Algorithmic Foundations of Learning **Lecture 4**

VC Dimension. Covering and Packing Numbers

Lecturer: Patrick Rebeschini *Version: November 1, 2022*

4.1 Introduction

In the last lecture we established bounds for the Rademacher complexity in various examples of regression. In this lecture we investigate the setting of binary classification, where the Rademacher complexity can be bounded by combinatorial quantities, namely, the growth function and the VC-dimension.

The binary classification setting is represented by the choice $\mathcal{X} \subseteq \mathbb{R}^d$ for a given dimension d, $\mathcal{Y} = \{-1, 1\}$, and $\mathcal{A} \subseteq \mathcal{B} = \{a : \mathcal{X} \to \{-1, 1\}\}$. In this case:

$$\mathcal{A} \circ \{x_1, \dots, x_n\} := \{(a(x_1), \dots, a(x_n)) \in \{-1, 1\}^n : a \in \mathcal{A}\}.$$

If we use the zero-one loss function to evaluate the quality of our action, we can directly relate the Rademacher complexity of the set $\mathcal{L} \circ \{(x_1, y_1), \dots, (x_n, y_n)\}$ with the Rademacher complexity of the set $\mathcal{A} \circ \{x_1, \dots, x_n\}$, without using the contraction property that we used for regression.

Proposition 4.1 *Choose the loss function* $\hat{y} \to \phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$. *Then, for any* $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$,

$$\boxed{\mathtt{Rad}(\mathcal{L} \circ \{(x_1, y_1), \dots, (x_n, y_n)\}) = \frac{1}{2} \mathtt{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\})}$$

Proof: As $\phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y} = (1 - y\hat{y})/2$, using that $\mathbf{E}\Omega_i = 0$ and that $y_i\Omega_i$ has the same distribution as Ω_i, we get

$$\begin{aligned} \mathtt{Rad}(\mathcal{L} \circ s) &= \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i \phi(a(x_i), y_i) = \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i \frac{1 - y_i a(x_i)}{2} \\ &= \frac{1}{2} \mathbf{E} \frac{1}{n} \sum_{i=1}^{n} \Omega_i + \frac{1}{2} \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i y_i a(x_i) = \frac{1}{2} \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i a(x_i) = \frac{1}{2} \mathtt{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\}). \end{aligned}$$

■

4.2 Growth function

For any $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, we have $\mathcal{A} \circ x \subseteq \{-1, 1\}^n$ and the set $\mathcal{A} \circ x$ is finite even if the class $\mathcal{A}$ has infinitely many elements. In fact, we have $|\mathcal{A} \circ x| \leq 2^n$. The maximal cardinality of this set over the choice of n points $x_1, \dots, x_n \in \mathcal{X}$ is called the growth function of $\mathcal{A}$ (evaluated at n).

Definition 4.2 *The* growth function *of* $\mathcal{A}$ *is defined as follows, for any integer* $n \geq 1$*:*

$$\boxed{\tau_{\mathcal{A}}(n) := \sup_{x \in \mathcal{X}^n} |\mathcal{A} \circ x|}$$

The quantity $\tau_{\mathcal{A}}(n)$ is the maximal cardinality of the set of distinct labelings of n points in $\mathcal{X}$ that can be obtained using classifiers from $\mathcal{A}$. As the growth function of $\mathcal{A}$ if finite, an immediate application of Massart's lemma, Lemma 2.9, shows that $\tau_{\mathcal{A}}(n)$ can be used to upper bound the Rademacher complexity.

