

## 26 Rademacher Complexities

---

In Chapter 4 we have shown that uniform convergence is a sufficient condition for learnability. In this chapter we study the Rademacher complexity, which measures the rate of uniform convergence. We will provide generalization bounds based on this measure.

### 26.1 The Rademacher Complexity

Recall the definition of an  $\epsilon$ -representative sample from Chapter 4, repeated here for convenience.

**DEFINITION 26.1** ( $\epsilon$ -Representative Sample) A training set  $S$  is called  $\epsilon$ -representative (w.r.t. domain  $Z$ , hypothesis class  $\mathcal{H}$ , loss function  $\ell$ , and distribution  $\mathcal{D}$ ) if

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon.$$

We have shown that if  $S$  is an  $\epsilon/2$  representative sample then the ERM rule is  $\epsilon$ -consistent, namely,  $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ .

To simplify our notation, let us denote

$$\mathcal{F} \stackrel{\text{def}}{=} \ell \circ \mathcal{H} \stackrel{\text{def}}{=} \{z \mapsto \ell(h, z) : h \in \mathcal{H}\},$$

and given  $f \in \mathcal{F}$ , we define

$$L_{\mathcal{D}}(f) = \mathbb{E}_{z \sim \mathcal{D}}[f(z)] \quad , \quad L_S(f) = \frac{1}{m} \sum_{i=1}^m f(z_i).$$

We define the *representativeness* of  $S$  with respect to  $\mathcal{F}$  as the largest gap between the true error of a function  $f$  and its empirical error, namely,

$$\text{Rep}_{\mathcal{D}}(\mathcal{F}, S) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)). \quad (26.1)$$

Now, suppose we would like to estimate the representativeness of  $S$  using the sample  $S$  only. One simple idea is to split  $S$  into two disjoint sets,  $S = S_1 \cup S_2$ ; refer to  $S_1$  as a validation set and to  $S_2$  as a training set. We can then estimate the representativeness of  $S$  by

$$\sup_{f \in \mathcal{F}} (L_{S_1}(f) - L_{S_2}(f)). \quad (26.2)$$

This can be written more compactly by defining  $\sigma = (\sigma_1, \dots, \sigma_m) \in \{\pm 1\}^m$  to be a vector such that  $S_1 = \{z_i : \sigma_i = 1\}$  and  $S_2 = \{z_i : \sigma_i = -1\}$ . Then, if we further assume that  $|S_1| = |S_2|$  then Equation (26.2) can be rewritten as

$$\frac{2}{m} \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i). \quad (26.3)$$

The Rademacher complexity measure captures this idea by considering the expectation of the above with respect to a random choice of  $\sigma$ . Formally, let  $\mathcal{F} \circ S$  be the set of all possible evaluations a function  $f \in \mathcal{F}$  can achieve on a sample  $S$ , namely,

$$\mathcal{F} \circ S = \{(f(z_1), \dots, f(z_m)) : f \in \mathcal{F}\}.$$

Let the variables in  $\sigma$  be distributed i.i.d. according to  $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$ . Then, the Rademacher complexity of  $\mathcal{F}$  with respect to  $S$  is defined as follows:

$$R(\mathcal{F} \circ S) \stackrel{\text{def}}{=} \frac{1}{m} \mathbb{E}_{\sigma \sim \{\pm 1\}^m} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]. \quad (26.4)$$

More generally, given a set of vectors,  $A \subset \mathbb{R}^m$ , we define

$$R(A) \stackrel{\text{def}}{=} \frac{1}{m} \mathbb{E}_{\sigma} \left[ \sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right]. \quad (26.5)$$

The following lemma bounds the expected value of the representativeness of  $S$  by twice the expected Rademacher complexity.

LEMMA 26.2

$$\mathbb{E}_{S \sim \mathcal{D}^m} [\text{Rep}_{\mathcal{D}}(\mathcal{F}, S)] \leq 2 \mathbb{E}_{S \sim \mathcal{D}^m} R(\mathcal{F} \circ S).$$

*Proof* Let  $S' = \{z'_1, \dots, z'_m\}$  be another i.i.d. sample. Clearly, for all  $f \in \mathcal{F}$ ,  $L_{\mathcal{D}}(f) = \mathbb{E}_{S'}[L_{S'}(f)]$ . Therefore, for every  $f \in \mathcal{F}$  we have

$$L_{\mathcal{D}}(f) - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f)] - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f) - L_S(f)].$$

