13.1 Rademacher Complexity

Given a function class $G : \mathcal{X} \rightarrow \mathbb{R}$, let $\sigma_1, \ldots, \sigma_m$ be i.i.d. Rademacher random variables, that is $\sigma_i \in \{-1, 1\}$ with $P(\sigma_i = 1) = 1/2$, and let $S = (x_1, \ldots, x_m)$ be a sample from $\mathcal{X}$. Then the empirical Rademacher complexity is defined as:

\[ \hat{R}_S(G) = \mathbb{E}_{\sigma} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i) \right], \]

and the Rademacher complexity is defined as:

\[ R_m(G) = \mathbb{E}_{S \sim D^m} \left[ \hat{R}_S(G) \right]. \]

We note that the Rademacher complexity is distribution-specific.

Based on Rademacher complexity, we can show the following generalization bound:

**Theorem 13.1.** Let $G$ be a function class mapping $\mathcal{X}$ to $[0, 1]$. Then, with probability at least $1 - \delta$ and for all $g \in G$,

\[ \mathbb{E}_{x \sim D}[g(x)] \leq \frac{1}{m} \sum_{i=1}^{m} g(x_i) + 2R_m(G) + \sqrt{\frac{\log(1/\delta)}{2m}}, \]

where $S = (x_1, \ldots, x_m) \sim D^m$.

The proof of the above theorem requires McDiarmid’s inequality, which is presented as following:

**Theorem 13.2 (McDiarmid’s inequality).** Let $D$ be a distribution on $\mathcal{X}$, and let $f$ be a function taking finite subsets of $\mathcal{X}$ as input. Suppose that $f$ satisfies bounded difference condition with the uniform constant $c$, i.e.,

\[ |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c \]

. Then with probability at least $1 - \delta$,

\[ f(S) - \mathbb{E}_{S \sim D^m}[f(S)] \leq \sqrt{\frac{mc^2}{2} \log(1/\delta)}, \]

where $S \sim D^m$.

**Proof:** (Sketch) Let $S = (x_1, \ldots, x_m) \sim D^m$. We define a martingale $Z_i = \mathbb{E}[f(S) - f(S)|x_1, \ldots, x_{i-1}]$. It is easy to see that $|Z_i - Z_{i-1}| \leq c$ for all $i$. Then applying Azuma’s inequality to the martingale difference sequence $\{Z_i\}$ yields the desired result. See Appendix D of the textbook *Foundation of Machine Learning* for a full proof.

We are ready to prove Theorem 13.1.
Proof of Theorem 13.1: To ease some notations, we define: $E := E_{x \sim D}[g(x)]$, $\hat{E}g := \frac{1}{|S|} \sum_{x \in S} g(x)$, and $\Phi(S) := \sup_{g \in G}(Eg - \hat{E}g)$.

The proof is composed of two parts:

1. $\Phi(S) \leq E_{S \sim D^{m}}[\Phi(S)] + \sqrt{\frac{\log(1/\delta)}{2m}}$.
2. $E_{S \sim D^{m}}[\Phi(S)] \leq 2R_{m}(G)$.

For part 1, we begin with showing that $|\Phi(S) - \Phi(S')| \leq \frac{1}{m}$ when $S$ and $S'$ differ by one element (and let it be the $i^{th}$ one):

$$\Phi(S) - \Phi(S') = \sup_{g \in G}(Eg - \hat{E}g) - \sup_{g \in G}(Eg - \hat{E}g')$$

$$\leq \sup_{g \in G}(Eg - \hat{E}g - Eg + \hat{E}g')$$

$$= \sup_{g \in G}\frac{g(x'_{i}) - g(x_{i})}{m}$$

$$\leq \frac{1}{m}$$

The first inequality holds since supremum of difference is greater than difference of supremum.

By symmetry, we have $|\Phi(S) - \Phi(S')| \leq \frac{1}{m}$. Then, by setting $c = \frac{1}{m}$, applying McDiarmid’s inequality yields the desired inequality.

