that can correctly classify all instances, and such a learning problem is said to be *non-separable* or *inconsistent* with respect to the learning algorithm  $\mathcal{L}$ .

Given a training set  $D$ , we wish the learning algorithm  $\mathcal{L}$  can learn a hypothesis  $h$  that is close to the target concept  $c$ . Readers may wonder why not learn the exact target concept  $c$ ? The reason is that the machine learning process is subject to many factors. For example, since the training set  $D$  usually contains finite samples, there often exist many *equivalent hypotheses* that cannot be distinguished by learning algorithms on  $D$ . Also, there exists some randomness when sampling  $D$  from  $\mathcal{D}$ , and hence the hypotheses learned from different equal-sized training sets could be different. Therefore, instead of learning the exact target concept  $c$ , we wish to learn a hypothesis  $h$  with an error bounded by a given value with high confidence, that is, a hypothesis that is probably approximately correct (i.e., PAC). Let  $1 - \delta$  denote the confidence, and we have the formal definition as follows:

**Definition 12.1 (PAC Identify)** A learning algorithm  $\mathcal{L}$  is said to PAC identify the concept class  $\mathcal{C}$  from the hypothesis space  $\mathcal{H}$  if, for any  $c \in \mathcal{C}$  and distribution  $\mathcal{D}$ , and  $\epsilon, \delta \in (0, 1)$ , the learning algorithm  $\mathcal{L}$  outputs a hypothesis  $h \in \mathcal{H}$  satisfying

$$P(E(h) \leq \epsilon) \geq 1 - \delta. \quad (12.9)$$

Such a learning algorithm  $\mathcal{L}$  has a probability of at least  $1 - \delta$  of learning an approximation of the target concept  $c$  with an error of at most  $\epsilon$ . Following Definition 12.1, we can further define the following:

**Definition 12.2 (PAC Learnable)** A target concept class  $\mathcal{C}$  is said to be PAC learnable with respect to the hypothesis space  $\mathcal{H}$  if there exists a learning algorithm  $\mathcal{L}$  such that, for any  $\epsilon, \delta \in (0, 1)$  and distribution  $\mathcal{D}$ , the learning algorithm  $\mathcal{L}$  can PAC identify the concept class  $\mathcal{C}$  from the hypothesis space  $\mathcal{H}$  for any  $m \geq \text{poly}(1/\epsilon, 1/\delta, \text{size}(\mathbf{x}), \text{size}(c))$ , where  $\text{poly}(\cdot, \cdot, \cdot, \cdot)$  is a polynomial function and  $m$  is the number of *i.i.d.* training samples drawn from the distribution  $\mathcal{D}$ .

See Sect. 1.4.

In general, the fewer the training samples, the higher the randomness.

The sample size  $m$  is related to the error  $\epsilon$ , the confidence  $1 - \delta$ , the complexity of data  $\text{size}(\mathbf{x})$ , and the complexity of target concept  $\text{size}(c)$ .

For learning algorithms, it is necessary to consider the running time complexity. Hence, we further define:

**Definition 12.3** (*PAC Learning Algorithm*) A concept class  $\mathcal{C}$  is said to be efficiently PAC learnable by its PAC learning algorithm  $\mathcal{L}$  if  $\mathcal{C}$  is PAC learnable by  $\mathcal{L}$  within a polynomial time  $\text{poly}(1/\epsilon, 1/\delta, \text{size}(\mathbf{x}), \text{size}(c))$ .

Suppose the learning algorithm  $\mathcal{L}$  processes each sample with a constant time, then the running time complexity is equivalent to the sample complexity, and we could focus only on the sample complexity:

**Definition 12.4** (*Sample Complexity*) The sample complexity of a PAC learning algorithm  $\mathcal{L}$  is the smallest sample size  $m \geq \text{poly}(1/\epsilon, 1/\delta, \text{size}(\mathbf{x}), \text{size}(c))$  required by  $\mathcal{L}$ .

PAC learning provides a formal framework for describing the learning ability of learning algorithms, and many important questions can be discussed theoretically under this framework. For example, what are the requirements for learning a good model for a given learning problem? What are the conditions for an algorithm to learn effectively? How many training samples are required to learn a good model?

A hypothesis space  $\mathcal{H}$  includes all possible output hypotheses of a learning algorithm  $\mathcal{L}$ , and a key element of PAC learning is the complexity of  $\mathcal{H}$ . If the hypothesis space is the same as the concept class (i.e.,  $\mathcal{H} = \mathcal{C}$ ), then  $\mathcal{C}$  is said to be *properly PAC learnable* with respect to  $\mathcal{H}$ . Intuitively, it means the ability of the learning algorithm properly matches the learning problem. However, it is impractical to assume that  $\mathcal{H} = \mathcal{C}$  since we do not know the concept class for real problems, let alone some learning algorithm  $\mathcal{L}$  with  $\mathcal{H}$  is exact  $\mathcal{C}$ . Therefore, it is more realistic to study the cases when the hypothesis space and the concept class are different (i.e.,  $\mathcal{H} \neq \mathcal{C}$ ). In general, a larger  $\mathcal{H}$  is more likely to contain the target concept we are looking for, though the larger hypothesis space also makes it more difficult to find the target concept.  $\mathcal{H}$  is called a *finite hypothesis space* if  $|\mathcal{H}|$  is finite, and an *infinite hypothesis space* otherwise.

## 12.3 Finite Hypothesis Space

### 12.3.1 Separable Case

In separable cases, the target concept  $c$  is in the hypothesis space  $\mathcal{H}$  (i.e.,  $c \in \mathcal{H}$ ). Then, given a training set  $D$  with size  $m$ , how can we find a hypothesis from  $\mathcal{H}$  satisfying the constraint of a given error parameter?

It is natural to come up with the following learning strategy. Since the labels of the samples in  $D$  are assigned by the target concept  $c \in \mathcal{H}$ , any hypotheses that misclassify any samples in  $D$  must not be the target concept  $c$ . Hence, we simply eliminate all hypotheses that are inconsistent with  $D$  and keep the rest. When the training set  $D$  is sufficiently large, we can keep eliminating inconsistent hypotheses from  $\mathcal{H}$  until there is only one hypothesis left, which must be the target concept  $c$ . In practice, however, since the training data is usually limited, we may end up with more than one hypothesis that is consistent with  $D$ , and we cannot distinguish them without additional information.

Given that the training data is limited, how many samples do we need to learn a good approximation of the target concept  $c$ ? For PAC learning, we say a training set  $D$  is sufficient for a learning algorithm  $\mathcal{L}$  if  $\mathcal{L}$  can find an  $\epsilon$ -approximation of the target concept with a probability of at least  $1 - \delta$ .

