1 Rademacher Complexity

Recall that in Theorem 2.1, we analyzed empirical risk minimization with a finite hypothesis class \( \mathcal{F} \), i.e., \(|\mathcal{F}| < +\infty\). Here as in Theorem 4.1, we will prove results for a possibly infinite hypothesis class \( \mathcal{F} \).

We will make a generalization with respect to the previous results. In this lecture, we have \( z \in \mathcal{Z} \) and we use a hypothesis \((\forall h \in \mathcal{F}) \ h : \mathcal{Z} \to \mathbb{R}\). This setting is more general than for prediction problems as in Theorems 2.1 and 4.1, but it should be clear than previous results can be generalized as well. For prediction, we have pairs \((x,y)\) and we try to predict \( y \in \mathcal{Y} \) from \( x \in \mathcal{X} \) by using functions \((\forall f \in \mathcal{F}) \ f : \mathcal{X} \to \mathcal{Y}\). Then, we used a distortion function \( d : \mathcal{Y} \times \mathcal{Y} \to [0,1] \) and defined the risk in terms of \( d(y, f(x)) \). Let \( z = (x, y) \) and thus \( \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \). We can define \( h(z) = d(y, f(x)) \) and obtain the prediction problem.

Next, we present several definitions that will naturally come later from our analysis in Theorem 6.1.

**Definition 6.1.** A random variable \( x \in \{-1, +1\} \) is called Rademacher if and only if \( P[x = -1] = P[x = +1] = 1/2 \).

**Definition 6.2.** The empirical Rademacher complexity of the hypothesis class \( \mathcal{F} \) with respect to a data set \( S = \{z_1 \ldots z_n\} \) is defined as:

\[
\hat{R}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right) \right]
\]

where \( \sigma = \{\sigma_1 \ldots \sigma_n\} \) are \( n \) independent Rademacher random variables.

Intuitively speaking, the inner quantity \( \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \) measures the correlation between \( \sigma_i \) and \( h(z_i) \).

**Definition 6.3.** The Rademacher complexity of the hypothesis class \( \mathcal{F} \) with respect to \( n \) samples is defined as:

\[
R_n(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^n} \left[ \hat{R}_S(\mathcal{F}) \right]
\]
Let \( z \) be a random variable of support \( Z \) and distribution \( D \). Let \( S = \{ z_1 \ldots z_n \} \) be a data set of \( n \) samples drawn from \( D \). Let \( F \) be a hypothesis class. For every \( h \in F \), we define the following shorthand notation:

\[
\mathbb{E}_D[h] = \mathbb{E}_{z \sim D}[h(z)] \\
\hat{\mathbb{E}}_S[h] = \frac{1}{n} \sum_{i=1}^{n} h(z_i)
\]

Next, we show an important function for the result in Theorem 6.1 which fulfills the condition in McDiarmid’s inequality (Theorem 5.1).

**Lemma 6.1.** Let \( z \) be a random variable of support \( Z \) and distribution \( D \). Let \( S = \{ z_1 \ldots z_n \} \) be a data set of \( n \) samples. Let \( F \) be a hypothesis class satisfying \( F \subseteq \{ f \mid f : Z \to [0,1] \} \). The function:

\[
\varphi(S) = \sup_{h \in F} \left( \mathbb{E}_D[h] - \hat{\mathbb{E}}_S[h] \right)
\]

satisfies the following condition:

\[
(\forall i, \forall z_1 \ldots z_i, \tilde{z}_i \in Z) \ |\varphi(z_1, \ldots, z_i, \ldots, z_n) - \varphi(z_1, \ldots, \tilde{z}_i, \ldots, z_n)| \leq 1/n
\]

(The proof of the above might be left for homework very soon.)

Next, we show that the empirical Rademacher complexity fulfills the condition in McDiarmid’s inequality (Theorem 5.1).