Proposition 4.3 *For any $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$ we have*

$$\boxed{\texttt{Rad}(\mathcal{A} \circ x) \le \sqrt{\frac{2 \log \tau_{\mathcal{A}}(n)}{n}}}$$

Proof: By Massart's lemma we immediately get

$$\texttt{Rad}(\mathcal{A} \circ x) = \frac{1}{n} \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{i=1}^{n} \Omega_i a(x_i) \le \max_{t \in \{-1,1\}^n} \|t\|_2 \frac{\sqrt{2 \log |\mathcal{A} \circ x|}}{n} \le \sqrt{\frac{2 \log \tau_{\mathcal{A}}(n)}{n}}.$$

■

As mentioned above, $\tau_{\mathcal{A}}(n) \le 2^n$. However, if the class $\mathcal{A}$ is such that $\tau_{\mathcal{A}}(n) = 2^n$ for all n, the bound in Proposition 4.3 does not imply that the Rademacher complexity goes to zero as the sample size n goes to infinity. For this to happen, we need the class $\mathcal{A}$ to have a growth function that grows *polynomially* in n, not exponentially.

If $|\mathcal{A}| < \infty$, then $\tau_{\mathcal{A}}(n) \le |\mathcal{A}|$, and below we give a few examples where $\mathcal{A}$ is infinite but $\tau_{\mathcal{A}}(n)$ growths polynomially.

Example 4.4 (Half spaces over the real line) *Let $d = 1$ and consider the class of half-line classifiers given by $\mathcal{A} = \{x \in \mathbb{R} \to a(x) = 2\mathbf{1}_{x \le w} - 1 : w \in \mathbb{R}\}$. Clearly, $\mathcal{A}$ is of infinite size. This family of classifiers can produce $n+1$ distinct labelings for any set of n distinct points $\{x_1, \dots, x_n\}$, corresponding to the $n+1$ patterns (here we use the label 0 to denote the label -1)*

$$\begin{array}{l} 0000\cdots0 \\ 1000\cdots0 \\ 1100\cdots0 \\ \vdots \\ 1111\cdots1 \end{array}$$

For any set $\{x_1, \dots, x_n\}$, possibly with repetitions, we have $|\mathcal{A} \circ \{x_1, \dots, x_n\}| \le n+1$ so $\tau_{\mathcal{A}}(n) = n+1$.

Example 4.5 (Intervals over the real line) *Let $d = 1$ and consider the class of interval classifiers given by $\mathcal{A} = \{x \in \mathbb{R} \to a(x) = 2\mathbf{1}_{w^- \le x \le w^+} - 1 : w^- \le w^+\}$. Clearly, $\mathcal{A}$ is of infinite size. This family of classifiers can produce $1 + n(n+1)/2$ distinct labelings for any set of n distinct points $\{x_1, \dots, x_n\}$, corresponding to the $1 + n(n+1)/2$ patterns (here we use the label 0 to denote the label -1)*

$$\begin{array}{llllll} 0000\cdots00 & & & & & \\ 1000\cdots00 & 0100\cdots00 & 0010\cdots00 & \cdots & 00000\cdots10 & 00000\cdots01 \\ 1100\cdots00 & 0110\cdots00 & 0011\cdots00 & \cdots & 0000\cdots11 & \\ \vdots & & & & & \\ 1111\cdots11 & & & & & \end{array}$$

There is 1 *pattern with zero* 1*'s,* n *patterns with one* 1*,* $n-1$ *patterns with two* 1*'s, etc, namely,*

$$1 + \sum_{k=0}^{n-1} (n-k) = 1 + n(n+1)/2.$$

For any set $\{x_1, \ldots, x_n\}$*, possibly with repetitions, we have* $|\mathcal{A} \circ \{x_1, \ldots, x_n\}| \leq 1 + n(n+1)/2$ *so* $\tau_{\mathcal{A}}(n) = 1 + n(n+1)/2$.

4.3 VC dimension

The growth function provides a notion of complexity for the class of classifiers $\mathcal{A}$ in binary classification. However, as the examples above already attest, it is not always easy to compute the growth function. One would like to relate the notion of growth function to a quantity that is more amenable to computations. This notion is given by the Vapnik-Chervonenkis (VC) dimension.

Recall that $\tau_{\mathcal{A}}(n) \leq 2^n$.