We first estimate the probability of having a hypothesis that performs perfectly on the training set but still with a generalization error greater than  $\epsilon$ . Suppose the generalization error of a hypothesis  $h$  is greater than  $\epsilon$ , then, for any *i.i.d.* sample  $(\mathbf{x}, y)$  drawn from the distribution  $\mathcal{D}$ , we have

$$\begin{aligned} P(h(\mathbf{x}) = y) &= 1 - P(h(\mathbf{x}) \neq y) \\ &= 1 - E(h) \\ &< 1 - \epsilon. \end{aligned} \tag{12.10}$$

Since  $D$  contains  $m$  samples independently drawn from  $\mathcal{D}$ , the probability that  $h$  and  $D$  are consistent is given by

$$\begin{aligned} P((h(\mathbf{x}_1) = y_1) \wedge \dots \wedge (h(\mathbf{x}_m) = y_m)) &= (1 - P(h(\mathbf{x}) \neq y))^m \\ &< (1 - \epsilon)^m. \end{aligned} \tag{12.11}$$

Though we do not know which hypothesis  $h \in \mathcal{H}$  will be the output by the learning algorithm  $\mathcal{L}$ , we only need to ensure that the total probability of having any hypotheses that are consistent with  $D$  and have generalization errors greater than  $\epsilon$  is not greater than  $\delta$ . That is, ensuring the total probability

## 12.3 Finite Hypothesis Space

$$\begin{aligned} P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0) &< |\mathcal{H}| (1 - \epsilon)^m \\ &< |\mathcal{H}| e^{-m\epsilon} \end{aligned} \quad (12.12)$$

is not greater than  $\delta$ , that is,

$$|\mathcal{H}| e^{-m\epsilon} \leq \delta. \quad (12.13)$$

Hence, we have

$$m \geq \frac{1}{\epsilon} \left( \ln |\mathcal{H}| + \ln \frac{1}{\delta} \right), \quad (12.14)$$

which shows that every finite hypothesis space  $\mathcal{H}$  is PAC learnable, and the required sample size is given by (12.14). As the number of samples increases, the generalization error of the output hypothesis  $h$  converges toward 0 at a convergence rate of  $O(\frac{1}{m})$ .

### 12.3.2 Non-separable Case

For difficult learning problems, the target concept  $c$  is usually not in the hypothesis space  $\mathcal{H}$ . Suppose  $\widehat{E}(h) \neq 0$  for any  $h \in \mathcal{H}$ , that is, every hypothesis in  $\mathcal{H}$  misclassifies at least one training example, then, from Hoeffding's inequality, we have:

**Lemma 12.1** *Let  $D$  be a training set containing  $m$  samples independently drawn from a distribution  $\mathcal{D}$ . Then, for any  $h \in \mathcal{H}$  and  $0 < \epsilon < 1$ , we have*

$$P(\widehat{E}(h) - E(h) \geq \epsilon) \leq \exp(-2m\epsilon^2), \quad (12.15)$$

$$P(E(h) - \widehat{E}(h) \geq \epsilon) \leq \exp(-2m\epsilon^2), \quad (12.16)$$

$$P(|E(h) - \widehat{E}(h)| \geq \epsilon) \leq 2 \exp(-2m\epsilon^2). \quad (12.17)$$

**Corollary 12.1** *Let  $D$  be a training set containing  $m$  samples independently drawn from a distribution  $\mathcal{D}$ . Then, for any  $h \in \mathcal{H}$  and  $0 < \epsilon < 1$ , the following holds with a probability of at least  $1 - \delta$ :*

$$\widehat{E}(h) - \sqrt{\frac{\ln(2/\delta)}{2m}} \leq E(h) \leq \widehat{E}(h) + \sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (12.18)$$

Corollary 12.1 shows that, for a large  $m$ , the empirical error of  $h$  is a good approximation to its generalization error. For finite hypothesis spaces, we have

**Theorem 12.1** *Let  $\mathcal{H}$  be a finite hypothesis space. Then, for any  $h \in \mathcal{H}$  and  $0 < \delta < 1$ , we have*

$$P\left(|E(h) - \widehat{E}(h)| \leq \sqrt{\frac{\ln |\mathcal{H}| + \ln(2/\delta)}{2m}}\right) \geq 1 - \delta. \quad (12.19)$$

**Proof** Let  $h_1, h_2, \dots, h_{|\mathcal{H}|}$  denote the hypotheses in  $\mathcal{H}$ , and we have

$$\begin{aligned} & P(\exists h \in \mathcal{H} : |E(h) - \widehat{E}(h)| > \epsilon) \\ &= P((|E_{h_1} - \widehat{E}_{h_1}| > \epsilon) \vee \dots \vee (|E_{h_{|\mathcal{H}|}} - \widehat{E}_{h_{|\mathcal{H}|}}| > \epsilon)) \\ &\leq \sum_{h \in \mathcal{H}} P(|E(h) - \widehat{E}(h)| > \epsilon). \end{aligned}$$

From (12.17), we have

$$\sum_{h \in \mathcal{H}} P(|E(h) - \widehat{E}(h)| > \epsilon) \leq 2|\mathcal{H}| \exp(-2m\epsilon^2),$$

which proves (12.19) by letting  $\delta = 2|\mathcal{H}| \exp(-2m\epsilon^2)$ .  $\square$

That is to find the best hypothesis in  $\mathcal{H}$ .

A learning algorithm  $\mathcal{L}$  cannot learn an  $\epsilon$ -approximation of the target concept  $c$  if  $c \notin \mathcal{H}$ . However, for a given hypothesis space  $\mathcal{H}$ , the hypothesis  $h \in \mathcal{H}$  with the smallest generalization error is still a reasonably good target. In other words, instead of targeting at  $c$ , we find an  $\epsilon$ -approximation of  $h$ , i.e.,  $\arg \min_{h \in \mathcal{H}} E(h)$ . This approach generalizes PAC learning to *agnostic learning* in which  $c \notin \mathcal{H}$ . Accordingly, we define

**Definition 12.5** (*Agnostic PAC learnable*) A hypothesis space  $\mathcal{H}$  is said to be agnostic PAC learnable if there exists a learning algorithm  $\mathcal{L}$  such that, for any  $\epsilon, \delta \in (0, 1)$  and distribution  $\mathcal{D}$ , the learning algorithm  $\mathcal{L}$  outputs a hypothesis  $h \in \mathcal{H}$  satisfying

$$P(E(h) - \min_{h' \in \mathcal{H}} E(h') \leq \epsilon) \geq 1 - \delta, \quad (12.20)$$

for any  $m \geq \text{poly}(1/\epsilon, 1/\delta, \text{size}(\mathbf{x}), \text{size}(c))$ , where  $m$  is the number of *i.i.d.* training samples drawn from the distribution  $\mathcal{D}$ .

Similar to PAC learnable, a hypothesis space  $\mathcal{H}$  is said to be efficiently agnostic PAC learnable by its agnostic PAC learning algorithm  $\mathcal{L}$  if  $\mathcal{H}$  is agnostic PAC learnable by  $\mathcal{L}$  within a polynomial time  $\text{poly}(1/\epsilon, 1/\delta, \text{size}(\mathbf{x}), \text{size}(c))$ . The sample complexity of the learning algorithm  $\mathcal{L}$  is the smallest sample size  $m$  satisfying the above requirements.



## 12.4 VC Dimension

Hypothesis spaces in real-world applications are usually infinite, such as all intervals in the real domain and all hyperplanes in the  $\mathbb{R}^d$  space. To study the learnability of such cases, we need to measure the complexity of hypothesis spaces. A general approach is to consider the *Vapnik–Chervonenkis dimension* (VC dimension) (Vapnik and Chervonenkis 1971). We first introduce three concepts: *growth function*, *dichotomy*, and *shattering*.

