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 ϵ-representative sample from Chapter 4, repeated here for convenience.

DEFINITION 26.1 (ϵ-Representative Sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, 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 ϵ-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

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

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,

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

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

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

This can be written more compactly by defining $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m) \in \{\pm 1\}^m$ 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

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

The Rademacher complexity measure captures this idea by considering the expectation of the above with respect to a random choice of $\boldsymbol{\sigma}$. 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,

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

Let the variables in $\boldsymbol{\sigma}$ 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:

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

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

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

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

LEMMA 26.2

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

Proof Let $S' = \{z'_1, \ldots, z'_m\}$ 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

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

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

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

Taking expectation over S on both sides we obtain

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

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:

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

Let σ_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

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

Repeating this for all j we obtain that

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

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

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

□

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*

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

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

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

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

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

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

$$L_{\mathcal{D}}(h^\star) = \mathop{\mathbb{E}}_{S}[L_S(h^\star)] \geq \mathop{\mathbb{E}}_{S}[L_S(\mathrm{ERM}_{\mathcal{H}}(S))].$$

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

Next, we derive bounds similar to the bounds in Theorem 26.3 with a better dependence on the confidence parameter δ. 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* $f : V^m \to \mathbb{R}$ *be a function of* m *variables such that for some* $c > 0$*, for all* $i \in [m]$ *and for all* $x_1, \ldots, x_m, x_i' \in V$ *we have*

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

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

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

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* $h \in \mathcal{H}$*,*

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

   In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$*.*
2. *With probability of at least* $1-\delta$*, for all* $h \in \mathcal{H}$*,*

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

   In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$*.*
3. *For any* $h^\star$*, with probability of at least* $1-\delta$*,*

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

Proof First note that the random variable $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$ 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 $1 - \delta$,

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

The first inequality of the theorem follows from the definition of $\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)$. 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 $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$ and note that

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

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 $h^\star$ does not depend on S; hence by using Hoeffding's inequality we obtain that with probaility of at least $1 - \delta/2$,

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

Combining this with the union bound we conclude our proof. □

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,

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

and we conclude our proof. □

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 $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ 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 $A' = \{\lambda \mathbf{a}_1, \ldots, \lambda \mathbf{a}_N\}$. We upper bound the Rademacher complexity as follows:

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

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

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

and therefore

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

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

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

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

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 $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function, namely for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \le \rho\,|\alpha - \beta|$. For $\mathbf{a} \in \mathbb{R}^m$ let $\boldsymbol{\phi}(\mathbf{a})$ denote the vector $(\phi_1(a_1), \ldots, \phi_m(y_m))$. Let $\boldsymbol{\phi} \circ A = \{\boldsymbol{\phi}(\mathbf{a}) : a \in A\}$. Then,*

$$R(\boldsymbol{\phi} \circ A) \le \rho\, R(A).$$

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 $A_i = \{(a_1, \ldots, a_{i-1}, \phi_i(a_i), a_{i+1}, \ldots, a_m) : \mathbf{a} \in A\}$. Clearly, it suffices to prove that for any set A and all i we have $R(A_i) \le R(A)$. Without loss of generality we will prove the latter claim for $i = 1$ and to simplify notation we omit the subscript from ϕ_1. We have

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

where in the last inequality we used the assumption that ϕ 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,

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

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 $m\,R(A)$, which concludes our proof. □

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:

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

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

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

Next, using Jensen's inequality we have that

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

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

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

Combining this with Equation (26.15) and Equation (26.16) we conclude our proof. □

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

LEMMA 26.11 *Let $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ 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,

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

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

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.

We shall consider the following general constraint-based formulation. Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq B\}$ be our hypothesis class, and let $Z = \mathcal{X} \times \mathcal{Y}$ be the examples domain. Assume that the loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}$ is of the form

$$\ell(\mathbf{w}, (\mathbf{x}, y)) = \phi(\langle \mathbf{w}, \mathbf{x} \rangle, y), \tag{26.18}$$

where $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is such that for all $y \in \mathcal{Y}$, the scalar function $a \mapsto \phi(a, y)$ is ρ-Lipschitz. For example, the hinge-loss function, $\ell(\mathbf{w}, (\mathbf{x}, y)) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\}$, can be written as in Equation (26.18) using $\phi(a, y) = \max\{0, 1 - ya\}$, and note that ϕ is 1-Lipschitz for all $y \in \{\pm 1\}$. Another example is the absolute loss function, $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\langle \mathbf{w}, \mathbf{x} \rangle - y|$, which can be written as in Equation (26.18) using $\phi(a, y) = |a - y|$, which is also 1-Lipschitz for all $y \in \mathbb{R}$.