Definition 4.6 (VC dimension) *The* Vapnik-Chervonenkis dimension*, or VC-dimension, of* $\mathcal{A}$ *is the largest integer* n *such that* $\tau_{\mathcal{A}}(n) = 2^n$*, namely,*

$$\boxed{\mathtt{VC}(\mathcal{A}) := \max\{n \in \mathbb{N} : \tau_{\mathcal{A}}(n) = 2^n\}}$$

If $\tau_{\mathcal{A}}(n) = 2^n$ *for all integer* n*, then* $\mathtt{VC}(\mathcal{A}) = \infty$.

The quantities $\tau_{\mathcal{A}}(1), \tau_{\mathcal{A}}(2), \ldots$ are also called *shatter coefficients*. We say that $\mathcal{A}$ *shatters* the set of points $\{x_1, \ldots, x_n\}$ if $|\mathcal{A} \circ \{x_1, \ldots, x_n\}| = 2^n$. By definition, the n-th shatter coefficient $\tau_{\mathcal{A}}(n)$ is equal to 2^n if there exists a set of points $\{x_1, \ldots, x_n\}$ that is shattered by $\mathcal{A}$. The VC dimension is the maximum number of different elements that is shattered by $\mathcal{A}$.

Example 4.7 (Half spaces over the real line) *As* $\tau_{\mathcal{A}}(n) = n + 1$*, we have* $\tau_{\mathcal{A}}(1) = 2^1$ *and* $\tau_{\mathcal{A}}(2) = 3 < 2^2$*. Thus,* $\mathtt{VC}(\mathcal{A}) = 1$.

Example 4.8 (Intervals over the real line) *As* $\tau_{\mathcal{A}}(n) = 1 + n(n+1)/2$*, we have* $\tau_{\mathcal{A}}(1) = 2^1$*,* $\tau_{\mathcal{A}}(2) = 2^2$*, and* $\tau_{\mathcal{A}}(3) = 7 < 2^3$*. Thus,* $\mathtt{VC}(\mathcal{A}) = 2$.

The previous examples show how to compute the VC dimension having knowledge of the growth function $\tau_{\mathcal{A}}$. However, as announced above, the convenience of the VC dimension stems from the fact that we can compute it even if we do not known $\tau_{\mathcal{A}}$. To prove that $\mathtt{VC}(A) = k$ *it suffices* to find a set of distinct points $\{x_1, \ldots, x_k\}$ that are shattered by $\mathcal{A}$ (i.e., classifiers in $\mathcal{A}$ can assign all possible 2^k labelings to these points) and to prove that any set of points $\{x_1, \ldots, x_{k+1}\}$ can not be shattered by $\mathcal{A}$ (i.e., for any set of $k+1$ points there is a label that can *not* be assigned by classifiers in $\mathcal{A}$). To make this point, we revisit the examples given above using this approach.

Example 4.9 (Half spaces over the real line) *Given the set* $\{x_1\}$*, we can find* $a \in \mathcal{A}$ *such that* $a(x_1) = -1$ *and* $a(x_1) = 1$*, i.e., both patterns* 0 *and* 1 *can be reproduced. Given the set* $\{x_1, x_2\}$ *with distinct elements, any* $a \in \mathcal{A}$ *can not reproduce the pattern* 01*. Thus,* $\mathtt{VC}(\mathcal{A}) = 1$.

Example 4.10 (Intervals over the real line) *Given the set* $\{x_1, x_2\}$ *with distinct elements, the patterns* 00, 10, 01, *and* 11 *can be replicated. Given the set* $\{x_1, x_2, x_3\}$ *with distinct elements, any* $a \in \mathcal{A}$ *can not reproduce the pattern* 101. *Thus,* $\texttt{VC}(\mathcal{A}) = 2$.

See **Problem 2.5** in the Problem Sheets for further examples that emphasize the convenience of the VC dimension over the growth function.

The following lemma shows how the growth function is related to the VC dimension. In particular, this lemma shows that if the VC dimension if finite, then the growth function eventually growths at most polynomially in n.

