EECS 598: Statistical Learning Theory, Winter 2014 **Topic 10**

Rademacher Complexity

Lecturer: Clayton Scott *Scribe: Yan Deng, Kevin Moon*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

Rademacher complexity is a measure of the richness of a class of real-valued functions. In this sense, it is similar to the VC dimension. In fact, we will establish a uniform deviation bound in terms of Rademacher complexity, and then use this result to prove the VC inequality. Unlike VC dimension, however, Rademacher complexity is not restricted to binary functions, and will also prove useful later in the analysis of other learning algorithms such as kernel-based algorithms.

2 Rademacher Complexity

Let $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$ be a set of functions $\mathcal{Z} \to [a,b]$ where $a, b \in \mathbb{R}, a < b$. Let $Z_1, \ldots, Z_n$ be i.i.d. random variables on $\mathcal{Z}$ following some distribution P. Denote the sample $S = (Z_1, \ldots, Z_n)$.

The *empirical Rademacher complexity* of $\mathcal{G}$ with respect to the sample S is

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i) \right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ with $\sigma_i \overset{iid}{\sim} \text{unif}\{-1, 1\}$. Here $\sigma_1, \ldots, \sigma_n$ are known as Rademacher random variables. The complexity $\widehat{\mathfrak{R}}_S(\mathcal{G})$ is random because of the randomness of S.

The *Rademacher complexity* of $\mathcal{G}$ is

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{G})].$$

Rademacher complexity is sometimes called “Rademacher average.”

An interpretation in the context of binary classification is that $\mathcal{G}$ is rich, equivalently, $\widehat{\mathfrak{R}}_S(\mathcal{G})$ or $\mathfrak{R}_n(\mathcal{G})$ is high, if we can choose functions g to accurately match different random sign combinations reflected by σ. Note that the complexity is bounded, since elements of $\mathcal{G}$ are bounded within the interval $[a, b]$.

Theorem 1 (One-sided Rademacher complexity bound)**.** *Let $Z, Z_1, \ldots, Z_n$ be iid random variables taking values in a set $\mathcal{Z}$. Consider a set of functions $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$. $\forall \delta > 0$, with probability $\geq 1 - \delta$, we have with respect to the draw of sample S that:*

$$\forall g \in \mathcal{G},\ \mathbb{E}[g(Z)] \leq \frac{1}{n} \sum_{i=1}^n g(Z_i) + 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 1/\delta}{2n}}. \quad (1)$$

In addition, $\forall \delta > 0$, with probability $\geq 1 - \delta$, we have with respect to the draw of S that:

$$\forall g \in \mathcal{G},\ \mathbb{E}[g(\mathcal{Z})] \leq \frac{1}{n} \sum_{i=1}^n g(Z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 2/\delta}{2n}}. \quad (2)$$

The final term in both (1) and (2) is typically much smaller than the Rademacher complexity. Note that (1) and (2) are one-sided uniform deviation bounds, and that (2) is a *data-dependent* bound.

Before proving the theorem, we first review the following useful facts.

Fact 1: For any real-valued functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$, $\sup_x f_1(x) - \sup_x f_2(x) \leq \sup_x (f_1(x) - f_2(x))$.

To see this, let $\epsilon > 0$ and let x_1 be such that $f_1(x_1) \geq \sup_x f_1(x) - \epsilon$. Then,

$$\sup_x f_1(x) - \sup_x f_2(x) \leq \sup_x f_1(x) - f_2(x_1) \leq f_1(x_1) - f_2(x_1) + \epsilon \leq \sup_x (f_1(x) - f_2(x)) + \epsilon.$$

$\epsilon > 0$ was arbitrary, so the result follows.

Fact 2: For any real-valued functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$, $\sup_x (f_1(x) + f_2(x)) \leq \sup_x f_1(x) + \sup_x f_2(x)$.

Fact 3: $\sup(\cdot)$ is a convex function, i.e., if $(x_\lambda)_{\lambda \in \Lambda}$ and $(x'_\lambda)_{\lambda \in \Lambda}$ are two sequences (where Λ is possibly uncountable), then $\forall \alpha \in [0, 1]$,

$$\sup_{\lambda \in \Lambda} (\alpha x_\lambda + (1 - \alpha) x'_\lambda) \leq \alpha \sup_{\lambda \in \Lambda} x_\lambda + (1 - \alpha) \sup_{\lambda \in \Lambda} x'_\lambda$$

This is an immediate consequence of Fact 2.

