Instructor: Adarsh Barik
Scribed by: Mohit Chhabra
Lecture 5, COV878
August 27, 2025

LECTURE 5: RADEMACHER COMPLEXITY

1 Introduction

Rademacher complexity framework allows for dealing with very large and infinite hypothesis classes. The hypothesis class $\mathcal{F}$ maps inputs from a domain $\mathcal{Z}$ to real-valued outputs, allowing evaluation of loss functions mapping to real numbers.

$$\mathcal{F} : \mathcal{Z} \to \mathbb{R}$$
$$z = (x, y)$$

The loss function $L(h(x), y)$ quantifies the error of predicting $h(x)$ when the true label is y. We require the loss to be bounded to measure risk properly.

Specific example Previously, we have encountered a hypothesis class, $\mathcal{H}$, that contained hypothesis of the form $h : \mathcal{X} \to \mathcal{Y}$. One can construct another hypothesis class of form $F \triangleq L \circ \mathcal{H} \triangleq \{z \to L(h, z) : h \in \mathcal{H}\}$. Essentially, for $f \in \mathcal{F}$, we compute

$$f(x, y) = L(h(x), y)$$

Like before, risk $R_D(f)$ measures the expected loss over the data distribution, representing the true generalization error, while empirical risk $\hat{R}_S(f)$ is the average loss on a finite samples, serving as an estimate of the risk.

$$R_D(f) = \mathbb{E}_{z \sim D}[f(z)]$$

Empirical Risk :

$$\hat{R}_S(f) = \frac{1}{n} \sum_{i=1}^{n} f(z_i)$$

Representativeness The function $\phi(S)$ measures how representative a sample set S is, by capturing the worst-case difference between true risk and empirical risk across all hypotheses in $\mathcal{F}$. Smaller values indicate samples are a good representative of the distribution.

$$\phi(S) = \sup_{f \in \mathcal{F}} (R_D(f) - \hat{R}_S(f))$$

The quantity $\phi(S)$ is called the representativeness of S with respect to $\mathcal{F}$.

In practice, we often do not have access to the data distribution D. Therefore, one natural approach to estimate $\phi(S)$ will be to divide S into training and validation sets. Splitting S into training and validation sets S_1 and S_2 allows estimation of the generalization gap. Consider,

$$S = S_1 \cup S_2, S_1 \cap S_2 = \emptyset$$

where S_1 is the validation set and S_2 is the training set. Let

$$|S_1| = |S_2| = \frac{n}{2}$$

We can use S_1 and S_2 to estimate $\phi(S)$. Consider the maximum difference in empirical risks over the hypothesis class:

$$\sup_{f\in\mathcal{F}}(\hat{R}_{S_1}(f) - \hat{R}_{S_2}(f))$$

which measures the largest discrepancy in empirical performance on these two samples. To analyze this difference, we introduce variables $\{\sigma_i\}_{i=1}^n$, which take values in $\{-1,1\}^n$ as follows:

$$\sigma_i = \begin{cases} +1, & \text{if } z_i \in S_1, \\ -1, & \text{if } z_i \in S_2. \end{cases}$$

Using this notation, we can rewrite the supremum difference as

$$\sup_{f\in\mathcal{F}}\left(\frac{2}{n}\sum_{z\in S_1} f(z) - \frac{2}{n}\sum_{z\in S_2} f(z)\right) = \frac{2}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i). \tag{1}$$

The quantity in equation (1) acts as a motivation for defining Rademacher complexity.

Definition 1.1 (Radmacher random variable)**.** *A random variable $X \in \{-1,1\}$ is called Rademacher random variable if :*

$$\mathbb{P}[x=1] = \mathbb{P}[x=-1] = \frac{1}{2}$$

Definition 1.2 (Empirical Rademacher Complexity)**.** *The Empirical Rademacher Complexity of the hypothesis class $\mathcal{F}$ with respect to a dataset $s = \{z_1, z_2, \cdots z_n\}$ is defined as :*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i))]$$

where $\sigma = \{\sigma_1, \sigma_2 \cdots \sigma_n\}$ are independent Rademacher random variable.

