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

Unit 6

Rademacher Complexity

1 Introduction

In the previous unit we saw how the VC dimension controls the growth function, which in turn determines the sample complexity of uniform convergence. In this unit we provide somewhat different perspective on uniform convergence, via *Rademacher complexity*. The Rademacher complexity depends both on the hypothesis class $\mathcal{H}$ and on the unknown distribution $\mathcal{D}$, in contrast to the VC dimension that depends solely on the hypothesis class. Thus, the Rademacher complexity can be understood as an average-case analysis, in contrast to the worst-case analysis offered by the VC dimension. As we will see in the next Unit, the more nuanced Rademacher complexity analysis also lends itself to a technique called chaining, which will allow us to eliminate the unnecessary logarithmic factor the that appeared in our derivation of the fundamental theorem in the previous unit.

2 Concentration of Measure for Uniform Convergence

The following notation will be handy for our analysis of uniform convergence.

▶ **Notation 1.** We write

$$\Delta_S(\mathcal{H}) = \sup_{h\in\mathcal{H}} \left|L_S(h) - L_{\mathcal{D}}(h)\right|,$$

$$\Delta_S^+(\mathcal{H}) = \sup_{h\in\mathcal{H}} L_S(h) - L_{\mathcal{D}}(h),$$

$$\Delta_S^-(\mathcal{H}) = \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h),$$

where $\mathcal{H}$ is a class of functions, S is a sample, $\mathcal{D}$ is a distribution, and the loss function should be understood from the context. ⌟

Uniform convergence means that if S is a large i.i.d. sample then with high probability $\Delta_S(\mathcal{H})$ is small. The following claim says that $\Delta_S(\mathcal{H})$ is close to its expectation. This implies that, in order to show that $\Delta_S(\mathcal{H})$ is small, it will suffice to bound its expectation.

▷ Claim 2. Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$, let $\ell : \mathcal{Y}^2 \to [0,1]$ be a loss function, and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Then for any $m \in \mathbb{N}$ and $\varepsilon > 0$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S(\mathcal{H}) \geq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_S(\mathcal{H})\right] + \varepsilon\right] \leq \exp\left(-2m\varepsilon^2\right),$$

and similarly,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H}) \geq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H})\right] + \varepsilon\right] \leq \exp\left(-2m\varepsilon^2\right),$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S^-(\mathcal{H}) \geq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_S^-(\mathcal{H})\right] + \varepsilon\right] \leq \exp\left(-2m\varepsilon^2\right).$$ ⌟

This claim follows from the following concentration-of-measure theorem.

▶ **Theorem 3** (McDiarmid's Inequality). *Let Ω be a set and let $f : \Omega^m \to \mathbb{R}$ be a function. Assume there exist $c_1, \ldots, c_m \in \mathbb{R}$ such that f satisfies the following* bounded differences property*:*

$$\forall z_1, \ldots, z_m, z'_1, \ldots, z'_m \in \Omega \ \forall i \in [m] : \\ |f(z_1, \ldots, z_i, \ldots, z_m) - f(z_1, \ldots, z'_i, \ldots, z_m)| \leq c_i.$$

Let $Z_1, \ldots, Z_m$ be independent random variables taking values in Ω. Then for any $\varepsilon > 0$,

$$\mathbb{P}\left[f(Z_1, \ldots, Z_m) - \mathbb{E}\left[f(Z_1, \ldots, Z_m)\right] \geq \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}\left[\mathbb{E}\left[f(Z_1, \ldots, Z_m)\right] - f(Z_1, \ldots, Z_m) \geq \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

▶ Remark 4.