For part 2, we use the two-sample trick. Let $S' = (x'_{1}, \ldots, x'_{n}) \sim D^{m}$.

$$E_{S \sim D^{m}}[\Phi(S)] = E_{S, S' \sim D^{m}}\left[\sup_{g \in G}(Eg - \hat{E}g)\right]$$

$$\leq E_{S, S' \sim D^{m}}\left[\sup_{g \in G}(\hat{E}g' - \hat{E}g)\right]$$

$$= E_{S, S' \sim D^{m}}\left[\sup_{g \in G}\frac{1}{m} \sum_{i=1}^{m}(g(x'_{i}) - g(x_{i}))\right]$$

$$= E_{S', S' \sim D^{m}, \sigma}\left[\sup_{g \in G}\frac{1}{m} \sum_{i=1}^{m} \sigma_{i}(g(x'_{i}) - g(x_{i}))\right]$$

$$\leq E_{S', S' \sim D^{m}, \sigma}\left[\sup_{g \in G}\frac{1}{m} \sum_{i=1}^{m} \sigma_{i}g(x'_{i})\right] + E_{S \sim D^{m}, \sigma}\left[\sup_{g \in G}\frac{1}{m} \sum_{i=1}^{m} -\sigma_{i}g(x_{i})\right]$$

$$= 2R_{m}(G)$$

Combining the two parts gives us

$$E_{x \sim D}[g(x)] \leq \frac{1}{m} \sum_{i=1}^{m} g(x_{i}) + 2R_{m}(G) + \sqrt{\frac{\log(1/\delta)}{2m}}$$

The proof is complete.

13.2 Generalization Bound for Binary Classification

Given a hypothesis class $\mathcal{H}$ with functions taking $\pm 1$ values, the associated loss class of $\mathcal{H}$ is defined as:

$$G := \{g_{h}(x, y) = 1| h(x) \neq y | h \in \mathcal{H}\}.$$
Lemma 13.3. For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, we have $\hat{R}_S(G) = \frac{1}{2} \hat{R}_{S|X}(\mathcal{H})$, where $S \upharpoonright X = (x_1, \ldots, x_m)$.

Proof: The proof is easy. See Lemma 3.1 in the textbook.

The following theorem demonstrates an application of Rademacher complexity that provides us a generalization bound for binary classification.

Theorem 13.4. For binary classification with 0-1 loss, let $\mathcal{H}$ be a class hypothesis mapping $X$ to $\{-1, 1\}$. Then with probability $\geq 1 - \delta$, for any $h \in \mathcal{H}$, we have:

$$R(h) \leq \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $S \sim \mathcal{D}^m$.

Proof: This directly follows from Theorem 13.1 and Lemma 13.3.

13.3 Massart’s Lemma

Lastly, we present Massart’s lemma, which gives us a better expression of $\mathcal{R}_m(\cdot)$.

Theorem 13.5 (Massart’s lemma). Let $A \subseteq \mathbb{R}^m$ be a finite set of points with $r = \max_{x \in A} \|x\|_2$. Then we have

$$\mathbb{E}_\sigma \left[ \max_{x \in A} \sum_{i=1}^m x_i \sigma_i \right] \leq r \sqrt{2 \log(|A|)},$$

where $(x_1, \ldots, x_n)$ is a vector in $A$.

Proof: Let $t > 0$ be a number to be chosen later.

$$\exp\left(t \mathbb{E}_\sigma \left[ \max_{x \in A} x^\top \sigma \right] \right) \leq \mathbb{E}_\sigma \left[ \exp(t \max_{x \in A} x^\top \sigma) \right]$$

(Jensen’s inequality)

$$\leq \mathbb{E}_\sigma \left[ \sum_{x \in A} \exp(tx^\top \sigma) \right]$$

(summation $\geq$ maximum)

$$= \sum_{x \in A} \mathbb{E}_\sigma \left[ \exp(tx^\top \sigma) \right]$$

$$= \sum_{x \in A} \prod_{i=1}^m \mathbb{E}_\sigma \left[ \exp(tx_i \sigma_i) \right]$$

$$= \sum_{x \in A} \prod_{i=1}^m \mathbb{E}_\sigma \left[ \exp(tx_i \sigma_i) \right]$$

(applying Hoeffding’s lemma)

$$\leq \prod_{x \in A} \exp\left(\frac{2tx_i^2}{8}\right)$$

$$= \sum_{x \in A} \exp\left(\frac{t^2}{2} \sum_{i=1}^m x_i^2\right)$$

(recall that $r = \max_{x \in A} \|x\|_2$)

Taking logarithm, and dividing by $t$ on both sides, we get
\begin{align*}
\mathbb{E}_\sigma \left[ \max_{x \in A} x^\top \sigma \right] & \leq \frac{\log(|A|)}{t} + \frac{tr^2}{2}.
\end{align*}

It is minimized when taking \( t = \sqrt{\frac{\log(|A|)}{r^2/2}} = \frac{\sqrt{2\log(|A|)}}{r} \), and it leads to the bound:

\begin{align*}
\mathbb{E}_\sigma \left[ \max_{x \in A} x^\top \sigma \right] & \leq r \sqrt{2 \log(|A|)}.
\end{align*}