The following theorem bounds the generalization error of all predictors in $\mathcal{H}$ using their empirical error.

THEOREM 26.12 *Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that with probability 1 we have that $\|\mathbf{x}\|_2 \leq R$. Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq B\}$ and let $\ell : \mathcal{H} \times Z \to \mathbb{R}$ be a loss function of the form given in Equation (26.18) such that for all $y \in \mathcal{Y}$, $a \mapsto \phi(a, y)$ is a ρ-Lipschitz function and such that $\max_{a \in [-BR, BR]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample of size m,*

$$\forall \mathbf{w} \in \mathcal{H}, \quad L_{\mathcal{D}}(\mathbf{w}) \leq L_S(\mathbf{w}) + \frac{2\rho BR}{\sqrt{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Proof Let $F = \{(\mathbf{x}, y) \mapsto \phi(\langle \mathbf{w}, \mathbf{x} \rangle, y) : \mathbf{w} \in \mathcal{H}\}$. We will show that with probability 1, $R(F \circ S) \leq \rho BR/\sqrt{m}$ and then the theorem will follow from Theorem 26.5. Indeed, the set $F \circ S$ can be written as

$$F \circ S = \{(\phi(\langle \mathbf{w}, \mathbf{x}_1 \rangle, y_1), \ldots, \phi(\langle \mathbf{w}, \mathbf{x}_m \rangle, y_m)) : \mathbf{w} \in \mathcal{H}\},$$

and the bound on $R(F \circ S)$ follows directly by combining Lemma 26.9, Lemma 26.10, and the assumption that $\|\mathbf{x}\|_2 \leq R$ with probability 1. □

We next derive a generalization bound for hard-SVM based on the previous theorem. For simplicity, we do not allow a bias term and consider the hard-SVM problem:

$$\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i,\ y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \tag{26.19}$$

THEOREM 26.13 *Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$ such that there exists some vector $\mathbf{w}^\star$ with $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\langle \mathbf{w}^\star, \mathbf{x} \rangle \geq 1] = 1$ and such that $\|\mathbf{x}\|_2 \leq R$ with probability 1. Let $\mathbf{w}_S$ be the output of Equation (26.19). Then, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, we have that*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq sign(\langle \mathbf{w}_S, \mathbf{x} \rangle)] \leq \frac{2\,R\,\|\mathbf{w}^\star\|}{\sqrt{m}} + (1 + R\,\|\mathbf{w}^\star\|)\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Proof Throughout the proof, let the loss function be the ramp loss (see Section 15.2.3). Note that the range of the ramp loss is $[0,1]$ and that it is a 1-Lipschitz function. Since the ramp loss upper bounds the zero-one loss, we have that

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq \text{sign}(\langle \mathbf{w}_S, \mathbf{x}\rangle)] \leq L_{\mathcal{D}}(\mathbf{w}_S).$$

Let $B = \|\mathbf{w}^\star\|_2$ and consider the set $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq B\}$. By the definition of hard-SVM and our assumption on the distribution, we have that $\mathbf{w}_S \in \mathcal{H}$ with probability 1 and that $L_S(\mathbf{w}_S) = 0$. Therefore, using Theorem 26.12 we have that

$$L_{\mathcal{D}}(\mathbf{w}_S) \leq L_S(\mathbf{w}_S) + \frac{2BR}{\sqrt{m}} + \sqrt{\frac{2\ln(2/\delta)}{m}}.$$

□

Remark 26.1 Theorem 26.13 implies that the sample complexity of hard-SVM grows like $\frac{R^2\|\mathbf{w}^\star\|^2}{\epsilon^2}$. Using a more delicate analysis and the separability assumption, it is possible to improve the bound to an order of $\frac{R^2\|\mathbf{w}^\star\|^2}{\epsilon}$.

The bound in the preceding theorem depends on $\|\mathbf{w}^\star\|$, which is unknown. In the following we derive a bound that depends on the norm of the output of SVM; hence it can be calculated from the training set itself. The proof is similar to the derivation of bounds for structure risk minimization (SRM).

