Algorithmic Foundations of Learning **Lecture 5**

Covering Numbers Bounds for Rademacher Complexity. Chaining

Lecturer: Patrick Rebeschini *Version: December 8, 2021*

5.1 Introduction

In the case of classification we established that only a finite number of elements in the hypothesis class $\mathcal{A}$ really matter as far as establishing a notion of complexity of $\mathcal{A}$ that can be used to bound uniform deviations in expectation: only the classifiers yielding different labelings matter. We did so using combinatorial arguments, leading to the notion of complexity given by the growth function, which measures the maximal size of $\mathcal{A}$ when restricted to a given number of points. This quantity, in turn, can be upper-bounded in terms of the VC dimension.

We will now apply the same idea in the setting of regression, where we consider real-valued predictors. We will isolate a few (finitely many) predictors of interest, bound the Rademacher complexity of the set of restrictions to samples in terms of the Rademacher complexity of these representative predictors, and control the error the we commit by only considering a subset of $\mathcal{A}$. Our goal is to find a finite set that explains "most of" the deviation in expectation, up to a certain precision parameter ε. To do so, we will use metric arguments and the notion of covering numbers. This analysis, in fact, will yield improvements also in the setting of binary classification, allowing to remove the term $\log(en/\texttt{VC}(\mathcal{A}))$ in the bound of Proposition 4.13.

Recall that the regression setting is represented by the choice $\mathcal{X} = \mathbb{R}^d$ for a given dimension d, $\mathcal{Y} = \mathbb{R}$, and

$$\mathcal{A} \subseteq \mathcal{B} = \{a : \mathcal{X} \to \mathcal{Y}\}.$$

5.2 Covering Numbers Bounds for Rademacher Complexity

Given $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, define the following pseudonorms on the space $\mathcal{A}$: for any $a \in \mathcal{A}$,

$$\|a\|_{p,x} := \Big(\frac{1}{n}\sum_{i=1}^n |a(x_i)|^p\Big)^{1/p} \qquad \text{for any } p \in [1,\infty),$$
$$\|a\|_{\infty,x} := \max_i |a(x_i)|.$$

These pseudonorms induce the following pseudometrics on the space $\mathcal{A}$, for any $a, b \in \mathcal{A}$,

$$\|a-b\|_{p,x} := \Big(\frac{1}{n}\sum_{i=1}^n |a(x_i)-b(x_i)|^p\Big)^{1/p} \qquad \text{for any } p \in [1,\infty),$$
$$\|a-b\|_{\infty,x} := \max_i |a(x_i)-b(x_i)|.$$

There is a monotone behavior of covering and packing numbers on the space $(\mathcal{A}, \|\cdot\|_{p,x})$ with respect to p.

Proposition 5.1 *For any $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, $1 \le p \le q$, and $\varepsilon > 0$, we have*

$$\boxed{\texttt{Cov}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \le \texttt{Cov}(\mathcal{A}, \|\cdot\|_{q,x}, \varepsilon)}$$

$$\boxed{\texttt{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \le \texttt{Pack}(\mathcal{A}, \|\cdot\|_{q,x}, \varepsilon)}$$

Proof: See **Problem 2.6** in the Problem Sheets. ■

Proposition 5.1 shows that once we obtain results (i.e. upper bounds) involving covering numbers for the pseudometric space $(\mathcal{A}, \|\cdot\|_{p,x})$, for $p \geq 1$, then we can immediately derive results involving covering *and* packing numbers (recall the duality property, Proposition 4.15) for the norm $\|\cdot\|_{q,x}$, for any $q \geq p$.

Covering numbers of the pseudometric space $(\mathcal{A}, \|\cdot\|_{1,x})$ can be used to bound the empirical Rademacher complexity.