- In the special case where $c_i = 1/m$ for all $i \in [m]$, the bound specifies to $\exp\left(-2m\varepsilon^2\right)$.
- McDiarmid's inequality is a generalization of Hoeffding's inequality. Hoeffding's is a concentration of measure result for the average of independent random variables, whereas McDiarmid's is a concentration of measure result for any function of independent random variables that satisfies the bounded differences property (including the average).
- McDiarmid's inequality is very powerful, because the function f can be arbitrarily complex so long as it satisfies the bounded differences property. Below we will apply McDiarmid's inequality to the function $f(S) = \Delta_S(\mathcal{H})$, which is a non-trivial function that involves a supremum over a possibly infinite class $\mathcal{H}$.

▶ **Exercise 5.** *Prove Theorem 3.*

Proof of Claim 2. We prove the first inequality, the proof for the other two inequalities is similar.

Fix $m \in \mathbb{N}$. First, we claim that $\Delta_S(\mathcal{H})$ satisfies the following bounded differences property. For any $S, S' \in (\mathcal{X} \times \mathcal{Y})^m$ with $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$, $S' = \big((x'_1, y'_1), \ldots, (x'_m, y'_m)\big)$, if there exists $j \in [m]$ such that such that $(x_i, y_i) = (x'_i, y'_i)$ for all $i \neq j$, then

$$|\Delta_S(\mathcal{H}) - \Delta_{S'}(\mathcal{H})| \leq \frac{1}{m}. \tag{1}$$

To see this, notice that

$$\begin{aligned}
\Delta_{S'}(\mathcal{H}) &= \sup_{h \in \mathcal{H}} \left|L_{S'}(h) - L_{\mathcal{D}}(h)\right| = \sup_{h \in \mathcal{H}} \left|L_{S'}(h) - L_S(h) + L_S(h) - L_{\mathcal{D}}(h)\right| \\
&\leq \sup_{h \in \mathcal{H}} \left|L_{S'}(h) - L_S(h)\right| + \left|L_S(h) - L_{\mathcal{D}}(h)\right| \\
&= \sup_{h \in \mathcal{H}} \left|\frac{\ell(h(x'_j), y'_j) - \ell(h(x_j), y_j)}{m}\right| + \left|L_S(h) - L_{\mathcal{D}}(h)\right| \\
&\leq 1/m + \sup_{h \in \mathcal{H}} \left|L_S(h) - L_{\mathcal{D}}(h)\right| \\
&= 1/m + \Delta_S(\mathcal{H}).
\end{aligned}$$

Applying the same argument with roles of S and S' reversed implies that $|\Delta_S(\mathcal{H}) - \Delta_{S'}(\mathcal{H})| \leq 1/m$. Hence, by McDiarmid's inequality,

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[\Delta_S(\mathcal{H}) \geq \mathop{\mathbb{E}}_{S' \sim \mathcal{D}^m}\left[\Delta_{S'}(\mathcal{H})\right] + \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m (1/m)^2}\right) = \exp\left(-2m\varepsilon^2\right). \qquad \blacktriangleleft$$

3 Definition of Rademacher Complexity

▶ **Definition 6.** *Let $A \subseteq \mathbb{R}^m$ be a bounded set of vectors.*[1] *The Rademacher average of A is*

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

where $\sigma = (\sigma_1,\ldots,\sigma_m)$ is a vector of random variables (called Rademacher variables) chosen independently and uniformly from $\{1,-1\}$.

The Rademacher average quantifies how well A correlates with a random vector $\sigma \in \{\pm1\}^m$. Similarly, the following definition quantifies the *richness* or *complexity* of a class of functions by measuring how well the functions can correlate with random labels.