THEOREM 26.14 *Assume that the conditions of Theorem 26.13 hold. Then, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, we have that*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq \mathit{sign}(\langle \mathbf{w}_S, \mathbf{x}\rangle)] \leq \frac{4R\|\mathbf{w}_S\|}{\sqrt{m}} + \sqrt{\frac{\ln(\frac{4\log_2(\|\mathbf{w}_S\|)}{\delta})}{m}}.$$

Proof For any integer i, let $B_i = 2^i$, $\mathcal{H}_i = \{\mathbf{w} : \|\mathbf{w}\| \leq B_i\}$, and let $\delta_i = \frac{\delta}{2i^2}$. Fix i, then using Theorem 26.12 we have that with probability of at least $1-\delta_i$

$$\forall \mathbf{w} \in \mathcal{H}_i, \quad L_{\mathcal{D}}(\mathbf{w}) \leq L_S(\mathbf{w}) + \frac{2B_iR}{\sqrt{m}} + \sqrt{\frac{2\ln(2/\delta_i)}{m}}$$

Applying the union bound and using $\sum_{i=1}^{\infty}\delta_i \leq \delta$ we obtain that with probability of at least $1-\delta$ this holds for all i. Therefore, for all $\mathbf{w}$, if we let $i = \lceil \log_2(\|\mathbf{w}\|)\rceil$ then $\mathbf{w} \in \mathcal{H}_i$, $B_i \leq 2\|\mathbf{w}\|$, and $\frac{2}{\delta_i} = \frac{(2i)^2}{\delta} \leq \frac{(4\log_2(\|\mathbf{w}\|))^2}{\delta}$. Therefore,

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}) &\leq L_S(\mathbf{w}) + \frac{2B_iR}{\sqrt{m}} + \sqrt{\frac{2\ln(2/\delta_i)}{m}} \\ &\leq L_S(\mathbf{w}) + \frac{4\|\mathbf{w}\|R}{\sqrt{m}} + \sqrt{\frac{4(\ln(4\log_2(\|\mathbf{w}\|)) + \ln(1/\delta))}{m}}. \end{aligned}$$

In particular, it holds for $\mathbf{w}_S$, which concludes our proof. □

Remark 26.2 Note that all the bounds we have derived do not depend on the dimension of $\mathbf{w}$. This property is utilized when learning SVM with kernels, where the dimension of $\mathbf{w}$ can be extremely large.

26.4 Generalization Bounds for Predictors with Low ℓ_1 Norm

In the previous section we derived generalization bounds for linear predictors with an ℓ_2-norm constraint. In this section we consider the following general ℓ_1-norm constraint formulation. Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_1 \leq B\}$ be our hypothesis class, and let $Z = \mathcal{X} \times \mathcal{Y}$ be the examples domain. Assume that the loss function, $\ell : \mathcal{H} \times Z \to \mathbb{R}$, is of the same form as in Equation (26.18), with $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ being ρ-Lipschitz w.r.t. its first argument. The following theorem bounds the generalization error of all predictors in $\mathcal{H}$ using their empirical error.

THEOREM 26.15 *Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that with probability 1 we have that $\|\mathbf{x}\|_\infty \leq R$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 \leq B\}$ and let $\ell : \mathcal{H} \times Z \to \mathbb{R}$ be a loss function of the form given in Equation (26.18) such that for all $y \in \mathcal{Y}$, $a \mapsto \phi(a, y)$ is an ρ-Lipschitz function and such that $\max_{a \in [-BR, BR]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample of size m,*

$$\forall \mathbf{w} \in \mathcal{H}, \quad L_{\mathcal{D}}(\mathbf{w}) \leq L_S(\mathbf{w}) + 2\rho BR\sqrt{\frac{2\log(2d)}{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Proof The proof is identical to the proof of Theorem 26.12, while relying on Lemma 26.11 instead of relying on Lemma 26.10. □

It is interesting to compare the two bounds given in Theorem 26.12 and Theorem 26.15. Apart from the extra $\log(d)$ factor that appears in Theorem 26.15, both bounds look similar. However, the parameters B, R have different meanings in the two bounds. In Theorem 26.12, the parameter B imposes an ℓ_2 constraint on $\mathbf{w}$ and the parameter R captures a low ℓ_2-norm assumption on the instances. In contrast, in Theorem 26.15 the parameter B imposes an ℓ_1 constraint on $\mathbf{w}$ (which is stronger than an ℓ_2 constraint) while the parameter R captures a low ℓ_∞-norm assumption on the instance (which is weaker than a low ℓ_2-norm assumption). Therefore, the choice of the constraint should depend on our prior knowledge of the set of instances and on prior assumptions on good predictors.

26.5 Bibliographic Remarks

The use of Rademacher complexity for bounding the uniform convergence is due to (Koltchinskii & Panchenko 2000, Bartlett & Mendelson 2001, Bartlett & Mendelson 2002). For additional reading see, for example, (Bousquet 2002, Boucheron, Bousquet & Lugosi 2005, Bartlett, Bousquet & Mendelson 2005).

Our proof of the concentration lemma is due to Kakade and Tewari lecture notes. Kakade, Sridharan & Tewari (2008) gave a unified framework for deriving bounds on the Rademacher complexity of linear classes with respect to different assumptions on the norms.