

# Lecture 3: Rademacher Complexity and Vapnik-Chervonenkis Dimension

Tianpei Xie

Dec. 18th., 2022

## Contents

<b>1</b>	<b>PAC Learnability for Infinite Hypotheses Set</b>	<b>2</b>
<b>2</b>	<b>Rademacher Complexity</b>	<b>2</b>
<b>3</b>	<b>Vapnik-Chervonenkis Dimension</b>	<b>8</b>
3.1	Definition of Vapnik-Chervonenkis Dimension . . . . .	8
3.2	Growth Function . . . . .	9
3.3	Relate Growth Function to Rademacher Complexity . . . . .	12
3.4	Generalization Bounds via Growth Function and VC-Dimension . . . . .	13
3.5	Examples . . . . .	14
3.6	Lower Bounds . . . . .	18
3.7	Fundamental Theorem of PAC Learning . . . . .	19

# 1 PAC Learnability for Infinite Hypothesis Set

- **Remark** (*Bounding Excess Risk via Uniform Deviation*)

The definition of *agnostic PAC learnability* requires that *the excess risk* would be bounded above uniformly over all distributions.

$$L(h_n) - \inf_{h \in \mathcal{H}} L(h) \leq 2 \sup_{h \in \mathcal{H}} |L(h) - \hat{L}(h)| \quad (1)$$

where  $h^* = \operatorname{argmin}_{h \in \mathcal{H}} L(h)$  and  $\hat{L}(h)$  is the training error of  $g$ . The second last inequality is due to the fact that  $h_n$  minimizes the training error. Thus *the estimation error* can be bounded uniformly by *the generalization error bound*  $|L(h) - \hat{L}(h)|$  for any  $h \in \mathcal{H}$ .

In this chapter, we discuss various ways to bound the uniform deviation:

$$\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| := \|\hat{\mathcal{P}}_n - \mathcal{P}\|_{\mathcal{H}}$$

- **Remark** (*Universal Consistency for Infinite Hypothesis Set*)

When  $|\mathcal{H}| < \infty$ , we can use sample complexity bounds that involve  $\log |\mathcal{H}|$  for *universal consistency* of ERM. Obviously, we cannot do so when  $|\mathcal{H}| = \infty$ . A general idea of analyzing infinite hypothesis set consists of *reducing the infinite case to the analysis of finite sets of hypotheses* and then proceed as in the previous chapter.

There are different techniques for that reduction, each relying on a *different notion of complexity* for the family of hypotheses.

# 2 Rademacher Complexity

- **Remark** (*Notations*)

We will continue to use  $\mathcal{H}$  to denote a *hypothesis set* as in the previous chapters, and  $h \in \mathcal{H}$  an *element* of  $\mathcal{H}$ . Many of the results of this section are general and hold for an arbitrary *loss function*  $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ . To each  $h : \mathcal{X} \rightarrow \mathcal{Y}$ , we can associate a function  $g$ :

$$g : (x, y) \in \mathcal{X} \times \mathcal{Y} \rightarrow L(h(x), y)$$

without explicitly describing the specific loss  $L$  used. In what follows  $\mathcal{G}$  will generally be interpreted as *the family of loss functions associated to  $\mathcal{H}$* .

- **Definition** (*Empirical Rademacher Complexity*)

Let  $\mathcal{G}$  be a family of functions mapping from  $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$  to  $[a, b]$  and  $\mathcal{D} = (z_1, \dots, z_n)$  a fixed *sample* of size  $n$  with elements in  $\mathcal{Z}$ . Then, the empirical Rademacher complexity of  $\mathcal{G}$  with respect to the sample  $\mathcal{D}$  is defined as:

$$\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) = \mathbb{E}_{\sigma} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right] \quad (2)$$

where  $\sigma := (\sigma_1, \dots, \sigma_n)$  are *independent uniform random variables* taking values in  $\{-1, +1\}$ . The random variables  $\sigma_i$  are called Rademacher variables.

- **Definition (*Rademacher Random Variable and Rademacher Process*)**

A Rademacher variables is a **symmetric Bernoulli** random variable on  $\{-1, +1\}$  with equal probability  $P(\sigma_i = +1) = P(\sigma_i = -1) = \frac{1}{2}$ . A **Rademacher sequence**  $\sigma := (\sigma_1, \dots, \sigma_n)$  is seen as a *uniform distribution* on binary hypercube  $\{-1, 1\}^n$ .

A Rademacher process indexed by the function class  $\mathcal{G}$  is defined as

$$\mathfrak{R}_n := \sum_{i=1}^n \sigma_i \delta_{g(Z_i)},$$

where  $\sigma$  is a *Rademacher sequence* and is independent from  $Z$ . It is a **sub-gaussian stochastic process**.  $\{\sigma_i g(Z_i)\}$  **conditioning** on  $\{Z_i\}$  is a *Rademacher process*.

- **Remark (*How Well to Fit Random Noise*)**

The *Rademacher complexity* captures the richness of a family of functions by measuring **the degree to which a hypothesis set can fit random noise**. The richer or more complex families  $\mathcal{H}$  can generate more vectors  $h_{\mathcal{D}}$  and thus better correlate with random noise, on  $\mathcal{D}_n$ .

The intuition is that if a hypothesis set can fit arbitrary noise, then it is too large to bound the performance of ERM, i.e. it is very likely to have overfitting (zero empirical error but arbitrary bad generalization error).

- **Remark**

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}_{\sigma} \left[ \sup_{g \in \mathcal{G}} \frac{\langle \sigma_n, g_n \rangle}{n} \right]$$

which measures the **correlation** between the random noise  $\sigma_n := \{\sigma_i\}$  and  $g_n \equiv \{g(X_i, Y_i)\}_{i=1}^n$ . The supremum  $\sup_{g \in \mathcal{G}} \langle \sigma_n, g_n \rangle / n$  is a measure of how well the function class  $\mathcal{H}$  correlates with  $\sigma$  over the sample  $\mathcal{D}_n$ . Thus, **the empirical Rademacher complexity** measures on average how well **the function class  $\mathcal{H}$  correlates with random noise** on  $\mathcal{D}_n$ .

- **Definition (*Rademacher Complexity*)**

Let  $\mathcal{P}$  denote the distribution according to which samples are drawn. For any integer  $n \geq 1$ , **the Rademacher complexity** of  $\mathcal{G}$  is defined as the *expectation of the empirical Rademacher complexity* over all samples of size  $n$  drawn according to  $\mathcal{P}$ :

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{\mathcal{P}^n} \left[ \hat{\mathfrak{R}}_{\mathcal{D}_n}(\mathcal{G}) \right].$$

- **Proposition 2.1 (*Uniform Bound via Rademacher Complexity*)** [Mohri et al., 2018]

Let  $\mathcal{G}$  be a family of functions mapping from  $\mathcal{Z}$  to  $[0, 1]$ . Then, for any  $\delta > 0$ , **with probability at least  $1 - \delta$** , each of the following holds for all  $g \in \mathcal{G}$ :

$$\mathbb{E}[g(Z)] \leq \frac{1}{n} \sum_{i=1}^n g(Z_i) + 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}} \quad (3)$$

and

$$\mathbb{E}[g(Z)] \leq \frac{1}{n} \sum_{i=1}^n g(Z_i) + 2\hat{\mathfrak{R}}_n(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \quad (4)$$

- **Proof:** Define a function  $\Phi$  on  $\mathcal{D}_n$  by

$$\Phi(\mathcal{D}_n) := \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)] \right].$$

This function has *bounded difference* if *only one sample makes changes*. In particular, let  $\mathcal{D}_n \equiv (Z_1, \dots, Z_n)$  and  $\tilde{\mathcal{D}}_n^{(i)} \equiv (Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n)$ . Then

$$\left| \Phi(\mathcal{D}_n) - \Phi(\tilde{\mathcal{D}}_n^{(i)}) \right| \leq \frac{1}{n} \sup_{g \in \mathcal{G}} |g(Z_i) - g(Z'_i)| \leq \frac{1}{n}.$$