Proposition 5.2 *For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, let $\sup_{a\in\mathcal{A}} \|a\|_{2,x} \leq c_x$. Then,*

$$\boxed{\mathtt{Rad}(\mathcal{A} \circ x) \leq \inf_{\varepsilon>0} \left\{ \varepsilon + \frac{\sqrt{2}\, c_x}{\sqrt{n}} \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)} \right\}}$$

Proof: Fix $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ and $\varepsilon > 0$. Let $\mathcal{C} \subseteq \mathcal{A}$ be a minimal ϵ-cover of $(\mathcal{A}, \|\cdot\|_{1,x})$, and for any $a \in \mathcal{A}$ let $\tilde{a} \in \mathcal{C}$ be such that $\|a - \tilde{a}\|_{1,x} \leq \varepsilon$. We have

$$\begin{aligned}
\mathtt{Rad}(\mathcal{A} \circ x) &= \mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i a(x_i) \leq \mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i (a(x_i) - \tilde{a}(x_i)) + \mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i \tilde{a}(x_i) \\
&\leq \varepsilon + \mathbf{E} \sup_{a\in\mathcal{C}} \frac{1}{n} \sum_{i=1}^n \Omega_i a(x_i) \leq \varepsilon + \sup_{a\in\mathcal{C}} \sqrt{\sum_{i=1}^n a(x_i)^2} \frac{\sqrt{2 \log |\mathcal{C}|}}{n} \\
&\leq \varepsilon + c_x \sqrt{\frac{2 \log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)}{n}},
\end{aligned}$$

where the last inequality follows by Massart's lemma, and $|\mathcal{C}| = \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)$ by definition. The result follows by taking the infimum over $\varepsilon > 0$. ■

The bound in Proposition 5.2 establishes a tradeoff with respect to the parameter ε, as when ε decreases $\mathtt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)$ increases. This bound is data-dependent, as the right-hand side depends on $x \in \mathcal{X}^n$.

5.3 Chaining

Proposition 5.2 is established by only using one fixed level of granularity ($\varepsilon > 0$) at a time, and taking the infimum over $\varepsilon > 0$ to obtain the final bound. An improved version of this result can be established by integrating over different levels of granularity. In this case, we need to work with covering numbers for the pseudometric space $(\mathcal{A}, \|\cdot\|_{2,x})$.[1]

Proposition 5.3 *For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ and $\sup_{a\in\mathcal{A}} \|a\|_{2,x} \leq c_x$ we have*

$$\boxed{\mathtt{Rad}(\mathcal{A} \circ x) \leq \inf_{\varepsilon\in[0, c_x/2]} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^{c_x/2} \mathrm{d}\nu \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)} \right\}}$$

Proof: Fix $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$. For each $j \in \mathbb{N}_+$ let $\varepsilon_j := c_x/2^j$ and let $\mathcal{C}_j \subseteq \mathcal{A}$ be a minimal ε_j-cover of $(\mathcal{A}, \|\cdot\|_{2,x})$. We have $|\mathcal{C}_j| = \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon_j)$. For any $a \in \mathcal{A}$ and $j \in \mathbb{N}_+$ let $a_j \in \mathcal{C}_j$ such that

[1] The chaining argument relies on using the triangle inequality on the application of Massart's lemma, where the $\|\cdot\|_2$ norm naturally appears. This is the reason why we can not work with $(\mathcal{A}, \|\cdot\|_{1,x})$ but we need to consider $(\mathcal{A}, \|\cdot\|_{2,x})$.

$\|a - a_j\|_{2,x} \leq \varepsilon_j$. The sequence $a_1, a_2, \ldots$ (of elements of covers with decreasing radius) converges towards a. This sequence can be used to define the following telescoping sum, for a given $m \in \mathbb{N}$ to be chosen later:

$$a = a - a_m + \sum_{j=1}^{m}(a_j - a_{j-1})$$

with $a_0 := 0$. This telescoping sum can be thought of as a "chain" connecting $a_0 = 0$ to a. This is the reason why the technique we are going to describe is called *chaining*. We have

$$\mathtt{Rad}(\mathcal{A} \circ x) = \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i a(x_i) \leq \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i (a(x_i) - a_m(x_i)) + \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i \sum_{j=1}^{m} (a_j(x_i) - a_{j-1}(x_i)).$$

