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*

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

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.