Fact 4: Jensen's inequality, i.e., if f is convex, then $f(\mathbb{E}[U]) \leq \mathbb{E}[f(U)]$.

Now, we are ready to prove Theorem 1.

Proof. For notational brevity, denote $\mathbb{E}[g] = \mathbb{E}[g(\mathcal{Z})]$ and $\widehat{\mathbb{E}}_S[g] = \frac{1}{n} \sum_{i=1}^n g(Z_i)$. The idea is to apply the bounded difference inequality (BDI) to

$$\phi(S) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right).$$

First, we verify the bounded difference assumption. Denoting $S'_i = (Z_1, \ldots, Z_{i-1}, Z'_i, Z'_{i+1}, \ldots, Z_n)$, we have

$$\begin{aligned}
\phi(S) - \phi(S'_i) &= \sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right) - \sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_{S'}[g] \right) \\
&\leq \sup_{g \in \mathcal{G}} \left(\widehat{\mathbb{E}}_{S'}[g] - \widehat{\mathbb{E}}_S[g] \right) && \text{(by Fact 1)} \\
&= \sup_{g \in \mathcal{G}} \frac{1}{n} \left(g(Z'_i) - g(Z_i) \right) \leq \frac{b - a}{n}.
\end{aligned}$$

Similarly, we can prove $\phi(S') - \phi(S_i) \leq (b - a)/n$ and therefore $|\phi(S') - \phi(S_i)| \leq (b - a)/n$. By the BDI, we have that with probability $\geq 1 - \delta$,

$$\phi(S) - \mathbb{E}_S[\phi(S)] \leq (b - a) \sqrt{\frac{\log(1/\delta)}{2n}}.$$

To establish (1), it remains to show that $\mathbb{E}_S[\phi(S)] \leq 2\mathfrak{R}_n(\mathcal{G})$.

Thus let us introduce another random sample (called a *ghost sample*) $S' = (Z'_1, \ldots, Z'_n)$ with $Z'_i \overset{iid}{\sim} P$, independent of S. Then

$$\begin{aligned}
\mathbb{E}_S[\phi(S)] &= \mathbb{E}_S \left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right) \right] \\
&= \mathbb{E}_S \left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}_{S'} \left[\widehat{\mathbb{E}}_{S'}[g] - \widehat{\mathbb{E}}_S[g] \right] \right) \right] && \text{(by } \mathbb{E}[g] = \mathbb{E}_{S'} \widehat{\mathbb{E}}_{S'}[g])
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\left(\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right)\right] && \text{(by Facts 3 and 4)}\\
&= \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\left(g(Z_i')-g(Z_i)\right)\right]\\
&= \mathbb{E}_{\sigma,S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g(Z_i')-g(Z_i)\right)\right] && (*)\\
&\leq \mathbb{E}_{\sigma,S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i')+\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)g(Z_i)\right] && \text{(by Fact 2)}\\
&= \mathbb{E}_{\sigma,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i')\right]+\mathbb{E}_{\sigma,S}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right] && (\sigma_i \text{ is symmetric}, \forall i)\\
&= 2\mathfrak{R}_n(\mathcal{G}).
\end{aligned}
$$

The equality $(*)$ holds because *(i)* $(Z_i)_{i=1}^n$ and $(Z_i')_{i=1}^n$ are i.i.d. (hence $g(Z_i')-g(Z_i)$ and $g(Z_i)-g(Z_i')$ have the same distribution), and *(ii)* σ_i is symmetric.

To establish (2), we apply the BDI again to $\phi(S)=\widehat{\mathfrak{R}}_S(\mathcal{G})$. Observe that

$$
\begin{aligned}
\phi(S)-\phi(S_i') &= \widehat{\mathfrak{R}}_S(\mathcal{G})-\widehat{\mathfrak{R}}_{S'}(\mathcal{G})\\
&= \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)-\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{j\neq i}\sigma_j g(Z_j)+\frac{1}{n}\sigma_i g(Z_i')\right)\right]\\
&\leq \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sigma_i(g(Z_i)-g(Z_i'))\right]\leq\frac{b-a}{n}. && \text{(by Fact 1).}
\end{aligned}
$$

Similarly, we can prove $\phi(S')-\phi(S_i)\leq(b-a)/n$ and thus $|\phi(S')-\phi(S_i)|\leq(b-a)/n$. Applying the BDI, we have that with probability $\geq 1-\delta/2$,

$$
\mathfrak{R}_n(\mathcal{G})\leq\widehat{\mathfrak{R}}_S(\mathcal{G})+(b-a)\sqrt{\frac{\log 2/\delta}{2n}}. \tag{3}
$$

Combining (1) (with δ replaced by $\delta/2$) and the inequality above, we then establish (2), because

$$
\begin{aligned}
\Pr(\text{violating (2)}) &\leq \Pr(\text{violating (1)})+\Pr(\text{violating (3)})\\
&\leq \delta/2+\delta/2=\delta.
\end{aligned}
$$

□

The following two-sided bound also holds.