▶ **Definition 7.** *Fix $m \in \mathbb{N}$. Let $\mathcal{X}$ be a nonempty set, and let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [-1,1]$. For any set $S = (x_1,\ldots,x_m) \in \mathcal{X}^m$, let*

$$\mathcal{F}(S) = \big\{\big(f(x_1),\ldots,f(x_m)\big) :\ f\in\mathcal{F}\big\} \subseteq \mathbb{R}^m.$$

(i) *Fix $S \in \mathcal{X}^m$. The empirical Rademacher complexity of $\mathcal{F}$ with respect to S is*

$$\mathsf{Rad}_S(\mathcal{F}) = \mathsf{Rad}(\mathcal{F}(S)) = \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right].$$

(ii) *Let $\mathcal{D}$ be a distribution over $\mathcal{X}$. The Rademacher complexity of size m of $\mathcal{F}$ with respect to $\mathcal{D}$ is*

$$\mathsf{Rad}_{\mathcal{D},m}(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathsf{Rad}(\mathcal{F}(S))\right] = \mathbb{E}_{S\sim\mathcal{D}^m,\sigma\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right].$$

We will simply write Rad_m when $\mathcal{D}$ is understood from context.

4

▶ **Definition 8.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets, let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$, and let $\ell : \mathcal{Y}^2 \to [0,1]$ be a loss function. The loss class of $\mathcal{H}$ with respect to ℓ is*

$$\mathcal{L} = \{(x,y) \mapsto \ell(h(x),y) :\ h\in\mathcal{H}\}.$$

(Namely, the loss class $\mathcal{L}$ is a set of functions $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$.)

▷ Claim 9. Let $\mathcal{X}$ be a set, let $\mathcal{H}$ be a class of functions $\mathcal{X}\to\{1,-1\}$, and let

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

be the loss class of $\mathcal{H}$ with respect to the 0-1 loss. Let $S = \big((x_1,y_1),\ldots,(x_m,y_m)\big) \in (\mathcal{X}\times\{\pm1\})^m$ be a sample, and let $S_x = (x_1,\ldots,x_m)$. Then

$$\mathsf{Rad}_S(\mathcal{L}) = {}^1\!/_2 \cdot \mathsf{Rad}_{S_x}(\mathcal{H}).$$

[1] Namely, there exists $M\in\mathbb{R}$ such that $\|v\|_2 \le M$ for all $v\in A$.

Proof.

$$\begin{aligned}
\mathsf{Rad}_S(\mathcal{L}) &= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}(h(x_i)\neq y_i)\right] \\
&= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{(1-y_i h(x_i))}{2}\right] \\
&= \frac{1}{2}\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i + \frac{1}{m}\sum_{i=1}^m \sigma_i(-y_i)h(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i + \sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i(-y_i)h(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i(-y_i)h(x_i)\right] \qquad (\mathbb{E}\left[\Sigma_{i=1}^m\sigma_i\right]=0) \\
&\overset{(\star)}{=} \frac{1}{2}\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right] = \frac{1}{2}\cdot\mathsf{Rad}_{S_x}(\mathcal{H}),
\end{aligned}$$

where $(\star)$ holds because σ_i and $\sigma_i(-y_i)$ have the same distribution. ◀

▶ **Exercise 10.** *Show that if $\mathcal{H}$ is a class of functions $\mathcal{X}\to\{0,1\}$ then $\mathsf{Rad}_S(\mathcal{L}) = \mathsf{Rad}_{S_x}(\mathcal{H})$.*

▶ **Lemma 11.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, and let $\mathcal{H}$ be a class of functions $\mathcal{X}\to\mathcal{Y}$. Let $\mathcal{L}$ be the loss class of $\mathcal{H}$ with respect to some loss function $\ell:\ \mathcal{Y}^2\to[0,1]$. Then for any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ and $m\in\mathbb{N}$,*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H})\right] \le 2\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}), \text{ and}$$

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\Delta_S^-(\mathcal{H})\right] \le 2\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}).$$

Proof. We will present the proof for the first inequality, the proof for the second inequality is similar. Note that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(h)\right] = \frac{1}{m}\sum_{i\in[m]}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\ell(h(x_i),y_i)\right] = \frac{1}{m}\sum_{i\in[m]}L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h). \tag{2}$$

Hence, we can use the double sampling technique to express $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H})\right]$ as an expectation concerning a finite sample:

