CS861: Theoretical Foundations of Machine Learning Lecture 4 - 09/11/2024
University of Wisconsin–Madison, Fall 2024

Lecture 04: Rademacher Complexity & Growth Function

Lecturer: Kirthevasan Kandasamy *Scribed by: Song Bian, Minghao Yan*

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.*

In this lecture, we first introduce the **Rademacher complexity**, which can be applied to derive an upper bound for $\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(\hat{R}_S(h)-R(h))\right]$. After that, we will state a bound for PAC learning. Then, we will take a digression to introduce **sub-Gaussian random variables** and **maximal inequality**. Finally, we will introduce the **growth function**.

1 Rademacher Complexity

We begin by introducing the formal definition of Rademacher complexity.

Definition 1. *Given a hypothesis class $\mathcal{H}$ and a sample size $n\in\mathbb{N}$, the Rademacher complexity of $\mathcal{H}$ is defined as:*

$$\mathrm{Rad}_n(\mathcal{H}) = \mathbb{E}_{S\sim\mathbb{P}_{xy}^n}\left[\widehat{\mathrm{Rad}}(S,\mathcal{H})\right] = \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(h(x_i),y_i)\right]$$

where $\sigma=(\sigma_1,\ldots,\sigma_n)$ is a vector of independent Rademacher random variables, and $\ell(h(x_i),y_i)$ is the loss associated with hypothesis h on the example (x_i,y_i).

Lemma 1. *Given a hypothesis class $\mathcal{H}$ and a sample size $n\in\mathbb{N}$, the following inequality holds:*

$$\mathbb{E}_{S\sim\mathbb{P}_{xy}^n}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_S(h)-R(h)\right)\right]\le 2\mathrm{Rad}_n(\mathcal{H})$$

where $\hat{R}_S(h)$ is the empirical risk over the sample S and $R(h)$ is the true risk of hypothesis h.

Lemma 1 will be useful in deriving the PAC learning bound for ERM (Empirical Risk Minimization), which we will discuss in the next section. Below, we provide the proof of Lemma 1.

Proof

$$\begin{aligned}
LHS &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_S(h)-\mathbb{E}_{S'}[\hat{R}_{S'}(h)]\right)\right] \quad \text{(where } S' \text{ is independent from } S \text{ and drawn from } \mathbb{P}_{xy}^n\text{)}\\
&= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}\left[\hat{R}_S(h)-\hat{R}_{S'}(h)\right]\right] \quad \text{(by linearity of expectation)}\\
&\le \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_S(h)-\hat{R}_{S'}(h)\right)\right] \quad \text{(applying subadditivity: } \sup\mathbb{E}\le\mathbb{E}\sup\text{)}\\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i),y_i)-\frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i'),y_i')\right)\right] \quad \text{(the definition of } \hat{R}_S(h) \text{ and } \hat{R}_{S'}(h)\text{)}\\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\left[\ell(h(x_i),y_i)-\ell(h(x_i'),y_i')\right]\right] \quad \text{(writing the difference between empirical risks as a sum)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(\sigma_i \ell(h(x_i),y_i) - \sigma_i \ell(h(x_i'),y_i'))\right] \quad (\forall \sigma, \text{ we have equality in expectation}) \\
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(h(x_i),y_i) + \frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\ell(h(x_i'),y_i')\right)\right] \quad (\text{splitting the sum into two parts}) \\
&\le \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(h(x_i),y_i)\right] + \mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(h(x_i'),y_i')\right] \quad (\sup(a+b) \le \sup a + \sup b) \\
&= \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(h(x_i),y_i)\right] + \mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(h(x_i'),y_i')\right] \quad (\text{since } \sigma_i \text{ is symmetric}) \\
&= 2\mathrm{Rad}_n(\mathcal{H}) \quad (\text{by the definition of Rademacher complexity}).
\end{aligned}
$$

Thus, we have shown that the left-hand side is bounded by $2\mathrm{Rad}_n(\mathcal{H})$. □

1.1 PAC Learning Bound for ERM

Theorem 1. *Let $\mathcal{H}$ be a hypothesis class with finite Rademacher complexity $\mathrm{Rad}_n(\mathcal{H})$. Let $\hat{h}$ be the hypothesis obtained via ERM using an i.i.d dataset of n samples. For any $\varepsilon > 0$, there exist universal constants c_1 and c_2 such that with probability at least $1 - 2e^{-2n\varepsilon^2}$, we have:*

$$R(\hat{h}) \le \inf_{h\in\mathcal{H}} R(h) + c_1 \mathrm{Rad}_n(\mathcal{H}) + c_2\varepsilon$$

You will prove this theorem in the next homework. The following ideas may help prove this.