Theorem 2 (Two-sided Rademacher complexity bound)**.** *Consider a set of classifiers $\mathcal{G}\subseteq[a,b]^{\mathcal{Z}}$. $\forall\delta>0$, with probability $\geq 1-\delta$, we have with respect to the draw of sample S that:*

$$
\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(\mathcal{Z})]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right|\leq 2\mathfrak{R}_n(\mathcal{G})+(b-a)\sqrt{\frac{\log 2/\delta}{2n}}. \tag{4}
$$

In addition, $\forall\delta>0$, with probability $\geq 1-\delta$, we have with respect to the draw of S that:

$$
\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(\mathcal{Z})]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right|\leq 2\widehat{\mathfrak{R}}_S(\mathcal{G})+3(b-a)\sqrt{\frac{\log 4/\delta}{2n}}. \tag{5}
$$

The proof is left as an exercise.

3 Bounds for Binary Classification

Consider a set of binary classifiers $\mathcal{H} \subset \{-1,1\}^{\mathcal{X}}$. Let $\mathcal{Z} = \mathcal{X} \times \{-1,1\}$. Define another set $\mathcal{G}$ based on $\mathcal{H}$ as $\mathcal{G} = \{(x,y) \to \mathbf{1}_{\{h(x)\neq y\}} : h \in \mathcal{H}\}$. Let $S = \{Z_1, \ldots, Z_n\} = \{(X_1,Y_1), \ldots, (X_n,Y_n)\}$, and also let $T = \{X_1, \ldots, X_n\}$, which is the projection of S on the domain $\mathcal{X}$. The empirical Rachemacher complexity of $\mathcal{H}$ should be written $\widehat{\mathfrak{R}}_T(\mathcal{H})$, however, we will follow convention and write it as $\widehat{\mathfrak{R}}_S(\mathcal{H})$. There should be no confusion since the domain of elements of $\mathcal{H}$ is $\mathcal{X}$, so only the X_is in the sample can be used when evaluating the empirical Rademacher complexity. Thus we have

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right].$$

Lemma 1. $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_S(\mathcal{H})$

Proof. From the definitions, we have

$$\begin{aligned}
\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{G}) &= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{\{h(X_i)\neq Y_i\}}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \frac{1-Y_i h(X_i)}{2}\right] \\
&= \mathbb{E}_\sigma\left[\frac{1}{2n}\sum_{i=1}^n \sigma_i + \frac{1}{2}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i(-Y_i)\,h(X_i)\right] \\
&= \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right] \\
&= \frac{1}{2}\widehat{\mathcal{R}}_S(\mathcal{H}),
\end{aligned}$$

where the second to last step follows from the facts that $\mathbb{E}_\sigma[\sigma_i] = 0$ and σ_i and $\sigma_i(-Y_i)$ have the same distribution. □

Now observe that $\mathbb{E}[g] = \mathbb{E}\left[\mathbf{1}_{\{h(X)\neq Y\}}\right] = R(h)$ when $g \in \mathcal{G}$ is defined in terms of $h \in \mathcal{H}$. Note also that $\frac{1}{n}\sum_{i=1}^n g(Z_i) = \widehat{R}_n(h)$. This gives the following corollary:

Corollary 1. $\forall\delta > 0$, *with probability* $\geq 1-\delta$,

$$\sup_{h\in\mathcal{H}}\left(R(h) - \widehat{R}_n(h)\right) \leq \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2n}},$$

and with probability $\geq 1-\delta$,

$$\sup_{h\in\mathcal{H}}\left(R(h) - \widehat{R}_n(h)\right) \leq \widehat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Remark. A two-sided version of this corollary also holds, with $\delta \to \delta/2$.