**Lemma 6.2.** Let \( F \) be a hypothesis class satisfying \( F \subseteq \{ f \mid f : Z \to [0,1] \} \). The empirical Rademacher complexity satisfies the following condition:

\[
(\forall i, \forall z_1 \ldots z_i, \tilde{z}_i \in Z) \ \left| \hat{\mathbb{R}}_{\{z_1 \ldots z_i \ldots z_n\}}(F) - \hat{\mathbb{R}}_{\{\tilde{z}_1 \ldots z_i \ldots z_n\}}(F) \right| \leq 1/n
\]

(The proof of the above might be left for homework very soon.)

Next, we show a generalization bound that holds for any class of bounded real-valued functions.

**Theorem 6.1** (Rademacher-based uniform convergence). Let \( z \) be a random variable of support \( Z \) and distribution \( D \). Let \( S = \{ z_1 \ldots z_n \} \) be a data set of \( n \) i.i.d. samples drawn from \( D \). Let \( F \) be a hypothesis class satisfying \( F \subseteq \{ f \mid f : Z \to [0,1] \} \). Fix \( \delta \in (0,1) \). With probability at least \( 1 - \delta \) over the choice of \( S \), we have:

\[
(\forall h \in F) \ \mathbb{E}_D[h] \leq \hat{\mathbb{E}}_S[h] + 2\mathcal{R}_n(F) + \sqrt{\frac{\log (1/\delta)}{2n}}
\]

In addition, with probability at least \( 1 - \delta \) over the choice of \( S \), we have:

\[
(\forall h \in F) \ \mathbb{E}_D[h] \leq \hat{\mathbb{E}}_S[h] + 2\hat{\mathcal{R}}_S(F) + 3\sqrt{\frac{\log (2/\delta)}{2n}}
\]
Proof. Given a function $f : \mathbb{Z}^n \to \mathbb{R}$, we define the following shorthand notation:

$$E_S[f(S)] = E_{S \sim D^n}[f(S)]$$

By Lemma 6.1, the function $\varphi : \mathbb{Z}^n \to \mathbb{R}$ fulfills the condition in McDiarmid's inequality (Theorem 5.1) and therefore:

$$P[\varphi(S) - E_S[\varphi(S)] \geq \varepsilon] \leq e^{-\frac{\varepsilon^2}{2n}}$$

Setting $e^{-\frac{\varepsilon^2}{2n}} = \delta$, we get $\varepsilon = \sqrt{\log \frac{1}{\delta} / 2n}$. Thus:

$$P[\varphi(S) < E_S[\varphi(S)] + \sqrt{\frac{\log (1/\delta)}{2n}}] = 1 - P[\varphi(S) - E_S[\varphi(S)] \geq \sqrt{\frac{\log (1/\delta)}{2n}}] \geq 1 - \delta \quad (2)$$

Note that by the definition of the supremum, by the definition of the function $\varphi : \mathbb{Z}^n \to \mathbb{R}$ in eq.(1), and by eq.(2), with probability at least $1 - \delta$, we have:

$$\left( \forall h \in \mathcal{F} \right) E_D[h] - \hat{E}_S[h] \leq \sup_{h \in \mathcal{F}} \left( E_D[h] - \hat{E}_S[h] \right) = \varphi(S)$$

$$< E_S[\varphi(S)] + \sqrt{\frac{\log (1/\delta)}{2n}} \quad (3)$$

The final step is to bound $E_S[\varphi(S)]$ in eq.(3) in terms of the Rademacher complexity of $\mathcal{F}$. In order to do this, we introduce a “ghost sample” $T = \{\bar{z}_1 \ldots \bar{z}_n\}$ of $n$ i.i.d. samples drawn from $\mathcal{D}$. Note that:

$$E_T \left[ \hat{E}_T[h] \bigg| S \right] = E_T \left[ \hat{E}_T[h] \right] = E_D[h]$$

$$E_T \left[ \hat{E}_S[h] \bigg| S \right] = E_S[h] \quad (4)$$

Furthermore, since $S \cup T = \{z_1, \bar{z}_1 \ldots z_n, \bar{z}_n\}$ is a set $2n$ i.i.d. random variables, for every $j \in \{1 \ldots n\}$ we have:

$$E_{S,T} \left[ \sup_{h \in \mathcal{F}} \left( \frac{1}{n} \left( h(\bar{z}_j) - h(z_j) + \sum_{i=1,i \neq j}^n (h(\bar{z}_i) - h(z_i)) \right) \right) \right]$$

$$= E_{S,T} \left[ \sup_{h \in \mathcal{F}} \left( \frac{1}{n} \left( h(z_j) - h(\bar{z}_j) + \sum_{i=1,i \neq j}^n (h(\bar{z}_i) - h(z_i)) \right) \right) \right] \quad (5)$$