- In the case where there exists an $h^* \in \mathcal{H}$ such that $R(h^*) = \inf_{h\in\mathcal{H}} R(h)$, we can use the following decomposition:

$$
\begin{aligned}
R(\hat{h}) - R(h^*) &= \left(R(\hat{h}) - \hat{R}(\hat{h})\right) + \left(\hat{R}(\hat{h}) - R(h^*)\right) \\
&\le \underbrace{R(\hat{h}) - \hat{R}(\hat{h})}_{T_1} + \underbrace{\hat{R}(h^*) - R(h^*)}_{T_2}
\end{aligned}
$$

  Using McDiarmid's inequality, we can bound both terms T_1 and T_2. McDiarmid's inequality gives us

$$\mathbb{P}(\forall h \in \mathcal{H}, \hat{R}(h) - R(h) \le \mathbb{E}_S[\sup_{h\in\mathcal{H}}(\hat{R}(h) - R(h))] + \varepsilon) \le 2\mathrm{Rad}_n(\mathcal{H})$$

- If there is no $h^* \in \mathcal{H}$ such that $R(h^*) = \inf_{h\in\mathcal{H}} R(h)$, we need to handle this case separately. This part will not be shown here.

Now we have a PAC bound in terms of the Rademacher complexity, why don't we stop here?

1. It is difficult to directly bound the Rademacher complexity for a given hypothesis class $\mathcal{H}$. On the other hand, the VC dimension is easier to bound.
2. The VC dimension also makes it more straightforward to prove lower bounds on the sample complexity.

2 Sub-Gaussian Random Variables and the Maximal Inequality

Definition 2. *A random variable X is said to be a σ-sub-Gaussian random variable (denoted as σ-sub-Gaussian) if it satisfies the following condition for all $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

Intuitively, a σ-sub-Gaussian random variable has tails that decay at least as fast as the tails of a Gaussian random variable with mean 0 and variance σ^2.

Lemma 2. *If X is a σ-sub-Gaussian random variable, then for any constant a, aX is $(a\sigma)$-sub-Gaussian.*

Proof We start with the expectation for the scaled random variable aX:

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(aX-\mathbb{E}[aX])}\right] &= \mathbb{E}\left[e^{\lambda a(X-\mathbb{E}[X])}\right] \quad \text{(factoring out the constant } a\text{)}\\ &= \mathbb{E}\left[\left(e^{\lambda a(X-\mathbb{E}[X])}\right)\right] \quad \text{(same expectation as above)}\\ &\leq e^{\frac{\lambda^2(a\sigma)^2}{2}} \quad \text{(since } X \text{ is } \sigma\text{-sub-Gaussian).}\end{aligned}$$

This shows that aX is $(a\sigma)$-sub-Gaussian. □

Lemma 3. *If X_1 and X_2 are independent σ_1-sub-Gaussian and σ_2-sub-Gaussian random variables, respectively, then $X_1 + X_2$ is $\left(\sqrt{\sigma_1^2+\sigma_2^2}\right)$-sub-Gaussian.*

Lemma 4 (Tail Bound)**.** *If X is a σ-sub-Gaussian random variable, then for any $\varepsilon > 0$, the following tail bounds hold:*

$$\mathbb{P}(X - \mathbb{E}[X] > \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}},$$
$$\mathbb{P}(X - \mathbb{E}[X] < -\varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}},$$

and as a consequence,

$$\mathbb{P}(|X - \mathbb{E}[X]| > \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}}.$$

Proof We will prove the first inequality $\mathbb{P}(X - \mathbb{E}[X] > \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$.

Assume without loss of generality that $\mathbb{E}[X] = 0$. Then:

$$\begin{aligned}\mathbb{P}(X > \varepsilon) &= \mathbb{P}(e^{\lambda X} > e^{\lambda\varepsilon}) \quad \text{(apply exponential to both sides)}\\ &\leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda\varepsilon}} \quad \text{(apply Markov's inequality)}\\ &\leq e^{\frac{\lambda^2\sigma^2}{2}-\lambda\varepsilon} \quad \text{(since } X \text{ is } \sigma\text{-sub-Gaussian).}\end{aligned}$$

This bound holds for any $\lambda \in \mathbb{R}$. Choosing $\lambda = \frac{\varepsilon}{\sigma^2}$ optimizes the bound and yields:

$$\mathbb{P}(X > \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}.$$

□

Example 2 (Gaussian Random Variables are Sub-Gaussian)**.** Let $X \sim N(\mu, \sigma^2)$. Then X is σ-sub-Gaussian. This holds because the moment generating function of a Gaussian random variable $X \sim N(\mu, \sigma^2)$ is exactly:

$$\mathbb{E}[e^{\lambda(X-\mu)}] = e^{\frac{\lambda^2\sigma^2}{2}}$$

which satisfies the sub-Gaussian definition.
Example 3 (Bounded Random Variables are Sub-Gaussian)**.** Let X be a bounded random variable with support $\mathrm{supp}(X) \subseteq [a, b]$, where $-\infty < a \leq b < \infty$. Then X is $\left(\frac{b-a}{2}\right)$-sub-Gaussian.
This holds because Hoeffding's lemma implies that for any bounded random variable, the tail decay is at least as fast as a Gaussian distribution with variance proportional to the range of the variable.
Properties:

1. Let $a \in \mathbb{R}$. If X is σ-sub-Gaussian, then aX is $|a|\sigma$-sub-Gaussian.
   (This follows directly from the definition of sub-Gaussian random variables, where scaling by a multiplies the variance by a^2.)