Lemma 4.11 (Sauer-Shelah's Lemma) *For any* $n \ge \texttt{VC}(\mathcal{A})$, *we have*

$$\boxed{\tau_{\mathcal{A}}(n) \le \sum_{k=0}^{\texttt{VC}(\mathcal{A})} \binom{n}{k} \le \left(\frac{en}{\texttt{VC}(\mathcal{A})}\right)^{\texttt{VC}(\mathcal{A})}}$$

In particular,

$$\boxed{\tau_{\mathcal{A}}(n) \begin{cases} = 2^n & \text{if } n \le \texttt{VC}(\mathcal{A}) \\ \le \left(\frac{en}{\texttt{VC}(\mathcal{A})}\right)^{\texttt{VC}(\mathcal{A})} & \text{if } n \ge \texttt{VC}(\mathcal{A}) \end{cases}}$$

Proof: See **Problem 2.4** in the Problem Sheets. ■

Remark 4.12 (On the importance of the VC dimension) *In binary classification, the VC dimension* characterises *learnable problems. It can be shown that if the VC dimension is finite, then for any distribution there exists a classifier that achieves arbitrary small error with a* polynomial *number of samples. On the other hand, if the VC dimension is infinite, then for any classifier there exists a distribution where the classifier requires an exponential number of samples to reach an arbitrary small error. The VC theory is typically covered in courses in computer science, and has many connections to combinatorics. In this lecture notes we will limit to emphasize some key ideas, in particular relating Rademacher bounds to VC-dimension bounds.*

4.4 VC dimension bound for Rademacher complexity

We now have the ingredients to bound the Rademacher complexity in terms of the VC dimension.

Proposition 4.13 *For any* $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ *we have*

$$\boxed{\texttt{Rad}(\mathcal{A} \circ x) \le \sqrt{\frac{2\,\texttt{VC}(\mathcal{A}) \log(en/\texttt{VC}(\mathcal{A}))}{n}}}$$

Proof: It follows from Proposition 4.3 and Lemma 4.11. ■

Recall that Proposition 2.11 and Proposition 4.1 yield the following bound:

$$\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \le \mathbf{E}\, \texttt{Rad}(\mathcal{A} \circ \{X_1, \ldots, X_n\})$$

Hence, the bound on the Rademacher complexity given in Proposition 4.13 is useful as it reduces the problem of establishing generalisation bounds to the problem of computing the VC dimension for the chosen class of classifiers $\mathcal{A}$. However, this bound is "data-independent" as it holds for *any* $x \in \mathcal{X}^n$. As such, this bound is crude as it does not allow to exploit the *statistical* nature of the data.

The term $\log(en/\texttt{VC}(\mathcal{A}))$ can be removed using covering numbers and a technique called chaining, as we will see in the next lecture.

4.5 Covering and Packing Numbers

We now introduce covering numbers and packing numbers in general terms. These quantities are related and play a key role in various settings beyond maximal inequalities (that we are currently investigating), including minimax lower bounds as we will see later on in the course.

Recall that a *pseudometric space* $(\mathcal{S}, \rho)$ is a set $\mathcal{S}$ and a function $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ (called a *pseudometric*) such that, for any $x, y, z \in \mathcal{S}$ we have:

- $\rho(x, y) = \rho(y, x)$ (symmetry);
- $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (triangle inequality);
- $\rho(x, x) = 0$.

A *metric space* is obtained if one further assumes that $\rho(x, y) = 0$ implies $x = y$. Covering and packing numbers are defined on pseudometric spaces.

