UC Berkeley **Jonathan Shafer**
CS 294-220, Spring 2021

Unit 7

Covering Numbers and Chaining

1 Definitions

▶ **Definition 1.** *A pseudo-metric space is a tuple (Ω, ρ) where Ω is a set and $\rho : \Omega \times \Omega \to [0, \infty)$ is a function such that for every $x, y, z \in \Omega$ the following properties hold:*

1. *Identity:* $\rho(x, x) = 0$.
2. *Symmetry:* $\rho(x, y) = \rho(y, x)$.
3. *Triangle inequality:* $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

▶ Remark 2. Pseudo-metric spaces differ from metric spaces in that a metric space satisfies the stronger *identity of indiscernibles* property, $\rho(x, y) = 0 \iff x = y$. In other word, in a pseudo-metric space it is possible that $\rho(x, y) = 0$ for $x \neq y$, but that is not possible in metric space. Additionally, note that the definition does not change if we allow $\rho : \Omega \times \Omega \to \mathbb{R}$, because it is possible to deduce that ρ is non-negative from the other assumptions: $\forall x, y \in \Omega : \ 0 = \rho(x, x) \leq \rho(x, y) + \rho(y, x) = 2\rho(x, y)$. ⌟

(a) An ε-cover of the gray square. The cover is not internal.

(b) The set of blue points forms an ε-packing of the gray disk. The $\varepsilon/2$-balls around them are non-intersecting.

Figure 1 Illustrations of covering and packing. Source: Michael Wolf [3].

▶ **Definition 3.** *Let (Ω, ρ) be a pseudo-metric space, let $X, C, P \subseteq \Omega$, and let $\varepsilon > 0$.*

- *We say that C is an ε-cover of X if*

$$\forall x \in X \ \exists c \in C : \ \rho(x, c) \leq \varepsilon.$$

- *We say that C is an internal ε-cover of X if $C \subseteq X$ and C is an ε-cover of X.*
- *The ε-cover number of X is*

$$N(X, \varepsilon, \rho) = \inf \{|C| : \ C \subseteq \Omega \ \wedge \ C \text{ is an } \varepsilon\text{-cover of } X\}.$$

- *The internal ε-cover number of X is*

$$N_{\text{in}}(X, \varepsilon, \rho) = \inf \{|C| : \ C \subseteq X \ \wedge \ C \text{ is an } \varepsilon\text{-cover of } X\},$$

- *We say that P is an ε-packing of X if* $P \subseteq X$

$$\forall x, y \in P : \ \rho(x,y) > \varepsilon.$$

- *The ε-packing number of X is*

$$M(X, \varepsilon, \rho) = \sup\{|P| : \ P \subseteq X \ \wedge \ P \text{ is an } \varepsilon\text{-packing of } X\}.$$

When ρ is understood from context we will simply write $N(X,\varepsilon)$, $N_{\text{in}}(X,\varepsilon)$, and $M(X,\varepsilon)$.

All these numbers are closely related.

▷ Claim 4. Let (Ω, ρ) be a pseudo-metric space, let $X \subseteq \Omega$, and let $\varepsilon > 0$. Then

$$N(X,\varepsilon) \le N_{\text{in}}(X,\varepsilon) \le M(X,\varepsilon) \le N(X, \varepsilon/2).$$

▶ **Definition 5.** *Let (Ω,ρ) be a pseudo-metric space, let $x \in \Omega$ and $\varepsilon \ge 0$. The ε-ball centered at x is $B_\varepsilon(x) = \{y \in \Omega : \ \rho(x,y) \le \varepsilon\}$.*

Proof of Claim 4. We prove this claim for the case where all the numbers are finite. The first inequality is immediate from the definitions.

For the second inequality, let $P \subseteq X$ be an ε-packing of X such that $|P| = M(X,\varepsilon)$. Then P is maximal in the sense that for any point $x \in X \setminus P$, the set $P \cup \{x\}$ is not an ε-packing of X. Namely, for any $x \in X$ there exists $p \in P$ such that $\rho(x,p) \le \varepsilon$. Hence, P is an internal ε-cover of X, and so $N_{\text{in}}(X,\varepsilon) \le |P|$.

For the last inequality, let C be an $\varepsilon/2$-cover of X such that $|C| = N(X, \varepsilon/2)$. Let P be any ε-packing of X. We constructs an injective function $f : P \to C$, and this completes the proof because it implies that $|P| \le |C|$. For each $p \in P$, we define $f(p)$ to be an arbitrary $c \in C$ such that $p \in B_{\varepsilon/2}(c)$ (such a c always exists because C is an $\varepsilon/2$-cover). To see that f is injective, assume for contradiction that there exist $p_1, p_2 \in P$ such that $p_1 \ne p_2$ and $f(p_1) = f(p_2) = c$. Then $\rho(p_1,p_2) \le \rho(p_1,c) + \rho(c,p_2) \le \varepsilon/2 + \varepsilon/2 = \varepsilon$, which is a contradiction to P being an ε-packing. ◀