We bound the two summands separately. The first summand is bounded by ε_m, as

$$\sum_{i=1}^{n} \Omega_i (a(x_i) - a_m(x_i)) \leq \sum_{i=1}^{n} |a(x_i) - a_m(x_i)| = n\|a - a_m\|_{1,x} \leq n\|a - a_m\|_{2,x} \leq n\varepsilon_m.$$

As the are at most $|\mathcal{C}_j||\mathcal{C}_{j-1}|$ different ways to create a vector in $\mathbb{R}^n$ of the form

$$\begin{pmatrix} a_j(x_1) - a_{j-1}(x_1) \\ \vdots \\ a_j(x_n) - a_{j-1}(x_n) \end{pmatrix}$$

with $a_j \in \mathcal{C}_j$ and $a_{j-1} \in \mathcal{C}_{j-1}$, using Massart's Lemma the second summand can be upper bounded by

$$\begin{aligned} \sum_{j=1}^{m} \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i (a_j(x_i) - a_{j-1}(x_i)) &\leq \sum_{j=1}^{m} \sup_{a \in \mathcal{A}} \sqrt{\sum_{i=1}^{n} (a_j(x_i) - a_{j-1}(x_i))^2} \frac{\sqrt{2 \log |\mathcal{C}_j||\mathcal{C}_{j-1}|}}{n} \\ &= \sum_{j=1}^{m} \sup_{a \in \mathcal{A}} \|a_j - a_{j-1}\|_{2,x} \frac{\sqrt{2 \log |\mathcal{C}_j||\mathcal{C}_{j-1}|}}{\sqrt{n}}. \end{aligned}$$

By the triangle inequality for the pseudonorm $\|\cdot\|_{2,x}$ we have (using that $\varepsilon_{k-1} = 2\varepsilon_k$)

$$\|a_j - a_{j-1}\|_{2,x} \leq \|a_j - a\|_{2,x} + \|a - a_{j-1}\|_{2,x} \leq \varepsilon_j + \varepsilon_{j-1} = 3\varepsilon_j = 6(\varepsilon_j - \varepsilon_{j+1}).$$

Also, $|\mathcal{C}_j| = \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon_j)$ and $|\mathcal{C}_{j-1}| \leq |\mathcal{C}_j|$. Putting everything together we find

$$\begin{aligned} \mathtt{Rad}(\mathcal{A} \circ x) &\leq \varepsilon_m + \frac{12}{\sqrt{n}} \sum_{j=1}^{m} (\varepsilon_j - \varepsilon_{j+1}) \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon_j)} \\ &\leq 2\varepsilon_{m+1} + \frac{12}{\sqrt{n}} \int_{\varepsilon_{m+1}}^{c_x/2} \mathrm{d}\nu \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)}, \end{aligned}$$

where the last inequality follows as the integral is lower-bounded by its lower Riemann sum as the function $\nu \to \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)$ is non-increasing. For any $\varepsilon \in [0, c_x/2]$, choose m such that $\varepsilon < \varepsilon_{m+1} \leq 2\varepsilon$. The statement of the proposition follows by taking the infimum over $\varepsilon \in [0, c_x/2]$. ∎

The integral in Proposition 5.3 is called *Dudley Entropy Integral*. This integral allows to exploit the decay of the cover number $\mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)$ as ν increases. As the function $\nu \to \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)$ is non-increasing, we clearly have

$$\int_{\varepsilon}^{c_x/2} \mathrm{d}\nu \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)} \leq \frac{c_x}{2} \sqrt{\log \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon)},$$

which shows how (modulo constants) Proposition 5.3 yields an improvement over the bound obtained by Proposition 5.2 using that $\mathtt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon) \leq \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon)$.

5.4 Linear Predictors ℓ_∞/ℓ_1 Constraints

As an application of Proposition 5.3, we consider the case of linear predictors with ℓ_∞/ℓ_1 constraints, which adds to the examples covered in Lecture 3. The following result (modulo constants) can also be obtained via a direct application of Massart's Lemma, using the fact that the supremum of a linear function over the ℓ_∞ ball is achieved at one of the extreme points, and there are 2^d of them (so that the final dependence growths as $\sqrt{d}$).