Example. Let $\Pi = \{A_1, \ldots, A_k\}$ be a fixed partition of $\mathcal{X}$, such as a regular partition or a recursive dyadic partition. Let $\mathcal{H} = \{$classifiers that are constant on cells in $\Pi\}$. Then $|\mathcal{H}| = 2^k$. We'll obtain a bound on the empirical Rademacher complexity of $\mathcal{H}$. Let $\ell(A)$ denote the label assigned to $A \in \Pi$. Then

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\mathcal{H}) &= \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h\left(X_i\right) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{j=1}^{k} \sum_{i : X_i \in A_j} \sigma_i h\left(X_i\right) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sum_{A \in \Pi} \sup_{\ell(A)} \sum_{i : X_i \in A} \sigma_i \ell(A) \right] \\
&= \frac{1}{n} \sum_{A \in \Pi} \mathbb{E}_\sigma \left[\sup_{\ell(A)} \sum_{i : X_i \in A} \sigma_i \ell(A) \right].
\end{aligned}$$

Manipulating the terms inside the expectation gives

$$\begin{aligned}
\mathbb{E}_\sigma \left[\sup_{\ell(A)} \sum_{i : X_i \in A} \sigma_i \ell(A) \right] &= \mathbb{E}_\sigma \left[\sup_{\ell(A)} \ell(A) \sum_{i : X_i \in A} \sigma_i \right] \\
&= \mathbb{E}_\sigma \left[\left| \sum_{i : X_i \in A} \sigma_i \right| \right] && (\ell(A) \in \{-1, 1\}) \\
&= \mathbb{E}_\sigma \left[\sqrt{ \left(\sum_{i : X_i \in A} \sigma_i \right)^2 } \right] \\
&\le \sqrt{ \mathbb{E}_\sigma \left[\left(\sum_{i : X_i \in A} \sigma_i \right)^2 \right] } && \text{(Jensen's inequality)} \\
&= \sqrt{\#\{i : X_i \in A\}},
\end{aligned}$$

where the last line follows because $\mathbb{E}_\sigma\left(\sigma_i \sigma_j\right) = \begin{cases} 0, & i \neq j, \\ 1, & i = j \end{cases}$. If $n_j = \#\{i : X_i \in A_j\}$, then

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\mathcal{H}) &= \frac{1}{n} \sum_{j=1}^{k} \sqrt{n_j} \\
&= \sum_{j=1}^{k} \sqrt{\frac{\widehat{P}(A_j)}{n}},
\end{aligned}$$

where $\widehat{P}(A_j) = \frac{n_j}{n}$. The only inequality in the above derivation was Jensen's inequality, and by the Kintchine-Kahane inequality the reverse inequality holds if we include a multiplicative factor of $\sqrt{2}$, so the calculation is tight up to this factor. The Rachemacher complexity in this example can actually be computed exactly in terms of binomal probabilities. This is left as an exercise.

4 Proof of VC Inequality

To prove the VC inequality, we will focus on bounding $\widehat{R}_n(\mathcal{H})$ in terms of the shatter coefficient.

Theorem 3. *(Massart's Lemma) Let $A \subseteq \mathbb{R}^n$, $|A| \leq \infty$. Set $r = \max_{u\in A} \|u\|_2$. Then*

$$\mathbb{E}_\sigma\left[\frac{1}{n}\sup_{u\in A}\sum_{i=1}^n \sigma_i u_i\right] \leq \frac{r\sqrt{2\ln|A|}}{n},$$

where $u = (u_1, \dots, u_n)^T$.

Proof. $\forall\, t \geq 0$, we have that

$$\begin{aligned}
\exp\left(t\mathbb{E}_\sigma\left[\sup_{u\in A}\sum_{i=1}^n \sigma_i u_i\right]\right) &= \exp\left(\mathbb{E}_\sigma\left[t\sup_{u\in A}\sum_{i=1}^n \sigma_i u_i\right]\right) \\
&\leq \mathbb{E}_\sigma\left[\exp\left(t\sup_{u\in A}\sum_{i=1}^n \sigma_i u_i\right)\right] && \text{(Jensen's inequality)} \\
&= \mathbb{E}_\sigma\left[\sup_{u\in A}\exp\left(t\sum_{i=1}^n \sigma_i u_i\right)\right] && \text{(exponential is strictly increasing)} \\
&\leq \sum_{u\in A}\mathbb{E}_\sigma\left[\exp\left(t\sum_{i=1}^n \sigma_i u_i\right)\right].
\end{aligned}$$