2 Intuition for Covering and Packing Numbers

▶ **Example 6.** *Consider the metric space $(\mathbb{R}^d, \rho)$ where $\rho(x,y) = \|x-y\|$ and $\|\cdot\|$ is some ℓ_p norm. For any ball $B_r(x) \subseteq \mathbb{R}^d$ of radius r centered at $x \in \mathbb{R}^d$, the volume (Lebesgue measure) of $B_r(x)$ is given by the formula $V(B_r(x)) = C_{d,p} \cdot r^d$, where $C_{d,p} \in \mathbb{R}$ is some constant that depends on p and d. For example, if $d = p = 2$ then $C_{d,p} = \pi$, yielding the familiar formula πr^2 for the area of a circle in the Euclidean plane.*

In this metric space we can obtain the following bounds for the the packing and covering numbers of a ball $B_r(x)$:

- *If $\varepsilon \le r$ then $M(B_r(x),\varepsilon) \le \left(\frac{3r}{\varepsilon}\right)^d$. To see this, let $P \subseteq B_r(x)$ be an ε-packing of $B_r(x)$. Then for every $p_1, p_2 \in P$, the balls $B_{\varepsilon/2}(p_1)$ and $B_{\varepsilon/2}(p_2)$ are disjoint, and contained in $B_{r+\varepsilon/2}(x)$. Hence*

$$V\left(B_{r+\varepsilon/2}(x)\right) \ge \sum_{p \in P} V(B_{\varepsilon/2}(p)) = |P| \cdot V(B_{\varepsilon/2}(0)),$$

so

$$|P| \le \frac{V\left(B_{r+\varepsilon/2}(x)\right)}{V(B_{\varepsilon/2}(0))} = \frac{C_{d,p}(r+\varepsilon/2)^d}{C_{d,p}(\varepsilon/2)^d} \le \left(\frac{r+r/2}{\varepsilon/2}\right)^d = \left(\frac{3r}{\varepsilon}\right)^d.$$

- $N(B_r(x), \varepsilon) \geq \left(\frac{r}{\varepsilon}\right)^d$. *Indeed, if* $C \subseteq \Omega$ *is an* ε*-cover of* $B_r(x)$*, then*

$$V(B_r(x)) \leq \sum_{c \in C} V(B(c, \varepsilon)) = |C| \cdot V(B(0, \varepsilon)),$$

and this implies that

$$|C| \geq \frac{V(B_r(x))}{V(B(0,\varepsilon))} = \frac{C_{d,p} r^d}{C_{d,p} \varepsilon^d} = \left(\frac{r}{\varepsilon}\right)^d.$$

Thus, for $\varepsilon \leq r$ *we have that* $\left(\frac{r}{\varepsilon}\right)^d \leq N(B_r(x), \varepsilon) \leq M(B_r(x), \varepsilon) \leq \left(\frac{3r}{\varepsilon}\right)^d$.

The above example demonstrates a fairly general phenomena. For sets in $\mathbb{R}^d$ bounded by a constant, the log of the packing and covering numbers, $\ln M(A, \varepsilon)$ and $\ln N(A, \varepsilon)$, tend to scale like $d \ln\left(\frac{1}{\varepsilon}\right)$.[1] Perhaps surprisingly, a similar phenomena also holds for metric spaces of functions, if we replace the algebraic dimension with the VC dimension, as we will see in the next section.

3 Packing Numbers for VC Classes

▶ **Definition 7.** *Let* V *be a vector space over* $\mathbb{R}$*. A seminorm is a function* $p : V \to \mathbb{R}$ *satisfying*

- *Triangle inequality.* $\forall u, v \in V : \; p(u + v) \leq p(u) + p(v)$.
- *Absolute homogeneity.* $\forall v \in V \; \forall c \in \mathbb{R} : \; p(cv) = |c| p(v)$.

▶ Remark 8. The definition of seminorm also implies *non-negativity*, namely $p(v) \geq 0 \; \forall v \in V$. To see this, note that $0 = |0| p(u) = p(0u) = p(0) = p(v - v) \leq p(v) + p(-v) = 2p(v)$. A *norm* is a semi-norm where $p(v) = 0 \implies v = 0$. ⌟

▶ **Definition 9.** *Let* Ω *be a set, let* $\mathcal{F}$ *be a class of functions* $\Omega \to \mathbb{R}$*, let* $S = (z_1, \ldots, z_m) \in \Omega^m$ *and let* $p > 0$*. The empirical* p*-semi-norm of* $\mathcal{F}$ *with respect to* S *is a function* $\|\cdot\|_{S,p} : \mathcal{F} \to \mathbb{R}$ *such that*

$$\|f\|_{S,p} = \|1/m \cdot f(S)\|_p = \left(\frac{1}{m} \sum_{i=1}^{m} |f(z_i)|^p\right)^{1/p}.$$

Additionally, for $p = \infty$ *we define* $\|f\|_{S,\infty} = \|f(S)\|_\infty = \max_{i \in [m]} |f(z_i)|$.

▶ **Notation 10.** Let Ω be a set, let $\mathcal{F}$ be a class of functions $\Omega \to \mathbb{R}$, $S \in \Omega^m$, and $p \in (0, \infty]$. We write $\rho_{S,p} : \mathcal{F}^2 \to \mathbb{R}$ to denote $\rho_{S,p}(f_1, f_2) = \|f_1 - f_2\|_{S,p}$.