The proof consists of three parts:

1. Bound the *probability* of **tail event of  $\Phi$**

$$\Phi(\mathcal{D}_n) - \mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)]$$

This part follows from the *bounded difference inequality*: for any  $\delta > 0$ , with probability at least  $1 - \delta/2$ ,

$$\Phi(\mathcal{D}_n) - \mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)] \leq \sqrt{\frac{\log(2/\delta)}{2n}} \quad (5)$$

2. Bound the *expectation*  $\mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)]$  by *Rademacher complexity* on  $\mathcal{G}$

$$\mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)] \leq 2\mathfrak{R}_n(\mathcal{G}).$$

Recall that  $Z' = (Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n)$  and  $Z = (Z_1, \dots, Z_n)$ . We know that  $\mathbb{E}_{Z'} \left[ \frac{1}{n} \sum_{i=1}^n g(Z'_i) \right] = \mathbb{E}[g(Z)]$  since  $Z'$  and  $Z$  are independent identically distributed.

$$\begin{aligned} \mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)] &= \mathbb{E}_Z \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)] \right] \right] \\ &= \mathbb{E}_Z \left[ \sup_{g \in \mathcal{G}} \mathbb{E}_{Z'} \left[ \frac{1}{n} \sum_{i=1}^n g(Z_i) - \frac{1}{n} \sum_{i=1}^n g(Z'_i) \right] \right] \\ &\quad (\text{Jenson's inequality and convexity of sup}) \\ &\leq \mathbb{E}_{Z, Z'} \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n g(Z_i) - \frac{1}{n} \sum_{i=1}^n g(Z'_i) \right] \right] \\ &\quad (\text{symmetrization}) \\ &= \mathbb{E}_{Z, Z', \sigma} \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n \sigma_i (g(Z_i) - g(Z'_i)) \right] \right] \\ &\quad (\text{sub-additivity of sup}) \\ &\leq \mathbb{E}_{Z, \sigma} \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i) \right] \right] + \mathbb{E}_{Z', \sigma} \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n -\sigma_i g(Z'_i) \right] \right] \\ &= 2\mathbb{E}_{Z, \sigma} \left[ \sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i) \right] \right] := 2\mathfrak{R}_n(\mathcal{G}) \quad (6) \end{aligned}$$

The ***symmetrization step*** works since  $\sigma_i = \pm 1$  with equal probability. If  $\sigma_i = +1$ , then the summand is unchanged. If  $\sigma_i = -1$ , the summand is swapped but the distribution is unchanged as well since  $Z'$  and  $Z$  are independent identically distributed. The last equality follows from the definition of Rademacher complexity and the fact that  $\sigma_i$  and  $-\sigma_i$  follow the same distribution since they are symmetric Bernoulli random variables. Finally, replacing  $\mathbb{E}_{\mathcal{P}^n} [\Phi(\mathcal{D}_n)]$  by its upper bound  $2\mathfrak{R}_n(\mathcal{G})$  and  $\delta/2 \rightarrow \delta$  gives us inequality (3).

3. Bound the probability of tail events for empirical Rademacher complexity:

$$\mathfrak{R}_n(\mathcal{G}) - \widehat{\mathfrak{R}}_n(\mathcal{G})$$

In this part, we use the *bounded difference inequality* again, noticing that the empirical Rademacher complexity is function with bounded difference  $1/n$ . Therefore,

$$\mathfrak{R}_n(\mathcal{G}) \leq \widehat{\mathfrak{R}}_n(\mathcal{G}) + \sqrt{\frac{\log(2/\delta)}{2n}} \quad (7)$$

Finally, we combine the inequalities in (3) and (7) using union bounds to obtain the final result. ■

- **Remark (*Rademacher Complexity as Uniform Bound*)**

The above proposition states that any positive integer  $n \geq 1$  and any scalar  $\delta \geq 0$ , with probability at least  $1 - \delta$ , we have

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)] \right| \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- **Remark (*Symmetrization*)**

The technique that develops the uniform deviation bound via *Rademacher complexity* is called ***symmetrization***. Consider ***the empirical process***

$$g \rightarrow (\widehat{\mathcal{P}}_n - \mathcal{P})g = \frac{1}{n} \sum_{i=1}^n (g(Z_i) - \mathcal{P}g)$$

where  $\mathcal{P}g := \int g d\mathcal{P} = \mathbb{E}_{\mathcal{P}}[g(Z)]$  is seen as an *operator* on function (by identification of the measure and functional). The *symmetrization* refers to the idea to replace ***the empirical process*** with ***a symmetrized process***

$$g \rightarrow \widehat{\mathcal{P}}_n^\epsilon g := \frac{1}{n} \sum_{i=1}^n \epsilon_i g(Z_i)$$

where  $\epsilon := (\epsilon_1, \dots, \epsilon_n)$  are ***i.i.d. Rademacher random variables*** (symmetric Bernoulli random variable on  $\{-1, 1\}$ ) and they are ***independent*** of  $(Z_1, \dots, Z_n)$ . The idea is that, for fixed  $(Z_1, \dots, Z_n)$ , the symmetrized empirical measure is a *Rademacher process*, hence a *sub-Gaussian process*.

**Lemma 2.2 (*Symmetrization*).** [Wellner et al., 2013]

For every ***nondecreasing, convex***  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$  and class of measurable functions  $\mathcal{G}$ ,

$$\mathbb{E} \left[ \Phi \left( \left\| \widehat{\mathcal{P}}_n - \mathcal{P} \right\|_{\mathcal{G}} \right) \right] \leq \mathbb{E} \left[ \Phi \left( 2 \left\| \widehat{\mathcal{P}}_n^\epsilon \right\|_{\mathcal{G}} \right) \right]$$

- **Lemma 2.3 (Rademacher Complexity for 0-1 Loss)** [Mohri et al., 2018]

Let  $\mathcal{H}$  be a family of functions taking values in  $\{-1, +1\}$  and let  $\mathcal{G}$  be the family of loss functions associated to  $\mathcal{H}$  for **the zero-one loss**:

$$\mathcal{G} := \{(x, y) \mapsto \mathbb{1}_{h(x) \neq y} : h \in \mathcal{H}\}.$$

For any sample  $\mathcal{D} = ((X_1, Y_1), \dots, (X_n, Y_n))$  of elements in  $\mathcal{X} \times \{-1, +1\}$ , let  $\mathcal{D}_{\mathcal{X}} := (X_1, \dots, X_n)$ . Then, the following relation holds between **the empirical Rademacher complexities** of  $\mathcal{G}$  and  $\mathcal{H}$ :

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) \leq \frac{1}{2} \widehat{\mathfrak{R}}_{\mathcal{D}_{\mathcal{X}}}(\mathcal{H})$$

This implies that  $\mathfrak{R}(\mathcal{G}) = \frac{1}{2} \mathfrak{R}(\mathcal{H})$ .

(Hint:  $\mathbb{1}_{h(x) \neq y} := \frac{1}{2}(1 - yh(x))$ . Also  $\sigma$  and  $-\sigma y$  have the same distribution.)

- This immediately gives us the result below:

**Proposition 2.4 (Rademacher Complexity Bound for Binary Classification)** [Mohri et al., 2018]

Let  $\mathcal{H}$  be a family of functions taking values in  $\{-1, +1\}$  and let  $\mathcal{P}$  be the distribution over the input space  $\mathcal{X}$ . Then, for any  $\delta > 0$ , **with probability at least  $1 - \delta$**  over a sample  $\mathcal{D}_n$  of size  $n$  drawn according to  $\mathcal{P}_X$ , each of the following holds for **any**  $h \in \mathcal{H}$ :

$$L(h) \leq \widehat{L}_n(h) + \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}} \quad (8)$$

and

$$L(h) \leq \widehat{L}_n(h) + \widehat{\mathfrak{R}}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (9)$$

- **Remark (Compute Empirical Rademacher Complexity)**

Note that the Rademacher variable  $\sigma_i$  and  $-\sigma_i$  have the same distribution. So

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{H}) = \mathbb{E}_{\sigma} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (-\sigma_i h(X_i)) \right] = -\mathbb{E}_{\sigma} \left[ \inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) \right]$$

Now, for a fixed value of  $\sigma$ , computing

$$\inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i)$$

is equivalent to an **empirical risk minimization problem**, which is known to be **computationally hard** for some hypothesis sets. Thus, in some cases, computing  $\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{H})$  could be computationally hard.