Taking supremum over  $f \in \mathcal{F}$  of both sides, and using the fact that the supremum of expectation is smaller than expectation of the supremum we obtain

$$\begin{aligned} \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) &= \sup_{f \in \mathcal{F}} \mathbb{E}_{S'}[L_{S'}(f) - L_S(f)] \\ &\leq \mathbb{E}_{S'} \left[ \sup_{f \in \mathcal{F}} (L_{S'}(f) - L_S(f)) \right]. \end{aligned}$$

Taking expectation over  $S$  on both sides we obtain

$$\begin{aligned} \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) \right] &\leq \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} (L_{S'}(f) - L_S(f)) \right] \\ &= \frac{1}{m} \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m (f(z'_i) - f(z_i)) \right]. \end{aligned} \quad (26.6)$$

Next, we note that for each  $j$ ,  $z_j$  and  $z'_j$  are i.i.d. variables. Therefore, we can replace them without affecting the expectation:

$$\begin{aligned} \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left( (f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i)) \right) \right] &= \\ \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left( (f(z_j) - f(z'_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i)) \right) \right]. \end{aligned} \quad (26.7)$$

Let  $\sigma_j$  be a random variable such that  $\mathbb{P}[\sigma_j = 1] = \mathbb{P}[\sigma_j = -1] = 1/2$ . From Equation (26.7) we obtain that

$$\begin{aligned} & \mathbb{E}_{S, S', \sigma_j} \left[ \sup_{f \in \mathcal{F}} \left( \sigma_j (f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i)) \right) \right] \\ &= \frac{1}{2} (\text{l.h.s. of Equation (26.7)}) + \frac{1}{2} (\text{r.h.s. of Equation (26.7)}) \\ &= \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left( (f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i)) \right) \right]. \end{aligned} \quad (26.8)$$

Repeating this for all  $j$  we obtain that

$$\mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m (f(z'_i) - f(z_i)) \right] = \mathbb{E}_{S, S', \sigma} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i (f(z'_i) - f(z_i)) \right]. \quad (26.9)$$

Finally,

$$\sup_{f \in \mathcal{F}} \sum_i \sigma_i (f(z'_i) - f(z_i)) \leq \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z'_i) + \sup_{f \in \mathcal{F}} \sum_i -\sigma_i f(z_i)$$

and since the probability of  $\sigma$  is the same as the probability of  $-\sigma$ , the right-hand side of Equation (26.9) can be bounded by

$$\begin{aligned} & \mathbb{E}_{S, S', \sigma} \left[ \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z'_i) + \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) \right] \\ &= m \mathbb{E}_{S'} [R(\mathcal{F} \circ S')] + m \mathbb{E}_S [R(\mathcal{F} \circ S)] = 2m \mathbb{E}_S [R(\mathcal{F} \circ S)]. \end{aligned}$$

□

The lemma immediately yields that, in expectation, the ERM rule finds a hypothesis which is close to the optimal hypothesis in  $\mathcal{H}$ .

**THEOREM 26.3** *We have*

$$\mathbb{E}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}}(S)) - L_S(\text{ERM}_{\mathcal{H}}(S))] \leq 2 \mathbb{E}_{S \sim \mathcal{D}^m} R(\ell \circ \mathcal{H} \circ S).$$

Furthermore, for any  $h^* \in \mathcal{H}$

$$\mathbb{E}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*)] \leq 2 \mathbb{E}_{S \sim \mathcal{D}^m} R(\ell \circ \mathcal{H} \circ S).$$

Furthermore, if  $h^* = \operatorname{argmin}_h L_{\mathcal{D}}(h)$  then for each  $\delta \in (0, 1)$  with probability of at least  $1 - \delta$  over the choice of  $S$  we have

$$L_{\mathcal{D}}(\operatorname{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*) \leq \frac{2 \mathbb{E}_{S' \sim \mathcal{D}^m} R(\ell \circ \mathcal{H} \circ S')}{\delta}.$$

*Proof* The first inequality follows directly from Lemma 26.2. The second inequality follows because for any fixed  $h^*$ ,

$$L_{\mathcal{D}}(h^*) = \mathbb{E}_S[L_S(h^*)] \geq \mathbb{E}_S[L_S(\operatorname{ERM}_{\mathcal{H}}(S))].$$

The third inequality follows from the previous inequality by relying on Markov's inequality (note that the random variable  $L_{\mathcal{D}}(\operatorname{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*)$  is nonnegative).  $\square$

Next, we derive bounds similar to the bounds in Theorem 26.3 with a better dependence on the confidence parameter  $\delta$ . To do so, we first introduce the following bounded differences concentration inequality.