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H})\right] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} L_S(h) - L_{\mathcal{D}}(h)\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} L_S(h) - \mathbb{E}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right]\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i\in[m]}\ell(h(x_i),y_i) - \ell(h(x_i'),y_i')\right]\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i\in[m]} Q_i(h)\right]\right] \\
&\le \mathbb{E}_{S\sim\mathcal{D}^m,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]} Q_i(h)\right],
\end{aligned}$$

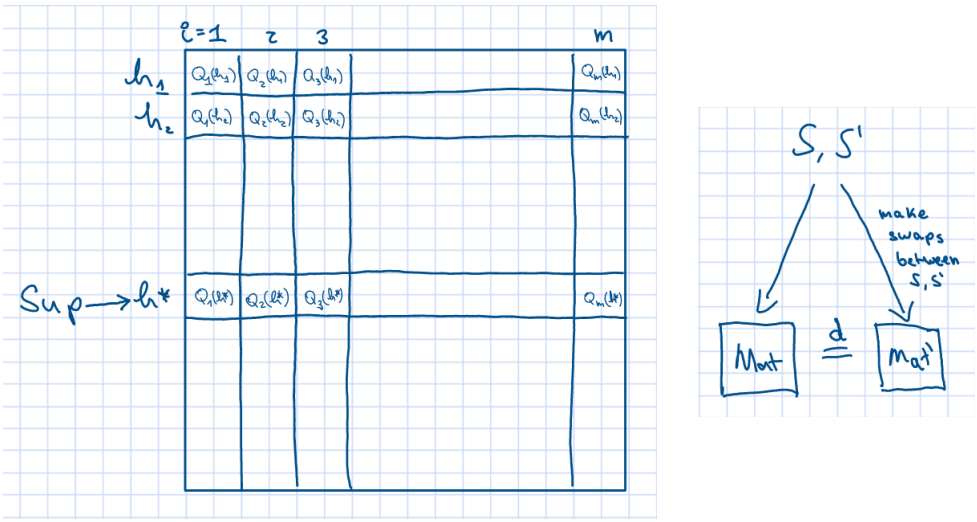

Figure 1 Intuition for the symmetrization technique. The supremum chooses one row from this matrix of random variables. The entire matrix has the same distribution whether or not we swap some subset of the examples in S and S'.

where we used the notation $Q_i(h) = \ell(h(x_i), y_i) - \ell(h(x'_i), y'_i)$.

S and S' are independent and have the same distribution, and therefore we can view the samples from S and S' as being interchangeable. Formally, we use the following *symmetrization* technique (we already saw a variant of this technique in the previous unit). Each random variable $Q_i(h) = \ell(h(x_i), y_i) - \ell(h(x'_i), y'_i)$ is equal in distribution to $-Q_i(h) = \ell(h(x'_i), y'_i) - \ell(h(x_i), y_i)$, because flipping the sign corresponds to swapping the names of (x_i, y_i) with (x'_i, y'_i) for some i. Moreover, if we introduce Rademacher variables σ_i (that are independent and uniform over $\{\pm 1\}$), then the entire matrix $(Q_i(h))_{i\in[m],h\in\mathcal{H}}$ is equal in distribution to the matrix $(\sigma_i Q_i(h))_{i\in[m],h\in\mathcal{H}}$, for the same reason (see also Figure 1). Hence,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m \\ S'\sim\mathcal{D}^m}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}Q_i\right] &= \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m \\ S'\sim\mathcal{D}^m \\ \sigma\sim\{\pm1\}^m}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}\sigma_i Q_i\right] \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}\sigma_i\left(\ell(h(x_i),y_i)-\ell(h(x'_i),y'_i)\right)\right] \\
&\le \mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}\sigma_i\ell(h(x_i),y_i)+\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}(-\sigma_i)\ell(h(x'_i),y'_i)\right] \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}\sigma_i\ell(h(x_i),y_i)\right]+\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i\in[m]}(-\sigma_i)\ell(h(x'_i),y'_i)\right] \\
&= 2\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}),
\end{aligned}$$

Where the last equality uses the fact that σ_i and $-\sigma_i$ have the same distribution. ◀

As a corollary, we obtain the following PAC learning bounds.