▶ **Lemma 11.** *Let* Ω *be a set, let* $\mathcal{F}$ *be a class of functions* $\Omega \to \{0, 1\}$ *with* $\mathsf{VC}(\mathcal{F}) = d < \infty$*. Let* $S \in \Omega^m$ *and* $p \in (0, \infty]$*. Then* $(\mathcal{F}, \rho_{S,p})$ *is a pseudo-metric space, and for all* $\varepsilon > 0$*,*

$$M(\mathcal{F}, \varepsilon) \leq \left(\frac{4e}{\varepsilon^p} \ln\left(\frac{2e}{\varepsilon^p}\right)\right)^d.$$

[1] One can in fact use this idea to define a notion of dimension in metric spaces that do not have an algebraic notion of dimension. This is called the Minkowski–Bouligand dimension and it is used, for example, to define the dimension of fractals.

Proof. To prove the lemma it suffices to show that

$$M(\mathcal{F},\varepsilon,\rho_{S,1}) \le \left(\frac{4e}{\varepsilon}\ln\left(\frac{2e}{\varepsilon}\right)\right)^d. \tag{1}$$

To see that this suffices, note that because the functions $f \in \mathcal{F}$ are binary,

$$\|f\|_{S,p}^p = \frac{1}{m}\sum_{i=1}^m |f(z_i)|^p = \frac{1}{m}\sum_{i=1}^m |f(z_i)| = \|f\|_{S,1},$$

so $\|f\|_{S,p} = \varepsilon \iff \|f\|_{S,1} = \varepsilon^p$. That is, a subset $P \subseteq \mathcal{F}$ is an ε-packing of $\mathcal{F}$ in $(\mathcal{F},\rho_{S,p})$ if and only if it is an ε^p-packing of $\mathcal{F}$ in $(\mathcal{F},\rho_{S,1})$. Together with (1) this implies that

$$M(\mathcal{F},\varepsilon,\rho_{S,p}) = M(\mathcal{F},\varepsilon^p,\rho_{S,1}) \le \left(\frac{4e}{\varepsilon^p}\ln\left(\frac{2e}{\varepsilon^p}\right)\right)^d.$$

We now prove (1). First, note that $M(\mathcal{F},\varepsilon,\rho_{S,1}) \le \left|\mathcal{F}|_S\right|$. To see this, assume for contradiction that there exists a suitable ε-packing P such that $|P| > \left|\mathcal{F}|_S\right|$. By the pigeonhole principle, this implies that there exist two functions $f, f' \in P$ such that $f \ne f'$ and $f|_S = f'|_S$. This implies that $\rho_{S,1}(f,f') = 0$ which is a contradiction to P being an ε-packing. As a consequence, $M(\mathcal{F},\varepsilon,\rho_{S,1})$ is finite and so there exists a finite set $P \subseteq \mathcal{F}$ such that P is an ε-packing of $\mathcal{F}$ with respect to $\rho_{S,1}$ and $|P| = M = M(\mathcal{F},\varepsilon,\rho_{S,1})$.

Second, note that for any $f, f' \in P$,

$$\begin{aligned}\varepsilon < \rho_{S,1}(f,f') &= \frac{1}{m}\sum_{i=1}^m |f(z_i) - f'(z_i)| = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\big(f(z_i) \ne f'(z_i)\big) \\ &= \mathbb{P}_{i\in[m]}\left[f(z_i) \ne f'(z_i)\right]. \end{aligned} \tag{2}$$

Third, let $\tilde{S} = (\tilde{z}_1,\ldots,\tilde{z}_m) \in \Omega^m$ be a vector of elements chosen independently and uniformly from S, namely $\tilde{S} \sim (U(S))^m$, where $U(S)$ is the uniform distribution on the elements in S. Then for any $f, f' \in P$,

$$\begin{aligned}\mathbb{P}_{\tilde{S}}\left[f|_{\tilde{S}} = f'|_{\tilde{S}}\right] &= \mathbb{P}_{\tilde{S}}\left[\bigcap_{i=1}^m \{f(\tilde{z}_i) = f'(\tilde{z}_i)\}\right] = \prod_{i=1}^m \mathbb{P}_{\tilde{S}}\left[f(\tilde{z}_i) = f'(\tilde{z}_i)\right] \\ &= \prod_{i=1}^m \mathbb{P}_{i\in[m]}\left[f(z_i) = f'(z_i)\right] = \prod_{i=1}^m \left(1 - \mathbb{P}_{i\in[m]}\left[f(z_i) \ne f'(z_i)\right]\right) \\ &< \prod_{i=1}^m (1-\varepsilon) \le e^{-\varepsilon m}. \qquad \text{(from (2))}\end{aligned}$$

Fourth, fix $m \ge \frac{2}{\varepsilon}\ln(M)$. Applying a union bound to the last inequality yields

$$\mathbb{P}_{\tilde{S}}\left[\exists f, f' \in P :\ f|_{\tilde{S}} = f'|_{\tilde{S}}\right] \le \binom{M}{2} e^{-\varepsilon m} < M^2 e^{-\varepsilon m} \le 1.$$