Proposition 5.4 *Let $\mathcal{A}_\infty := \{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_\infty \leq 1\}$. Then, for any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$,*

$$\boxed{\texttt{Rad}(\mathcal{A}_\infty \circ x) \leq 12\gamma \frac{\max_i \|x_i\|_1}{\sqrt{n}} \sqrt{d}}$$

where $\gamma := \int_0^{1/2} \mathrm{d}\nu \sqrt{\log(3/\nu)}$.

Proof: Fix $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$. Proposition 5.3 and the bound provided by Proposition 5.1 with $p = 2$ and $q = \infty$ yield

$$\texttt{Rad}(\mathcal{A}_\infty \circ x) \leq \inf_{\varepsilon \in [0, c_x/2]} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^{c_x/2} \mathrm{d}\nu \sqrt{\log \texttt{Cov}(\mathcal{A}_\infty, \|\cdot\|_{\infty,x}, \nu)} \right\}.$$

By Hölder's inequality, for any $a \in \mathcal{A}_\infty$ we have $a(x) = w^\top x \leq \|w\|_\infty \|x\|_1 \leq \|x\|_1$,

$$c_x = \sup_{a \in \mathcal{A}_\infty} \|a\|_{2,x} = \sup_{a \in \mathcal{A}_\infty} \sqrt{\frac{1}{n} \sum_{i=1}^n a(x_i)^2} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \|x_i\|_1^2} \leq c,$$

with $c := \max_i \|x_i\|_1$. To bound the ν-covering number $\texttt{Cov}(\mathcal{A}_\infty, \|\cdot\|_{\infty,x}, \nu)$, it is sufficient to find a ν-cover of the pseudometric space $(\mathcal{A}_\infty, \|\cdot\|_{\infty,x})$. Note that for any $a, b \in \mathcal{A}_\infty$ with $a(x) = w_a^\top x$ and $b(x) = w_b^\top x$, by Hölder inequality, we have

$$\|a - b\|_{\infty,x} = \max_i |a(x_i) - b(x_i)| = \max_i |(w_a - w_b)^\top x_i| \leq c \|w_a - w_b\|_\infty.$$

Hence, to find an ν-cover it suffices to find a set of vectors $\mathcal{C} \subseteq \mathbb{R}^d$ such that for any $w_a \in \{w \in \mathbb{R}^d : \|w\|_\infty \leq 1\}$ there exists $w \in \mathcal{C}$ with $\|w_a - w\|_\infty \leq \nu/c$. We can define $\mathcal{C}$ to be the set of vertices of the hypercubes that we obtain by dividing the hypercube with side length 2 into hypercubes with side length $2\nu/c$. Clearly, any $w_a \in \{w \in \mathbb{R}^d : \|w\|_\infty \leq 1\}$ must land in one of these hypercubes, and each coordinate is at most ν/c away from one of the vertices. There are at most $(\lceil c/\nu \rceil + 1)^d$ vertices, so $\texttt{Cov}(\mathcal{A}_\infty, d_\infty(x_1, \ldots, x_n), \nu) \leq (\lceil c/\nu \rceil + 1)^d \leq (c/\nu + 2)^d \leq (3c/\nu)^d$, where the last inequality holds for $c/\nu \geq 1$, or, equivalently, if $\nu \leq c$. We have

$$\begin{aligned}
\texttt{Rad}(\mathcal{A}_\infty \circ x) &\leq \inf_{\varepsilon \in [0, c_x/2]} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^{c/2} \mathrm{d}\nu \sqrt{d \log(3c/\nu)} \right\} \leq \frac{12\sqrt{d}}{\sqrt{n}} \int_0^{c/2} \mathrm{d}\nu \sqrt{\log(3c/\nu)} \\
&= \frac{12c\sqrt{d}}{\sqrt{n}} \int_0^{1/2} \mathrm{d}\nu \sqrt{\log(3/\nu)}.
\end{aligned}$$

■

5.5 Binary Classification

We now apply Proposition 5.3 to the case of binary classification, which allows us to remove the term $\log(en/\mathtt{VC}(\mathcal{A}))$ from the bound of Proposition 4.13.