▶ **Theorem 12.** *Let $\delta \in (0,1)$, let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$, and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\ell : \mathcal{Y}^2 \to [0,1]$ be a loss function. Let $S \sim \mathcal{D}^m$ for some $m \in \mathbb{N}$.*

(i) *With probability at least $1-\delta$,*

$$\forall h \in \mathcal{H} : \; L_{\mathcal{D}}(h) \le L_S(h) + 2\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

where $\mathcal{L}$ is the loss class of $\mathcal{H}$ w.r.t. ℓ.

(ii) *Assume $\mathcal{Y} = \{\pm 1\}$ and ℓ is the 0-1 loss. Then with probability at least $1-\delta$,*

$$\forall h \in \mathcal{H} : \; L_{\mathcal{D}}(h) \le L_S(h) + \mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

(iii) *Assume $\mathcal{Y} = \{\pm 1\}$, ℓ is the 0-1 loss, and let h be the output of an $\mathrm{ERM}_{\mathcal{H}}$ algorithm executed on S. Then with probability at least $1-\delta$,*

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

Proof.

(i) Denote $\varepsilon = \sqrt{\frac{\ln(1/\delta)}{2m}}$. From Claim 2,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S^-(\mathcal{H}) \ge \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_{S'}^-(\mathcal{H})\right] + \varepsilon\right] \le \exp\left(-2m\varepsilon^2\right) = \delta.$$

Hence, with probability at least $1-\delta$ it is the case that $\Delta_S(\mathcal{H}) \le \mathbb{E}\left[\Delta_{S'}(\mathcal{H})\right] + \varepsilon$, and then for all $h \in \mathcal{H}$,

$$\begin{aligned} L_{\mathcal{D}}(h) &= L_S(h) + (L_{\mathcal{D}}(h) - L_S(h)) \\ &\le L_S(h) + \Delta_S^-(\mathcal{H}) \\ &\le L_S(h) + \mathbb{E}\left[\Delta_{S'}^-(\mathcal{H})\right] + \varepsilon \\ &\le L_S(h) + 2\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}) + \varepsilon. && \text{(Lemma 11).} \end{aligned}$$

(ii) Follows from Item (i) and Claim 9.

(iii) Conceptually, this follows from (ii) together with the fact that uniform convergence implies that any ERM algorithm is a PAC learner (Lemma 11 in Lecture 3). More fully, choosing $\varepsilon = \sqrt{\frac{2\ln(2/\delta)}{m}}$, Claim 2 implies that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S^+(\mathcal{H}) \ge \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_{S'}^+(\mathcal{H})\right] + \varepsilon/2\right] \le \exp\left(-2m(\varepsilon/2)^2\right) = \delta/2, \tag{3}$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\Delta_S^-(\mathcal{H}) \ge \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_{S'}^-(\mathcal{H})\right] + \varepsilon/2\right] \le \exp\left(-2m(\varepsilon/2)^2\right) = \delta/2. \tag{4}$$

Hence, with probability at least $1-\delta$, for any $h^* \in \mathcal{H}$,

$$\begin{aligned} L_{\mathcal{D}}(h) &\le \underbrace{L_{\mathcal{D}}(h) - L_S(h)}_{\le \Delta_S^-(\mathcal{H})} + \underbrace{L_S(h) - L_S(h^*)}_{\le 0} + \underbrace{L_S(h^*) - L_{\mathcal{D}}(h^*)}_{\le \Delta_S^+(\mathcal{H})} + L_{\mathcal{D}}(h^*) \\ &\le L_{\mathcal{D}}(h^*) + \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_{S'}^-(\mathcal{H})\right] + \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\Delta_{S'}^+(\mathcal{H})\right] + \varepsilon && \text{(From (3) and (4))} \\ &\le L_{\mathcal{D}}(h^*) + 4\mathsf{Rad}_{\mathcal{D},m}(\mathcal{L}) + \varepsilon, && \text{(Lemma 11)} \\ &\le L_{\mathcal{D}}(h^*) + 2\mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) + \varepsilon, && \text{(Claim 9)} \end{aligned}$$