Namely,

$$\mathbb{P}_{\tilde{S}}\left[\forall f, f' \in P :\ f|_{\tilde{S}} \ne f'|_{\tilde{S}}\right] > 0,$$

and in particular there exists an assignment to $\tilde{S}$ such that $f|_{\tilde{S}} \ne f'|_{\tilde{S}}$ for all $f, f' \in P$. This

implies that

$$\begin{aligned} M &\le \left|\mathcal{F}|_{\tilde{S}}\right| \\ &\le \tau_{\mathcal{F}}(m) && \text{(definition of } \tau_{\mathcal{F}}\text{)} \\ &\le (em/d)^d && \text{(Sauer's lemma)} \\ &\le \left(\frac{2e}{\varepsilon d}\ln(M)\right)^d. && \text{(choice of } m\text{)} \end{aligned}$$

Equivalently, $M^{1/d} \le \frac{2e}{\varepsilon}\ln\left(M^{1/d}\right)$. Finally, invoking Claim 17 in Unit 5 with $x = M^{1/d}$ and $y = \frac{2e}{\varepsilon}$, we obtain

$$M^{1/d} \le \frac{4e}{\varepsilon}\ln\left(\frac{2e}{\varepsilon}\right),$$

and this completes the proof. ◀

4 Uniform Convergence via Covering Numbers

In Unit 5 Theorem 9 we saw that

$$\mathbb{P}\left[\exists h \in \mathcal{H} :\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \le 4\tau_{\mathcal{H}}(2m)\exp\left(-\frac{\varepsilon^2 m}{8}\right). \tag{3}$$

Hence, if $\tau_{\mathcal{H}}$ grows sub-exponentially, then the class will satisfy uniform convergence and therefore the ERM algorithm learns successfully. Observe that we can think of $\tau_{\mathcal{H}}$ as a covering number. Specifically, assume the set of labels is $\mathcal{Y} \subseteq \mathbb{R}$, $|\mathcal{Y}| < \infty$, and fix $\varepsilon > 0$ such that $\forall y, y' \in \mathcal{Y}$, if $y \ne y'$ then $|y - y'| > \varepsilon$. Let $X = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. Then

$$\begin{aligned} \left|\mathcal{H}|_X\right| &= \left|\{h|_X :\ h \in \mathcal{H}\}\right| \\ &= \max\{|H| :\ H \subseteq \mathcal{H} \ \wedge\ \forall h, h' \in H,\ h \ne h' \Rightarrow h|_X \ne h'|_X\} \\ &= \max\{|H| :\ H \subseteq \mathcal{H} \ \wedge\ \forall h, h' \in H,\ h \ne h' \Rightarrow \rho_{X',\infty}(h, h') > \varepsilon\} \\ &= N_{\text{in}}\left(\mathcal{H}, \varepsilon, \rho_{X',\infty}\right), \end{aligned}$$

where $X' = (x_1, \ldots, x_m)$. Hence,

$$\tau_{\mathcal{H}}(m) = \max_{\substack{X \subseteq \mathcal{X} \\ |X| = m}} \left|\mathcal{H}|_X\right| = \max_{X \in \mathcal{X}^m} N_{\text{in}}\left(\mathcal{H}, \varepsilon, \rho_{X,\infty}\right).$$

This motivates the following definition.

▶ **Definition 12.** *Let Ω be a set, $\mathcal{F}$ be a set of functions $\Omega \to \mathbb{R}$, $p \in (0, \infty]$, and $\varepsilon > 0$. The uniform ε-covering number for $\mathcal{F}$ with respect to the empirical p-semi-norm is*

$$N_p^{\text{uniform}}(\mathcal{F}, \varepsilon, m) = \max\{N_{\text{in}}(\mathcal{F}, \varepsilon, \rho_{S,p}) :\ S \in \Omega^m\}$$

The following theorem shows that we can generalize the result of Eq. (3) to other covering numbers beyond $\tau_{\mathcal{H}}$.

▶ **Theorem 13.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets, let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$, let $\ell : \mathcal{Y}^2 \to [0, c]$ be a loss function bounded by some positive $c \in \mathbb{R}$, and let $\mathcal{L}$ be the loss class of $\mathcal{H}$ with respect to ℓ. Then for any $\varepsilon > 0$, any $m \in \mathbb{N}$, and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\exists h \in \mathcal{H} : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \leq 4N_1^{\text{uniform}}\left(\mathcal{L}, \epsilon/8, 2m\right) \exp\left(-\frac{\epsilon^2 m}{32c^2}\right).$$

Proof. We modify the proof of Theorem 9 in Unit 5. Recall that in that proof, we considered two independent samples $S = (z_1, \ldots, z_m)$, and $S' \sim \mathcal{D}^m$, $S' = (z'_1, \ldots, z'_m)$, and showed that[2]

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\exists h \in \mathcal{H} : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \leq 2\mathbb{P}_{S\sim\mathcal{D}^m, S'\sim\mathcal{D}^m}\left[\exists h \in \mathcal{H} : \ \left|\frac{1}{m}\sum_{i\in[m]} \ell_h(z_i) - \ell_h(z'_i)\right| \geq \frac{\varepsilon}{2}\right], \tag{4}$$

where for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $\ell_h(z) = \ell(h(x), y)$ is the loss function in $\mathcal{L}$ associated with h. Let C be an internal $\frac{\varepsilon}{8}$-cover of $\mathcal{L}$ with respect to $\rho_{S\circ S',1}$ such that

$$|C| = N_1^{\text{uniform}}\left(\mathcal{L}, \varepsilon/8, 2m\right).$$

(A suitable cover C with this cardinality exists by the definition of N_1^{uniform}.) The idea of the next step is to replace each ℓ_h with an approximation $\ell'_h \in C$, that is, to quantize or approximate the functions in $\mathcal{L}$ using the coarser set of functions C. Specifically, for any $h \in \mathcal{H}$, let $\ell'_h \in C$ be a function such that $\rho_{S\circ S',1}(\ell_h, \ell'_h) \leq \varepsilon/8$. Then the expression in Eq. (4) can be rewritten as follows.