Given a hypothesis space  $\mathcal{H}$  and a set of instances  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ , where each hypothesis  $h \in \mathcal{H}$  can label every instance in  $D$ . The labeling result is denoted by

$$h|_D = \{(h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_m))\},$$

any element of which is called a *dichotomy*. The number of dichotomies generated by the hypotheses in  $\mathcal{H}$  over  $D$  increases as  $m$  increases.

**Definition 12.6** For  $m \in \mathbb{N}$ , the growth function  $\Pi_{\mathcal{H}}(m)$  of a hypothesis space  $\mathcal{H}$  is defined as

$$\Pi_{\mathcal{H}}(m) = \max_{\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subseteq \mathcal{X}} |\{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_m)) \mid h \in \mathcal{H}\}|. \quad (12.21)$$

The growth function  $\Pi_{\mathcal{H}}(m)$  gives the largest number of dichotomies that the hypothesis space  $\mathcal{H}$  can generate over  $m$  instances. The more dichotomies, the more representation power, that is, the better adaptability to learning problems. The growth function describes the representation power of a hypothesis space  $\mathcal{H}$ , which also reflects the complexity of the hypothesis space. We can now use a growth function to present the relationship between the empirical error and the generalization error:

**Theorem 12.2** For any  $m \in \mathbb{N}$ ,  $0 < \epsilon < 1$ , and  $h \in \mathcal{H}$ , we have

$$P(|E(h) - \widehat{E}(h)| > \epsilon) \leq 4\Pi_{\mathcal{H}}(2m) \exp\left(-\frac{m\epsilon^2}{8}\right). \quad (12.22)$$

Different hypotheses in  $\mathcal{H}$  may generate identical or different dichotomies over  $D$ . The number of dichotomies could be finite even for an infinite hypothesis space  $\mathcal{H}$ ; for example, there are at most  $2^m$  dichotomies over  $m$  instances. We say that a hypothesis

For example, in binary classification problems, there are at most 4 dichotomies given 2 instances, and 8 dichotomies given 3 instances.

$\mathbb{N}$  is the natural number domain.

The proof can be found in Vapnik and Chervonenkis (1971).

space  $\mathcal{H}$  can *shatter* a data set  $D$  if  $\mathcal{H}$  can generate all possible dichotomies of  $D$ , that is,  $\Pi_{\mathcal{H}}(m) = 2^m$ .

We can now formally define the VC dimension as follows:

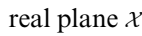
**Definition 12.7** The VC dimension of a hypothesis space  $\mathcal{H}$  is the size of the largest instance set  $D$  shattered by  $\mathcal{H}$ :

$$\text{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}. \quad (12.23)$$

$\text{VC}(\mathcal{H}) = d$  says that there exists an instance set  $D$  of size  $d$  that can be shattered by  $\mathcal{H}$ . However, it does not mean every instance set  $D$  of size  $d$  can be shattered by  $\mathcal{H}$ . Some readers may have recognized that the definition of the VC dimension does not involve the underlying data distribution  $\mathcal{D}$ ! In other words, the VC dimension of a hypothesis space  $\mathcal{H}$  can be calculated even if the data distribution is unknown.

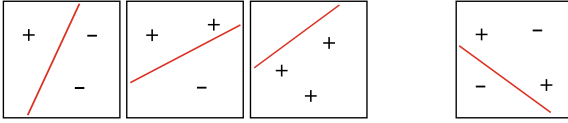
In general, we can calculate the VC dimension of  $\mathcal{H}$  as follows: the VC dimension of  $\mathcal{H}$  is  $d$  if there exists an instance set of size  $d$  shattered by  $\mathcal{H}$  while there is no instance set of size  $d + 1$  shattered by  $\mathcal{H}$ . We illustrate the calculation of the VC dimension with the following two examples:

**Example 12.1** (*Interval* $[a, b]$  in the real domain) Let  $\mathcal{H} = \{h_{[a,b]} : a, b \in \mathbb{R}, a \leq b\}$  denote the set of all closed intervals in the real domain  $\mathcal{X} = \mathbb{R}$ . For every  $x \in \mathcal{X}$ , we have  $h_{[a,b]}(x) = +1$  if  $x \in [a, b]$ ; otherwise,  $h_{[a,b]}(x) = -1$ . Letting  $x_1 = 0.5$  and  $x_2 = 1.5$ , then,  $\{x_1, x_2\}$  is shattered by the hypotheses  $\{h_{[0,1]}, h_{[0,2]}, h_{[1,2]}, h_{[2,3]}\}$  from  $\mathcal{H}$ , hence the VC dimension of  $\mathcal{H}$  is at least 2. However, there is no hypothesis  $h_{[a,b]} \in \mathcal{H}$  that can generate the dichotomy  $\{(x_3, +), (x_4, -), (x_5, +)\}$  for a data set containing any 3 instances  $\{x_3, x_4, x_5\}$ , where  $x_3 < x_4 < x_5$ . Hence, the VC dimension of  $\mathcal{H}$  is 2.

**Example 12.2** (*Linear separators in the 2-dimensional real plane*) Let  $\mathcal{H}$  denote the set of all linear separators in the 2-dimensional real plane  $\mathcal{X} = \mathbb{R}^2$ . From  Figure 12.1 we see that there exists a data set of size 3 shattered by  $\mathcal{H}$ , whereas there is no instance set of size 4 shattered by  $\mathcal{H}$ . Hence, the VC dimension of the hypothesis space  $\mathcal{H}$  of all linear separators in the 2-dimensional real plane is 3.

From Definition 12.7, we see the following relationship between the VC dimension and the growth function (Sauer 1972):

**Lemma 12.2** If the VC dimension of a hypothesis space  $\mathcal{H}$  is  $d$ , then, for any  $m \in \mathbb{N}$ , we have



All of the  $2^3=8$  dichotomies can be made by linear separators

(a) 3 instances.

At least one of the  $2^4=16$  dichotomies cannot be made by linear separators

(b) 4 instances.

**Fig. 12.1** The VC dimension of the hypothesis space of all linear separators in the 2-dimensional real plane is 3

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^d \binom{m}{i}. \quad (12.24)$$

Also known as *Sauer's Lemma*.