**LEMMA 26.4** (McDiarmid's Inequality) *Let  $V$  be some set and let  $f : V^m \rightarrow \mathbb{R}$  be a function of  $m$  variables such that for some  $c > 0$ , for all  $i \in [m]$  and for all  $x_1, \dots, x_m, x'_i \in V$  we have*

$$|f(x_1, \dots, x_m) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c.$$

*Let  $X_1, \dots, X_m$  be  $m$  independent random variables taking values in  $V$ . Then, with probability of at least  $1 - \delta$  we have*

$$|f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots, X_m)]| \leq c \sqrt{\ln\left(\frac{2}{\delta}\right) m/2}.$$

On the basis of the McDiarmid inequality we can derive generalization bounds with a better dependence on the confidence parameter.

**THEOREM 26.5** *Assume that for all  $z$  and  $h \in \mathcal{H}$  we have that  $|\ell(h, z)| \leq c$ . Then,*

1. *With probability of at least  $1 - \delta$ , for all  $h \in \mathcal{H}$ ,*

$$L_{\mathcal{D}}(h) - L_S(h) \leq 2 \mathbb{E}_{S' \sim \mathcal{D}^m} R(\ell \circ \mathcal{H} \circ S') + c \sqrt{\frac{2 \ln(2/\delta)}{m}}.$$

*In particular, this holds for  $h = \operatorname{ERM}_{\mathcal{H}}(S)$ .*

2. *With probability of at least  $1 - \delta$ , for all  $h \in \mathcal{H}$ ,*

$$L_{\mathcal{D}}(h) - L_S(h) \leq 2 R(\ell \circ \mathcal{H} \circ S) + 4c \sqrt{\frac{2 \ln(4/\delta)}{m}}.$$

*In particular, this holds for  $h = \operatorname{ERM}_{\mathcal{H}}(S)$ .*

3. *For any  $h^*$ , with probability of at least  $1 - \delta$ ,*

$$L_{\mathcal{D}}(\operatorname{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*) \leq 2 R(\ell \circ \mathcal{H} \circ S) + 5c \sqrt{\frac{2 \ln(8/\delta)}{m}}.$$

*Proof* First note that the random variable  $\text{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$  satisfies the bounded differences condition of Lemma 26.4 with a constant  $2c/m$ . Combining the bounds in Lemma 26.4 with Lemma 26.2 we obtain that with probability of at least  $1 - \delta$ ,

$$\text{Rep}_{\mathcal{D}}(\mathcal{F}, S) \leq \mathbb{E} \text{Rep}_{\mathcal{D}}(\mathcal{F}, S) + c \sqrt{\frac{2 \ln(2/\delta)}{m}} \leq 2 \mathbb{E}_{S'} R(\ell \circ \mathcal{H} \circ S') + c \sqrt{\frac{2 \ln(2/\delta)}{m}}.$$

The first inequality of the theorem follows from the definition of  $\text{Rep}_{\mathcal{D}}(\mathcal{F}, S)$ . For the second inequality we note that the random variable  $R(\ell \circ \mathcal{H} \circ S)$  also satisfies the bounded differences condition of Lemma 26.4 with a constant  $2c/m$ . Therefore, the second inequality follows from the first inequality, Lemma 26.4, and the union bound. Finally, for the last inequality, denote  $h_S = \text{ERM}_{\mathcal{H}}(S)$  and note that

$$\begin{aligned} L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) &= L_{\mathcal{D}}(h_S) - L_S(h_S) + L_S(h_S) - L_S(h^*) + L_S(h^*) - L_{\mathcal{D}}(h^*) \\ &\leq (L_{\mathcal{D}}(h_S) - L_S(h_S)) + (L_S(h^*) - L_{\mathcal{D}}(h^*)). \end{aligned} \quad (26.10)$$

The first summand on the right-hand side is bounded by the second inequality of the theorem. For the second summand, we use the fact that  $h^*$  does not depend on  $S$ ; hence by using Hoeffding's inequality we obtain that with probability of at least  $1 - \delta/2$ ,

$$L_S(h^*) - L_{\mathcal{D}}(h^*) \leq c \sqrt{\frac{\ln(4/\delta)}{2m}}. \quad (26.11)$$

Combining this with the union bound we conclude our proof.  $\square$

The preceding theorem tells us that if the quantity  $R(\ell \circ \mathcal{H} \circ S)$  is small then it is possible to learn the class  $\mathcal{H}$  using the ERM rule. It is important to emphasize that the last two bounds given in the theorem depend on the specific training set  $S$ . That is, we use  $S$  both for learning a hypothesis from  $\mathcal{H}$  as well as for estimating the quality of it. This type of bound is called a *data-dependent bound*.