3
Let $\sigma = \{\sigma_1 \ldots \sigma_n\}$ be $n$ independent Rademacher random variables. By the definition of the function $\varphi : \mathbb{Z}^n \to \mathbb{R}$ in eq.(1), we have:

$$\mathbb{E}_S[\varphi(S)] = \mathbb{E}_S \left[ \sup_{h \in F} \left( \mathbb{E}_T[h] - \hat{\mathbb{E}}_S[h] \right) \right]$$

$$= \mathbb{E}_S \left[ \sup_{h \in F} \left( \mathbb{E}_T \left[ \frac{1}{n} \sum_{i=1}^n h(\bar{z}_i) - \frac{1}{n} \sum_{i=1}^n h(z_i) \right] \right) \right]$$

$$= \mathbb{E}_S \left[ \sup_{h \in F} \left( \mathbb{E}_T \left[ \frac{1}{n} \sum_{i=1}^n (h(\bar{z}_i) - h(z_i)) \right] \right) \right]$$

$$\leq \mathbb{E}_S \left[ \mathbb{E}_T \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n (h(\bar{z}_i) - h(z_i)) \right) \right] \right]$$

$$= \mathbb{E}_{S,T} \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n (h(\bar{z}_i) - h(z_i)) \right) \right]$$

$$= \frac{1}{2} \mathbb{E}_{S,T} \left[ \sup_{h \in F} \left( \frac{1}{n} \left( h(\bar{z}_1) - h(z_1) + \sum_{i=2}^n (h(\bar{z}_i) - h(z_i)) \right) \right) \right] +$$

$$\frac{1}{2} \mathbb{E}_{S,T} \left[ \sup_{h \in F} \left( \frac{1}{n} \left( h(z_1) - h(\bar{z}_1) + \sum_{i=2}^n (h(z_i) - h(\bar{z}_i)) \right) \right) \right]$$

$$= \mathbb{E}_{S,T,\sigma_1} \left[ \sup_{h \in F} \left( \frac{1}{n} \left( \sigma_1(h(\bar{z}_1) - h(z_1)) + \sum_{i=2}^n (h(\bar{z}_i) - h(z_i)) \right) \right) \right]$$

$$\vdots$$

$$= \mathbb{E}_{S,T,\sigma} \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i(h(\bar{z}_i) - h(z_i)) \right) \right]$$

$$\leq \mathbb{E}_{S,T,\sigma} \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i h(\bar{z}_i) \right) + \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n -\sigma_i h(z_i) \right) \right]$$

$$= \mathbb{E}_T \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i h(\bar{z}_i) \right) \right] + \mathbb{E}_S \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i h(z_i) \right) \right]$$

$$= \mathbb{E}_T \left[ \mathbb{G}_T(F) \right] + \mathbb{E}_S \left[ \mathbb{G}_S(F) \right]$$

where the step in eq.(6.a) follows from eq.(4). The step in eq.(6.b) follows from Jensen’s inequality and convexity of sup. The step in eq.(6.c) follows from the law of total expectation. The step in eq.(6.d) follows from eq.(5). (Note that we could have used any other weights that sum up to 1 instead of $1/2 + 1/2$, but $1/2$ becomes useful for the next step.) The step in eq.(6.e) uses the fact that
since $\sigma_1$ is a Rademacher random variable, then $E_{\sigma_1}[f(\sigma_1)] = \frac{1}{2}f(+1) + \frac{1}{2}f(-1)$ for any arbitrary function $f : \mathbb{R} \to \mathbb{R}$. The step in eq.(6.f) follows from the fact that $\sup_{x \in \mathcal{X}} f(x) + g(x) \leq \sup_{x \in \mathcal{X}} f(x) + \sup_{x \in \mathcal{X}} g(x)$ for any arbitrary pair of functions $f, g : \mathcal{X} \to \mathbb{R}$ and any domain $\mathcal{X}$. By eq.(3) and the above, we prove the first claim.