**Proof** We will proceed by induction. The theorem holds when  $m = 1$ , and  $d = 0$  or  $d = 1$ . Hypothesizing that the theorem holds for  $(m - 1, d - 1)$  and  $(m - 1, d)$ . Letting  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  and  $D' = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{m-1}\}$ , we have

$$\mathcal{H}_{|D} = \{(h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_m)) \mid h \in \mathcal{H}\},$$

$$\mathcal{H}_{|D'} = \{(h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_{m-1})) \mid h \in \mathcal{H}\}.$$

Since every hypothesis  $h \in \mathcal{H}$  classifies  $\mathbf{x}_m$  as either  $+1$  or  $-1$ , every sequence appeared in  $\mathcal{H}_{|D'}$  will appear in  $\mathcal{H}_{|D}$  once or twice. Let  $\mathcal{H}_{D'|D}$  denote the set of sequences from  $\mathcal{H}_{|D'}$  that appear twice in  $\mathcal{H}_{|D}$ , that is,

$$\mathcal{H}_{D'|D} = \{(y_1, y_2, \dots, y_{m-1}) \in \mathcal{H}_{|D'} \mid \exists h, h' \in \mathcal{H}, \\ (h(\mathbf{x}_i) = h'(\mathbf{x}_i) = y_i) \wedge (h(\mathbf{x}_m) \neq h'(\mathbf{x}_m)), 1 \leq i \leq m-1\}.$$

Since the sequences in  $\mathcal{H}_{D'|D}$  appear twice in  $\mathcal{H}_{|D}$  but once in  $\mathcal{H}_{|D'}$ , we have

$$|\mathcal{H}_{|D}| = |\mathcal{H}_{|D'}| + |\mathcal{H}_{D'|D}|. \quad (12.25)$$

For the data set  $D'$  of size  $m - 1$ , we have, from the induction assumption,

$$|\mathcal{H}_{|D'}| \leq \Pi_{\mathcal{H}}(m-1) \leq \sum_{i=0}^d \binom{m-1}{i}. \quad (12.26)$$

Let  $Q$  denote the set of instances shattered by  $\mathcal{H}_{D'|D}$ . From the definition of  $\mathcal{H}_{D'|D}$ , we know that  $\mathcal{H}_{|D}$  can shatter  $Q \cup \{\mathbf{x}_m\}$ . Since the VC dimension of  $\mathcal{H}$  is  $d$ , the largest possible VC dimension of  $\mathcal{H}_{D'|D}$  is  $d - 1$ . Therefore, we have

$$|\mathcal{H}_{D'|D}| \leq \Pi_{\mathcal{H}}(m-1) \leq \sum_{i=0}^{d-1} \binom{m-1}{i}. \quad (12.27)$$

From (12.25)–(12.27), we have

$$\begin{aligned} |\mathcal{H}_{|D}| &\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\ &= \sum_{i=0}^d \left( \binom{m-1}{i} + \binom{m-1}{i-1} \right) \\ &= \sum_{i=0}^d \binom{m}{i}. \end{aligned}$$

$\binom{m-1}{-1} = 0$ .

From the arbitrariness of data set  $D$ , Lemma 12.2 follows.  $\square$

From Lemma 12.2, we can calculate the upper bound of the growth function:

**Corollary 12.2** *If the VC dimension of a hypothesis space  $\mathcal{H}$  is  $d$ , then, for any integer  $m \geq d$ , we have*

$$\Pi_{\mathcal{H}}(m) \leq \left( \frac{e \cdot m}{d} \right)^d. \quad (12.28)$$

$e$  is Euler's number.

## 12

**Proof**

$$\begin{aligned} \Pi_{\mathcal{H}}(m) &\leq \sum_{i=0}^d \binom{m}{i} \\ &\leq \sum_{i=0}^d \binom{m}{i} \left( \frac{m}{d} \right)^{d-i} \\ &= \left( \frac{m}{d} \right)^d \sum_{i=0}^d \binom{m}{i} \left( \frac{d}{m} \right)^i \\ &\leq \left( \frac{m}{d} \right)^d \sum_{i=0}^m \binom{m}{i} \left( \frac{d}{m} \right)^i \\ &= \left( \frac{m}{d} \right)^d \left( 1 + \frac{d}{m} \right)^m \\ &\leq \left( \frac{e \cdot m}{d} \right)^d. \end{aligned}$$

$m \geq d$ .

$\square$

From Corollary 12.2 and Theorem 12.2, we have the generalization error bound in terms of the VC dimension, also known as the VC bound:

**Theorem 12.3** *If the VC dimension of a hypothesis space  $\mathcal{H}$  is  $d$ , then, for any  $m > d$ ,  $\delta \in (0, 1)$ , and  $h \in \mathcal{H}$ , we have*

$$P \left( |E(h) - \widehat{E}(h)| \leq \sqrt{\frac{8d \ln \frac{2em}{d} + 8 \ln \frac{4}{\delta}}{m}} \right) \geq 1 - \delta. \quad (12.29)$$

**Proof** Setting  $4\Pi_{\mathcal{H}}(2m) \exp(-\frac{m\epsilon^2}{8}) \leq 4(\frac{2em}{d})^d \exp(-\frac{m\epsilon^2}{8}) = \delta$ , we have

$$\epsilon = \sqrt{\frac{8d \ln \frac{2em}{d} + 8 \ln \frac{4}{\delta}}{m}},$$

which completes the proof by substituting the above equation into Theorem 12.2.  $\square$

From Theorem 12.3, the generalization error bound in (12.29) is dependent only on the sample size  $m$  and converges toward 0 at a convergence rate of  $O(\frac{1}{\sqrt{m}})$ . Since the VC bound is independent of the data distribution  $\mathcal{D}$  and the data set  $D$ , it is *distribution-free* and *data-independent*.

Let  $h$  denote the hypothesis output by a learning algorithm  $\mathcal{L}$ . Then, we say  $\mathcal{L}$  satisfies the Empirical Risk Minimization (ERM) principle if

$$\widehat{E}(h) = \min_{h' \in \mathcal{H}} \widehat{E}(h'). \quad (12.30)$$

Then, we have the following theorem:

**Theorem 12.4** *Every hypothesis space  $\mathcal{H}$  with a finite VC dimension is (agnostic) PAC learnable.*

**Proof** Suppose  $\mathcal{L}$  is a learning algorithm satisfying the ERM principle, and  $h$  is the hypothesis output by  $\mathcal{L}$ . Let  $g$  be the hypothesis with the smallest generalization error in  $\mathcal{H}$ , that is,

$$E(g) = \min_{h \in \mathcal{H}} E(h). \quad (12.31)$$

Letting

$$\delta' = \frac{\delta}{2},$$

$$\sqrt{\frac{(\ln 2/\delta')}{2m}} = \frac{\epsilon}{2}. \quad (12.32)$$

From Corollary 12.1, the following holds with a probability of at least  $1 - \delta/2$ :

$$\widehat{E}(g) - \frac{\epsilon}{2} \leq E(g) \leq \widehat{E}(g) + \frac{\epsilon}{2}. \quad (12.33)$$

Setting

$$\sqrt{\frac{8d \ln \frac{2em}{d} + 8 \ln \frac{4}{\delta'}}{m}} = \frac{\epsilon}{2}, \quad (12.34)$$

then, from Theorem 12.3, we have

$$P\left(E(h) - \widehat{E}(h) \leq \frac{\epsilon}{2}\right) \geq 1 - \frac{\delta}{2}. \quad (12.35)$$

Hence, the following holds with a probability of at least  $1 - \delta$ :

$$\begin{aligned} E(h) - E(g) &\leq \widehat{E}(h) + \frac{\epsilon}{2} - \left(\widehat{E}(g) - \frac{\epsilon}{2}\right) \\ &= \widehat{E}(h) - \widehat{E}(g) + \epsilon \\ &\leq \epsilon. \end{aligned}$$

We can solve  $m$  from (12.32) and (12.34). Then, from the arbitrariness of  $\mathcal{H}$ , we have Theorem 12.4.  $\square$

## 12.5 Rademacher Complexity

From Sect. 12.4, we see that the VC bound is distribution-free and data-independent (i.e., it is valid for any data distribution), which makes the analysis of generalization error bound “universal”. However, since it does not take the data set into account, the VC bound is generally loose, especially for “poor” data distributions that are far from the typical situation in learning problems.