### 26.1.1 Rademacher Calculus

Let us now discuss some properties of the Rademacher complexity measure. These properties will help us in deriving some simple bounds on  $R(\ell \circ \mathcal{H} \circ S)$  for specific cases of interest.

The following lemma is immediate from the definition.

**LEMMA 26.6** For any  $A \subset \mathbb{R}^m$ , scalar  $c \in \mathbb{R}$ , and vector  $\mathbf{a}_0 \in \mathbb{R}^m$ , we have

$$R(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A\}) \leq |c| R(A).$$

The following lemma tells us that the convex hull of  $A$  has the same complexity as  $A$ .

LEMMA 26.7 Let  $A$  be a subset of  $\mathbb{R}^m$  and let  $A' = \{\sum_{j=1}^N \alpha_j \mathbf{a}^{(j)} : N \in \mathbb{N}, \forall j, \mathbf{a}^{(j)} \in A, \alpha_j \geq 0, \|\boldsymbol{\alpha}\|_1 = 1\}$ . Then,  $R(A') = R(A)$ .

*Proof* The main idea follows from the fact that for any vector  $\mathbf{v}$  we have

$$\sup_{\boldsymbol{\alpha} \geq \mathbf{0}: \|\boldsymbol{\alpha}\|_1 = 1} \sum_{j=1}^N \alpha_j v_j = \max_j v_j.$$

Therefore,

$$\begin{aligned} m R(A') &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha} \geq \mathbf{0}: \|\boldsymbol{\alpha}\|_1 = 1} \sup_{\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(N)}} \sum_{i=1}^m \sigma_i \sum_{j=1}^N \alpha_j a_i^{(j)} \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha} \geq \mathbf{0}: \|\boldsymbol{\alpha}\|_1 = 1} \sum_{j=1}^N \alpha_j \sup_{\mathbf{a}^{(j)}} \sum_{i=1}^m \sigma_i a_i^{(j)} \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{a} \in A} \sum_{i=1}^m \sigma_i a_i \\ &= m R(A), \end{aligned}$$

and we conclude our proof.  $\square$

The next lemma, due to Massart, states that the Rademacher complexity of a finite set grows logarithmically with the size of the set.

LEMMA 26.8 (Massart lemma) Let  $A = \{\mathbf{a}_1, \dots, \mathbf{a}_N\}$  be a finite set of vectors in  $\mathbb{R}^m$ . Define  $\bar{\mathbf{a}} = \frac{1}{N} \sum_{i=1}^N \mathbf{a}_i$ . Then,

$$R(A) \leq \max_{\mathbf{a} \in A} \|\mathbf{a} - \bar{\mathbf{a}}\| \frac{\sqrt{2 \log(N)}}{m}.$$

*Proof* Based on Lemma 26.6, we can assume without loss of generality that  $\bar{\mathbf{a}} = \mathbf{0}$ . Let  $\lambda > 0$  and let  $A' = \{\lambda \mathbf{a}_1, \dots, \lambda \mathbf{a}_N\}$ . We upper bound the Rademacher complexity as follows:

$$\begin{aligned} m R(A') &= \mathbb{E}_{\boldsymbol{\sigma}} \left[ \max_{\mathbf{a} \in A'} \langle \boldsymbol{\sigma}, \mathbf{a} \rangle \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[ \log \left( \max_{\mathbf{a} \in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a} \rangle} \right) \right] \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}} \left[ \log \left( \sum_{\mathbf{a} \in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a} \rangle} \right) \right] \\ &\leq \log \left( \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sum_{\mathbf{a} \in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a} \rangle} \right] \right) \quad // \text{ Jensen's inequality} \\ &= \log \left( \sum_{\mathbf{a} \in A'} \prod_{i=1}^m \mathbb{E}_{\sigma_i} [e^{\sigma_i a_i}] \right), \end{aligned}$$

where the last equality occurs because the Rademacher variables are independent. Next, using Lemma A.6 we have that for all  $a_i \in \mathbb{R}$ ,

$$\mathbb{E}_{\sigma_i} e^{\sigma_i a_i} = \frac{\exp(a_i) + \exp(-a_i)}{2} \leq \exp(a_i^2/2),$$

and therefore

$$\begin{aligned} mR(A') &\leq \log \left( \sum_{\mathbf{a} \in A'} \prod_{i=1}^m \exp \left( \frac{a_i^2}{2} \right) \right) = \log \left( \sum_{\mathbf{a} \in A'} \exp (\|\mathbf{a}\|^2/2) \right) \\ &\leq \log \left( |A'| \max_{\mathbf{a} \in A'} \exp (\|\mathbf{a}\|^2/2) \right) = \log(|A'|) + \max_{\mathbf{a} \in A'} (\|\mathbf{a}\|^2/2). \end{aligned}$$