- **Proposition 2.5 (Talagrand's Contraction Principle)** [Boucheron et al., 2013, Vershynin, 2018]

Let  $x_1, \dots, x_n$  be vectors whose real-valued components are indexed by  $T$ , that is,  $x_i = (x_{i,s})_{s \in T}$ . For each  $i = 1, \dots, n$ , let  $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$  be a **1-Lipschitz function** such that

$\varphi_i(0) = 0$ . Let  $\epsilon_1, \dots, \epsilon_n$  be independent Rademacher random variables, and let  $\Psi : [0, \infty) \rightarrow \mathbb{R}$  be a **non-decreasing convex** function. Then

$$\mathbb{E} \left[ \Psi \left( \sup_{s \in T} \sum_{i=1}^n \epsilon_i \varphi_i(x_{i,s}) \right) \right] \leq \mathbb{E} \left[ \Psi \left( \sup_{s \in T} \sum_{i=1}^n \epsilon_i x_{i,s} \right) \right] \quad (10)$$

and

$$\mathbb{E} \left[ \Psi \left( \frac{1}{2} \sup_{s \in T} \left| \sum_{i=1}^n \epsilon_i \varphi_i(x_{i,s}) \right| \right) \right] \leq \mathbb{E} \left[ \Psi \left( \sup_{s \in T} \left| \sum_{i=1}^n \epsilon_i x_{i,s} \right| \right) \right]. \quad (11)$$

The proof of this result need the following lemma

**Lemma 2.6** [Boucheron et al., 2013, Vershynin, 2018]

Let  $\Psi : \mathbb{R} \rightarrow \mathbb{R}$  denote a **convex non-decreasing function**. Let  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  be a 1-Lipschitz function such that  $\varphi_i(0) = 0$ . Let  $T \subset \mathbb{R}^2$ . Then

$$\Psi \left( \sup_{s \in T} \{s_1 + \varphi(s_2)\} \right) + \Psi \left( \sup_{s \in T} \{s_1 - \varphi(s_2)\} \right) \leq \Psi \left( \sup_{s \in T} \{s_1 + s_2\} \right) + \Psi \left( \sup_{s \in T} \{s_1 - s_2\} \right)$$

- **Corollary 2.7 (Talagrand's Lemma)** [Mohri et al., 2018]

Let  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  be an  $L$ -Lipschitz. Then, for any hypothesis set  $\mathcal{H}$  of real-valued functions, the following inequality holds:

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\varphi \circ \mathcal{H}) \leq L \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{H}) \quad (12)$$

- **Proposition 2.8 (Rademacher Complexity of a Convex Hull of Function Class)**

Let  $\mathcal{H}$  be a set of functions mapping from  $\mathcal{X}$  to  $\mathbb{R}$ . Then, for any sample  $\mathcal{D}$ , we have

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\text{conv}(\mathcal{H})) = \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{H}) \quad (13)$$

where  $\text{conv}(\mathcal{H})$  is **the convex hull** of set  $\mathcal{H}$

$$\text{conv}(\mathcal{H}) := \left\{ \sum_{k=1}^p \lambda_k h_k : p \geq 1, \forall k \in [1, p], \lambda_k \geq 0, h_k \in \mathcal{H}, \sum_{k=1}^p \lambda_k \leq 1 \right\}.$$

**Proof:**

$$\begin{aligned} \widehat{\mathfrak{R}}_{\mathcal{D}}(\text{conv}(\mathcal{H})) &= \mathbb{E}_{\sigma} \left[ \sup_{g \in \text{conv}(\mathcal{H})} \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right] \\ &= \frac{1}{n} \mathbb{E}_{\sigma} \left[ \sup_{h_1, \dots, h_p \in \mathcal{H}} \sup_{\{\lambda_k\} \in \Delta_p} \sum_{i=1}^n \sigma_i \sum_{k=1}^p \lambda_k h_k(X_i) \right] \\ &= \frac{1}{n} \mathbb{E}_{\sigma} \left[ \sup_{h_1, \dots, h_p \in \mathcal{H}} \sup_{\{\lambda_k\} \in \Delta_p} \sum_{k=1}^p \lambda_k \sum_{i=1}^n \sigma_i h_k(X_i) \right] \end{aligned}$$

Note that extreme point of a convex hull is its vertex, so the supremum of  $\sum_{k=1}^p \lambda_k (\sum_{i=1}^n \sigma_i h_k(X_i))$  over  $\lambda$  can only be reached at one of  $\{\sum_{i=1}^n \sigma_i h_k(X_i)\}_{k=1}^p$ . Thus above equality becomes

$$\begin{aligned} \widehat{\mathfrak{R}}_{\mathcal{D}}(\text{conv}(\mathcal{H})) &= \frac{1}{n} \mathbb{E}_{\sigma} \left[ \sup_{h_1, \dots, h_p \in \mathcal{H}} \max_{k \in [1, p]} \sum_{i=1}^n \sigma_i h_k(X_i) \right] \\ &= \frac{1}{n} \mathbb{E}_{\sigma} \left[ \sup_{h \in \mathcal{H}} \sum_{i=1}^n \sigma_i h(X_i) \right] = \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{H}). \quad \blacksquare \end{aligned}$$

### 3 Vapnik-Chervonenkis Dimension

#### 3.1 Definition of Vapnik-Chervonenkis Dimension

- **Remark** (*Restriction based on Behavior of Functions on Data  $\mathcal{D}$* )

Recall the No-Free-Lunch theorem and its proof [Shalev-Shwartz and Ben-David, 2014]. There, 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 will succeed on the same distribution*. To do so, the adversary used a finite set  $\mathcal{D} \subset \mathcal{X}$  and considered a family of distributions that are **concentrated on elements of  $\mathcal{D}$** . Each distribution was *derived from a “true” target function* from  $\mathcal{D}$  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  $\mathcal{D}$  to  $\{0, 1\}$ .

When considering *PAC learnability of a hypothesis class  $\mathcal{H}$* , the adversary is restricted to *constructing distributions* for which *some hypothesis  $h \in \mathcal{H}$  achieves a zero risk*. Since we are considering **distributions that are concentrated on elements of  $\mathcal{D}$** , we should study how  $\mathcal{H}$  behaves on  $\mathcal{D}$

- **Definition** (*Restriction of  $\mathcal{H}$  to  $\mathcal{D}$* ).

Let  $\mathcal{H}$  be a class of functions from  $\mathcal{X}$  to  $\{0, 1\}$  and let  $\mathcal{D} = \{x_1, \dots, x_n\} \subset \mathcal{X}$ .

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,

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

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

- **Remark** (*What You See Is All You Know*)

Using the output of functions on *a finite set of samples*, we can define an **equivalence relationship**:  $f \sim g$  if and only if their **outputs** vector on given finite set  $\mathcal{D}_m$  are **the same**. Thus, unlike the original space  $\mathcal{H}$ , **the quotient space  $\mathcal{H} / \sim$**  is a much *simpler function space* with **finite dimensional representation**.

In other word, **what you see is all you know**, i.e. there is no way to distinguish  $f$  and  $g$  beyond their answers to given limited set of questions in  $\mathcal{D}$ .

- **Definition** (*Shattering*).