*Rademacher complexity* presents another characterization of the complexity of the hypothesis space, and the difference from the VC dimension lies in consideration of data distribution in some sense.

Given a data set  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ , the empirical error of a hypothesis  $h$  is given by

*Rademacher complexity* is named after the German mathematician H. Rademacher (1892–1969).

$$\begin{aligned}
\widehat{E}(h) &= \frac{1}{m} \sum_{i=1}^m \mathbb{I}(h(\mathbf{x}_i) \neq y_i) \\
&= \frac{1}{m} \sum_{i=1}^m \frac{1 - y_i h(\mathbf{x}_i)}{2} \\
&= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^m y_i h(\mathbf{x}_i),
\end{aligned} \tag{12.36}$$

where  $\frac{1}{m} \sum_{i=1}^m y_i h(\mathbf{x}_i)$  represents the consistency between the predicted values  $h(\mathbf{x}_i)$  and the ground-truth labels  $y_i$ . It takes the maximum value 1 if  $h(\mathbf{x}_i) = y_i$  for all  $i \in \{1, 2, \dots, m\}$ . In other words, the hypothesis with the smallest empirical error is

$$\arg \max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m y_i h(\mathbf{x}_i). \tag{12.37}$$

In practice, however, the data set may have been corrupted by some noises, that is, the label  $y_i$  of sample  $(\mathbf{x}_i, y_i)$  is affected by some random factors and is no longer the ground-truth label of  $\mathbf{x}_i$ . In such cases, sometimes it is better to select a hypothesis that has considered the influence of random noises, rather than the best hypothesis over the training set.

We introduce the Rademacher random variable  $\sigma_i$ , which takes value  $+1$  or  $-1$  with an equal probability of 0.5. With  $\sigma_i$ , we rewrite (12.37) as

$$\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i). \tag{12.38}$$

It is likely that we cannot find the maximum value since  $\mathcal{H}$  is infinite. Hence, we replace the maximum by the supremum.

We consider all hypotheses in  $\mathcal{H}$  and take the expectation over (12.38) as

$$\mathbb{E}_{\sigma} \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i) \right], \tag{12.39}$$

where  $\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$ . Equation (12.39) takes value in  $[0, 1]$  and expresses the representation power of the hypothesis  $\mathcal{H}$ . For example, (12.39) equals to 0 when  $|\mathcal{H}| = 1$ , that is, there is only one hypothesis in  $\mathcal{H}$ ; (12.39) equals to 1 when  $|\mathcal{H}| = 2^m$  and  $\mathcal{H}$  shatters  $D$ , that is, for any  $\sigma$ , there exists a hypothesis such that  $h(\mathbf{x}_i) = \sigma_i$  ( $i = 1, 2, \dots, m$ ).

Let  $\mathcal{F} : \mathcal{Z} \rightarrow \mathbb{R}$  be a real-valued function space, and  $Z = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m\}$  be a set of *i.i.d.* instances, where  $\mathbf{z}_i \in \mathcal{Z}$ . By replacing  $\mathcal{X}$  and  $\mathcal{H}$  in (12.39) with  $\mathcal{Z}$  and  $\mathcal{F}$ , respectively, we have

**Definition 12.8** The empirical Rademacher complexity of a function space  $\mathcal{F}$  with respect to  $Z$  is defined as

$$\widehat{R}_Z(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right]. \quad (12.40)$$

The empirical Rademacher complexity measures the correlation between the function space  $\mathcal{F}$  and the random noise in the data set  $Z$ . To analyze the correlation between  $\mathcal{F}$  and  $\mathcal{D}$  over  $Z$ , we can take the expectation over the data set  $Z$  with  $m$  i.i.d. samples drawn from  $\mathcal{D}$ :

**Definition 12.9** The Rademacher complexity of a function space  $\mathcal{F}$  with respect to a distribution  $\mathcal{D}$  over  $Z$  is defined as

$$R_m(\mathcal{F}) = \mathbb{E}_{Z \subseteq \mathcal{Z}; |Z|=m} [\widehat{R}_Z(\mathcal{F})]. \quad (12.41)$$

Based on the Rademacher complexity, we can define the generalization error bound of function space  $\mathcal{F}$  (Mohri et al. 2012):

**Theorem 12.5** Let  $\mathcal{F} : \mathcal{Z} \rightarrow [0, 1]$  be a real-valued function space, and  $Z = \{z_1, z_2, \dots, z_m\}$  be a set of i.i.d. samples drawn from  $\mathcal{D}$  over  $\mathcal{Z}$ . Then, for any  $\delta \in (0, 1)$  and  $f \in \mathcal{F}$ , the following holds with a probability of at least  $1 - \delta$ :

$$\mathbb{E}[f(z)] \leq \frac{1}{m} \sum_{i=1}^m f(z_i) + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}, \quad (12.42)$$

$$\mathbb{E}[f(z)] \leq \frac{1}{m} \sum_{i=1}^m f(z_i) + 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (12.43)$$

**Proof** Letting

$$\begin{aligned} \widehat{E}_Z(f) &= \frac{1}{m} \sum_{i=1}^m f(z_i), \\ \Phi(Z) &= \sup_{f \in \mathcal{F}} \mathbb{E}[f] - \widehat{E}_Z(f), \end{aligned}$$

and let  $Z'$  be another data set that is the same as  $Z$  except for one instance. Suppose that  $z_m \in Z$  and  $z'_m \in Z'$  are the two different instances. Then, we have



## 12.5 Rademacher Complexity

$$\begin{aligned}
\Phi(Z') - \Phi(Z) &= \left( \sup_{f \in \mathcal{F}} \mathbb{E}[f] - \widehat{E}_{Z'}(f) \right) - \left( \sup_{f \in \mathcal{F}} \mathbb{E}[f] - \widehat{E}_Z(f) \right) \\
&\leq \sup_{f \in \mathcal{F}} \widehat{E}_Z(f) - \widehat{E}_{Z'}(f) \\
&= \sup_{f \in \mathcal{F}} \frac{f(z_m) - f(z'_m)}{m} \\
&\leq \frac{1}{m}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\Phi(Z) - \Phi(Z') &\leq \frac{1}{m}, \\
|\Phi(Z) - \Phi(Z')| &\leq \frac{1}{m}.
\end{aligned}$$

According to McDiarmid's inequality (12.7), for any  $\delta \in (0, 1)$ , the following holds with a probability of at least  $1 - \delta$ :

$$\Phi(Z) \leq \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(1/\delta)}{2m}}, \quad (12.44)$$

where the upper bound of  $\mathbb{E}_Z[\Phi(Z)]$  is given by

$$\begin{aligned}
\mathbb{E}_Z[\Phi(Z)] &= \mathbb{E}_Z \left[ \sup_{f \in \mathcal{F}} \mathbb{E}[f] - \widehat{E}_Z(f) \right] \\
&= \mathbb{E}_Z \left[ \sup_{f \in \mathcal{F}} \mathbb{E}_{Z'} [\widehat{E}_{Z'}(f) - \widehat{E}_Z(f)] \right] \\
&\leq \mathbb{E}_{Z, Z'} \left[ \sup_{f \in \mathcal{F}} \widehat{E}_{Z'}(f) - \widehat{E}_Z(f) \right] \\
&= \mathbb{E}_{Z, Z'} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z'_i) - f(z_i)) \right] \\
&= \mathbb{E}_{\sigma, Z, Z'} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z'_i) - f(z_i)) \right] \\
&\leq \mathbb{E}_{\sigma, Z'} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] + \mathbb{E}_{\sigma, Z} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m -\sigma_i f(z_i) \right] \\
&= 2\mathbb{E}_{\sigma, Z} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] \\
&= 2R_m(\mathcal{F}).
\end{aligned}$$

Using Jensen's inequality (12.4) and the convexity of the supremum function.