For proving the second claim. By Lemma 6.1, the function $\varphi : \mathbb{Z}^n \to \mathbb{R}$ fulfills the condition in McDiarmid’s inequality (Theorem 5.1). By Lemma 6.2, the empirical Rademacher complexity fulfills the condition in McDiarmid’s inequality (Theorem 5.1). Additionally, since $\mathcal{R}_n(\mathcal{F}) = E_S[\hat{\mathcal{R}}_S(\mathcal{F})]$ and by the union bound, we have:

$$P[\varphi(S) - E_S[\varphi(S)] \geq \varepsilon \text{ or } \mathcal{R}_n(\mathcal{F}) - \hat{\mathcal{R}}_S(\mathcal{F}) \geq \varepsilon]$$

$$\leq P[\varphi(S) - E_S[\varphi(S)] \geq \varepsilon] + P[\mathcal{R}_n(\mathcal{F}) - \hat{\mathcal{R}}_S(\mathcal{F}) \geq \varepsilon]$$

$$\leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(1/n)^2}} + e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(1/n)^2}}$$

$$= 2e^{-2n\varepsilon^2}$$

Setting $2e^{-2n\varepsilon^2} = \delta$, we get $\varepsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$. Thus:

$$P[\varphi(S) + 2\mathcal{R}_n(\mathcal{F}) < E_S[\varphi(S)] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + 3\varepsilon]$$

$$\geq 1 - P[\varphi(S) - E_S[\varphi(S)] \geq \varepsilon \text{ or } \mathcal{R}_n(\mathcal{F}) - \hat{\mathcal{R}}_S(\mathcal{F}) \geq \varepsilon]$$

$$\geq 1 - \delta$$

The proof then continues as from eq.(2). □

2 Elementary Properties

The key in applying Rademacher-based uniform convergence in practice, is to bound the empirical Rademacher complexity for specific problems (i.e., for specific hypothesis classes). In what follows, we review several useful properties of the Rademacher complexity.

Claim 6.1 (Elementary properties). Let $\mathcal{F}$ and $\mathcal{G}$ be two hypothesis classes and let $a \in \mathbb{R}$ be a constant. Define the shorthand notation:

$$a\mathcal{F} = \{af \mid f \in \mathcal{F}\}$$

$$\mathcal{F} + \mathcal{G} = \{f + g \mid f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}$$

We have:

i. $\mathcal{F} = \{h\} \Rightarrow \hat{\mathcal{R}}_S(\mathcal{F}) = 0$

ii. $\mathcal{R}_n(a\mathcal{F}) = |a| \mathcal{R}_n(\mathcal{F})$

iii. $\mathcal{F} \subseteq \mathcal{G} \Rightarrow \hat{\mathcal{R}}_S(\mathcal{F}) \leq \hat{\mathcal{R}}_S(\mathcal{G})$

iv. $\hat{\mathcal{R}}_S(\mathcal{F} + \mathcal{G}) \leq \hat{\mathcal{R}}_S(\mathcal{F}) + \hat{\mathcal{R}}_S(\mathcal{G})$
Proof. Recall that $\sigma_1 \ldots \sigma_n$ are $n$ independent Rademacher random variables. Let $\sigma_i^j \equiv \sigma_i \ldots \sigma_j$ and $\sigma \equiv \sigma_1^n$. Let $S = \{z_1 \ldots z_n\}$.