A hypothesis class  $\mathcal{H}$  **shatters** a finite set  $\mathcal{D} \subset \mathcal{X}$  if **the restriction of  $\mathcal{H}$  to  $\mathcal{D}$**  is the set of **all functions** from  $\mathcal{D}$  to  $\{0, 1\}$ . That is,

$$|\mathcal{H}_{\mathcal{D}}| = 2^{|\mathcal{D}|}.$$

- **Remark** Whenever some set  $\mathcal{D}$  is **shattered** by  $\mathcal{H}$ , the **adversary** is **not restricted** by  $\mathcal{H}$ , as they can **construct a distribution over  $\mathcal{D}$**  based on **any** target function from  $\mathcal{D}$  to  $\{0, 1\}$ , while still maintaining the realizability assumption.
- The following is the corollary of the *No Free Lunch Theorem*:

**Corollary 3.1** [Shalev-Shwartz and Ben-David, 2014]

Let  $\mathcal{H}$  be a hypothesis class of functions from  $\mathcal{X}$  to  $\{0, 1\}$ . Let  $n$  be a training set size.



Assume that there exists a set  $\mathcal{D} \subset X$  of size  $2n$  that is **shattered** by  $\mathcal{H}$ . Then, for any learning algorithm,  $\mathcal{A}$ , there exist a **distribution**  $\mathcal{P}$  over  $\mathcal{X} \times \{0,1\}$  and a predictor  $h \in \mathcal{H}$  such that

$$L_{\mathcal{P}}(h) = 0$$

but **with probability of at least**  $1/7$  over the choice of  $\mathcal{D} \sim \mathcal{P}^n$  we have that

$$L_{\mathcal{P}}(\mathcal{A}(\mathcal{D})) \geq 1/8.$$

- **Remark (A Model that can Explain Everything is Worthless)**

If  $\mathcal{H}$  **shatters** some set  $\mathcal{D}$  of size  $2m$  then we cannot learn  $\mathcal{H}$  using  $m$  examples.

Intuitively, if a set  $\mathcal{D}$  is **shattered** by  $\mathcal{H}$ , and we receive a sample containing half the instances of  $\mathcal{D}$ , the labels of these instances give us **no information** about the labels of the **rest** of the instances in  $\mathcal{D}$  - **every possible labeling of the rest of the instances can be explained by some hypothesis in  $\mathcal{H}$ .**

Philosophically,

**If someone can explain every phenomenon, his explanations are worthless.**

- **Definition (VC-Dimension).**

The Vapnik-Chervonenkis (VC) dimension of a hypothesis class  $\mathcal{H}$ , denoted  $VCdim(\mathcal{H})$  or simply  $v(\mathcal{H})$ , is the maximal size of a set  $\mathcal{D} \subset \mathcal{X}$  that can be **shattered** by  $\mathcal{H}$ .

If  $\mathcal{H}$  can shatter sets of **arbitrarily large size** we say that  $\mathcal{H}$  has **infinite VC-dimension**.

- **Theorem 3.2 (No Free Lunch, VC Dimension)** [Shalev-Shwartz and Ben-David, 2014]  
Let  $\mathcal{H}$  be a class of **infinite VC-dimension**. Then,  $\mathcal{H}$  is **not PAC learnable**.

### 3.2 Growth Function

- **Remark** We defined the notion of **shattering**, by considering **the restriction of  $\mathcal{H}$  to a finite set of instances**. The **growth function** measures the **maximal “effective” size of  $\mathcal{H}$  on a set of  $n$  examples**. Formally:

- **Definition (Growth Function).**

Let  $\mathcal{H}$  be a hypothesis class. Then **the growth function of  $\mathcal{H}$** , denoted  $\tau_{\mathcal{H}} : \mathbb{N} \rightarrow \mathcal{N}$ , is defined as

$$\tau_{\mathcal{H}}(n) := \max_{\mathcal{D} \subset \mathcal{X} : |\mathcal{D}|=n} |\mathcal{H}_{\mathcal{D}}|.$$

In words,  $\tau_{\mathcal{H}}(n)$  is **the number of different functions** from a set  $\mathcal{D}$  of **size  $n$**  to  $\{0,1\}$  that can be obtained by **restricting  $\mathcal{H}$  to  $\mathcal{D}$** .

- **Remark** if  $VCdim(\mathcal{H}) = d$  then for any  $n \leq d$  we have  $\tau_{\mathcal{H}}(n) = 2^n$ . In such cases,  $\mathcal{H}$  induces **all possible functions** from  $\mathcal{D}$  to  $\{0,1\}$ .
- **Lemma 3.3 (Sauer’s Lemma).** [Shalev-Shwartz and Ben-David, 2014, Mohri et al., 2018]  
Let  $\mathcal{H}$  be a hypothesis class with  $VCdim(\mathcal{H}) \leq d < \infty$ . Then, for all  $n$ ,

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^d \binom{n}{i} \tag{14}$$

In particular, if  $n > d + 1$  then

$$\tau_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d. \quad (15)$$

**Proof:** To prove the lemma it suffices to prove the following *stronger claim*: For any  $\mathcal{D} = \{x_1, \dots, x_n\}$  we have

$$\forall \mathcal{H}, |\mathcal{H}_{\mathcal{D}}| \leq |\{\mathcal{B} \subseteq \mathcal{D} : \mathcal{H} \text{ shatters } \mathcal{B}\}| \quad (16)$$

The reason why Equation (16) is sufficient to prove the lemma is that if  $VCdim(\mathcal{H}) \leq d$  then no set whose size is *larger than*  $d$  is *shattered* by  $\mathcal{H}$  and therefore

$$|\{\mathcal{B} \subseteq \mathcal{D} : \mathcal{H} \text{ shatters } \mathcal{B}\}| \leq \sum_{i=0}^d \binom{n}{i}.$$

When  $n > d + 1$  the right-hand side of the preceding is *at most*  $\left(\frac{en}{d}\right)^d$  since

$$\begin{aligned} \sum_{i=0}^d \binom{n}{i} 1^{d-i} &\leq \sum_{i=0}^d \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\ &\leq \sum_{i=0}^n \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\ &= \left(\frac{n}{d}\right)^d \sum_{i=0}^n \binom{n}{i} \left(\frac{d}{n}\right)^i \\ &= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \leq \left(\frac{n}{d}\right)^d e^d \end{aligned}$$

where  $1 + x \leq e^x$ .

We are left with proving Equation (16) and we do it using an *inductive argument*.

1. For  $n = 1$ , no matter what  $\mathcal{H}$  is, either both sides of Equation (16) equal 1 or both sides equal 2 (*the empty set* is always considered to be shattered by  $\mathcal{H}$ ).
2. Assume Equation (16) holds for sets of size  $k < n$  and let us prove it for sets of size  $n$ . Fix  $\mathcal{H}$  and  $\mathcal{D} = \{x_1, \dots, x_n\}$ . Denote  $\mathcal{D}_{-1} = \{x_2, \dots, x_n\}$  and in addition, define the following two sets:

$$Y_0 = \{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_{\mathcal{D}} \vee (1, y_2, \dots, y_n) \in \mathcal{H}_{\mathcal{D}}\},$$

and

$$Y_1 = \{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_{\mathcal{D}} \wedge (1, y_2, \dots, y_n) \in \mathcal{H}_{\mathcal{D}}\}.$$

It is easy to verify that  $|\mathcal{H}_{\mathcal{D}}| = |Y_0| + |Y_1|$  (See Figure 1). Additionally, since  $Y_0 = \mathcal{H}_{\mathcal{D}_{-1}}$ , using *the induction assumption* (applied on  $\mathcal{H}$  and  $\mathcal{D}_{-1}$ ) we have that

$$\begin{aligned} |Y_0| &= |\mathcal{H}_{\mathcal{D}_{-1}}| \leq |\{\mathcal{B} \subseteq \mathcal{D}_{-1} : \mathcal{H} \text{ shatters } \mathcal{B}\}| \\ &= |\{\mathcal{B} \subseteq \mathcal{D} : x_1 \notin \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}|. \end{aligned}$$

	x1	x2	x3	
	h1	0	0	0
	h2	0	0	1
	h3	0	1	0
	h4	0	1	1
	h5	1	0	0
	h6	1	0	1
	h7	1	1	0
	h8	1	1	1