The summand is an MGF. Due to independence,

$$\begin{aligned}
\sum_{u\in A}\mathbb{E}_\sigma\left[\exp\left(t\sum_{i=1}^n \sigma_i u_i\right)\right] &= \sum_{u\in A}\prod_{i=1}^n \mathbb{E}_{\sigma_i}\left[\exp\left(t\sigma_i u_i\right)\right] \\
&\leq \sum_{u\in A}\prod_{i=1}^n \exp\left(t^2\left(2u_i\right)^2/8\right),
\end{aligned}$$

where the bound comes from the following lemma:

Lemma 2. *Let V be a random variable on $\mathbb{R}$ with $\mathbb{E}[V] = 0$ and $V \in [a, b]$ with probability one. Then for all $t > 0$,*

$$\mathbb{E}\left[e^{tV}\right] \leq e^{t^2(b-a)^2/8}.$$

This lemma was given and proved as Lemma 1 in the notes on Hoeffding's inequality. It was used to prove Hoeffding's inequality. In our case, we used $V = \sigma_i u_i$, $a = -u_i$, and $b = u_i$. Continuing with the proof of Massart's lemma,

$$\begin{aligned}
\sum_{u\in A}\prod_{i=1}^n \exp\left(t^2\left(2u_i\right)^2/8\right) &= \sum_{u\in A}\exp\left(\frac{t^2}{2}\sum_{i=1}^n u_i^2\right) \\
&= \sum_{u\in A}\exp\left(\frac{t^2\|u\|_2^2}{2}\right) \\
&\leq \sum_{u\in A}\exp\left(\frac{t^2 r^2}{2}\right) \\
&= |A|\exp\left(\frac{t^2 r^2}{2}\right).
\end{aligned}$$

Taking the log of both sides and dividing by t gives

$$\begin{aligned}
\mathbb{E}_\sigma\left[\sup_{u\in A}\sum_{i=1}^n \sigma_i u_i\right] &\leq \frac{\ln|A|}{t} + \frac{tr^2}{2} \\
&= r\sqrt{2\ln|A|},
\end{aligned}$$

where the last step follows from choosing $t = \frac{\sqrt{2 \ln |A|}}{r}$. Dividing both sides by n completes the proof. □

This theorem is the key result that bridges the gap between VC theory and Rademacher complexity. We first state and prove a one-sided version of the VC inequality.

Theorem 4 (One-sided VC Inequality)**.** *For $0 < \delta < 1$, with probability $\geq 1 - \delta$,*

$$\sup_{h \in \mathcal{H}} \left(R(h) - \widehat{R}_n(h) \right) \leq \sqrt{\frac{8 \left(\ln S_{\mathcal{H}}(n) + \ln \left(1/\delta \right) \right)}{n}}.$$

Equivalently, for any $\epsilon > 0$,

$$\Pr\left(\sup_{h \in \mathcal{H}} \left(R(h) - \widehat{R}_n(h) \right) \geq \epsilon \right) \leq S_{\mathcal{H}}(n) e^{-n\epsilon^2/8}. \tag{6}$$

Proof. Let $\mathcal{H} = \{-1, 1\}^{\mathcal{X}}$ and $S = (X_1, \ldots, X_n) \in \mathcal{X}^n$. Denote $\mathcal{H}|_S = \{(h(X_1), \ldots, h(X_n)) : h \in \mathcal{H}\}$. If $u \in \mathcal{H}|_S$, then $\|u\|_2 = \sqrt{n}$. By Massart's lemma,

$$\begin{aligned} R_n(\mathcal{H}) &= \mathbb{E}_S \left[\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}|_S} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) \right] \right] \\ &\leq \mathbb{E}_S \left[\frac{\sqrt{n 2 \ln |\mathcal{H}|_S|}}{n} \right] \\ &\leq \sqrt{\frac{2 \ln \mathbb{E} \, |\mathcal{H}|_S|}{n}} \qquad (7) \\ &\text{(Jensen's inequality)} \\ &\leq \sqrt{\frac{2 \ln S_{\mathcal{H}}(n)}{n}}, \qquad (8) \end{aligned}$$

where the last step follows from the fact that $|\mathcal{H}|_S| \leq S_{\mathcal{H}}(n)$. From Corollary 1 we deduce that with probability $\geq 1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \widehat{R}_n(h) \right) \leq \sqrt{\frac{2 \ln S_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\ln \left(1/\delta \right)}{2n}}.$$

