1 Rademacher Averages

Recall that we are interested in bounding the difference between empirical and true expectations uniformly over some function class \( G \). In the context of classification or regression, we are typically interested in a class \( G \) that is the loss class associated with some function class \( F \). That is, given a bounded loss function \( \ell : \mathcal{D} \times \mathcal{Y} \to [0, 1] \), we consider the class

\[
\ell_F := \{ (x, y) \mapsto \ell(f(x), y) \mid f \in \mathcal{F} \}.
\]

Rademacher averages give us a powerful tool to obtain uniform convergence results. We begin by examining the quantity

\[
\mathbb{E} \left[ \sup_{g \in G} \left( \mathbb{E} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \right],
\]

where \( Z, \{Z_i\}_{i=1}^{m} \) are i.i.d. random variables taking values in some space \( Z \) and \( G \subseteq [a, b]^{Z} \) is a set of bounded functions. We will later show that the random quantity we are interested in, namely

\[
\sup_{g \in G} \left( \mathbb{E} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right),
\]

will be close to the above expectation with high probability.

Let \( \epsilon_1, \ldots, \epsilon_m \) be i.i.d. \( \{\pm\}\)-valued random variables with \( \mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = 1/2 \). These are also independent of the sample \( Z_1, \ldots, Z_m \). Define the empirical Rademacher average of \( G \) as

\[
\hat{\mathcal{R}}_m(G) := \mathbb{E} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i g(Z_i) \mid Z_1^m \right].
\]

The Rademacher average of \( G \) is defined as

\[
\mathcal{R}_m(G) := \mathbb{E} \left[ \hat{\mathcal{R}}_m(G) \right].
\]

**Theorem 1.1.** We have,

\[
\mathbb{E} \left[ \sup_{g \in G} \left( \mathbb{E} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \right] \leq 2\mathcal{R}_m(G).
\]

**Proof.** Introduce the ghost sample \( Z_1', \ldots, Z_m' \). By that we mean that \( Z_i' \)'s are independent of each other and of \( Z_i \)'s.
and have the same distribution as the latter. Then we have,

\[
\mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \mathbb{E}[g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \right] = \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{m} \sum_{i=1}^{m} \left( \mathbb{E}[g(Z)] - g(Z_i) \right) \right) \right] = \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{m} \sum_{i=1}^{m} \left( g(Z'_i) - g(Z_i) \right) \right) \right] \leq \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{m} \sum_{i=1}^{m} \epsilon_i (g(Z'_i) - g(Z_i)) \right) \right] = \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{m} \sum_{i=1}^{m} \epsilon_i g(Z'_i) \right) \right] + \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i g(Z_i) \right] = 2 \mathcal{R}_m(\mathcal{G}).
\]

Since \( \mathcal{R}_m(-\mathcal{G}) = \mathcal{R}_m(\mathcal{G}) \), we have the following corollary.

**Corollary 1.2.** We have,

\[
\mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{m} \sum_{i=1}^{m} g(Z_i) - \mathbb{E}[g(Z)] \right) \right] \leq 2 \mathcal{R}_m(\mathcal{G}).
\]

Since \( g(X_i) \in [a, b] \),

\[
\sup_{g \in \mathcal{G}} \left( \mathbb{E}[g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right)
\]

does not change by more than \( (b - a)/m \) if some \( Z_i \) is changed to \( Z'_i \). Applying the bounded differences inequality, we get the following corollary.

**Corollary 1.3.** With probability at least \( 1 - \delta \),

\[
\sup_{g \in \mathcal{G}} \left( \mathbb{E}[g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \leq 2 \mathcal{R}_m(\mathcal{G}) + (b - a) \sqrt{\frac{\ln(1/\delta)}{2m}}
\]

Recall that we denote the empirical \( \ell \)-loss minimizer by \( \hat{f}^\ell \). We refer to \( L_\ell(\hat{f}^\ell) - \min_{f \in \mathcal{F}} L_\ell(f) \) as the estimation error. The next theorem bounds the estimation error using Rademacher averages.
2 Expected Regret

Now let us examine the expected regret of the empirical risk minimizer (e.g. analogous to the statistical risk). Let

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} g(Z_i)$$

where $\tau$ is the training set and

$$g^* = \arg\min_{g \in \mathcal{G}} \mathbb{E}[g(Z)]$$

which is true minimizer.

**Lemma 2.1.** The expected regret is:

$$\mathbb{E}[\mathbb{E}[\hat{g}(Z)] - \mathbb{E}[g^*(Z)]] \leq 2\mathcal{R}_m(\mathcal{G}) + \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} g^*(Z_i) - \mathbb{E}[g^*(Z)]\right]$$

where the expectation is with respect $\hat{g}$ (due to randomness in the training set).

**Proof.** Let $\hat{g}$

The expected regret is:

$$\mathbb{E}[\mathbb{E}[\hat{g}(Z)] - \mathbb{E}[g^*(Z)]] \leq \mathbb{E}\left[\mathbb{E}[\hat{g}(Z)] - \frac{1}{m} \sum_{i=1}^{m} \hat{g}(Z_i) + \frac{1}{m} \sum_{i=1}^{m} \hat{g}(Z_i) - \mathbb{E}[g^*(Z)]\right]$$

$$\leq \mathbb{E}\left[\mathbb{E}[\hat{g}(Z)] - \frac{1}{m} \sum_{i=1}^{m} \hat{g}(Z_i) + \frac{1}{m} \sum_{i=1}^{m} g^*(Z_i) - \mathbb{E}[g^*(Z)]\right]$$

$$\leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}[\hat{g}(Z)] - \frac{1}{m} \sum_{i=1}^{m} \hat{g}(Z_i)\right)\right] + \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} g^*(Z_i) - \mathbb{E}[g^*(Z)]\right]$$

$$\leq 2\mathcal{R}_m(\mathcal{G}) + \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} g^*(Z_i) - \mathbb{E}[g^*(Z)]\right]$$

The final claim is straightforward. $\square$

3 Growth function

Consider the case $\mathcal{Y} = \{\pm 1\}$ (classification). Let $\ell$ be the 0-1 loss function and $\mathcal{F}$ be a class of $\pm 1$-valued functions. We can relate the Rademacher average of $\ell_{\mathcal{X}}$ to that of $\mathcal{F}$ as follows.

