1 Bounded Differences Inequality

Suppose $Z_1, \ldots, Z_m$ are independent random variables taking values in some space $\mathcal{Z}$ and $f: \mathcal{Z}^m \rightarrow \mathbb{R}$ is a function that satisfies, for all $i$,

$$\sup_{z_1, \ldots, z_i, z_i', z_{i+1}, \ldots, z_m} |f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq c_i$$

for some constants $c_1, \ldots, c_m$. Then we have,

$$\mathbb{P} \left( f(Z_1^m) - \mathbb{E} [f(Z_1^m)] \geq t \right) \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^m c_i^2} \right).$$

2 Rademacher Averages

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

$$\phi_{\mathcal{F}} := \{(x, y) \mapsto \phi(f(x), y) \mid f \in \mathcal{F}\}.$$ 

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

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

where $Z, \{Z_i\}_{i=1}^m$ are i.i.d. random variables taking values in some space $\mathcal{Z}$ and $\mathcal{G} \subseteq [a, b]^\mathcal{Z}$ is a set of bounded functions. By the bounded differences inequality, the random quantity we are interested in, namely

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

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 $\mathcal{G}$ as

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

**Theorem 2.1.** We have,
\[
\mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left( \mathbb{E} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \right] \leq 2\mathcal{R}_m(\mathcal{G}).
\]

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

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

**Corollary 2.2.** 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 2.3.** With probability at least \( 1 - \delta \),
\[
\sup_{g \in \mathcal{G}} \left( \mathbb{E} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \leq 2\mathcal{R}_m(\mathcal{G}) + (b - a) \sqrt{\frac{\ln(1/\delta)}{2m}}
\]

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

**Theorem 2.4.** Let \( \phi_{\mathcal{F}} \) denote the loss class associated with \( F \). Then, we have, with probability at least \( 1 - 2\delta \),

\[
L_{\phi}(\hat{f}_{\phi}^*) - \min_{f \in F} L_{\phi}(f) \leq 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}}.
\]

**Proof.** Denote the function in \( F \) with minimum risk by \( f_{\mathcal{F}}^* \). Since the loss function takes values in the interval \([0,1]\), applying the previous corollary to the class \( \phi_{\mathcal{F}} \), we get, with probability at least \( 1 - 2\delta \),

\[
L_{\phi}(\hat{f}_{\phi}^*) - \hat{L}_{\phi}(\hat{f}_{\phi}^*) \leq 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + \sqrt{\frac{\ln(1/\delta)}{2m}}.
\]

Also, by the bounded differences inequality, we have with probability at least \( 1 - \delta \),

\[
\hat{L}_{\phi}(f_{\mathcal{F}}^*) - L_{\phi}(f_{\mathcal{F}}^*) \leq \sqrt{\frac{\ln(1/\delta)}{2m}}.
\]

Thus we have, with probability at least \( 1 - 2\delta \),

\[
L_{\phi}(\hat{f}_{\phi}^*) - L_{\phi}(f_{\mathcal{F}}^*) \leq \hat{L}_{\phi}(\hat{f}_{\phi}^*) - L_{\phi}(f_{\mathcal{F}}^*) + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + \sqrt{\frac{\ln(1/\delta)}{2m}}
\]

\[
\leq \hat{L}_{\phi}(\hat{f}_{\phi}^*) - \hat{L}_{\phi}(\hat{f}_{\phi}^*) + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}}
\]

\[
\leq 0 + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}}
\]

\( \square \)

### 3 Expected Regret and Generalization

**Lemma 3.1.** Let \( \mathcal{F} \) be the class of linear predictors, with the \( L_1 \)-norm of the weights bounded by \( W_1 \). Also assume that with probability one that \( ||x||_\infty \leq X_\infty \). Then

\[
\mathfrak{R}(\mathcal{F}) \leq X_\infty W_1 \sqrt{\frac{2\log d}{m}}
\]

where \( d \) is the dimensionality of \( x \).

**Proof.** Let \( \mathcal{F}_{x_1, x_2, \ldots, x_m} \) be the class:

\[
\{(w \cdot x_1, w \cdot x_2, \ldots, w \cdot x_m) : ||w||_1 \leq W_1\}
\]
Using the definition of the dual norms, we now bound this empirical Rademacher complexity:

$$\mathcal{R}(\mathcal{F}) = \frac{1}{m} \mathbb{E} \left[ \sup_{\|w\|_1 \leq W_1} \sum_{i=1}^{m} \epsilon_i w \cdot x_i \right]$$

$$= \frac{1}{m} \mathbb{E} \left[ \sup_{\|w\|_1 \leq W_1} w \cdot \sum_{i=1}^{m} \epsilon_i x_i \right]$$

$$= \frac{W_1}{m} \mathbb{E} \left[ \left\| \sum_{i=1}^{m} \epsilon_i x_i \right\|_\infty \right]$$

$$= \frac{W_1}{m} \mathbb{E} \left[ \sup_{j} \sum_{i=1}^{m} \epsilon_i [x_i]_j \right]$$

$$\leq \frac{W_1 \sqrt{2 \log d}}{m} \sup_{j} \sqrt{\sum_{i=1}^{m} [x_i]_j^2}$$

$$\leq X_\infty W_1 \sqrt{\frac{2 \log d}{m}}$$

where we have used Massart’s finite lemma.

3.1 Generalization

**Corollary 3.2.** Under the assumptions above, for the L2 case, we have:

$$\mathcal{L}(\hat{w}_2) - \arg \min_{w: \|w\|_2 \leq W_2} \mathcal{L}(w) \leq 2L_\phi X_2 W_2 \frac{X_\infty}{\sqrt{m}} + 2 \sqrt{\frac{\log 2/\delta}{2m}}$$

and for the L1 case, we have:

$$\mathcal{L}(\hat{w}_1) - \arg \min_{w: \|w\|_1 \leq W_1} \mathcal{L}(w) \leq 2L_\phi X_\infty W_1 \sqrt{\frac{2 \log d}{m}} + 2 \sqrt{\frac{\log 2/\delta}{2m}}$$

The proof just follow from the previous lemmas, along with our Rademacher bound for loss classes.