Definition 4.14 (Covering and packing numbers) *Let $(\mathcal{S}, \rho)$ be a pseudometric space. Let $\varepsilon > 0$.*

- *The set $\mathcal{C} \subseteq \mathcal{S}$ is a* ε-cover *of $(\mathcal{S}, \rho)$ if for every $x \in \mathcal{S}$ there exists $y \in \mathcal{C}$ such that $\rho(x, y) \leq \varepsilon$. The set $\mathcal{C} \subseteq \mathcal{S}$ is a* minimal ε-cover *if there is no other ε-cover with lower cardinality. The cardinality of any minimal ε-cover is the* ε-covering number*, denoted by $\texttt{Cov}(\mathcal{S}, \rho, \varepsilon)$.*
- *The set $\mathcal{P} \subseteq \mathcal{S}$ is a* ε-packing *of $(\mathcal{S}, \rho)$ if for every $x, x' \in \mathcal{P}$ we have $\rho(x, x') > \varepsilon$. The set $\mathcal{P} \subseteq \mathcal{S}$ is a* maximal ε-packing *if there is no other ε-packing with greater cardinality. The cardinality of any maximal ε-packing is the* ε-packing number*, denoted by $\texttt{Pack}(\mathcal{S}, \rho, \varepsilon)$.*

Note that the notion of covering numbers involves a *minimization* problem while the notion of packing numbers involves a *maximization* problem. These two quantities are related to each others, as the next result attests.

Proposition 4.15 (Duality between covering and packing) *Let $(\mathcal{S}, \rho)$ be a pseudometric space. Let $\varepsilon > 0$. Then,*

$$\boxed{\texttt{Cov}(\mathcal{S}, \rho, \varepsilon) \leq \texttt{Pack}(\mathcal{S}, \rho, \varepsilon) \leq \texttt{Cov}(\mathcal{S}, \rho, \varepsilon/2)}$$

Proof: The inequality on the left follows as any *maximal* ε-packing is also a ε-cover. Let $\mathcal{P}$ be a maximal ε-packing. By the maximality property, we know that for any $y \in \mathcal{S}, y \notin \mathcal{P}$ there exits $x \in \mathcal{P}$ such that $\rho(x, y) \leq \varepsilon$ (otherwise we could add y to $\mathcal{P}$ and get a bigger ε-packing, contradicting the maximality assumption). Hence, $\mathcal{P}$ is also a ε-cover. We have

$$\texttt{Cov}(\mathcal{S}, \rho, \varepsilon) \leq |\mathcal{P}| = \texttt{Pack}(\mathcal{S}, \rho, \varepsilon).$$

The inequality on the right follows by noticing that *any* $\varepsilon/2$-cover (hence also any minimal cover) has a cardinality greater than *any* ε-packing (hence also the maximal). Let $\mathcal{C}$ be a $\varepsilon/2$-cover and $\mathcal{P}$ be a ε-packing. By the triangle inequality, the $\varepsilon/2$-ball centered around any $x \in \mathcal{C}$ contains at most one $y \in \mathcal{P}$ (otherwise, if there were $y, y' \in \mathcal{P}$ contained in the $\varepsilon/2$-ball centered at $x \in \mathcal{C}$, we would get $\rho(y, y') \leq \rho(y, x) + \rho(x, y') \leq \varepsilon$, contradicting the ε-packing assumption) so that $|\mathcal{P}| \leq |\mathcal{C}|$. If we now choose $\mathcal{C}$ to be a minimal $\varepsilon/2$-cover and $\mathcal{P}$ to be a maximal ε-packing, we have

$$\texttt{Pack}(\mathcal{S}, \rho, \varepsilon) = |\mathcal{P}| \leq |\mathcal{C}| = \texttt{Cov}(\mathcal{S}, \rho, \varepsilon/2).$$

■

Recall that a *pseudonormed space* $(\mathcal{S}, \|\cdot\|)$ is a vector space $\mathcal{S}$ and a function $\|\cdot\| : \mathcal{S} \to \mathbb{R}_+$ (called a *pseudonorm*) such that, for any $x, y \in \mathcal{S}$ and $c \in \mathbb{R}$, we have

- $\|cx\| = |c|\|x\|$;
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality);

A *normed space* is obtained if one further assumes that $\|x\| = 0$ implies $x = 0$, the zero vector. A pseudonorm $\|\cdot\|$ naturally induces a pseudometric by $\rho(x, y) = \|x - y\|$.