$$\begin{aligned}
\frac{\varepsilon}{2} &\leq \left|\frac{1}{m}\sum_{i\in[m]} \ell_h(z_i) - \ell_h(z'_i)\right| \qquad (5)\\
&= \left|\frac{1}{m}\sum_{i\in[m]} \ell_h(z_i) - \underbrace{\ell'_h(z_i) + \ell'_h(z_i)}_{=0} - \underbrace{\ell'_h(z'_i) + \ell'_h(z'_i)}_{=0} - \ell_h(z'_i)\right| \\
&\leq \frac{1}{m}\sum_{i\in[m]} |\ell_h(z_i) - \ell'_h(z_i)| + \left|\frac{1}{m}\sum_{i\in[m]} \ell'_h(z_i) - \ell'_h(z'_i)\right| + \frac{1}{m}\sum_{i\in[m]} |\ell'_h(z'_i) - \ell_h(z'_i)| \\
&\leq \frac{\varepsilon}{8} + \left|\frac{1}{m}\sum_{i\in[m]} \ell'_h(z_i) - \ell'_h(z'_i)\right| + \frac{\varepsilon}{8},
\end{aligned}$$

and so Eq. (5) implies $\left|\frac{1}{m}\sum_{i\in[m]} \ell'_h(z_i) - \ell'_h(z'_i)\right| \geq \varepsilon/4$. Thus,

[2] In the proof of Theorem 9 in Unit 5, this follows from Claim I together with the beginning of the proof of Claim II.

$$
\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^m, S'\sim\mathcal{D}^m}\left[\exists h\in\mathcal{H}:\ |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon/2\right] \\
&\quad\le 2\mathbb{P}_{S,S'}\left[\exists h\in\mathcal{H}:\ \left|\frac{1}{m}\sum_{i\in[m]}\ell_h(z_i)-\ell_h(z_i')\right|\ge\frac{\varepsilon}{2}\right] \\
&\quad= 2\mathbb{P}_{S,S'}\left[\exists h\in\mathcal{H}:\ \left|\frac{1}{m}\sum_{i\in[m]}\ell'_h(z_i)-\ell'_h(z_i')\right|\ge\varepsilon/4\right] \\
&\quad= 2\mathbb{E}_{S,S'}\mathbb{P}_{\sigma\sim\{\pm1\}^m}\left[\exists h\in\mathcal{H}:\ \left|\frac{1}{m}\sum_{i\in[m]}\sigma_i(\ell'_h(z_i)-\ell'_h(z_i'))\right|\ge\varepsilon/4\right]. \qquad (6)
\end{aligned}
$$

In the last equality we applied the symmetrization technique which we have seen in previous lectures, using the fact that $\left(\ell'_h(z_i)-\ell'_h(z_i')\right)_{i\in[m]} \overset{d}{=} \left(\sigma_i(\ell'_h(z_i)-\ell'_h(z_i'))\right)_{i\in[m]}$.

Next, for each S, S',

$$
\begin{aligned}
&\mathbb{P}_{\sigma\sim\{\pm1\}^m}\left[\bigcup_{h\in\mathcal{H}}\left\{\left|\frac{1}{m}\sum_{i\in[m]}\sigma_i(\ell'_h(z_i)-\ell'_h(z_i'))\right|\ge\varepsilon/4\right\}\right] \\
&\quad= \mathbb{P}_{\sigma}\left[\bigcup_{\ell'\in C}\left\{\left|\frac{1}{m}\sum_{i\in[m]}\sigma_i(\ell'(z_i)-\ell'(z_i'))\right|\ge\varepsilon/4\right\}\right] && (C=\{\ell'_h:\ h\in\mathcal{H}\}) \\
&\quad\le \sum_{\ell'\in C}\mathbb{P}_{\sigma}\left[\left|\frac{1}{m}\sum_{i\in[m]}\sigma_i(\ell'(z_i)-\ell'(z_i'))\right|\ge\varepsilon/4\right] && (\text{union bound}) \\
&\quad= \sum_{\ell'\in C}\mathbb{P}_{\sigma}\left[\left|\frac{1}{m}\sum_{i\in[m]}Z_i\right|\ge\varepsilon/4\right] && (\text{let } Z_i=\sigma_i(\ell'(z_i)-\ell'(z_i'))) \\
&\quad\le \sum_{\ell'\in C}2\exp\left(-\frac{2\left(\varepsilon/4\right)^2 m}{(2c)^2}\right) && (\text{Hoeffding's: } Z_i \text{ are i.i.d.,} \\
&&& \quad \mathbb{E}[Z_i]=0,\ Z_i\in[-c,c]) \\
&\quad= 2|C|\exp\left(-\frac{\varepsilon^2 m}{32c^2}\right) \\
&\quad\le 2\cdot N_1^{\text{uniform}}(\mathcal{L},\varepsilon/8,2m)\cdot\exp\left(-\frac{\varepsilon^2 m}{32c^2}\right). && (\text{choice of } C) \qquad (7)
\end{aligned}
$$

Combining Eq. (4), (6) and (7) yields the theorem. ◀

5 Chaining

Chaining is a technique for bounding the Rademacher average. Roughly speaking, we can think of it as a sophisticated union bound. In this section we loosely follow Nelson [2] and present a sequence of steps that gradually build up to the full technique.