2. If X_1 and X_2 are independent sub-Gaussian random variables with constants σ_1 and σ_2, respectively, then $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-sub-Gaussian.
   (This result follows from the property that the sum of independent sub-Gaussian random variables is also sub-Gaussian, with the variance being the sum of the individual variances.)

3. Let $X_1, \dots, X_n$ be independent σ-sub-Gaussian random variables. Then, for any $\varepsilon > 0$, the following concentration inequality holds:

   $$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

   (This is a consequence of Hoeffding's inequality, which provides a bound for the probability that the sum of independent sub-Gaussian random variables deviates from its expectation.)

2.1 Maximal Inequality for Sub-Gaussian Random Variables

Lemma 5 (Maximal Inequality)**.** *Let $Z_1, \dots, Z_n$ be σ-sub-Gaussian random variables (not necessarily independent). Then, the expected maximum of these random variables is bounded as follows:*

$$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \leq \sigma\sqrt{2\log(n)}.$$

Proof We begin by noting that for any $t > 0$, the following inequality holds:

$$\exp\left(t\mathbb{E}\left[\max_{i\in[n]} Z_i\right]\right) \leq \mathbb{E}\left[\exp\left(t\max_{i\in[n]} Z_i\right)\right].$$

This inequality is due to Jensen's inequality, which holds because the exponential function is convex.

Next, we use the fact that the maximum of exponentials is less than or equal to the sum of exponentials:

$$\mathbb{E}\left[\exp\left(t\max_{i\in[n]} Z_i\right)\right] = \mathbb{E}\left[\max_{i\in[n]} \exp(tZ_i)\right] \leq \mathbb{E}\left[\sum_{i=1}^{n}\exp(tZ_i)\right].$$

Since each Z_i is σ-sub-Gaussian, we know that:

$$\mathbb{E}\left[\exp(tZ_i)\right] \leq \exp\left(\frac{t^2\sigma^2}{2}\right).$$

Therefore, summing over n independent random variables gives:

$$\mathbb{E}\left[\sum_{i=1}^{n}\exp(tZ_i)\right] \leq n\exp\left(\frac{t^2\sigma^2}{2}\right).$$

Thus, we have the following inequality:

$$\exp\left(t\mathbb{E}\left[\max_{i\in[n]} Z_i\right]\right) \leq n\exp\left(\frac{t^2\sigma^2}{2}\right).$$

Taking the logarithm of both sides gives:

$$t\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \leq \log(n) + \frac{t^2\sigma^2}{2}.$$

Dividing both sides by t:

$$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \leq \frac{\log(n)}{t} + \frac{t\sigma^2}{2}.$$

Finally, we minimize the right-hand side by choosing $t = \frac{\sqrt{2\log(n)}}{\sigma}$. Substituting this value of t into the inequality gives:

$$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \leq \sigma\sqrt{2\log(n)}.$$

This concludes the proof. □

3 Growth Function

Definition 3. *Let $\mathcal{H}$ be a hypothesis space and $A = \{x_1, \ldots, x_n\} \subseteq X$ be a set of input points from the input space X. We will define the following quantities.*

- *$\mathcal{H}(A) := \{(h(x_1), \ldots, h(x_n)) \mid h \in \mathcal{H}\}$ is the set of all possible labelings of A induced by the hypothesis class $\mathcal{H}$.*
- *Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq X \times Y$ be a set of input-output pairs from $X \times Y$. Then:*

  $$L(S, \mathcal{H}) := \{(\ell(h(x_1), y_1), \ldots, \ell(h(x_n), y_n)) \mid h \in \mathcal{H}\}$$

  is the set of all possible loss values for the hypotheses in $\mathcal{H}$ evaluated on the sample S, where $\ell(h(x_i), y_i)$ is the loss incurred by hypothesis h on the example (x_i, y_i).

Exercise 4. Consider the hypothesis class $\mathcal{H} = \{h_a(x) = \mathbb{I}(x \geq a) \mid a \in \mathbb{R}\}$, which consists of all one-sided threshold classifiers.
Case 1: $A = \{x_1 = -1, x_2 = 1\}$
Case 2: $A = \{x_1 = 0, x_2 = 0\}$

Exercise 5. Now consider $L(S, \mathcal{H})$
Case 1: $S = \{(x_1 = -1, y_1 = 0), (x_2 = 1, y_2 = 1)\}$
Case 2: $S = \{(x_1 = 0, y_1 = 0), (x_2 = 0, y_2 = 1)\}$

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Yixuan Zhang and Elliot Pickens.