as desired. ◀

5 Bounding the Rademacher Complexity

Theorem 12 shows that to obtain PAC learning bounds, it suffices to bound the Rademacher complexity.

5.1 Estimating the Rademacher Complexity

In some cases, it is possible to show that the Rademacher complexity is small by estimating it empirically. Namely, one can take samples from the unknown distribution and compute the empirical Rademacher complexity. By McDiarmid's inequality, the empirical Rademacher complexity is close to the Rademacher complexity. This yields a version of Theorem 12 that contains the empirical Rademacher complexity instead of the Rademacher complexity.

Note that for a fixed sample S, the empirical Rademacher complexity is defined as an average over all possible assignments to the Rademacher variables, so computing it would appear to require exponential time in the number of samples in S. To over come this, one can instead estimate the empirical Rademacher complexity by sampling a small number of vectors of Rademacher variables uniformly at random, and taking the average only over these vectors. By Hoeffding's inequality, this estimate converges exponentially fast to the empirical Rademacher complexity, and this can again yield a bound similar to Theorem 12, that involves only the estimate of the empirical Rademacher complexity.

Unfortunately, the computational complexity can be prohibitive even if we attempt only to estimate the empirical Rademacher complexity as outlined above. This is because for any fixed sample S and vector of Rademacher variables σ, computing $\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)$ is a combinatorial optimization problem that involves searching over the entire class $\mathcal{F}$. For many hypothesis classes, this optimization problem can be NP-hard.

5.2 Combinatorial Bound on the Rademacher Complexity

Another approach is to bound the Rademacher Complexity using the VC dimension. This is tantamount to saying that the average-case analysis offered by the Rademacher complexity is upper bounded by the worst-case analysis offered by the VC dimension.

5.2.1 Massart's Lemma

▶ **Lemma 13** (Maximal Inequality). *Let $n \in \mathbb{N}$, let $v > 0$, let $Z_1, \ldots, Z_n$ be real-valued random variables, and assume that for all $i \in [n]$ and $\lambda > 0$,*

$$\psi_{Z_i}(\lambda) \leq \frac{\lambda^2 v}{2}.$$

Then

$$\mathbb{E}\left[\max\left\{Z_1, \ldots, Z_n\right\}\right] \leq \sqrt{2v \ln(n)}.$$

▶ Remark 14. The variables $Z_1, \ldots, Z_n$ in the lemma might not be independent.

Proof. For any $\lambda > 0$,

$$\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\max\left\{Z_1, \ldots, Z_n\right\}\right]\right) &= \exp\left(\mathbb{E}\left[\lambda \max_{i \in [n]} Z_i\right]\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda \max_{i \in [n]} Z_i\right)\right] \qquad \text{(Jensen's inequality)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\max_{i\in[n]} e^{\lambda Z_i}\right] \\
&\leq \mathbb{E}\left[\sum_{i\in[n]} e^{\lambda Z_i}\right] \\
&= \sum_{i\in[n]} \mathbb{E}\left[e^{\lambda Z_i}\right] \\
&\leq \sum_{i\in[n]} e^{\frac{\lambda^2 v}{2}} = n e^{\frac{\lambda^2 v}{2}}.
\end{aligned}
$$

Taking logarithms on both sides yields

$$\mathbb{E}\left[\max\left\{Z_1,\ldots,Z_n\right\}\right] \leq \frac{\ln(n)}{\lambda} + \frac{\lambda v}{2}.$$