Let $A\subseteq\mathbb{R}^m$ be a finite set such that $\|a\|_2\le r$ for all $a\in A$. We are interested in upper bounding the Rademacher average

$$
\mathsf{Rad}(A)=\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{1}{m}\sum_{i=1}^m\sigma_i a_i\right].
$$

Method 0: Union Bound

For any $t > 0$,

$$\mathbb{P}_{\sigma\sim\{\pm1\}^m}\left[\sup_{a\in A}\frac{1}{m}\sum_{i\in[m]}a_i\sigma_i > t\right] \le |A|\sup_{a\in A}\mathbb{P}_{\sigma\sim\{\pm1\}^m}\left[\frac{1}{m}\sum_{i\in[m]}a_i\sigma_i > t\right] \quad \text{(union bound)}$$

$$\le 2|A|\exp\left(-\frac{mt^2}{2r^2}\right). \quad \text{(Hoeffding's inequality)}$$

Hence,

$$\begin{aligned}
\mathsf{Rad}(A) &= \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^m}[Z] && (Z = \sup_{a\in A}\frac{1}{m}\sum_{i=1}^m\sigma_i a_i)\\
&= \mathbb{P}[Z\le t]\,\mathbb{E}[Z \mid Z\le t] + \mathbb{P}[Z>t]\,\mathbb{E}[Z\mid Z>t] && \text{(law of total expectation)}\\
&\le t + \mathbb{P}[Z>t]\,\frac{r}{\sqrt{m}} && (Z \le \sup_{a,\sigma}\frac{\langle\sigma,a\rangle}{m} \le \sup_{a,\sigma}\frac{\|a\|\|\sigma\|}{m} \le \frac{r}{\sqrt{m}}\text{ a.s.})\\
&\le t + \frac{2r}{\sqrt{m}}|A|\exp\left(-\frac{mt^2}{2r^2}\right). && \text{(from the previous inequality)}
\end{aligned}$$

Choosing $t = \frac{2r}{\sqrt{m}}$, we obtain

$$\mathsf{Rad}(A) \le \frac{2r}{\sqrt{m}} + \frac{2r}{\sqrt{m}}|A|\exp(-2) = O\left(|A|\cdot\frac{r}{\sqrt{m}}\right).$$

This bound is very weak, but it was also very simple to prove – the tools we used were the union bound and Hoeffding's inequality.

Method 1: Massart's Lemma

In Unit 6 we proved Massart's lemma, which states that

$$\mathsf{Rad}(A) \le \frac{r\sqrt{2\ln(|A|)}}{m}.$$

This is considerably stronger than the bound from Method 0 above. The proof we presented for Massart's lemma used the Maximal Inequality lemma, which was based on the Chernoff method for proving concentration bounds. However, in the homework we will see that Massart's lemma can also be proved in a manner very similar to Method 0 above. From that point of view, Massart's lemma is basically a clever application of the union bound with Hoeffding's inequality.

Method 2: ε-Cover

If $|A| = \infty$ (or A is a finite but very large set) then the bound from Massart's lemma is not useful. We can overcome this problem by approximating the large set $|A|$ with a much smaller set C, which is an ε-cover for A. On the one hand, C is small and so Massart's lemma gives a good bound on $\mathsf{Rad}(C)$, and on the other hand C is a "good enough" approximation of A, such that a good bound for $\mathsf{Rad}(C)$ implies a good bound for $\mathsf{Rad}(A)$.

More fully, let $C \subseteq A$ be an internal ε-cover of A such that $|C| = N_{\text{in}}(A,\varepsilon,\rho)$, where $\rho(x,y) = \|x-y\|_2$. For each $a\in A$, let $\pi(a) = c$ for $c\in C$ such that $\rho(a,c)\le\varepsilon$. By linearity of the inner product, for any $a\in A$ and $\sigma\in\{\pm1\}^m$,

$$\langle\sigma,a\rangle = \langle\sigma,\pi(a)\rangle + \langle\sigma,a-\pi(a)\rangle \le \langle\sigma,\pi(a)\rangle + \|\sigma\|_2\|a-\pi(a)\|_2 \le \langle\sigma,\pi(a)\rangle + \varepsilon\sqrt{m}. \quad (8)$$

Therefore,

$$\begin{aligned}
\mathsf{Rad}(A) &= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{\langle\sigma,a\rangle}{m}\right] \\
&\le \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{\langle\sigma,\pi(a)\rangle+\varepsilon\sqrt{m}}{m}\right] && \text{(by Eq. (8))} \\
&= \frac{\varepsilon}{\sqrt{m}} + \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{\langle\sigma,\pi(a)\rangle}{m}\right] \\
&= \frac{\varepsilon}{\sqrt{m}} + \mathsf{Rad}(C) \\
&\le \frac{\varepsilon}{\sqrt{m}} + \frac{r\sqrt{2\ln(|C|)}}{m}. && \text{(Massart's lemma)} \\
&= \frac{\varepsilon}{\sqrt{m}} + \frac{r\sqrt{2\ln(N_{\text{in}}(A,\varepsilon,\rho))}}{m}. && (9)
\end{aligned}$$

If $\varepsilon = 0$ this bound is the same as in Method 1. However, if we choose a value $\varepsilon > 0$ that minimizes Eq. (9), we can get a better bound.