**Lemma 3.1.** Suppose $\mathcal{F} \subseteq \{\pm 1\}^X$ and let $\ell(y', y) = 1[y' \neq y]$ be the 0-1 loss function. Then we have,

$$\mathcal{R}_m(\ell_{\mathcal{X}}) = \frac{1}{2} \mathcal{R}_m(\mathcal{F}).$$

3
Proof. Note that we can write \( \ell(y', y) \) as \((1 - yy')/2\). Then we have,

\[
\mathcal{R}_m(\ell_{\mathcal{F}}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \epsilon_i \frac{1 - Y_i f(X_i)}{2} \right] X_{1}^m, Y_{1}^m
\]

\[= \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \epsilon_i Y_i f(X_i) \right] X_{1}^m, Y_{1}^m \quad (1)\]

\[= \frac{1}{2} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (-\epsilon_i Y_i) f(X_i) \right] X_{1}^m, Y_{1}^m \]

\[= \frac{1}{2} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \epsilon_i f(X_i) \right] X_{1}^m, Y_{1}^m \quad (2)\]

\[= \frac{1}{2} \mathcal{R}_m(\mathcal{F}).\]

Equation (1) follows because \( \mathbb{E} [\epsilon_i | X_{1}^m, Y_{1}^m] = 0 \). Equation (2) follows because \( -\epsilon_i Y_i \)'s jointly have the same distribution as \( \epsilon_i \)'s.

Note that the Rademacher average of the class \( \mathcal{F} \) can also be written as

\[
\mathcal{R}_m(\mathcal{F}) = \mathbb{E} \left[ \sup_{a \in \mathcal{F}|_{X_{1}^m}} \frac{1}{m} \sum_{i=1}^m \epsilon_i a_i \right],
\]

where \( \mathcal{F}|_{X_{1}^m} \) is the function class \( \mathcal{F} \) restricted to the set \( X_1, \ldots, X_m \). That is,

\[
\mathcal{F}|_{X_{1}^m} := \{ (f(X_1), \ldots, f(X_m)) | f \in \mathcal{F} \}.
\]

Note that \( \mathcal{F}|_{X_{1}^m} \) is finite and

\[
|\mathcal{F}|_{X_{1}^m} | \leq \min\{|\mathcal{F}|, 2^m\}.
\]

Thus we can define the growth function as

\[
\Pi_{\mathcal{F}}(m) := \max_{x_{1}^m \in \mathbb{R}^m} |\mathcal{F}|_{X_{1}^m} |.
\]

The following lemma due to Massart allows us to bound the Rademacher average in terms of the growth function.

**Lemma 3.2.** *(Finite Class Lemma)* Let \( \mathcal{A} \) be some finite subset of \( \mathbb{R}^m \) and \( \epsilon_1, \ldots, \epsilon_m \) be independent Rademacher random variables. Let \( r = \sup_{a \in \mathcal{A}} \|a\| \). Then, we have,

\[
\mathbb{E} \left[ \sup_{a \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m \epsilon_i a_i \right] \leq r \frac{\sqrt{2 \ln |\mathcal{A}|}}{m}.
\]

**Proof.** Let

\[
\mu = \mathbb{E} \left[ \sup_{a \in \mathcal{A}} \sum_{i=1}^m \epsilon_i a_i \right].
\]
We have, for any $\lambda > 0$,

$$e^{\lambda \mu} \leq \mathbb{E} \left[ \exp \left( \lambda \sup_{a \in A} \sum_{i=1}^{m} \epsilon_i a_i \right) \right]$$

Jensen's inequality

$$= \mathbb{E} \left[ \sup_{a \in A} \exp \left( \lambda \sum_{i=1}^{m} \epsilon_i a_i \right) \right]$$

$$\leq \mathbb{E} \left[ \sum_{a \in A} \exp \left( \lambda \sum_{i=1}^{m} \epsilon_i a_i \right) \right]$$

$$= \sum_{a \in A} \mathbb{E} \left[ \exp \left( \lambda \sum_{i=1}^{m} \epsilon_i a_i \right) \right]$$

$$= \sum_{a \in A} \prod_{i=1}^{m} \mathbb{E} \left[ \exp \left( \lambda \epsilon_i a_i \right) \right]$$

$$\leq \sum_{a \in A} \prod_{i=1}^{m} e^{\lambda^2 a_i^2 / 2} \quad \because \text{Hoeffding's lemma}$$

$$= \sum_{a \in A} e^{\lambda^2 \|a\|^2 / 2}$$

$$\leq |A| e^{\lambda^2 r^2 / 2}$$

Taking logs and dividing by $\lambda$, we get that, for any $\lambda > 0$,

$$\mu \leq \ln |A| \frac{1}{\lambda} + \frac{\lambda r^2}{2}.$$ 

Setting $\lambda = \sqrt{2 \ln |A| / r^2}$ gives,

$$\mu \leq r \sqrt{2 \ln |A|},$$

which proves the lemma. \qed