Since  $R(A) = \frac{1}{\lambda} R(A')$  we obtain from the equation that

$$R(A) \leq \frac{\log(|A|) + \lambda^2 \max_{\mathbf{a} \in A} (\|\mathbf{a}\|^2/2)}{\lambda m}.$$

Setting  $\lambda = \sqrt{2 \log(|A|) / \max_{\mathbf{a} \in A} \|\mathbf{a}\|^2}$  and rearranging terms we conclude our proof.  $\square$

The following lemma shows that composing  $A$  with a Lipschitz function does not blow up the Rademacher complexity. The proof is due to Kakade and Tewari.

**LEMMA 26.9 (Contraction lemma)** *For each  $i \in [m]$ , let  $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$  be a  $\rho$ -Lipschitz function, namely for all  $\alpha, \beta \in \mathbb{R}$  we have  $|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho |\alpha - \beta|$ . For  $\mathbf{a} \in \mathbb{R}^m$  let  $\phi(\mathbf{a})$  denote the vector  $(\phi_1(a_1), \dots, \phi_m(a_m))$ . Let  $\phi \circ A = \{\phi(\mathbf{a}) : \mathbf{a} \in A\}$ . Then,*

$$R(\phi \circ A) \leq \rho R(A).$$

*Proof* For simplicity, we prove the lemma for the case  $\rho = 1$ . The case  $\rho \neq 1$  will follow by defining  $\phi' = \frac{1}{\rho} \phi$  and then using Lemma 26.6. Let  $A_i = \{(a_1, \dots, a_{i-1}, \phi_i(a_i), a_{i+1}, \dots, a_m) : \mathbf{a} \in A\}$ . Clearly, it suffices to prove that for any set  $A$  and all  $i$  we have  $R(A_i) \leq R(A)$ . Without loss of generality we will prove the latter claim for  $i = 1$  and to simplify notation we omit the subscript from  $\phi_1$ . We have

$$\begin{aligned} mR(A_1) &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{a} \in A_1} \sum_{i=1}^m \sigma_i a_i \right] \\ &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{a} \in A} \sigma_1 \phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right] \\ &= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{\mathbf{a} \in A} \left( \phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) + \sup_{\mathbf{a} \in A} \left( -\phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) \right] \\ &= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{\mathbf{a}, \mathbf{a}' \in A} \left( \phi(a_1) - \phi(a'_1) + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right] \\ &\leq \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{\mathbf{a}, \mathbf{a}' \in A} \left( |a_1 - a'_1| + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right], \quad (26.12) \end{aligned}$$

where in the last inequality we used the assumption that  $\phi$  is Lipschitz. Next, we note that the absolute value on  $|a_1 - a'_1|$  in the preceding expression can

be omitted since both  $\mathbf{a}$  and  $\mathbf{a}'$  are from the same set  $A$  and the rest of the expression in the supremum is not affected by replacing  $\mathbf{a}$  and  $\mathbf{a}'$ . Therefore,

$$mR(A_1) \leq \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{\mathbf{a}, \mathbf{a}' \in A} \left( a_1 - a'_1 + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right]. \quad (26.13)$$

But, using the same equalities as in Equation (26.12), it is easy to see that the right-hand side of Equation (26.13) exactly equals  $mR(A)$ , which concludes our proof.  $\square$

## 26.2 Rademacher Complexity of Linear Classes

In this section we analyze the Rademacher complexity of linear classes. To simplify the derivation we first define the following two classes:

$$\mathcal{H}_1 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_1 \leq 1\}, \quad \mathcal{H}_2 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \leq 1\}. \quad (26.14)$$

The following lemma bounds the Rademacher complexity of  $\mathcal{H}_2$ . We allow the  $\mathbf{x}_i$  to be vectors in any Hilbert space (even infinite dimensional), and the bound does not depend on the dimensionality of the Hilbert space. This property becomes useful when analyzing kernel methods.

