# Lecture 3: Rademacher Complexity and VC-Dimension

Tianpei Xie

Dec. 18th., 2022

## Contents

# 1 PAC Learnability for Infinite Hypothese 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***.

$$\begin{aligned} L(h_n) - \inf_{h\in\mathcal{H}} L(h) &= L(h_n) - \widehat{L}(h_n) + \widehat{L}(h_n) - L(h^*) \\ &\leq L(h_n) - \widehat{L}(h_n) + \widehat{L}(h^*) - L(h^*) \\ &\leq 2\sup_{h\in\mathcal{H}} \left| L(h) - \widehat{L}(h) \right| \end{aligned} \tag{1}$$

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

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

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

- **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 cannod 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* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. To each $h : \mathcal{X} \to \mathcal{Y}$, we can associate a function $g$:

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

  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 $\mathcal{D}_n = (z_1, \ldots, z_n)$ 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:

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

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

- **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**

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

  which measures the correlation between the random noise $\sigma_m := \{\sigma_i\}$ and $g_{\mathcal{D}_n} := \{g(X_i, Y_i)\}_{i=1}^n$. The supremum $\sup_{g \in \mathcal{G}} \langle \sigma_n \, , \, g_{\mathcal{D}_n} \rangle / n$ 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 $\mathcal{D}_n$.

- **Remark** (***Rademacher Variables*** $=$ ***Symmetric Bernoulli Random Variable***)
  *A Rademacher variables* is a *random variable* on $\{-1, +1\}$ with ***symmetric probability*** $P(\sigma_i = +1) = P(\sigma_i = -1) = \frac{1}{2}$, i.e. a set of i.i.d. symmetric Bernoulli random variables.

  *A Rademacher process* is a stochastic process $\{\sigma_i g(X_i)\}$ conditioning on $\{X_i\}$, which is subgaussian process.

- **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 $\mathcal{P}$:

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

- **Proposition 2.1** ***(Consistency Bound with respect to 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* $g \in \mathcal{G}$*:*

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

  *and*

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

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

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

Note that $\Phi(\mathcal{D}_n)$ is a *random variable*. This function has *bounded difference* if *only one sample makes changes*. The proof consists of three parts:

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

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

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

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

   This obtain the inequality (3).

3. Bound *the probability* of *tail events* for *empircal Radmatcher complexity*:

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

Combines the difference in (1) and (3) using union bounds.

- **Remark** 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

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

- **Lemma 2.2** *(**Radmatcher 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**:*

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

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

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

- **Proposition 2.3** *(**Rademacher Complexity Generalization Bounds**)[Mohri et al., 2018]*
*Let $\mathcal{H}$ be a family of functions taking values in $\{-1,+1\}$ and let $\mathcal{P}_X$ 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** $h \in \mathcal{H}$:*

$$L(h) \leq \widehat{L}_n(h) + \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{5}$$

*and*

$$L(h) \leq \widehat{L}_n(h) + \widehat{\mathfrak{R}}_{\mathcal{D}_n}(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{6}$$

- **Remark** (***Compute Emipirical Radmatcher Complexity***)
  Note that the Radmatcher 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.

# 3 VC-Dimension

## 3.1 Definition of VC-Dimension

- **Remark** Recall the decomposition of generalization error:

$$L(g_n^*) - L^* = \underbrace{\left(L(g_n^*) - \inf_{g\in\mathcal{H}} L(g)\right)}_{\text{estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{H}} L(g) - L^*\right)}_{\text{approximation error}}.$$

  where the first term is called *estimation error* and the second term is called approximation error. The definition of ***PAC learnability*** requires that *the estimation error* would be bounded ***uniformly*** *over* ***all distributions***.

  Can *infinite-size classes* be learnable, and, if so, what determines their *sample complexity* ?

- **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 $\mathcal{D} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$.

***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,

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

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 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 $\tau_{\mathcal{H}} : \mathbb{N} \to \mathcal{N}$, is defined as

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

  In words, $\tau_{\mathcal{H}}(m)$ 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* $n$*,*

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \tag{7}$$

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

$$\tau_{\mathcal{H}}(n) \leq \left(\frac{e\, n}{d}\right)^d. \tag{8}$$

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

$$\forall\, \mathcal{H},\ |\mathcal{H}_{\mathcal{D}}| \leq |\{\mathcal{B} \subseteq \mathcal{D} : \mathcal{H} \text{ shatters } \mathcal{B}\}| \tag{9}$$

  The reason why Equation (9) 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* $\left(\frac{e\, n}{d}\right)^d$.

  We are left with proving Equation (9) and we do it using an ***inductive argument***.

1. For $n = 1$, no matter what $\mathcal{H}$ is, either both sides of Equation (9) equal 1 or both sides equal 2 (*the empty set* is always considered to be shattered by $\mathcal{H}$).

2. *Assume* Equation (9) holds for sets of size $k < n$ and let us prove it for sets of size $n$.

   Fix $\mathcal{H}$ and $\mathcal{D} = \{x_1, \ldots, x_n\}$. Denote $\mathcal{D}_{-1} = \{x_2, \ldots, x_n\}$ and in addition, define the following two sets:

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

   and

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

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

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

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

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

   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

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

   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. ∎

### 3.3 Relate Growth Function to Radmatcher 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:*

  $$\mathbb{E}_\sigma \left[\frac{1}{n} \sup_{x \in A} \sum_{i=1}^{n} \sigma_i x_i\right] \leq \frac{r\sqrt{2 \log |A|}}{n} \tag{10}$$

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

- **Corollary 3.5** ***(Radmatcher 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:*

$$\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{\frac{2 \log \tau_{\mathcal{H}}(n)}{n}} \tag{11}$$

**Proof:** For a fixed sample $\mathcal{D} = (x_1, \ldots, x_n)$, we denote by $\mathcal{H}_{\mathcal{D}} \subset \{-1, +1\}^m$ the set of vectors of function values $(h(x_1), \ldots, h(x_n))$ 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:

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

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. ∎

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

- Combining Proposition 2.3 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 $h \in \mathcal{H}$,*

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{2 \log \tau_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{12}$$

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

$$\mathcal{P}\left\{ \left| L(h) - \widehat{L}_n(h) \right| > \epsilon \right\} \leq 4 \tau_{\mathcal{H}}(2n) \, \exp\left( -\frac{n\epsilon^2}{8} \right) \tag{13}$$

*which only differs from (12) 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 $h \in \mathcal{H}$:*

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{2d \log(en/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{14}$$

*Thus, the form of this generalization bound is*

$$L(h) \leq \widehat{L}_n(h) + O\left(\sqrt{\frac{\log(n/d)}{(n/d)}}\right), \tag{15}$$

*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 (13) leads to the following high-probability bound

$$L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{8d\log(2en/d) + 8\log(4/\delta)}{n}}, \tag{16}$$

  which has the general form of (15). *The log factor plays only a minor role in these bounds.* A finer analysis can be used in fact to eliminate that factor.

### 3.5 Examples

### 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 Erdös.

  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 3.8** *(**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 3.9** *(**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. 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***.

## References

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.