MATH/COMP 562 LECTURE NOTES RADEMACHER COMPLEXITY

ADAM M. OBERMAN

1. Rademacher Complexity

This section adapted from [MRT18, Section 3.1]

1.1. **Setup.**

Definition 1.1 (Empirical Rademacher complexity)**.** Let $\mathcal{G}$ be a family of functions mapping from z to $[a, b]$ and $S = (z_1, \ldots, z_m)$ a fixed sample of size m with elements in 2. Then, the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample S is defined as:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g\left(z_i\right)\right],$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$, with $\sigma_i s$ independent uniform random variables taking values in $\{-1, +1\}$.[3] The random variables σ_i are called Rademacher variables.

Definition 1.2 (Rademacher complexity)**.** Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:

$$\mathfrak{R}_m(\mathcal{G}) = \underset{S \sim \mathcal{D}^m}{\mathbb{E}}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$$

Definition 1.3. For any sample $S = (z_1, \ldots, z_m)$ and any $g \in \mathcal{G}$, we denote by $\widehat{\mathbb{E}}_S[g]$ the empirical average of g over S

$$\widehat{\mathbb{E}}_S[g] = \frac{1}{m} \sum_{i=1}^{m} g\left(z_i\right)$$

Lemma 1.4. *The function $H(S) = \widehat{\mathfrak{R}}_S(\mathcal{G})$ satisfies the bounded differences inequality,*

$$|H(S) - H\left(S'\right)| \leq \frac{b-a}{m} \tag{1}$$

Proof. By definition, changing one point in S changes $\widehat{\mathfrak{R}}_S(\mathcal{G})$ by at most $(b-a)/m$ □

Definition 1.5. Given any $m \geq 1$ and any dataset $S = S^m \subset \mathcal{X}^m$ define the function

$$\Phi(S) = \Phi(S, \mathcal{D}) = \sup_{g \in \mathcal{G}}\left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g]\right) \tag{2}$$

which is the worst generalization gap over for functions over the dataset S

Lemma 1.6 (Difference of sup)**.** *Let $f, g : \mathcal{X} \to \mathbb{R}$ be bounded. For any $I \subset \mathcal{X}$.*

$$\sup_{x \in I} f(x) - \sup_{x \in I} g(x) \leq \sup_{x \in I}\{f(x) - g(x)\} \tag{3}$$

Date: April 12, 2023.

Proof. Let $g^* = \sup_{x\in I} g(x)$. Then

$$\begin{aligned}
\sup_{x\in I} f(x) - \sup_{x\in I} g(x) &= \sup_{x\in I} f(x) - g^* \\
&\le \sup_{x\in I}\{f(x) - g(x)\} && \text{by definition of the supremum}
\end{aligned}$$

□

Lemma 1.7. *The function* Φ *defined by* (2) *satisfies the bounded differences inequality,*

$$|\Phi(S) - \Phi(S')| \le \frac{b-a}{m} \tag{4}$$

Proof. Let S and S' be two samples differing by exactly one point, say z_m in S and z'_m in S'. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\begin{aligned}
\Phi(S') - \Phi(S) &\le \sup_{g\in\mathcal{G}} \left(\widehat{\mathbb{E}}_S[g] - \widehat{\mathbb{E}}_{S'}[g]\right) && \text{by (3)} \\
&= \sup_{g\in\mathcal{G}} \frac{g(z_m) - g(z'_m)}{m} && \text{since } S, S' \text{ differ at one point} \\
&\le \frac{b-a}{m} && \text{since } g(z)\in[a,b]
\end{aligned}$$

Similarly, we can obtain $\Phi(S) - \Phi(S') \le (b-a)/m$, thus (4) holds. □

1.2. **Expectation of Phi.**

Theorem 1.8. *The function* Φ *defined by* (2) *satisfies*

$$\mathop{\mathbb{E}}_S[\Phi(S)] \le 2\mathfrak{R}_m(\mathcal{G})$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_S[\Phi(S)] &= \mathop{\mathbb{E}}_S\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S(g)\right)\right] && \text{by definition} \\
&= \mathop{\mathbb{E}}_S\left[\sup_{g\in\mathcal{G}} \mathop{\mathbb{E}}_{S'}\left[\widehat{\mathbb{E}}_{S'}(g) - \widehat{\mathbb{E}}_S(g)\right]\right] && \text{points in } S' \text{ sampled i.i.d. thus } \mathbb{E}[g] = \mathbb{E}_{S'}\left[\widehat{\mathbb{E}}_{S'}(g)\right] \\
&\le \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\left(\widehat{\mathbb{E}}_{S'}(g) - \widehat{\mathbb{E}}_S(g)\right)\right] && \text{sub-additivity of sup} \\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{S}} \frac{1}{m}\sum_{i=1}^m \left(g(z'_i) - g(z_i)\right)\right] && \text{by definition} \\
&= \mathop{\mathbb{E}}_{\sigma,S,S'}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \sigma_i\left(g(z'_i) - g(z_i)\right)\right]
\end{aligned}$$