A *typical* behavior of covering and packing numbers (at least for "small" spaces such as subsets of $\mathbb{R}^d$) is that they growth *exponentially* with the *algebraic* dimension. The proof of the following result uses a technique known as a *volume argument.*

Proposition 4.16 (Bounded balls) *Consider the normed space $(\mathbb{R}^d, \|\cdot\|)$, for a given positive integer d and a given norm $\|\cdot\|$. Let $\mathcal{B}^d_r(x) := \{y \in \mathbb{R}^d : \|y - x\| \leq r\}$ be the d-dimensional ball centered at $x \in \mathbb{R}^d$ with radius $r \geq 0$. If $\varepsilon \leq r$, for any $x \in \mathbb{R}^d$ we have*

$$\boxed{\left(\frac{r}{\varepsilon}\right)^d \leq \texttt{Cov}(\mathcal{B}^d_r(x), \|\cdot\|, \varepsilon) \leq \texttt{Pack}(\mathcal{B}^d_r(x), \|\cdot\|, \varepsilon) \leq \left(\frac{3r}{\varepsilon}\right)^d}$$

Proof: Without loss of generality, take $x = 0$. Let $\gamma := \mathrm{Vol}(\mathcal{B}^d_1(0))$ be the volume of a ball with radius 1. Performing a change of variables it is easy to see that $\mathrm{Vol}(\mathcal{B}^d_\ell(0)) = \ell^d \gamma$ for any $\ell \geq 0$.
To prove the upper bound, let $\mathcal{P}$ be a maximal ε-packing of the ball $\mathcal{B}^d_r(0)$. Note that $\{\mathcal{B}^d_{\varepsilon/2}(y) : y \in \mathcal{P}\}$ is a collection of disjoint balls contained in the ball $\mathcal{B}^d_{r+\varepsilon/2}(0)$. This yields, using $\varepsilon \leq r$,

$$\texttt{Pack}(\mathcal{B}^d_r(x), \|\cdot\|, \varepsilon) = |\mathcal{P}| \leq \frac{\mathrm{Vol}(\mathcal{B}^d_{r+\varepsilon/2}(0))}{\mathrm{Vol}(\mathcal{B}^d_{\varepsilon/2}(0))} = \frac{(r + \varepsilon/2)^d \gamma}{(\varepsilon/2)^d \gamma} \leq \left(\frac{3r}{\varepsilon}\right)^d.$$

To prove the lower bound, note that the volume of $\mathcal{B}^d_r(0)$ is upper bounded by the volume of a ball with radius ε times the ε-cover number, namely, $\mathrm{Vol}(\mathcal{B}^d_r(0)) \leq \mathrm{Vol}(\mathcal{B}^d_\varepsilon(0))\texttt{Cov}(\mathcal{B}^d_r(x), \|\cdot\|, \varepsilon)$.
The inequality in the middle follows by Proposition 4.15. ■

The example of bounded balls in $\mathbb{R}^d$ shows exponential growth of covering and packing numbers with respect to the Euclidean dimension, which is an algebraic notion of dimension. As we will see next, covering and packing numbers growth exponentially also with respect to the VC dimension, which is a *combinatorial* notion of dimension. Combined with the chaining technique, this is the main ingredient that will allow us to remove the term $\log(en/\texttt{VC}(\mathcal{A}))$ from the bound of Proposition 4.13.

Spaces where the logarithm of the covering number (a quantity known as the *metric entropy*, as we will see in the next lecture) grows as $m \log 1/\varepsilon$, for a given $m > 0$ (the "dimension") and for all ε in a certain range, e.g. $\varepsilon \in (0, 1)$, are known as *logarithmic metric entropy spaces.* There are also "bigger" spaces, such as *polynomial metric entropy spaces.* See [**?**], for instance.