$\sigma_i$  and  $-\sigma_i$  follow the same distribution.

The above gives the proof of (12.42). From Definition 12.9, we know that changing one instance in  $Z$  will change the value of  $\widehat{R}_Z(\mathcal{F})$  at most  $1/m$ . According to McDiarmid's inequality (12.7), the following holds with a probability of at least  $1 - \delta/2$ :

$$R_m(\mathcal{F}) \leq \widehat{R}_Z(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (12.45)$$

Then, from (12.44), the following holds with a probability of at least  $1 - \delta/2$ :

$$\Phi(Z) \leq \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

Hence, the following holds with a probability of at least  $1 - \delta$ :

$$\Phi(Z) \leq 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (12.46)$$

The above gives the proof of (12.43).  $\square$

Since  $\mathcal{F}$  in Theorem 12.5 is a real-valued function over the interval  $[0, 1]$ , Theorem 12.5 is applicable to regression problems only. For binary classification problems, we have the following theorem:

**Theorem 12.6** *Let  $\mathcal{H} : \mathcal{X} \rightarrow \{-1, +1\}$  be a hypothesis space and  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  be a set of i.i.d. instances drawn from  $\mathcal{D}$  over  $\mathcal{X}$ . Then, for any  $\delta \in (0, 1)$  and  $h \in \mathcal{H}$ , the following holds with a probability of at least  $1 - \delta$ :*

$$E(h) \leq \widehat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}, \quad (12.47)$$

$$E(h) \leq \widehat{E}(h) + \widehat{R}_D(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (12.48)$$

**Proof** Let  $\mathcal{H}$  be a hypothesis space of binary classification problems. By letting  $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$ ,  $h \in \mathcal{H}$  can be transformed to

$$f_h(\mathbf{z}) = f_h(\mathbf{x}, y) = \mathbb{I}(h(\mathbf{x}) \neq y), \quad (12.49)$$

which transforms the hypothesis space  $\mathcal{H}$  with an output domain of  $\{-1, +1\}$  to a function space  $\mathcal{F}_{\mathcal{H}} = \{f_h : h \in \mathcal{H}\}$  with an output domain of  $[0, 1]$ . From Definition 12.8, we have

$$\begin{aligned}
\widehat{R}_Z(\mathcal{F}_\mathcal{H}) &= \mathbb{E}_\sigma \left[ \sup_{f_h \in \mathcal{F}_\mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i f_h(\mathbf{x}_i, y_i) \right] \\
&= \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i \mathbb{I}(h(\mathbf{x}_i) \neq y_i) \right] \\
&= \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i \frac{1 - y_i h(\mathbf{x}_i)}{2} \right] \\
&= \frac{1}{2} \mathbb{E}_\sigma \left[ \frac{1}{m} \sum_{i=1}^m \sigma_i + \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (-y_i \sigma_i h(\mathbf{x}_i)) \right] \\
&= \frac{1}{2} \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (-y_i \sigma_i h(\mathbf{x}_i)) \right] \\
&= \frac{1}{2} \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (\sigma_i h(\mathbf{x}_i)) \right] && -y_i \sigma_i \text{ and } \sigma_i \text{ follow the same} \\
&= \frac{1}{2} \widehat{R}_D(\mathcal{H}). && \text{distribution.}
\end{aligned} \tag{12.50}$$

Taking the expectation of (12.50) gives

$$R_m(\mathcal{F}_\mathcal{H}) = \frac{1}{2} R_m(\mathcal{H}). \tag{12.51}$$

From (12.50), (12.51), and Theorem 12.5, we have Theorem 12.6 proved.  $\square$

Theorem 12.6 gives the generalization error bound based on the Rademacher complexity, also known as the Rademacher bound. In comparison with Theorem 12.3, the VC bound is distribution-free and data-independent, whereas the Rademacher bound depends on the distribution  $\mathcal{D}$  in (12.47) and the data set  $D$  in (12.48). In other words, the Rademacher bound depends on the data distribution of the specific learning problem. The Rademacher bound is generally tighter than the VC bound since it is “tailored” for the specific learning problem.

For the Rademacher complexity and the growth function, we have

**Theorem 12.7** *The Rademacher complexity  $R_m(\mathcal{H})$  and the growth function  $\Pi_\mathcal{H}(m)$  of a hypothesis space  $\mathcal{H}$  satisfy* See Mohri et al. (2012) for proof.

$$R_m(\mathcal{H}) \leq \sqrt{\frac{2 \ln \Pi_\mathcal{H}(m)}{m}}. \tag{12.52}$$

From (12.47), (12.52), and Corollary 12.2, we have

$$E(h) \leq \widehat{E}(h) + \sqrt{\frac{2d \ln \frac{em}{d}}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}. \quad (12.53)$$

In other words, we can derive the VC bound from the Rademacher complexity and the growth function.

## 12.6 Stability

The generalization error bound, based on either the VC dimension or Rademacher complexity, is independent of the specific learning algorithm. Hence, the analysis applies to all learning algorithms and enables us to study the nature of learning problems without considering specific design of learning algorithms. However, if we wish the analysis to be algorithm-dependent, then we need to take a different approach, and one direction is stability analysis.

As the name suggests, the *stability* of an algorithm concerns about whether a minor change of the input will cause a significant change in the output. The input of learning algorithms is a data set, so we need to define the changes on data sets.

Given a data set  $D = \{z_1 = (\mathbf{x}_1, y_1), z_2 = (\mathbf{x}_2, y_2), \dots, z_m = (\mathbf{x}_m, y_m)\}$ , where  $\mathbf{x}_i \in \mathcal{X}$  are *i.i.d.* instances drawn from distribution  $\mathcal{D}$  and  $y_i \in \{-1, +1\}$ . Let  $\mathcal{H} : \mathcal{X} \rightarrow \{-1, +1\}$  be a hypothesis space, and  $\mathcal{L}_D \in \mathcal{H}$  be the hypothesis learned by a learning algorithm  $\mathcal{L}$  on the training set  $D$ . Then, we consider the following changes on  $D$ :

- Let  $D^{\setminus i}$  denote the set  $D$  with the  $i$ th sample  $z_i$  excluded, that is,

$$D^{\setminus i} = \{z_1, z_2, \dots, z_{i-1}, z_{i+1}, \dots, z_m\},$$

- Let  $D^i$  denote the set  $D$  with the  $i$ th sample  $z_i$  replaced with  $z'_i$ , that is,

$$D^i = \{z_1, z_2, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m\},$$

where  $z'_i = (\mathbf{x}'_i, y'_i)$ , and  $\mathbf{x}'_i$  follows distribution  $\mathcal{D}$  and is independent of  $D$ .