Choosing $\lambda = \sqrt{\frac{2\ln(n)}{v}}$, which minimizes the right hand side, we obtain

$$\mathbb{E}\left[\max\left\{Z_1,\ldots,Z_n\right\}\right] \leq \sqrt{2v\ln(n)}.$$ ◀

▶ **Lemma 15** (Finite Class, Massart [9]). *Let $A \subseteq \mathbb{R}^m$ be a finite subset, and assume there exists $r \in \mathbb{R}$ such that $\forall x \in A: \ \|x\|_2 \leq r$. Then*

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

Proof. Write

$$\mathsf{Rad}(A) = \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^m}\left[\sup_{a\in A}\frac{\sigma\cdot a}{m}\right] = \frac{1}{m}\mathbb{E}\left[\max_{a\in A} Z_a\right], \tag{5}$$

where $Z_a = \sum_{i=1}^m \sigma_i a_i$, and we used the fact that A is finite. By Hoeffding's Lemma (Lemma 13 in Unit 3),

$$\psi_{\sigma_i a_i}(\lambda) \leq \frac{\lambda^2(2a_i)^2}{8} = \frac{\lambda^2 a_i^2}{2}.$$

Hence,

$$\psi_{Z_a}(\lambda) = \psi_{\left(\sum_{i=1}^m \sigma_i a_i\right)}(\lambda) = \sum_{i=1}^m \psi_{\sigma_i a_i}(\lambda) \leq \sum_{i=1}^m \frac{\lambda^2 a_i^2}{2} \leq \frac{\lambda^2 r^2}{2}.$$

Lemma 13 implies that

$$\mathbb{E}\left[\max_{a\in A} Z_a\right] \leq r\sqrt{2\ln(|A|)}. \tag{6}$$

Combining Eq. (5) and (6) implies the lemma. ◀

5.2.2 Learning Bounds for VC Classes from Rademacher Complexity

As a corollary from Lemma 15, the Rademacher complexity is bounded by the VC dimension.

▶ **Theorem 16.** *Let $\mathcal{X}$ be a nonempty set, and let $\mathcal{F}$ be a set of functions $\mathcal{X} \to \{0,1\}$, and let $\mathcal{D}$ be a distribution over $\mathcal{X}$. Then for all $m \in \mathbb{N}$,*

$$\mathsf{Rad}_{\mathcal{D},m}(\mathcal{F}) \leq \sup_{S \in \mathcal{X}^m} \mathsf{Rad}_S(\mathcal{F}) \leq \sqrt{\frac{2\ln\left(\tau_{\mathcal{F}}(m)\right)}{m}} \leq O\left(\sqrt{\frac{\ln\left(m/d\right)}{\left(m/d\right)}}\right),$$

where the last inequality holds if $\mathsf{VC}(\mathcal{F}) = d \leq \infty$.

Proof. The first inequality is immediate from the definition of $\mathsf{Rad}_{\mathcal{D},m}(\mathcal{F})$. For the second inequality, for any $S \in \mathcal{X}^m$,

$$\mathsf{Rad}_S(\mathcal{F}) \stackrel{\text{def}}{=} \mathsf{Rad}(\mathcal{F}(S)) \stackrel{\text{(Massart)}}{\leq} \left(\max_{f \in \mathcal{F}} \|f(S)\|_2\right) \cdot \frac{\sqrt{2\ln\left(|\mathcal{F}(S)|\right)}}{m} \leq \sqrt{\frac{2\ln\left(\tau_{\mathcal{F}}(m)\right)}{m}}.$$

We used the fact that

$$\max_{f \in \mathcal{F}} \|f(S)\|_2 \leq \|(1,\ldots,1)\|_2 = \sqrt{m}.$$

The final inequality in the statement follows from Sauer's Lemma, which states that $\tau_{\mathcal{F}}(m)$ is at most $\left(em/d\right)^d$, and so $\sqrt{\frac{2\ln\left(\tau_{\mathcal{F}}(m)\right)}{m}} \leq \sqrt{\frac{2(\ln(m/d)+1)}{(m/d)}}$. ◀