$\mathcal{H}$  (vertical arrow on the left)  
 $\mathcal{H}'$  (two vertical arrows on the left, one for h1-h4 and one for h5-h8)  
 $Y_0$  (blue box around rows h2-h4)  
 $Y_1$  (red box around rows h6-h8)  
 $\mathcal{D}$  (horizontal arrow at the bottom)  
 $\mathcal{D} \setminus \{x_1\}$  (horizontal arrow at the bottom, from x2 to x3)

**Figure 1: Proof of Sauer's lemma.**  $Y_0 = \mathcal{H}_{\mathcal{D} \setminus \{x_1\}}$  (blue) and  $Y_1 = \mathcal{H}'_{\mathcal{D} \setminus \{x_1\}}$  (red) where  $\mathcal{H}'$  are hypotheses  $h' \in \mathcal{H}$  so that there exists  $h \in \mathcal{H}$  with opposite labeling at  $x_1$ . [Shalev-Shwartz and Ben-David, 2014]

Next, define  $\mathcal{H}' \subseteq \mathcal{H}$  to be

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } (1 - h'(x_1), h'(x_2), \dots, h'(x_n)) = (h(x_1), h(x_2), \dots, h(x_n))\},$$

namely,  $\mathcal{H}'$  contains pairs of hypotheses  $(h, h')$  that *agree on  $\mathcal{D}_{-1}$  and differ on  $x_1$* . Using this definition, it is clear that if  $\mathcal{H}'$  *shatters* a set  $\mathcal{B} \subseteq \mathcal{D}$  then it *also shatters* the set  $\mathcal{B} \cup \{x_1\}$  and *vice versa*.

Combining this with the fact that  $Y_1 = \mathcal{H}'_{\mathcal{D}_{-1}}$  and using the inductive assumption (now applied on  $\mathcal{H}'$  and  $\mathcal{D}_{-1}$ ) we obtain that

$$\begin{aligned}
 |Y_1| &= |\mathcal{H}'_{\mathcal{D}_{-1}}| \leq |\{\mathcal{B} \subseteq \mathcal{D}_{-1} : \mathcal{H}' \text{ shatters } \mathcal{B}\}| \\
 &= |\{\mathcal{B} \subseteq \mathcal{D}_{-1} : \mathcal{H}' \text{ shatters } \mathcal{B} \cup \{x_1\}\}| \\
 &= |\{\mathcal{B} \subseteq \mathcal{D} : x_1 \in \mathcal{B} \wedge \mathcal{H}' \text{ shatters } \mathcal{B}\}| \\
 &= |\{\mathcal{B} \subseteq \mathcal{D} : x_1 \in \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}|.
 \end{aligned}$$

Overall, we have shown that

$$\begin{aligned}
 |\mathcal{H}_{\mathcal{D}}| &= |Y_0| + |Y_1| \\
 &\leq |\{\mathcal{B} \subseteq \mathcal{D} : x_1 \notin \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}| + |\{\mathcal{B} \subseteq \mathcal{D} : x_1 \in \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}| \\
 &= |\{\mathcal{B} \subseteq \mathcal{D} : \mathcal{H} \text{ shatters } \mathcal{B}\}|
 \end{aligned}$$

which concludes our proof.  $\blacksquare$

### 3.3 Relate Growth Function to Rademacher Complexity

- **Lemma 3.4 (Massart's Lemma)** [Mohri et al., 2018]

Let  $A \subseteq \mathbb{R}^n$  be a **finite set**, with  $r = \max_{x \in A} \|x\|_2$ , then the following holds:

$$\mathbb{E}_\sigma \left[ \frac{1}{n} \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right] \leq \frac{r \sqrt{2 \log |A|}}{n} \quad (17)$$

where  $\sigma_i$ 's are **independent uniform random variables** taking values in  $\{-1, +1\}$  and  $x_1, \dots, x_n$  are the components of vector  $x$ .

**Proof:** By Jensen's inequality

$$\begin{aligned} \exp \left( \lambda \mathbb{E}_\sigma \left[ \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right] \right) &\leq \mathbb{E}_\sigma \left[ \exp \left( \lambda \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right) \right] \\ &\leq \mathbb{E}_\sigma \left[ \sup_{x \in A} \exp \left( \lambda \sum_{i=1}^n \sigma_i x_i \right) \right] \\ &\leq \sum_{x \in A} \mathbb{E}_\sigma \left[ \exp \left( \lambda \sum_{i=1}^n \sigma_i x_i \right) \right] \\ &\leq \sum_{x \in A} \prod_{i=1}^n \mathbb{E}_\sigma [\exp (\lambda \sigma_i x_i)] \end{aligned}$$

We next use the independence of the  $\sigma_i$ 's, then apply *Hoeffding's lemma* since  $\sigma_i x_i$  are independent taking values in  $[-x_i, x_i]$ . And then apply the bound  $r$  on the norm  $\|x\|_2$ .

$$\mathbb{E}_\sigma [\exp (\lambda \sigma_i x_i)] \leq \exp \left( \frac{\lambda^2 (2x_i)^2}{8} \right) = \exp \left( \frac{\lambda^2 x_i^2}{2} \right)$$

Thus

$$\begin{aligned} \exp \left( \lambda \mathbb{E}_\sigma \left[ \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right] \right) &\leq \sum_{x \in A} \exp \left( \frac{\lambda^2 \sum_{i=1}^n (x_i)^2}{2} \right) = \sum_{x \in A} \exp \left( \frac{\lambda^2 \|x\|_2^2}{2} \right) \\ &\leq \sum_{x \in A} \exp \left( \frac{\lambda^2 r^2}{2} \right) = |A| \exp \left( \frac{\lambda^2 r^2}{2} \right) \end{aligned}$$

Taking the log of both sides and dividing by  $\lambda$  gives us:

$$\mathbb{E}_\sigma \left[ \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right] \leq \frac{\log |A|}{\lambda} + \frac{\lambda r^2}{2}.$$

Choosing optimal  $\lambda^* = \sqrt{2 \log |A|}/r$ , we have the Chernoff bound

$$\mathbb{E}_\sigma \left[ \sup_{x \in A} \sum_{i=1}^n \sigma_i x_i \right] \leq r \sqrt{2 \log |A|}.$$

Dividing both by  $n$  gives us the result. ■

- **Corollary 3.5** (*Rademacher Complexity Bounds by Growth Number*) [Mohri et al., 2018]

Let  $\mathcal{H}$  be a family of functions taking values in  $\{-1, +1\}$ . Then the following holds:

$$\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{\frac{2 \log \tau_{\mathcal{H}}(n)}{n}} \quad (18)$$

**Proof:** For a fixed sample  $\mathcal{D} = (x_1, \dots, x_n)$ , we denote by  $\mathcal{H}_{\mathcal{D}} \subset \{-1, +1\}^n$  the set of vectors of function values  $(h(x_1), \dots, h(x_n))$  where  $h$  is in  $\mathcal{H}$ . Since  $h \in \mathcal{H}$  takes values in  $\{-1, +1\}$ , the *norm* of these vectors is bounded by  $\sqrt{n}$ . We can then apply *Massart's lemma* as follows:

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{\mathcal{D}} \left[ \mathbb{E}_{\sigma} \left[ \sup_{u \in \mathcal{H}_{\mathcal{D}}} \frac{1}{n} \sum_{i=1}^n \sigma_i u_i \middle| \mathcal{D} \right] \right] \leq \mathbb{E}_{\mathcal{D}} \left[ \frac{\sqrt{n} \sqrt{2 \log |\mathcal{H}_{\mathcal{D}}|}}{n} \right]$$

By definition,  $|\mathcal{H}_{\mathcal{D}}|$  is bounded by *the growth function*, thus,

$$\mathfrak{R}_n(\mathcal{H}) \leq \mathbb{E}_{\mathcal{D}} \left[ \frac{\sqrt{n} \sqrt{2 \log \tau_{\mathcal{H}}(n)}}{n} \right] = \sqrt{\frac{2 \log \tau_{\mathcal{H}}(n)}{n}},$$

which concludes the proof.  $\blacksquare$

- **Definition** (*VC-Entropy*)

Let  $\mathcal{H}$  be an arbitrary collection of subsets in  $\mathcal{Y}$ . The quantity

$$h_{\mathcal{H}}(y_1, \dots, y_n) = \log |H \cap \{y_1, \dots, y_n\} : H \in \mathcal{H}|$$

is called *the Vapnik-Chervonenkis (or VC) entropy*. It is equal to  $\log \tau_{\mathcal{H}}(n)$ .