A loss function  $\ell(\mathcal{L}_D(\mathbf{x}), y) : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ , abbreviated as  $\ell(\mathcal{L}_D, z)$ , characterizes the difference between the predicted label  $\mathcal{L}_D(\mathbf{x})$  and the ground-truth label  $y$ . We now introduce several types of loss with respect to the hypothesis  $\mathcal{L}_D$  as follows:

- Generalization loss:

$$\ell(\mathcal{L}, D) = \mathbb{E}_{\mathbf{x} \in \mathcal{X}, z = (\mathbf{x}, y)} [\ell(\mathcal{L}_D, z)]; \quad (12.54)$$

- Empirical loss:

$$\widehat{\ell}(\mathcal{L}, D) = \frac{1}{m} \sum_{i=1}^m \ell(\mathcal{L}_D, \mathbf{z}_i); \quad (12.55)$$

- Leave-one-out loss:

$$\ell_{\text{loo}}(\mathcal{L}, D) = \frac{1}{m} \sum_{i=1}^m \ell(\mathcal{L}_{D \setminus i}, \mathbf{z}_i). \quad (12.56)$$

We define the *uniform stability* as follows:

**Definition 12.10** A learning algorithm  $\mathcal{L}$  is said to satisfy the  $\beta$ -uniform stability with respect to loss function  $\ell$  if, for any  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{z} = (\mathbf{x}, y)$ ,  $\mathcal{L}$  satisfies

$$|\ell(\mathcal{L}_D, \mathbf{z}) - \ell(\mathcal{L}_{D \setminus i}, \mathbf{z})| \leq \beta, \quad i = 1, 2, \dots, m. \quad (12.57)$$

If a learning algorithm  $\mathcal{L}$  satisfies the  $\beta$ -uniform stability with respect to a loss function  $\ell$ , then

$$\begin{aligned} & |\ell(\mathcal{L}_D, \mathbf{z}) - \ell(\mathcal{L}_{D^i}, \mathbf{z})| \\ & \leq |\ell(\mathcal{L}_D, \mathbf{z}) - \ell(\mathcal{L}_{D \setminus i}, \mathbf{z})| + |\ell(\mathcal{L}_{D^i}, \mathbf{z}) - \ell(\mathcal{L}_{D \setminus i}, \mathbf{z})| \\ & \leq 2\beta, \end{aligned}$$

which means that the stability of excluding an instance implies the stability of replacing an instance.

If the loss function  $\ell$  is bounded as  $0 \leq \ell(\mathcal{L}_D, \mathbf{z}) \leq M$  for all  $D$  and  $\mathbf{z} = (\mathbf{x}, y)$ , then, we have [Bousquet and Elisseeff (2002)]

**Theorem 12.8** Given a data set  $D$  with  $m$  i.i.d. instances drawn from the distribution  $\mathcal{D}$ . If a learning algorithm  $\mathcal{L}$  satisfies the  $\beta$ -uniform stability with respect to a loss function  $\ell$  upper bounded by  $M$ , then, for any  $m \geq 1$  and  $\delta \in (0, 1)$ , the following holds with a probability of at least  $1 - \delta$ :

See Bousquet and Elisseeff (2002) for proof.

$$\ell(\mathcal{L}, D) \leq \widehat{\ell}(\mathcal{L}, D) + 2\beta + (4m\beta + M) \sqrt{\frac{\ln(1/\delta)}{2m}}, \quad (12.58)$$

$$\ell(\mathcal{L}, D) \leq \ell_{\text{loo}}(\mathcal{L}, D) + \beta + (4m\beta + M) \sqrt{\frac{\ln(1/\delta)}{2m}}. \quad (12.59)$$

Theorem 12.8 shows the generalization error bound of the learning algorithm  $\mathcal{L}$  derived from the stability analysis. From (12.58), we see the convergence rate between the empirical error and the generalization error is  $\beta \sqrt{m}$ . When  $\beta = O(\frac{1}{m})$ , the con-

vergence rate becomes  $O(\frac{1}{\sqrt{m}})$ , which is consistent with those of VC bound and Rademacher bound in comparisons with Theorems 12.3 and 12.6.

The stability analysis of learning algorithm focuses on  $|\widehat{\ell}(\mathcal{L}, D) - \ell(\mathcal{L}, D)|$ , whereas the complexity analysis of the hypothesis space considers  $\sup_{h \in \mathcal{H}} |\widehat{E}(h) - E(h)|$ . In other words, the stability analysis does not necessarily consider every hypothesis in  $\mathcal{H}$ , but only analyzes the generalization error bound of the output hypothesis  $\mathcal{L}_D$  based on the properties (stability) of  $\mathcal{L}$ . So, what is the relationship between stability and learnability?

To ensure the generalization ability of a stable learning algorithm  $\mathcal{L}$ , we must assume  $\beta\sqrt{m} \rightarrow 0$ , that is, the empirical loss converges to the generalization loss; otherwise, learnability can hardly be discussed. For ease of computation, letting  $\beta = \frac{1}{m}$  and substituting into (12.58), we have

$$\ell(\mathcal{L}, D) \leq \widehat{\ell}(\mathcal{L}, D) + \frac{2}{m} + (4 + M)\sqrt{\frac{\ln(1/\delta)}{2m}}. \quad (12.60)$$

Given a loss function  $\ell$ , a learning algorithm  $\mathcal{L}$  is an ERM learning algorithm satisfying the ERM principle if its output hypothesis minimizes the empirical loss. We have the following theorem on stability and learnability:

**Theorem 12.9** *If an ERM learning algorithm  $\mathcal{L}$  is stable, then the hypothesis space  $\mathcal{H}$  is learnable.*

**Proof** Let  $g$  be the hypothesis with the minimum generalization loss in  $\mathcal{H}$ , that is,

$$\ell(g, D) = \min_{h \in \mathcal{H}} \ell(h, D).$$

Letting

$$\begin{aligned} \epsilon' &= \frac{\epsilon}{2}, \\ \frac{\delta}{2} &= 2 \exp(-2m(\epsilon')^2), \end{aligned}$$

then, from Hoeffding's inequality (12.6), the following holds with a probability of at least  $1 - \delta/2$  when  $m \geq \frac{2}{\epsilon^2} \ln \frac{4}{\delta}$ :

$$|\ell(g, D) - \widehat{\ell}(g, D)| \leq \frac{\epsilon}{2}.$$

For (12.60), by setting

Minimizing empirical error and minimizing empirical loss are sometimes different since there exist some poor loss functions  $\ell$  such that minimizing the loss does not minimize the empirical error. For ease of discussion, this chapter assumes that minimizing the loss always minimizes the empirical error.

$$\frac{2}{m} + (4 + M)\sqrt{\frac{\ln(2/\delta)}{2m}} = \frac{\epsilon}{2},$$

we have  $m = O(\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$ . Hence, the following holds with a probability of at least  $1 - \delta/2$ :

$$\ell(\mathcal{L}, D) \leq \widehat{\ell}(\mathcal{L}, D) + \frac{\epsilon}{2}.$$

Therefore, the following holds with a probability of at least  $1 - \delta$ :

$$\begin{aligned} \ell(\mathcal{L}, D) - \ell(g, D) &\leq \widehat{\ell}(\mathcal{L}, D) + \frac{\epsilon}{2} - \left(\widehat{\ell}(g, D) - \frac{\epsilon}{2}\right) \\ &\leq \widehat{\ell}(\mathcal{L}, D) - \widehat{\ell}(g, D) + \epsilon \\ &\leq \epsilon, \end{aligned}$$

which proves Theorem 12.9.  $\square$

Readers may wonder, why we can derive the learnability of a hypothesis space from the stability of a learning algorithm. Learning algorithm and hypothesis space are very different things. However, it is worth noting that stability is not irrelevant to hypothesis space as they are indeed connected by a loss function  $\ell$  according to the definition of stability.