We now observe that for $a, b \geq 0$, $\sqrt{a} + \sqrt{b} \leq \sqrt{a+b} + \sqrt{a+b} = 2\sqrt{a+b}$. Therefore, for $0 < \delta < 1$, with probability $\geq 1 - \delta$,

$$\begin{aligned} \sup_{h \in \mathcal{H}} \left(R(h) - \widehat{R}_n(h) \right) &\leq \sqrt{\frac{8 \left(\ln S_{\mathcal{H}}(n) + \ln \left(1/\delta \right)/4 \right)}{n}} \qquad (9) \\ &\leq \sqrt{\frac{8 \left(\ln S_{\mathcal{H}}(n) + \ln \left(1/\delta \right) \right)}{n}}. \end{aligned}$$

This establishes the first part of the theorem. To establish the second part, set the right-hand side equal to ϵ and solve for δ (if no such δ exists, the bound holds trivially). □

Remark. Note that step (7) is not necessary. We could have gone directly to (8) using the definition of the shatter coefficient. However, the intermediate result gives a uniform deviation bounds in terms of the expected cardinality of $\mathcal{H}|_S$, which we used to study monotone layers and convex sets in the lecture on VC Theory.

Finally, we state the standard two-sided VC inequality, whose proof is left as an exercise.

Theorem 5 (Two-sided VC Inequality)**.** *For* $0 < \delta < 1$*, with probability* $\geq 1 - \delta$*,*

$$\sup_{h \in \mathcal{H}} \left| R(h) - \widehat{R}_n(h) \right| \leq \sqrt{\frac{8\left(\ln S_{\mathcal{H}}(n) + \ln(2/\delta)\right)}{n}}.$$

Equivalently, for any $\epsilon > 0$*,*

$$\Pr\left(\sup_{h \in \mathcal{H}} \left| R(h) - \widehat{R}_n(h) \right| \geq \epsilon \right) \leq 2 S_{\mathcal{H}}(n) e^{-n\epsilon^2/8}.$$

Exercises

1. Can you improve the constants in the empirical Rademacher complexity bound (2) through a single, direct application of the bounded difference inequality?
2. Determine an exact formula for the empirical Rademacher complexity of the set of classifiers based on a fixed partition (see example above).
3. Let $\mathcal{G}$, $\mathcal{G}_1$, $\mathcal{G}_2$ denote arbitrary classes of functions $\mathcal{Z} \to [a, b]$, and let c, d be arbitrary real numbers. Show
   (a) $\widehat{\mathfrak{R}}_S(c\mathcal{G} + d) = |c|\widehat{\mathfrak{R}}(\mathcal{G})$, where $c\mathcal{G} + d := \{g'(z) = cg(z) + d \,|\, g \in \mathcal{G}\}$.
   (b) $\widehat{\mathfrak{R}}_S(\operatorname{conv}(\mathcal{G})) = \widehat{\mathfrak{R}}_S(\mathcal{G})$, where $\operatorname{conv}(\mathcal{G}) := \{\sum_{i=1}^n \alpha_i g_i \,|\, n \in \mathbb{N}, \alpha_i \geq 0, \sum_i \alpha_i = 1, g_i \in \mathcal{G}\}$.
   (c) $\widehat{\mathfrak{R}}_S(\mathcal{G}_1 + \mathcal{G}_2) = \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + \widehat{\mathfrak{R}}_S(\mathcal{G}_2)$, where $\mathcal{G}_1 + \mathcal{G}_2 := \{g(z) = g_1(z) + g_2(z) \,|\, g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$.
4. Two-sided uniform deviation bounds.
   (a) Prove Theorem 2. *Hint:* Apply the one-sided Rademacher bound again to $-\mathcal{G}$.
   (b) Prove Theorem 5. *Hint:* Observe that $R(-h) = 1 - R(h)$ and similarly for the empirical risk.
   (c) Show that if $\mathcal{G} = -\mathcal{G}$, then the two-sided Rademacher bound holds with the same constants as the one-sided version. In particular, the substitution $\delta \to \delta/2$ is unnecessary.
   (d) Show that if $\mathcal{H} = -\mathcal{H}$, then the two-sided VC inequality holds with the same constants as the one-sided version. In particular, the substitution $\delta \to \delta/2$ is unnecessary.
5. Use inequality (9) to improve the constant in the exponent of (6) at the expense of a larger term in front of the exponential.