CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 15: Rademacher Complexity (October 26)

Lecturer: Csaba Szepesvári *Scribes: Kushagra Chandak*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
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.*

Motivation. Using Rademacher complexity and chaining, we can get rid of the $\log n$ factor from the uniform convergence bounds. Using Rademacher complexity, we can bound the expected supremum of the underlying empirical process and then use concentration inequalities (like McDiarmid's) to get high probability bounds. First, we define the *uniform convergence complexity (one-sided)* or *expected maximum deviation (one-sided)*.

Definition 15.1 (Expected max deviation)**.** Let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and $P \in \mathcal{M}_1(\mathcal{Z})$. Then the expected maximum deviation for class $\mathcal{G}$ with respect to P is defined as

$$\varepsilon_n(\mathcal{G}, P) = \mathbb{E}[\sup_{g \in \mathcal{G}} Pg - P_n g] \,.$$

Proposition 15.2. *Let $g_n := \arg\min_{g \in \mathcal{G}} P_n g$ be the ERM map. Then*

$$\mathbb{E}[Pg_n] \leq \inf_{g \in \mathcal{G}} Pg + \varepsilon_n(\mathcal{G}, P) \,.$$

Proof. We start by adding and subtracting $P_n g_n$ to Pg_n:

$$\begin{aligned} Pg_n &= P_n g_n + (P - P_n) g_n \\ &\leq P_n g^* + (P - P_n) g_n && (g^* = \arg\min_{g \in \mathcal{G}} Pg) \\ &\leq P_n g^* + \sup_{g \in \mathcal{G}} (P - P_n) g \,. \end{aligned}$$

Taking expectation on both sides gives us the result. □

Now we are ready to define Rademacher complexity and relate it to $\varepsilon_n(\mathcal{G}, P)$.

Definition 15.3 (At sample Rademacher complexity)**.** Let $z_{1:n} \in \mathcal{Z}$ be a fixed sequence of length n in $\mathcal{Z}$ and $\sigma \sim \mathrm{Rad}(n)$ be a vector of $\{\pm 1\}$ random signs. Then the Rademacher complexity for class $\mathcal{G}$ at $z_{1:n}$ is defined as

$$R(\mathcal{G}, z_{1:n}) = \mathbb{E}[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)] \,.$$

For $Z_{1:n} \sim P^{\otimes n}$, the Rademacher complexity for $\mathcal{G}$ w.r.t to P is defined as

$$R_n(\mathcal{G}, P) = \mathbb{E} R(\mathcal{G}, Z_{1:n}) \,.$$

To relate $\varepsilon_n(\mathcal{G}, P)$ with $R_n(\mathcal{G}, P)$, we have the following nice theorem.

Theorem 15.4.

$$\varepsilon_n(\mathcal{G}, P) \leq 2 R_n(\mathcal{G}, P) \,.$$

Intuition for Rademacher complexity. Rademacher complexity measures the ability of $\mathcal{G}$ to fit to random symmetric noise (Rademacher noise). If $R_n(\mathcal{G}, P)$ is close to 0, the capacity of the class is bounded ($\mathcal{G}$ is less expressive) and if $R_n(\mathcal{G}, P)$ is close to 1 (if $\mathcal{G} = \{-1, 1\}$) then the capacity of $\mathcal{G}$ is unbounded. Consider the limiting case $\mathcal{G} = \{g\}$. In that case, $R(\mathcal{G}, z_{1:n}) = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)] = 0$.