Method 3: Chaining

Instead of committing to a particular ε-cover, chaining is a technique that uses a countable number of ε-covers with $\varepsilon \to 0$. We can think of this as a recursive application of Method 2, where we first approximate A by a coarse ε-cover, and then repeatedly improve the approximation with finer and finer covers. Formally, the result is as follows.

▶ **Theorem 14** (Dudley [1]). *Let $r > 0$ and let $A \subseteq \mathbb{R}^m$ be a set such that $\|a\|_2 \le r$ for all $a \in A$. Then*

$$\mathsf{Rad}(A) \le \frac{12}{m}\int_0^r \sqrt{\ln(N(A,\varepsilon,\rho))}\,\mathrm{d}\varepsilon.$$

Furthermore, if $r \ge 1$ then

$$\mathsf{Rad}(A) \le \frac{12r}{m}\int_0^1 \sqrt{\ln(N(A,\varepsilon,\rho))}\,\mathrm{d}\varepsilon.$$

▶ Remark 15.

- In particular, $r \ge 1$ if A is a set of boolean vectors, for instance if A is the 0-1 loss class for some class of hypotheses, .
- A small modification of the proof yields a bound of the form

$$\mathsf{Rad}(A) \le \inf_{\alpha\in[0,r/2]} 4\alpha + \frac{12}{m}\int_\alpha^r \sqrt{\ln(N(A,\varepsilon,\rho))}\,\mathrm{d}\varepsilon.$$

  In some cases, this bound has the advantage that the integral $\int_\alpha^r$ converges even though $\int_0^r$ does not converge (e.g., if $\sqrt{\ln N(A,\varepsilon,\rho)} > \frac{1}{\varepsilon}$ in some neighborhood of 0). See Theorem 1.19 in [3]. ⌟

Proof of Theorem 14. Let $\rho(x,y) = \|x-y\|_2$. For any $k \in \{0,1,2,\dots\}$, let $C_k \subseteq \mathbb{R}^m$ be an ε_k-cover of A where $\varepsilon_k = \frac{r}{2^k}$ and $|C_k| = N(A,\varepsilon_k,\rho)$. In particular, we can take $C_0 = \{0\}$,

because $\rho(a,0) = \|a\|_2 \le r = \varepsilon_0$ for all $a \in A$. For any k and any $a \in A$, let $\pi_k(a) = c$ such that $c \in C_k$ and $\rho(a,c) \le \varepsilon_k$. Furthermore, let $\Delta_k(a) = \pi_k(a) - \pi_{k-1}(a)$. For any $a \in A$,

$$a = \pi_0(a) + \sum_{k=1}^{\infty} \Delta_k(a) = \sum_{k=1}^{\infty} \Delta_k(a).$$

Hence,

$$\begin{aligned}
\mathsf{Rad}(A) &= \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{\langle a,\sigma\rangle}{m}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{a\in A}\frac{\langle\sum_{k=1}^{\infty}\Delta_k(a),\sigma\rangle}{m}\right] \\
&\le \sum_{k=1}^{\infty}\mathbb{E}_\sigma\left[\sup_{a\in A}\frac{\langle\Delta_k(a),\sigma\rangle}{m}\right] \qquad \text{(linearity of inner product, } \sup\Sigma \le \Sigma\sup\text{)} \\
&= \sum_{k=1}^{\infty}\mathsf{Rad}(\Delta_k), \qquad (10)
\end{aligned}$$

where $\Delta_k = \{\Delta_k(a) :\ a \in A\}$. Notice that:

$$\begin{aligned}
|\Delta_k| &= |\{\pi_k(a) - \pi_{k-1}(a) :\ a \in A\}| \\
&\le |\{\pi_k(a) :\ a \in A\}| \cdot |\{\pi_{k-1}(a) :\ a \in A\}| \\
&\le N(A,\varepsilon_k,\rho)\cdot N(A,\varepsilon_{k-1},\rho) \\
&\le N(A,\varepsilon_k,\rho)^2,
\end{aligned}$$

and for all $a \in A$,

$$\|\Delta_k(a)\|_2 \le \|\pi_k(a) - a\|_2 + \|a - \pi_{k-1}(a)\|_2 \le 3\varepsilon_k.$$

Thus, by Massart's lemma,

$$\mathsf{Rad}(\Delta_k) \le \frac{3\varepsilon_k\sqrt{2\ln\left(N(A,\varepsilon_k,\rho)^2\right)}}{m} = \frac{6\varepsilon_k\sqrt{\ln\left(N(A,\varepsilon_k,\rho)\right)}}{m}.$$