### 3.4 Generalization Bounds via Growth Function and VC-Dimension

- Combining Proposition 2.4 to Corollary 3.5, we have:

**Corollary 3.6** (*Growth Function Generalization Bound*) [Mohri et al., 2018]

Let  $\mathcal{H}$  be a family of functions taking values in  $\{-1, +1\}$ . Then, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , for any  $h \in \mathcal{H}$ ,

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{2 \log \tau_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \quad (19)$$

Growth function bounds can be also derived directly (without using Rademacher complexity bounds first). The resulting bound is then the following:

$$\mathcal{P} \left\{ \exists h \in \mathcal{H}, \left| L(h) - \widehat{L}_n(h) \right| > \epsilon \right\} \leq 4 \tau_{\mathcal{H}}(2n) \exp \left( -\frac{n \epsilon^2}{8} \right) \quad (20)$$

which only differs from (19) by constants.

- Applying *Sauer's Lemma* to Corollary 3.6, we have:

**Corollary 3.7 (VC-Dimension Generalization Bounds)** [Mohri et al., 2018]

Let  $\mathcal{H}$  be a family of functions taking values in  $\{-1, +1\}$  with **VC-dimension**  $d$ . Then, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , the following holds for all  $h \in \mathcal{H}$ :

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{2d \log(en/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \quad (21)$$

Thus, the form of this generalization bound is

$$L(h) \leq \widehat{L}_n(h) + O\left(\sqrt{\frac{\log(n/d)}{(n/d)}}\right), \quad (22)$$

which emphasizes **the importance of the ratio  $n/d$  for generalization**.

- **Remark** The theorem provides another instance of **Occam's razor principle** where simplicity is **measured** in terms of **smaller VC-dimension**.

VC-dimension bounds can be derived directly *without using an intermediate Rademacher complexity bound*: combining *Sauers lemma* with (20) leads to the following high-probability bound

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{8d \log(2en/d) + 8 \log(4/\delta)}{n}}, \quad (23)$$

which has the general form of (22). *The log factor plays only a minor role in these bounds*. A finer analysis via *metric entropy* and *Dudley's entropy integral* can be used to eliminate that factor.

### 3.5 Examples

- **Example (Threshold Functions over  $\mathbb{R}$ )**

Let  $\mathcal{H}$  be the **class of threshold functions** over  $\mathbb{R}$ ; i.e.  $h_a(x) := \mathbb{1}\{x \leq a\}$ . We claim that  $\text{VCdim}(\mathcal{H}) = 1$ .

Take a set  $\mathcal{D} = \{x_1\}$ . Now, if we take  $a = x_1 + 1$ , then we have  $h_a(x_1) = \mathbb{1}\{x_1 \leq a\} = 1$ , and if we take  $a = x_1 - 1$ , then we have  $h_a(x_1) = \mathbb{1}\{x_1 \leq a\} = 0$ . Therefore,  $\mathcal{H}_{\mathcal{D}}$  is the set of all functions from  $\mathcal{D}$  to  $\{0, 1\}$ , and  $\mathcal{H}$  *shatters*  $\mathcal{D}$ .

Now take a set  $\mathcal{D} = \{x_1, x_2\}$ , where  $x_1 \leq x_2$ . No  $h \in \mathcal{H}$  can account for the labeling  $(0, 1)$ , because any threshold that assigns the label 0 to  $x_1$  must assign the label 0 to  $x_2$  as well. Therefore not all functions from  $\mathcal{D}$  to  $\{0, 1\}$  are included in  $\mathcal{H}_{\mathcal{D}}$ ; hence  $\mathcal{D}$  is *not shattered* by  $\mathcal{H}$ . By definition, **VC dimension of  $\mathcal{H}$  is equal to 1 and  $\mathcal{H}$  is PAC learnable**.

- **Example (Intervals over  $\mathbb{R}$ )**

Let  $\mathcal{H}$  be the hypothesis class of **intervals over  $\mathbb{R}$** , namely,  $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ , where  $h_{a,b} : \mathbb{R} \rightarrow \{0, 1\}$  is a function such that  $h_{a,b}(x) = \mathbb{1}\{x \in (a, b)\}$ . We claim that  $\text{VCdim}(\mathcal{H}) = 2$ .

It is clear that the VC-dimension is *at least two*, since *all four dichotomies*  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$ ,  $(1, 1)$  can be realized. Now take an arbitrary set  $\mathcal{D} = \{x_1, x_2, x_3\}$  and assume without loss of generality that  $x_1 \leq x_2 \leq x_3$ . Then, the labeling  $(1, 0, 1)$  cannot be obtained by an interval and therefore  $\mathcal{H}$  does not shatter  $\mathcal{D}$ .



**Figure 6.1.** Left: 4 points that are shattered by axis aligned rectangles. Right: Any axis aligned rectangle cannot label  $c_5$  by 0 and the rest of the points by 1.

**Figure 2:** (left) There exists a 4-point set that can be shattered by axis aligned rectangle. (right). Any 5 point set cannot be shattered. [Shalev-Shwartz and Ben-David, 2014]

- **Example (*Axis Aligned Rectangles*)**

Let  $\mathcal{H}$  be the class of *axis aligned rectangles*, formally:

$$\mathcal{H} := \{h_{(a_1, a_2, b_1, b_2)} : a_1 \leq a_2 \wedge b_1 \leq b_2\},$$

where  $h_{(a_1, a_2, b_1, b_2)}(x_1, x_2) := \mathbb{1}\{(x_1, x_2) \in [a_1, a_2] \times [b_1, b_2]\}$ . We shall show in the following that  $\text{VCdim}(\mathcal{H}) = 4$ .

To prove this we need to *find a set of 4 points that are shattered by  $\mathcal{H}$* , and show that *no set of 5 points can be shattered by  $\mathcal{H}$* . Finding a set of 4 points that are shattered is easy (see Figure 2).

Now, consider any set  $\mathcal{D} \subset \mathbb{R}^2$  of 5 points. In  $\mathcal{D}$ , take a *leftmost point* (whose first coordinate is the smallest in  $\mathcal{D}$ ), a *rightmost point* (first coordinate is the largest), a *lowest point* (second coordinate is the smallest), and a *highest point* (second coordinate is the largest). Without loss of generality, denote  $\mathcal{D} = \{x_1, \dots, x_5\}$  and let  $x_5$  be the point that *was not selected*. Now, define the labeling  $(1, 1, 1, 1, 0)$ . It is *impossible to obtain this labeling* by an *axis aligned rectangle*. Indeed, such a rectangle must contain  $x_1, \dots, x_4$ ; but in this case the rectangle contains  $x_5$  as well, because its coordinates are within the intervals defined by the selected points. So,  $\mathcal{D}$  is not shattered by  $\mathcal{H}$ .