6 Discussion

Combining Theorems 16 and 12 (iii) yields the following learning bound for binary classification using an ERM algorithm with respect to the 0-1 loss:

$$\begin{aligned} L_{\mathcal{D}}(h) &\leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + 2\mathsf{Rad}_{\mathcal{D}_x,m}(\mathcal{H}) + \sqrt{\frac{2\ln(2/\delta)}{m}} \\ &\leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + O\left(\sqrt{\frac{\ln\left(m/d\right)}{\left(m/d\right)}}\right) + \sqrt{\frac{2\ln(2/\delta)}{m}} \\ &\leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + O\left(\sqrt{\frac{d\ln\left(m/d\right) + \ln(1/\delta)}{m}}\right). \end{aligned}$$

In particular this implies (via a direct calculation[2]) that taking

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

samples is sufficient for agnostic PAC learning a class of binary functions of VC dimension d with accuracy ε and confidence $1-\delta$. However, the fundamental theorem states a stronger bound of $O\left(\frac{d+\ln(1/\delta)}{\varepsilon^2}\right)$. In the next unit we will use connections between the Rademacher complexity and covering numbers to obtain that stronger bound.

---

[2] Sketch: clearly, this suffices to ensure that $\sqrt{\frac{2\ln(2/\delta)}{m}} \leq \frac{\varepsilon}{2}$. Furthermore, $\sqrt{\frac{\ln(x)}{(x)}} \leq \varepsilon \iff \frac{x}{\ln(x)} \geq \frac{1}{\varepsilon^2}$. Taking $x = \frac{3}{\varepsilon^2}\ln(\frac{1}{\varepsilon^2})$ implies $\frac{x}{\ln(x)} = \frac{1}{\varepsilon^2} \cdot \frac{\ln(\frac{1}{\varepsilon^2}\cdot\frac{1}{\varepsilon^2}\cdot\frac{1}{\varepsilon^2})}{\ln(3\cdot\frac{1}{\varepsilon^2}\cdot\ln(\frac{1}{\varepsilon^2}))} > \frac{1}{\varepsilon^2}$ for all $\varepsilon \leq \frac{1}{\sqrt{3}}$. Therefore, taking $\frac{m}{d} \geq O(\frac{1}{\varepsilon^2}\ln(\frac{1}{\varepsilon^2}))$ suffices to ensure that $O\left(\sqrt{\frac{\ln(m/d)}{(m/d)}}\right) \leq \frac{\varepsilon}{2}$.

6.1 Bibliographic Notes

The analysis of uniform convergence via Rademacher complexity was introduced by [6, 5, 1]. See also [2, 7]. Massart's Lemma is due to [9] (see also exposition in [3]).

Good expositions of the Rademacher complexity analysis are available in [10, Chapter 3], as well as [4, Chapters 6], [8, Section 3.8] and [11, Section 1.8].

References

1 Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Mach. Learn.*, 48(1-3):85–113, 2002. doi:10.1023/A:1013999503812.

2 Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002. URL: http://jmlr.org/papers/v3/bartlett02a.html.

3 Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities – A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.

4 Bruce Hajek and Maxim Raginsky. *ECE 543: Statistical Learning Theory*. University of Illinois at Urbana-Champaign, 2018. URL: https://web.archive.org/web/20210213043003/http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf.

5 Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, 47(5):1902–1914, 2001. doi:10.1109/18.930926.

6 Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

7 Vladimir Koltchinskii, Dmitry Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

8 Percy Liang. *CS229T/STAT231: Statistical Learning Theory*. Stanford University, 2016. URL: https://web.archive.org/web/20210114222356/https://web.stanford.edu/class/cs229t/notes.pdf.

9 Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

10 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

11 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.