For proving Claim i, since $F = \{h\}$, we have:

$$\hat{R}_S(F) = \mathbb{E}_{\sigma_1^n}\left[ \sup_{h \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right]$$

$$= \frac{1}{n} \mathbb{E}_{\sigma_1^n}\left[ \sum_{i=1}^{n} \sigma_i h(z_i) \right]$$

$$= \frac{1}{n} \mathbb{E}_{\sigma_2^n}\left[ \frac{1}{2} h(z_1) - \frac{1}{2} h(z_1) + \sum_{i=2}^{n} \sigma_i h(z_i) \right]$$

$$= \frac{1}{n} \mathbb{E}_{\sigma_2^n}\left[ \sum_{i=2}^{n} \sigma_i h(z_i) \right]$$

$$\vdots$$

$$= 0$$

For proving Claim ii, first assume that $a \neq 0$. Note that $\sigma_i$ is a Rademacher random variable if and only if $\text{sgn}(a)\sigma_i$ is a Rademacher random variable. Therefore:

$$\hat{R}_S(aF) = \mathbb{E}_{\sigma}\left[ \sup_{h \in aF} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right]$$

$$= \mathbb{E}_{\sigma}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i af(z_i) \right]$$

$$= \mathbb{E}_{\sigma}\left[ \sup_{f \in F} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i \text{sgn}(a)|a|f(z_i) \right) \right]$$

$$= |a| \mathbb{E}_{\sigma}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right]$$

$$= |a| \hat{R}_S(F)$$

For $a = 0$, trivially $\hat{R}_S(aF) = 0$ and we prove our claim.

For proving Claim iii, if $F \subseteq G$ then $\sup_{h \in F} \phi(h) \leq \sup_{h \in G} \phi(h)$ for any arbitrary function $\phi : G \rightarrow \mathbb{R}$, we have:

$$\hat{R}_S(F) = \mathbb{E}_{\sigma}\left[ \sup_{h \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right]$$

$$\leq \mathbb{E}_{\sigma}\left[ \sup_{h \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right]$$

$$= \hat{R}_S(G)$$
For proving Claim iv, since \( \sup_{f \in F, g \in G} (\phi(f) + \nu(g)) \leq \sup_{f \in F} \phi(f) + \sup_{g \in G} \nu(g) \)

for any arbitrary pair of functions \( \phi : F \to \mathbb{R} \) and \( \nu : G \to \mathbb{R} \), we have:

\[
\hat{R}_S(F + G) = \mathbb{E}_\sigma \left[ \sup_{h \in F+G} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right) \right] \\
= \mathbb{E}_\sigma \left[ \sup_{f \in F, g \in G} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f(z_i) + g(z_i)) \right) \right] \\
\leq \mathbb{E}_\sigma \left[ \sup_{f \in F} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right) + \sup_{g \in G} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right) \right] \\
= \hat{R}_S(F) + \hat{R}_S(G)
\]

\[\square\]

The next one is a straightforward conclusion from the above.

**Corollary 6.1.** Let \( F \) be a hypothesis class and let \( a, b \in \mathbb{R} \) be two constants. Define the shorthand notation:

\[ aF + b = \{ af + b \mid f \in F \} \]

We have:

\[ \hat{R}_S(aF + b) \leq |a| \ \hat{R}_S(F) \]

**Proof.** Define \( G = \{ b \} \). By Claim 6.1, we have:

\[
\hat{R}_S(aF + b) = \hat{R}_S(aF + G) \\
\leq \hat{R}_S(aF) + \hat{R}_S(G) \\
= |a| \ \hat{R}_S(F)
\]

\[\square\]