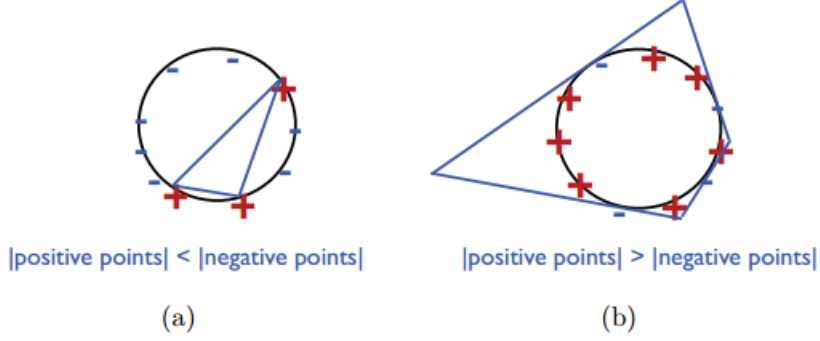
- **Example (*Hyperplanes in  $\mathbb{R}^d$* )**

Consider the set of *hyperplanes* in  $\mathbb{R}^d$ . A hyperplane can be represented as  $\langle w, x \rangle + b$ . Let  $\mathcal{H}$  be set of functions  $\text{sgn}(\langle w, x \rangle + b)$  for all  $w \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ . We claim that  $\text{VCdim}(\mathcal{H}) = d + 1$ , that is *the dimension of parameter set for hyperplane*.

We derive a *lower bound* by starting with a set of  $d + 1$  points in  $\mathbb{R}^d$ , setting  $x_0$  to be the *origin* and defining  $x_i$ , for  $i \in \{1, \dots, d\}$ , as the point whose  $i$ -th coordinate is 1 and all others are 0. Let  $y_0, y_1, \dots, y_d \in \{-1, +1\}$  be an arbitrary set of labels for  $x_0, x_1, \dots, x_d$ . Let  $w$  be the vector whose  $i$ -th coordinate is  $y_i$  and  $b = y_0/2$ . Then the classifier defined by the hyperplane of equation  $\langle w, x \rangle + y_0/2 = 0$  *shatters*  $x_0, x_1, \dots, x_d$  since for any  $i \in [0, d]$ ,

$$\text{sgn}\left(\langle w, x_i \rangle + \frac{y_0}{2}\right) = \text{sgn}\left(y_i + \frac{y_0}{2}\right) = y_i.$$

To obtain an *upper bound*, it suffices to show that *no set of  $d + 2$  points can be shattered by halfspaces*. We have the following *Radon's theorem*, which confirms that  $\mathcal{D}$  with  $d + 2$  points can be *partitioned into two subsets  $\mathcal{D}_1$  and  $\mathcal{D}_2$  whose convex hull intersect*. Observe that when two sets of points  $\mathcal{D}_1$  and  $\mathcal{D}_2$  are *separated by a hyperplane*, their convex



**Figure 3.4** Convex  $d$ -gons in the plane can shatter  $2d + 1$  points. (a)  $d$ -gon construction when there are more negative labels. (b)  $d$ -gon construction when there are more positive labels.

**Figure 3:** Convex  $d$ -gons in the plane can shatter  $2d + 1$  points. [Mohri et al., 2018]

hulls are also separated by that hyperplane. Thus,  $\mathcal{D}_1$  and  $\mathcal{D}_2$  **cannot be separated by a hyperplane** and  $\mathcal{D}$  is not shattered. Combining our lower and upper bounds, we have proven that the VC-dimension of hyperplanes in  $\mathbb{R}^d$  is equal to  $d + 1$ .

**Theorem 3.8 (Radon's Theorem)** [Mohri et al., 2018]

Any set  $\mathcal{X}$  of  $d + 2$  points in  $\mathbb{R}^d$  can be partitioned into two subsets  $\mathcal{X}_1$  and  $\mathcal{X}_2$  such that **the convex hulls** of  $\mathcal{X}_1$  and  $\mathcal{X}_2$  intersect.

- **Example (Finite Dimensional Vector Space of Real Functions on  $\mathbb{R}^d$ )**

Let  $\mathcal{F}$  be a *finite-dimensional vector space* of real functions on  $\mathbb{R}^d$ ,  $\dim(\mathcal{F}) = r < \infty$ . Let  $\mathcal{H}$  be the set of hypotheses, which we refer to as **the subgraph class** of  $\mathcal{F}$ .

$$\mathcal{G}(\mathcal{F}) := \left\{ \left\{ x \in \mathbb{R}^d : f(x) \leq 0 \right\} : f \in \mathcal{F} \right\}.$$

We claim that  $\text{VCdim}(\mathcal{G}(\mathcal{F})) \leq r$ . (Hint: select an arbitrary set of  $m = r + 1$  points and consider linear mapping  $u : \mathcal{F} \rightarrow \mathbb{R}^m$  defined by:  $u(f) = (f(x_1), \dots, f(x_m))$ .)

We summarize the result in the following proposition

**Proposition 3.9 (Finite-Dimensional Vector Spaces)** [Mohri et al., 2018, Wainwright, 2019]

Let  $\mathcal{F}$  be a vector space of functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  with dimension  $\dim(\mathcal{F}) = r < \infty$ . Then the subgraph class  $\mathcal{G}(\mathcal{F})$  has VC dimension **at most**  $r$ .

**Proof:** Let  $x := (x_1, \dots, x_m)$  where  $m = r + 1$ . Consider linear mapping  $u_x : \mathcal{F} \rightarrow \mathbb{R}^m$  induced by  $x$  as  $u_x(f) = (f(x_1), \dots, f(x_m))$ . By construction, the **range** of the mapping  $u_x$  is a *linear subspace* of  $\mathbb{R}^m$  with dimension at most  $\dim(\mathcal{F}) = r = m - 1 < m$ . Therefore, there must exist a *non-zero vector*  $\gamma \in \mathbb{R}^m$  such that  $\langle \gamma, u_x(f) \rangle = 0$  for all  $f \in \mathcal{F}$ . We may assume without loss of generality that at least one coordinate is positive, and then write

$$\sum_{i: \gamma_i > 0} \gamma_i f(x_i) = \sum_{j: \gamma_j \leq 0} (-\gamma_j) f(x_j), \quad \forall f \in \mathcal{F}. \quad (24)$$

Proceeding via proof by *contradiction*, suppose that there were to exist some  $f \in \mathcal{F}$  such that the associated subgraph set  $\{x \in \mathbb{R}^d : f(x) \leq 0\}$  included only the subset samples  $\{x_j : \gamma_j \leq 0\}$ .



For such a function  $f$ , the left-hand side of equation (24) would be *strictly positive* while the right-hand side would be *non-positive*, which is a *contradiction*. We conclude that  $\mathcal{H}(\mathcal{F})$  fails to shatter the set  $\{x_1, \dots, x_m\}$ , as claimed. ■

- **Example (Convex  $d$ -gons in  $\mathbb{R}^2$ )**

Consider  $\mathcal{H}$  the class of *convex  $d$ -gons in the plane*. We claim that  $\underline{\text{VCdim}}(\mathcal{H}) = 2d + 1$ ,

To get a lower bound, we show that *any set of  $2d + 1$  points can be fully shattered*. To do this, we select  $2d + 1$  points that *lie on a circle*, and for a particular labeling, if there are *more negative than positive labels*, then the *points with the positive labels are used as the polygon's vertices*, as in Figure 3 (a). Otherwise, the *tangents of the negative points* serve as the *edges of the polygon*, as shown in Figure 3 (b).

To derive an upper bound, it can be shown that *choosing points on the circle maximizes the number of possible dichotomies*, and thus  $\text{VCdim}(\mathcal{H}) = 2d + 1$ .

Note also that if  $\mathcal{H}$  is the set of *all convex polygons*, then  $\underline{\text{VCdim}}(\mathcal{H}) = \infty$ .

- **Example (Spheres in  $\mathbb{R}^d$ )**

Consider the space of all  $d-1$  dimensional sphere  $\mathcal{S}^d$ , with sphere  $\mathbb{S}_{a,b} := \{x \in \mathbb{R}^d : \|x - a\|_2 \leq b\}$ . Note that

$$x \in \mathbb{S}_{a,b} \Leftrightarrow f_{a,b}(x) := \|x\|_2^2 - 2\langle a, x \rangle + \|a\|_2^2 - b^2 \leq 0.$$

so that the sphere  $\mathbb{S}_{a,b}$  is a *subgraph* of the function  $f_{a,b}$ .