**LEMMA 26.10** *Let  $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$  be vectors in a Hilbert space. Define:  $\mathcal{H}_2 \circ S = \{(\langle \mathbf{w}, \mathbf{x}_1 \rangle, \dots, \langle \mathbf{w}, \mathbf{x}_m \rangle) : \|\mathbf{w}\|_2 \leq 1\}$ . Then,*

$$R(\mathcal{H}_2 \circ S) \leq \frac{\max_i \|\mathbf{x}_i\|_2}{\sqrt{m}}.$$

*Proof* Using Cauchy-Schwartz inequality we know that for any vectors  $\mathbf{w}, \mathbf{v}$  we have  $\langle \mathbf{w}, \mathbf{v} \rangle \leq \|\mathbf{w}\| \|\mathbf{v}\|$ . Therefore,

$$\begin{aligned} mR(\mathcal{H}_2 \circ S) &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{a} \in \mathcal{H}_2 \circ S} \sum_{i=1}^m \sigma_i a_i \right] \\ &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \sum_{i=1}^m \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right] \\ &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i \mathbf{x}_i \rangle \right] \\ &\leq \mathbb{E}_{\sigma} \left[ \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2 \right]. \end{aligned} \quad (26.15)$$

Next, using Jensen's inequality we have that

$$\mathbb{E}_{\sigma} \left[ \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2 \right] = \mathbb{E}_{\sigma} \left[ \left( \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2^2 \right)^{1/2} \right] \leq \left( \mathbb{E}_{\sigma} \left[ \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2^2 \right] \right)^{1/2} \quad (26.16)$$



Finally, since the variables  $\sigma_1, \dots, \sigma_m$  are independent we have

$$\begin{aligned} \mathbb{E}_{\sigma} \left[ \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2^2 \right] &= \mathbb{E}_{\sigma} \left[ \sum_{i,j} \sigma_i \sigma_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right] \\ &= \sum_{i \neq j} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \mathbb{E}_{\sigma} [\sigma_i \sigma_j] + \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{x}_i \rangle \mathbb{E}_{\sigma} [\sigma_i^2] \\ &= \sum_{i=1}^m \|\mathbf{x}_i\|_2^2 \leq m \max_i \|\mathbf{x}_i\|_2^2. \end{aligned}$$

Combining this with Equation (26.15) and Equation (26.16) we conclude our proof.  $\square$

Next we bound the Rademacher complexity of  $\mathcal{H}_1 \circ S$ .

LEMMA 26.11 *Let  $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$  be vectors in  $\mathbb{R}^n$ . Then,*

$$R(\mathcal{H}_1 \circ S) \leq \max_i \|\mathbf{x}_i\|_{\infty} \sqrt{\frac{2 \log(2n)}{m}}.$$

*Proof* Using Holder's inequality we know that for any vectors  $\mathbf{w}, \mathbf{v}$  we have  $\langle \mathbf{w}, \mathbf{v} \rangle \leq \|\mathbf{w}\|_1 \|\mathbf{v}\|_{\infty}$ . Therefore,

$$\begin{aligned} mR(\mathcal{H}_1 \circ S) &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{a} \in H_1 \circ S} \sum_{i=1}^m \sigma_i a_i \right] \\ &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq 1} \sum_{i=1}^m \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right] \\ &= \mathbb{E}_{\sigma} \left[ \sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq 1} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i \mathbf{x}_i \rangle \right] \\ &\leq \mathbb{E}_{\sigma} \left[ \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_{\infty} \right]. \end{aligned} \tag{26.17}$$

For each  $j \in [n]$ , let  $\mathbf{v}_j = (x_{1,j}, \dots, x_{m,j}) \in \mathbb{R}^m$ . Note that  $\|\mathbf{v}_j\|_2 \leq \sqrt{m} \max_i \|\mathbf{x}_i\|_{\infty}$ . Let  $V = \{\mathbf{v}_1, \dots, \mathbf{v}_n, -\mathbf{v}_1, \dots, -\mathbf{v}_n\}$ . The right-hand side of Equation (26.17) is  $mR(V)$ . Using Massart lemma (Lemma 26.8) we have that

$$R(V) \leq \max_i \|\mathbf{x}_i\|_{\infty} \sqrt{2 \log(2n)/m},$$

which concludes our proof.  $\square$

## 26.3 Generalization Bounds for SVM

In this section we use Rademacher complexity to derive generalization bounds for generalized linear predictors with Euclidean norm constraint. We will show how this leads to generalization bounds for hard-SVM and soft-SVM.