### 3 Ledoux-Talagrand Contraction Lemma

Next, we present a lemma that is very useful for decoupling the analysis of losses from the analysis of predictor functions. While the results are very general, first we illustrate its use. Consider a linear classification problem in which we predict \( y \in \{-1, +1\} \) from \( x \in \mathbb{R}^p \) by using a linear predictor. In this case, the loss function can be \( \phi(t) = \min (1, \max (0, 1 - t)) \), where \( t = y(\mathbf{w}, x) \). Similarly, consider a linear regression problem in which we predict \( y \in \mathbb{R} \) from \( x \in \mathbb{R}^p \) by using a linear predictor. In this case, the loss function can be \( \phi(t) = \min (1, t^2) \) or \( \phi(t) = \min (1, |t|) \), where \( t = y - \langle \mathbf{w}, x \rangle \).
Definition 6.4 (Lipschitz continuity). A function \( \phi : \mathbb{R} \to \mathbb{R} \) is called \( K \)-Lipschitz continuous if and only if there is a constant \( K < +\infty \) such that
\[
(\forall t, u \in \mathbb{R}) \ |\phi(t) - \phi(u)| \leq K |t - u|
\]
An everywhere differentiable (i.e., smooth) function \( \phi : \mathbb{R} \to \mathbb{R} \) is called \( K \)-Lipschitz continuous if and only if there is a constant \( K < +\infty \) such that
\[
(\forall t \in \mathbb{R}) \ |\phi'(t)| \leq K
\]
Lemma 6.3 (Ledoux-Talagrand contraction lemma). Assume that the hypothesis \( \mathcal{F} \subseteq \{ f \mid f : \mathcal{Z} \to \mathbb{R} \} \). Assume that the function \( \phi : \mathbb{R} \to \mathbb{R} \) is 1-Lipschitz continuous. Define the shorthand notation:
\[
\phi(\mathcal{F}) = \{ \phi(f) \mid f \in \mathcal{F} \}
\]
We have:
\[
\hat{\mathcal{R}}_S(\phi(\mathcal{F})) \leq \hat{\mathcal{R}}_S(\mathcal{F})
\]
Proof. Recall that \( \sigma_1 \ldots \sigma_n \) are independent Rademacher random variables. Let \( \sigma'_1 \equiv \sigma_1 \ldots \sigma_j \) and \( \sigma \equiv \sigma'_n \). Let \( S = \{ z_1 \ldots z_n \} \). Define the following shorthand notation:
\[
f_1^+ = \arg \sup_{f \in \mathcal{F}} \left( \phi(f(z_1)) + \sum_{i=2}^n \sigma_i \phi(f(z_i)) \right)
f_1^- = \arg \sup_{f \in \mathcal{F}} \left( -\phi(f(z_1)) + \sum_{i=2}^n \sigma_i \phi(f(z_i)) \right)
a_1 = \begin{cases} +1 & \text{if } f_1^+(z_1) - f_1^-(z_1) \geq 0 \\ -1 & \text{if } f_1^+(z_1) - f_1^-(z_1) < 0 \end{cases}
\]
By using the above notation, we have:
\[
\hat{\mathcal{R}}_S(\phi(\mathcal{F})) = \mathbb{E}_\sigma \left[ \sup_{h \in \phi(\mathcal{F})} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i h(z_i) \right) \right] 
\]
\[
= \frac{1}{n} \mathbb{E}_{\sigma_1^n} \left[ \sup_{f \in \mathcal{F}} \left( \sum_{i=1}^n \sigma_i \phi(f(z_i)) \right) \right] 
\]
\[
= \frac{1}{2n} \mathbb{E}_{\sigma_1^n} \left[ \sup_{f \in \mathcal{F}} \left( \phi(f(z_1)) + \sum_{i=2}^n \sigma_i \phi(f(z_i)) \right) + \sup_{f \in \mathcal{F}} \left( -\phi(f(z_1)) + \sum_{i=2}^n \sigma_i \phi(f(z_i)) \right) \right] 
\]
\[
= \frac{1}{2n} \mathbb{E}_{\sigma_1^n} \left[ \sup_{f, f' \in \mathcal{F}} \left( \phi(f(z_1)) + \sum_{i=2}^n \sigma_i \phi(f(z_i)) - \phi(f'(z_1)) + \sum_{i=2}^n \sigma_i \phi(f'(z_i)) \right) \right] 
\]
\[
\leq \frac{1}{2n} \mathbb{E}_{\sigma_1^n} \left[ \sup_{f, f' \in \mathcal{F}} \left( |f(z_1) - f'(z_1)| + \sum_{i=2}^n \sigma_i \phi(f(z_i)) + \sum_{i=2}^n \sigma_i \phi(f'(z_i)) \right) \right] \quad (7.a)
\]
\[\begin{align*}
&= \frac{1}{2n} \mathbb{E}_{\sigma_n} \left[ \sup_{f, f' \in F} \left( a_1 (f(z_1) - f'(z_1)) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) + \sum_{i=2}^{n} \sigma_i \phi (f'(z_i)) \right) \right] \\
&= \frac{1}{2n} \mathbb{E}_{\sigma_n} \left[ \sup_{f \in F} \left( a_1 f(z_1) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) \right) + \sup_{f \in F} \left( -a_1 f(z_1) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) \right) \right] \\
&= \frac{1}{2n} \mathbb{E}_{\sigma_n} \left[ \sup_{f \in F} \left( f(z_1) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) \right) + \sup_{f \in F} \left( -f(z_1) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) \right) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[ \sup_{f \in F} \left( \sigma_1 f(z_1) + \sum_{i=2}^{n} \sigma_i \phi (f(z_i)) \right) \right] \\
&= \ldots \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[ \left( \sum_{i=1}^{n} \sigma_i f(z_i) \right) \right] \\
&= \hat{R}_S(F)
\end{align*}\]