In order to compute the VC dimension of  $\mathcal{S}^d$ , we first define a feature map  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+2}$  via

$$\varphi(x) := (1, x_1, \dots, x_d, \|x\|_2^2)$$

and then consider functions of the form

$$g_c(x) := \langle c, \varphi(x) \rangle, \quad c \in \mathbb{R}^{d+2}$$

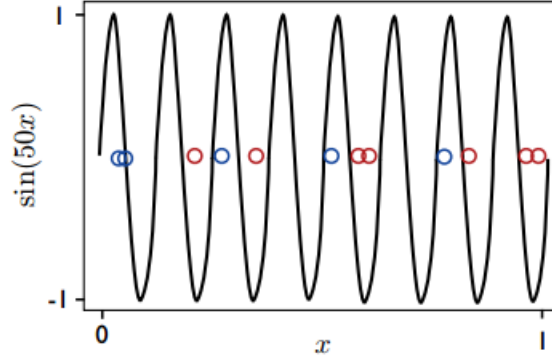
The family of functions  $\{g_c, c \in \mathbb{R}^{d+2}\}$  is a vector space of dimension  $d+2$ , and it contains the function class  $\{f_{a,b}, (a,b) \in \mathbb{R}^d \times \mathbb{R}_+\}$ . Consequently, by applying Proposition on vector space of real-valued functions to this larger vector space, we conclude that  $\underline{\text{VCdim}}(\mathcal{S}^d) \leq d + 2$ . *This bound is adequate for many purposes, but is **not sharp**: a more careful analysis shows that the VC dimension of spheres in  $\mathbb{R}^d$  is actually  $d + 1$ .*

- **Example (Sine Functions)**

The previous examples could suggest that *the VC-dimension of  $\mathcal{H}$  coincides with the number of free parameters defining  $\mathcal{H}$* . *However, this does not hold in general.*

The following provides a striking example from this point of view. Consider the following *family of sine functions*:  $\mathcal{H} := \{t \rightarrow \sin(\omega t) : \omega \in \mathbb{R}\}$ . One instance of this function class is shown in Figure 4. These sine functions can be used to *classify the points on the real line*: a point is labeled *positively* if it is *above the curve*, *negatively* otherwise.

Although this family of sine function is defined via *a single parameter*,  $\omega$ , it can be shown that  $\underline{\text{VCdim}}(\mathcal{H}) = \infty$ .



**Figure 3.5** An example of a sine function (with  $\omega = 50$ ) used for classification.

Figure 4: The class of sine functions has infinity VC dimension. [Mohri et al., 2018]

### 3.6 Lower Bounds

- **Remark (No-Free-Lunch via VC-Dimension)**

This section provides **lower bounds** on the *generalization error* of *any learning algorithm* in terms of the *VC-dimension* of the hypothesis set used.

These lower bounds are shown by finding *for any algorithm* a ‘**bad**’ distribution. Since the learning algorithm is *arbitrary*, it will be difficult to specify that particular distribution. Instead, it suffices to *prove its existence non-constructively*. At a high level, the proof technique used to achieve this is *the probabilistic method* of Paul Erdős.

In the context of the following proofs,

1. *first a lower bound* is given on the expected error over *the parameters defining the distributions*.
2. From that, *the lower bound* is shown to **hold for at least one set of parameters**, that is one distribution.

- **Proposition 3.10 (Lower Bound, Realizable Case)** [Mohri et al., 2018]

Let  $\mathcal{H}$  be a hypothesis set with **VC-dimension**  $d > 1$ . Then, for **any** learning algorithm  $\mathcal{A}$ , there **exist** a **distribution**  $\mathcal{P}$  over  $\mathcal{X}$  and a **target function**  $c \in \mathcal{H}$  such that

$$\mathcal{P}^n \left\{ L_{\mathcal{P},c}(\mathcal{A}_{\mathcal{H}}(\mathcal{D}_n)) > \frac{d-1}{32n} \right\} \geq \frac{1}{100}$$

- **Remark (Infinite VC-dimension  $\Rightarrow$  Non PAC-learnable)**

The theorem shows that for any algorithm  $\mathcal{A}$ , there exists a ‘*bad*’ distribution over  $\mathcal{X}$  and a target function  $f$  for which the error of the hypothesis returned by  $\mathcal{A}$  is

$$\Omega\left(\frac{d}{n}\right)$$

with some constant probability. This further demonstrates the key role played by the VC-dimension in learning. The result implies in particular that **PAC-learning in the non-realizable case is not possible when the VC-dimension is infinite**.

- **Proposition 3.11** (*Lower Bound, Non-Realizable Case*) [Mohri et al., 2018]  
Let  $\mathcal{H}$  be a hypothesis set with **VC-dimension**  $d > 1$ . Then, for any learning algorithm  $\mathcal{A}$ , there exists a distribution  $\mathcal{P}$  over  $\mathcal{X} \times \{0, 1\}$  such that:

$$\mathcal{P}^n \left\{ L_{\mathcal{P}}(\mathcal{A}_{\mathcal{H}}(\mathcal{D}_n)) - \inf_{h \in \mathcal{H}} L_{\mathcal{P}}(h) > \sqrt{\frac{d}{320n}} \right\} \geq \frac{1}{64}$$

Equivalently, for any learning algorithm, **the sample complexity** verifies

$$n \geq \frac{d}{320\epsilon^2}.$$

- **Remark** The theorem shows that for any algorithm  $\mathcal{A}$ , **in the non-realizable case** (i.e. the Bayes classifier is not in  $\mathcal{H}$ ), there exists a ‘bad’ distribution over  $\mathcal{X} \times \{0, 1\}$  such that the error of the hypothesis returned by  $\mathcal{A}$  is

$$\Omega\left(\sqrt{\frac{d}{n}}\right)$$

with some constant probability. This coincides with the upper bound  $\mathcal{O}(\sqrt{d/n})$ , which means that it is both *necessary and sufficient condition* for PAC learnability in agnostic setting.

The *VC-dimension* appears as a critical quantity in learning in *this general setting* as well. In particular, **with an infinite VC-dimension, agnostic PAC-learning is not possible.**

### 3.7 Fundamental Theorem of PAC Learning

- **Theorem 3.12** (*The Fundamental Theorem of Statistical Learning*). [Shalev-Shwartz and Ben-David, 2014]  
Let  $\mathcal{H}$  be a hypothesis class of functions from a domain  $\mathcal{X}$  to  $\{0, 1\}$  and let the loss function be the 0-1 loss. Then, the following are **equivalent**:

1.  $\mathcal{H}$  has the **uniform convergence property**.
2. Any **ERM** rule is a successful **agnostic PAC learner** for  $\mathcal{H}$ .
3.  $\mathcal{H}$  is **agnostic PAC learnable**.
4.  $\mathcal{H}$  is **PAC learnable**.
5. Any **ERM** rule is a successful **PAC learner** for  $\mathcal{H}$ .
6.  $\mathcal{H}$  has a **finite VC-dimension**.

- **Remark** From the above theorem, we see that  $\mathcal{H}$  is **PAC learnable**  $\Leftrightarrow$   $\mathcal{H}$  is **agnostic PAC learnable**.

If  $\mathcal{H}$  is (agnostic) PAC learnable, then the ERM rule is a successful (agnostic) PAC learner for  $\mathcal{H}$ .

- **Theorem 3.13** (*The Fundamental Theorem of Statistical Learning, Quantitative Version*). [Shalev-Shwartz and Ben-David, 2014]

Let  $\mathcal{H}$  be a hypothesis class of functions from a domain  $\mathcal{X}$  to  $\{0, 1\}$  and let the loss function be the 0-1 loss. Assume that  $VCdim(\mathcal{H}) = d < \infty$ . Then, there are absolute constants  $C_1, C_2$  such that

1.  $\mathcal{H}$  has the **uniform convergence property** with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2.  $\mathcal{H}$  is **agnostic PAC learnable** with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3.  $\mathcal{H}$  is **PAC learnable** with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}$$

## References

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.