For the last equation, we introduce Rademacher variables σ_i, which are uniformly distributed independent random variables taking values in $\{-1,+1\}$. This does not change the expectation appearing in (3.10): when $\sigma_i = 1$, the associated summand remains unchanged; when $\sigma_i = -1$, the associated summand flips signs, which is equivalent to swapping z_i and z_i' between S and S'. Since we are taking the expectation over all possible S and S', this swap does not affect the overall expectation; we are simply changing the order of the summands within the expectation.

In the next inequality, we will use the sub-additivity of the supremum function

$$\sup(U+V) \leq \sup(U) + \sup(V) \tag{5}$$

Continue from the last equation above,

$$\begin{aligned}
&\leq \underset{\sigma,S'}{\mathbb{E}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g\left(z_i'\right)\right] + \underset{\sigma,S}{\mathbb{E}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i g\left(z_i\right)\right] && \text{by (5)}\\
&= 2\underset{\sigma,S}{\mathbb{E}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g\left(z_i\right)\right]\\
&= 2\mathfrak{R}_m(\mathcal{G}) && \text{by defn}
\end{aligned}$$

stems from the definition of Rademacher complexity and the fact that the variables σ_i and $-\sigma_i$ are distributed in the same way. □

1.3. **Putting it together.**

Theorem 1.9. *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g\left(z_i\right) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{6}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g\left(z_i\right) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{7}$$

Proof. Using (2), we first established the bounded differences inequality for Φ, (4). This allows us to apply McDiarmid's inequality. For any $\delta > 0$,

$$\Phi(S) \leq \underset{S}{\mathbb{E}}[\Phi(S)] + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \quad \text{with probability at least } 1-\delta/2$$

using δ instead of $\delta/2$.

Using the Theorem 1.8, and the definition (2), this becomes

$$\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g]\right) \leq 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \quad \text{with probability at least } 1-\delta/2$$

since, this holds for any $g \in \mathcal{G}$, we obtain (6).

To derive a bound in terms of the empirical Rademacher complexity, $\widehat{\mathfrak{R}}_S(\mathcal{G})$, We use (1) from Lemma 1.4. This allows us to use McDiarmid's inequality. Thus,

$$\mathfrak{R}_m(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \quad \text{with probability } 1-\delta/2$$

Finally, we use the union bound to combine two inequalities above, which yields with probability at least $1-\delta$:

$$\Phi(S) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

which matches (7). ☐

2. Rademacher Complexity for Linear Hypotheses

Definition 2.1. Define $B_r = \{x \in \mathbb{R}^d \mid \|x\| \leq r\}$. Let $S = \{x_1, \dots x_m\} \subset B_r$ Consider the linear functions, $h(w,x) = w \cdot x$ and define

$$\mathcal{H}_\Lambda = \{h(x,w) = w \cdot x \mid x \in X, w \in B_\Lambda\}.$$

Theorem 2.2 (Theorem 5.10 of Mohri)**.** *The empirical Rademacher complexity of $\mathcal{H}_\Lambda$ is bounded as follows,*

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_\Lambda) \leq \frac{r\Lambda}{\sqrt{m}}$$

Proof. The proof follows through a series of inequalities:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|\mathbf{w}\|\leq\Lambda} \sum_{i=1}^m \sigma_i \mathbf{w}\cdot\mathbf{x}_i)\right] && \text{by defn}\\
&= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|\mathbf{w}\|\leq\Lambda} \mathbf{w}\cdot \sum_{i=1}^m \sigma_i \mathbf{x}_i\right] && \text{since } h \text{ is linear}\\
&\leq \frac{\Lambda}{m}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|\right] && \text{Cauchy-Schwarz and } \|w\| \leq \Lambda\\
&\leq \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|^2\right]\right]^{\frac{1}{2}} && \text{Jensen's inequality}\\
&= \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[\sum_{i,j=1}^m \sigma_i\sigma_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)\right]\right]^{\frac{1}{2}} &&\\
&\leq \frac{\Lambda}{m}\left[\sum_{i=1}^m \|\mathbf{x}_i\|^2\right]^{\frac{1}{2}} && \mathbb{E}\left[\sigma_i\sigma_j\right] = \mathbb{E}\left[\sigma_i\right]\mathbb{E}\left[\sigma_j\right] = 0 \text{ for } i \neq j\\
&\leq \frac{\Lambda\sqrt{mr^2}}{m} && \|\mathbf{x}_i\| \leq r\\
&= \frac{r\Lambda}{\sqrt{m}}
\end{aligned}$$

☐

References

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.