To this end, we first show that the packing number of the subspace of binary classifiers $\mathcal{A}$ equipped with the pseudonorm $\|\cdot\|_{1,x}$ is upper-bounded by a data-independent bound that grows exponentially with respect to the VC dimension (recall the exponential dependence on the dimension for packing and covering numbers on bounded balls, Proposition 4.16). We give a proof that uses a technique known as a *probabilistic method*.

Proposition 5.5 *Let $\mathcal{A} \subseteq \mathcal{B} = \{a : \mathcal{X} \to \{0,1\}\}$ with $\mathtt{VC}(\mathcal{A}) < \infty$. For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, $p \geq 1$, $\varepsilon > 0$ we have*

$$\boxed{\mathtt{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \leq \left(\frac{10}{\varepsilon^p} \log \frac{2e}{\varepsilon^p}\right)^{\mathtt{VC}(\mathcal{A})}}$$

Proof: Due to the nature of binary classification we have $\|a-b\|_{p,x}^p = \|a-b\|_{1,x}$ for any $a, b \in \mathcal{A}$, $p \geq 1$, $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, which implies that $\mathtt{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) = \mathtt{Pack}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon^p)$. Hence, it is sufficient to prove the statement for $p = 1$. Henceforth, fix $x \in \mathcal{X}^n$ and $\varepsilon > 0$.

The main idea of the proof is to relate the packing number of the subspace of binary classifiers $\mathcal{A}$ equipped with the pseudonorm $\|\cdot\|_{1,x}$ with the growth function. Let $\mathcal{P} \subseteq \mathcal{A}$ be a maximal ε-packing and let $\mu = |\mathcal{P}| = \mathtt{Pack}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)$. Recall that the growth function of $\mathcal{A}$ is defined as $m \to \tau_{\mathcal{A}}(m) := \sup_{z \in \mathcal{X}^m} |\mathcal{A} \circ z|$ with $\mathcal{A} \circ z = \{(a(z_1), \ldots, a(z_m)) : a \in \mathcal{A}\}$. That is, the growth function evaluated at m is the largest number of distinct labelings of m points in $\mathcal{X}$ that can be obtained using classifiers from $\mathcal{A}$. We use the *probabilistic method* to show that if $m \geq \frac{2}{\varepsilon} \log \mu$ then there exists $z \in \mathcal{X}^m$ such that $|\mathcal{P} \circ z| = |\mathcal{P}| = \mu$, so that

$$\mu = |\mathcal{P} \circ z| \leq |\mathcal{A} \circ z| \leq \tau_{\mathcal{A}}(m). \tag{5.1}$$

The connection with probability comes from the following observation: note that for any $a, b \in \mathcal{P}$ we have

$$\varepsilon < \|a-b\|_{1,x} = \frac{1}{n}\sum_{i=1}^n |a(x_i) - b(x_i)| = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a(x_i) \neq b(x_i)} = \mathbf{P}(a(Z) \neq b(Z)),$$

where Z is a uniform random variable taking values in $\{x_1, \ldots, x_n\}$. Let $Z_1, \ldots, Z_m$ be m i.i.d. random variables distributed as Z, and use the notation $a \circ \{Z_1, \ldots, Z_m\} = (a(Z_1), \ldots, a(Z_m))$ for any $a \in \mathcal{A}$. By the union bound we have