Plugging this into Eq. (10) yields

$$\begin{aligned}
\mathsf{Rad}(A) &\le \sum_{k=1}^{\infty}\frac{6\varepsilon_k\sqrt{\ln\left(N(A,\varepsilon_k,\rho)\right)}}{m} \\
&= \frac{6}{m}\sum_{k=1}^{\infty}\varepsilon_k\sqrt{\ln\left(N(A,\varepsilon_k,\rho)\right)} \\
&= \frac{12}{m}\sum_{k=1}^{\infty}\left(\frac{r}{2^k} - \frac{r}{2^{k+1}}\right)\sqrt{\ln\left(N(A,{}^{r}\!/_{2^k},\rho)\right)} \qquad (11) \\
&\le \frac{12}{m}\int_0^r \sqrt{\ln\left(N(A,\varepsilon,\rho)\right)}\,\mathrm{d}\varepsilon.
\end{aligned}$$

In the last inequality we used the fact the lower Riemann sum of the function $f(\varepsilon) = \sqrt{\ln\left(N(A,\varepsilon,\rho)\right)}$ is upper bounded by the integral of $f(\varepsilon)$, together with the fact that $f(\varepsilon)$ is monotone decreasing in $[0,r]$ with $f(r) = 0$. This completes the proof of the first part of the statement.

For the second part of the statement, notice that if $r \ge 1$ then $N(A,{}^{r}\!/_{2^k},\rho) \le N(A,{}^{1}\!/_{2^k},\rho)$, and so Eq. (11) is upper-bounded by

$$\frac{12r}{m}\int_0^1 \sqrt{\ln\left(N(A,\varepsilon,\rho)\right)}\,\mathrm{d}\varepsilon. \qquad \blacktriangleleft$$

6 Learning Bounds via Chaining

We now have all the ingredients to prove the tight sample complexity bound that appears in the Fundamental theorem (Theorem 13 in Unit 5).

▷ Claim 16. Let $\mathcal{X}$ be a set, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ with $\mathsf{VC}(\mathcal{H}) = d$. Then $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$m = O\left(\frac{d + \ln(1/\delta)}{\varepsilon^2}\right).$$

Proof. From Theorem 12 (iii) in Unit 6, the hypothesis h selected by an ERM algorithm that uses a sample of size m satisfies

$$L_{\mathcal{D}}(h) \le L_{\mathcal{D}}(\mathcal{H}) + 2\mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) + \sqrt{\frac{2\ln(2/\delta)}{m}}. \tag{12}$$

Let S be the sample and $A = \{h(S) : h \in \mathcal{H}\}$. Then $\|a\|_2 \le \sqrt{m}$ for all $a \in A$. Let $\rho(x,y) = \|x - y\|_2$. Dudley's theorem implies

$$\mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) \le \frac{12}{\sqrt{m}} \int_0^1 \sqrt{\ln\left(N(A,\varepsilon,\rho)\right)}\,\mathrm{d}\varepsilon = \frac{12}{\sqrt{m}} \int_0^1 \sqrt{\ln N(\mathcal{H},\varepsilon,\rho_{S,2})}\,\mathrm{d}\varepsilon. \tag{13}$$

Note that

$$\begin{aligned}
\ln N(\mathcal{H},\varepsilon,\rho_{S,2}) &\le \ln M(\mathcal{H},\varepsilon,\rho_{S,2}) && \text{(Claim 4)}\\
&\le d\ln\left(\frac{4e}{\varepsilon^2}\ln\left(\frac{2e}{\varepsilon^2}\right)\right) && \text{(by Lemma 11, } M(\mathcal{H},\varepsilon,\rho_{S,2}) \le \left(4e/\varepsilon \ln\left(2e/\varepsilon\right)\right)^d\text{)}\\
&\le d\ln\left(\frac{8e}{\varepsilon^4}\right) = d(\ln(8e) + 4\ln(1/\varepsilon)). && (\ln x \le x/e) \qquad (14)
\end{aligned}$$

Combining Eq. (13) and (14) and using numerical integration yields

$$\mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) \le 12\sqrt{\frac{d}{m}} \int_0^1 \sqrt{(\ln(8e) + 4\ln(1/\varepsilon))}\,\mathrm{d}\varepsilon \le 31\sqrt{\frac{d}{m}}.$$

Plugging this into Eq. (12), we obtain

$$L_{\mathcal{D}}(h) \le L_{\mathcal{D}}(\mathcal{H}) + 62\sqrt{\frac{d}{m}} + \sqrt{\frac{2\ln(2/\delta)}{m}} \le L_{\mathcal{D}}(\mathcal{H}) + O\left(\sqrt{\frac{d + \ln(1/\delta)}{m}}\right).$$

Thus, taking m as in the statement is sufficient to ensure that $L_{\mathcal{D}}(h) \le L_{\mathcal{D}}(\mathcal{H}) + \varepsilon$. ◀

7 Bibliographic Notes

The Chaining technique was introduced by Dudley [1], and has seen many subsequent developments and improvements.

We followed the expositions in [2] and in [3, Section 1.9]. See also this excellent video lecture. Many additional resources on chaining are available from the website (archived) of the conference on Chaining Methods and their Applications to Computer Science.

References

1 Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

2 Jelani Nelson. Chaining introduction with some computer science applications. *Bull. EATCS*, 120, 2016. URL: http://eatcs.org/beatcs/index.php/beatcs/article/view/450.

3 Michael M. Wolf. Mathematical foundations of supervised learning, 2020. URL: https://web.archive.org/web/20210114220017/https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2020S/ML_notes_main.pdf.