## 12.7 Further Reading

Valiant (1984) proposed PAC learning, which motivated a branch of machine learning research known as *Computational Learning Theory*. A good introductory textbook on this topic is Kearns and Vazirani (1994). The most important academic conference in this field is the Conference on Learning Theory (COLT).

Vapnik and Chervonenkis (1971) proposed the VC dimension, which makes it possible to study the complexity of infinite hypothesis spaces. Sauer's Lemma is named after Sauer (1972), while the same result was also derived in Vapnik and Chervonenkis (1971), Shelah (1972), respectively. This chapter mainly focuses on binary classification problems, and as for multiclass classification problems, the VC dimension can be extended to the Natarajan dimension (Natarajan 1989; Ben-David et al. 1995).

Rademacher complexity was introduced to machine learning by Koltchinskii and Panchenko (2000) and received more attention after Bartlett and Mendelson (2002). Bartlett et al. (2002) proposed the local Rademacher complexity, which can derive a tighter generalization error bound for noisy data.

The VC dimension is named after the surnames of the two authors.

Bousquet and Elisseeff (2002) introduced the stability analysis of machine learning algorithms, and motivated many studies on the relationship between stability and learnability. For example, Mukherjee et al. (2006), Shalev-Shwartz et al. (2010) showed the equivalence of ERM stability and ERM learnability. Since not all learning algorithms satisfy the ERM principle, Shalev-Shwartz et al. (2010) further studied the relationship between stability and learnability with respect to Asymptotical Empirical Risk Minimization (AERM).

This chapter mainly focuses on deterministic learning problems, that is, there is a deterministic label  $y$  for each sample  $x$ . Though most supervised learning problems are deterministic, there are also stochastic learning problems in which the label of an instance does not firmly belong to a single class but is decided by a posterior probability function conditioned on feature values. See Devroye et al. (1996) for more discussions on the generalization error bound in stochastic learning problems.



## Exercises

---

**12.1** Prove Jensen's inequality (12.4).

**12.2** Prove Lemma 12.1.

**12.3** Prove Corollary 12.1.

Hint: letting  $\delta = 2e^{-2m\epsilon^2}$ .

**12.4** Prove that the hypothesis space consisting of all linear hypothesis planes in  $\mathbb{R}^d$  has a VC dimension of  $d + 1$ .

**12.5** Calculate the VC dimension of the hypothesis space of decision stumps.

**12.6** Prove that the VC dimension of the hypothesis space of decision tree classifiers can be infinite.

**12.7** Prove that the VC dimension of the hypothesis space of  $k$ -nearest neighbors classifiers can be infinite.

**12.8** Prove that the Rademacher complexity of the constant function  $c$  is 0.

**12.9** Given function spaces  $\mathcal{F}_1$  and  $\mathcal{F}_2$ , prove that  $R_m(\mathcal{F}_1 + \mathcal{F}_2) \leq R_m(\mathcal{F}_1) + R_m(\mathcal{F}_2)$ , where  $R_m(\cdot)$  is the Rademacher complexity.

**12.10 \*** Considering Theorem 12.8, discuss the rationality of estimating an algorithm's generalization ability via cross-validation.

## Break Time

---

### Short Story: Leslie G. Valiant—The Father of Computational Learning Theory

Theoretical Computer Science (TCS) is an intersection of computer science and mathematics that focuses on mathematical topics of computing. A famous TCS problem is the “P versus NP problem”.

Computational learning theory, as a subfield of machine learning, is the intersection of machine learning and TCS. If we are talking about computational learning theory, we have to talk about the



British computer scientist Leslie G. Valiant (1949–). Valiant studied at King’s College, Cambridge, Imperial College London, and the University of Warwick, where he earned his Ph.D. degree in 1974. Before he became a professor at Harvard University in 1982, he taught at Carnegie Mellon University, the University of Leeds, and the University of Edinburgh. In 1984, *Communications of the ACM* published Valiant’s paper titled “A theory of the learnable”, in which PAC learning theory was proposed and laid the foundations of computational learning theory. In 2010, Valiant received the Turing Award for his seminal contributions to PAC learning theory, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing. The ACM Turing Award committee pointed out that Valiant’s paper published in 1984 created a new research area known as computational learning theory that puts machine learning on a sound mathematical footing. *ACM Computing News* also published an article titled “ACM Turing Award Goes to Innovator in Machine Learning” to emphasize the contributions of this first Turing Award recipient from machine learning.

## References

---

- Bartlett PL, Mendelson S (2002) Rademacher and Gaussian complexities: risk bounds and structural results. *J Mach Learn Res* 3:463–482
- Bartlett PL, Bousquet O, Mendelson S (2002). Localized rademacher complexities. Sydney, Australia, pp 44–58
- Ben-David S, Cesa-Bianchi N, Haussler D, Long PM (1995) Characterizations of learnability for classes of  $\{0, \dots, n\}$ -valued functions. *J Comput Syst Sci* 50(1):74–86
- Bousquet O, Elisseeff A (2002) Stability and generalization. *J Mach Learn Res* 2:499–526
- Devroye L, Györfi L, Lugosi G (eds) (1996) A probabilistic theory of pattern recognition. Springer, New York
- Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *J Am Stat Assoc* 58(301):13–30
- Kearns MJ, Vazirani UV (1994) An introduction to computational learning theory. MIT Press, Cambridge
- Koltchinskii V, Panchenko D (2000) Rademacher processes and bounding the risk of function learning. In: Gine E, Mason DM, Wellner JA (eds) High dimensional probability II. Birkhäuser Boston, Cambridge, pp 443–457
- McDiarmid C (1989) On the method of bounded differences. *Surv Comb* 141(1):148–188
- Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT Press, Cambridge
- Mukherjee S, Niyogi P, Poggio T, Rifkin RM (2006) Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv Comput Math* 25(1–3):161–193
- Natarajan BK (1989) On learning sets and functions. *Mach Learn* 4(1):67–97
- Sauer N (1972) On the density of families of sets. *J Comb Theory - Ser A* 13(1):145–147
- Shalev-Shwartz S, Shamir O, Srebro N, Sridharan K (2010) Learnability, stability and uniform convergence. *J Mach Learn Res* 11:2635–2670
- Shelah S (1972) A combinatorial problem; stability and order for models and theories in infinitary languages. *Pac J Math* 41(1):247–261
- Valiant LG (1984) A theory of the learnable. *Commun ACM* 27(11):1134–1142
- Vapnik VN, Chervonenkis A (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab Its Appl* 16(2):264–280