$$\begin{aligned}
&\mathbf{P}(|\mathcal{P} \circ \{Z_1, \ldots, Z_m\}| = \mu) \\
&\quad = \mathbf{P}(\text{Every } a \in \mathcal{P} \text{ produces a distinct label for } \{Z_1, \ldots, Z_m\}) \\
&\quad = \mathbf{P}(\text{For every } a, b \in \mathcal{P}, a \neq b, \text{ we have } a \circ \{Z_1, \ldots, Z_m\} \neq b \circ \{Z_1, \ldots, Z_m\}) \\
&\quad = 1 - \mathbf{P}(\text{There exists } a, b \in \mathcal{P}, a \neq b, \text{ such that } a \circ \{Z_1, \ldots, Z_m\} = b \circ \{Z_1, \ldots, Z_m\}) \\
&\quad = 1 - \mathbf{P}\Big(\bigcup_{a,b \in \mathcal{P}, a \neq b} \{a(Z_1) = b(Z_1), \ldots, a(Z_m) = b(Z_m)\}\Big) \\
&\quad \geq 1 - \sum_{a,b \in \mathcal{P}, a \neq b} \mathbf{P}(a(Z_1) = b(Z_1), \ldots, a(Z_m) = b(Z_m)) \\
&\quad = 1 - \sum_{a,b \in \mathcal{P}, a \neq b} \mathbf{P}(a(Z_1) = b(Z_1)) \cdots \mathbf{P}(a(Z_m) = b(Z_m)) \\
&\quad > 1 - \mu^2 (1-\varepsilon)^m \\
&\quad > 1 - \mu^2 e^{-m\varepsilon},
\end{aligned}$$

where we used $1 - \varepsilon < e^{-\varepsilon}$ for $\varepsilon > 0$. The lower bound is strictly greater than zero if $m \geq \frac{2}{\varepsilon} \log \mu$, in which case we are guaranteed that there exists $z \in \mathcal{X}^m$ such that $|\mathcal{P} \circ z| = \mu$ and (5.1) holds.

Using Sauer-Shelah's lemma, Lemma 4.11, we have $\tau_{\mathcal{A}}(m) \leq (em/\mathtt{VC}(\mathcal{A}))^{\mathtt{VC}(\mathcal{A})}$, and (5.1) with $m = \frac{2}{\varepsilon} \log \mu$ yields

$$\mu^{1/\mathtt{VC}(\mathcal{A})} \leq \frac{2e}{\varepsilon} \log(\mu^{1/\mathtt{VC}(\mathcal{A})}).$$

The proof follows as $a \leq b \log a$ implies $a \leq b \log b/(1 - 1/e)$ (see Lemma A.1 in [1], for instance), and $e/(1 - 1/e) \approx 4.86456 \leq 5$. ■

Theorem 5.6 *Let $\mathcal{A} \subseteq \mathcal{B} = \{a : \mathcal{X} \to \{0, 1\}\}$ with $\mathtt{VC}(\mathcal{A}) < \infty$. For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, we have*

$$\boxed{\mathtt{Rad}(\mathcal{A} \circ x) \leq 31 \sqrt{\frac{\mathtt{VC}(\mathcal{A})}{n}}}$$

Proof: Proposition 5.3 with $c_x = 1$ and $\varepsilon = 0$ yields

$$\mathtt{Rad}(\mathcal{A} \circ x) \leq \frac{12}{\sqrt{n}} \int_0^1 \mathrm{d}\nu \sqrt{\log \mathtt{Cov}(\mathcal{A}, \| \cdot \|_{2,x}, \nu)}.$$

By Proposition 4.15, the covering number is upper-bounded by the packing number. The bound in Proposition 5.5 yields, using $\log x \leq x/e$,

$$\log \mathtt{Cov}(\mathcal{A}, \| \cdot \|_{2,x}, \nu) \leq \log \mathtt{Pack}(\mathcal{A}, \| \cdot \|_{2,x}, \nu) \leq \mathtt{VC}(\mathcal{A}) \log \left(\frac{10}{\nu^2} \log \frac{2e}{\nu^2} \right) \leq \mathtt{VC}(\mathcal{A}) \log \left(\frac{20}{\nu^4} \right).$$

We get

$$\mathtt{Rad}(\mathcal{A} \circ x) \leq 12 \sqrt{\frac{\mathtt{VC}(\mathcal{A})}{n}} \int_0^1 \mathrm{d}\nu \sqrt{\log \left(\frac{20}{\nu^4} \right)} \leq 31 \sqrt{\frac{\mathtt{VC}(\mathcal{A})}{n}},$$

where we used that $\int_0^1 \mathrm{d}\nu \sqrt{\log \left(\frac{20}{\nu^4} \right)} \approx 2.55919$. ■

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.