where the step in eq. (7.a) follows from the 1-Lipschitz continuity of the function \( \phi : \mathbb{R} \rightarrow \mathbb{R} \) as in Definition 6.4. The step in eq. (7.b) follows from the fact that \( a_1 \in \{-1, +1\} \).

4 Application: Linear Prediction

As we mentioned before, the key in applying Rademacher-based uniform convergence in practice, is to bound the empirical Rademacher complexity for specific problems (i.e., for specific hypothesis classes). As shown by the Ledoux-Talagrand contraction lemma (Lemma 6.3), we can decouple the analysis of losses from the analysis of predictor functions. Next, we see a small example of how to upper bound the empirical Rademacher complexity for linear predictor functions.

Lemma 6.4 (Adapted from [1]). Let \( Z = \{z|z \in \mathbb{R}^p \text{ and } \|z\|_2 \leq Z\} \). Let \( F \) be the class of linear predictors, i.e.,

\[F = \{\langle w, z \rangle | w \in \mathbb{R}^p \text{ and } \|w\|_2 \leq W \}\]

Let \( S = \{z^{(1)}, \ldots, z^{(n)}\} \) be a data set of \( n \) samples. We have:

\[
\hat{R}_S(F) \leq \frac{ZW}{\sqrt{n}}
\]

Proof. Since \( F = \{\langle w, z \rangle | w \in \mathbb{R}^p \text{ and } \|w\|_2 \leq W \}, \) we have:

\[
\hat{R}_S(F) = \mathbb{E}_\sigma \left[ \sup_{h \in F} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z^{(i)}) \right) \right]
\]
\[
\frac{1}{n} \mathbb{E}_\sigma \left[ \sup_{\|w\|_2 \leq W} \left( \sum_{i=1}^n \sigma_i \langle w, z^{(i)} \rangle \right) \right] = \frac{1}{n} \mathbb{E}_\sigma \left[ \sup_{\|w\|_2 \leq W} (w, \sum_{i=1}^n \sigma_i z^{(i)}) \right] = \frac{W}{n} \mathbb{E}_\sigma \left[ \|\sum_{i=1}^n \sigma_i z^{(i)}\|_2 \right] \leq W \sqrt{\mathbb{E} \left[ \left( \sum_{i=1}^n \sigma_i z^{(i)} \right)^2 \right]} \]

where the step in eq.(8.a) follows from norm duality. The step in eq.(8.b) follows from Jensen’s inequality and concavity of the square root. The step in eq.(8.c) follows from the fact that \(\sigma_1 \ldots \sigma_n\) are independent. The step in eq.(8.d) follows from the fact that if \(\sigma_i\) is a Rademacher random variable, then \(\sigma_i^2 = 1\) and \(\mathbb{E} [\sigma_i] = 0\).

**References**