Definition 1.3 (Rademacher Complexity)**.**

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{S\sim D^n}[\hat{\mathfrak{R}}_S(\mathcal{F})]\,.$$

2 Main Results

Lemma 2.1. *Let $\mathcal{F}$ be a hypothesis class satisfying*

$$\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to [0,1]\},$$

then the function

$$\phi(S) = \sup_{f\in\mathcal{F}}(R_D(f) - \hat{R}_S(f))$$

satisfies the bounded difference property

$$|\phi(z_1, z_2, \cdots z_n) - \phi(z_1, z_2, \cdots z_i', z_{i+1}, \cdots z_n)| \le \frac{1}{n}$$

Exercise: Prove that Empirical Rademacher Complexity satisfies the bounded difference property

$$|\hat{\mathfrak{R}}_{\{z_1,z_2,\cdots z_n\}} - \hat{\mathfrak{R}}_{\{z_1,z_2,\cdots,z_i',z_{i+1},\cdots z_n\}}| \leq \frac{1}{n}$$

Theorem 2.2. *Let z be a random variable with support in $\mathcal{Z}$ and distribution D. Let $S = \{z_1, z_2, \cdots z_n\}$ be a dataset of n i.i.d samples drawn from D. Let $\mathcal{F}$ be a hypothesis class such that*

$$\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to [0,1]\},$$

Fix $\delta \in (0,1)$. With probability at least $1-\delta$ over the choice of S we have $\forall f \in \mathcal{F}$

$$R(f) \leq \hat{R}_S(f) + 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

In addition, with probability at least $1-\delta$ over choice of δ, we have $\forall f \in \mathcal{F}$

$$R(f) \leq \hat{R}_S + 2\hat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Proof. Recall that,

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i))]$$
$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{S\sim D^n}[\hat{\mathfrak{R}}_S(\mathcal{F})]$$

By the previous lemma :

$$\phi : \mathcal{X}^n \to \mathbb{R}$$

fulfills the bounded difference property. Therefore from McDiarmids' inequality, with $c_i = \frac{1}{n}$

$$\mathbb{P}[\phi(s) - \mathbb{E}_S[\phi(s)] \geq \epsilon] \leq e^{\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right)} = e^{-2n\epsilon^2} = \delta$$

$$\epsilon = \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\mathbb{P}\left[\phi(s)] \leq \mathbb{E}_S[\phi] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}\right] \geq 1-\delta$$

For all $f \in \mathcal{F}$

$$R_D(f) - \hat{R}_S(f) \leq \sup_{f\in\mathcal{F}}(R_D(f) - \hat{R}_S(f)) = \phi(s)$$

$$\leq \mathbb{E}_S[\phi] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

This theorem provides a high-probability uniform bound on the generalization gap:

$$R_D(f) - \hat{R}_S(f)\,.$$

It measures how well the empirical risk approximates the true risk across all hypotheses $f \in \mathcal{F}$. The key challenge is to control or estimate the term

$$\phi(S) = \sup_{f\in\mathcal{F}}(R_D(f) - \hat{R}_S(f))$$

which depends on the unknown true risk $R_D(f)$ involving the distribution D.

Ghost sampling approach Let $\{z'_1, z'_2, \cdots z'_n\}$ be i.i.d samples, the generalization gap for a hypothesis f is given by:

$$R_D(f) - \hat{R}_S(f)$$

measures the deviation of empirical risk from the true risk.

Using the ghost sampling, this difference can be equivalently expressed as

$$\mathbb{E}_{S'}[\hat{R}_{S'}(f)] - \hat{R}_S(f) = \mathbb{E}_{S' \sim D}[\hat{R}'_S(f) - \hat{R}_S(f)]$$

Taking the supremum and using the fact that the supremum of expectation is smaller than the expectation of the supremum:

$$\sup_{f \in \mathcal{F}}(R_D(f) - \hat{R}_S(f)) \leq \mathbb{E}_{S' \sim D}[\sup_{f \in \mathcal{F}}(R_D(f) - \hat{R}_S(f))]$$

Taking expectation over S over both sides:

$$\mathbb{E}_S[\phi(s)] \leq \mathbb{E}_{S \sim D}[\mathbb{E}_{S' \sim D}[\sup_{f \in \mathcal{F}}(R_D(f) - \hat{R}_S(f))]$$

$$= \frac{1}{n}\mathbb{E}_{S,S'}[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n}(f(z'_i) - f(z_i))]$$

We z_j and z'_j are i.i.d random variables and can be replaced with each other without affecting expectation:

$$\frac{1}{n}\mathbb{E}_{S,S'}[\sup_{f \in \mathcal{F}}(f(z'_j) - f(z_j)) + \mathbb{E}_{i \neq j}(f(z'_i) - f(z_i))] = \frac{1}{n}\mathbb{E}_{S,S'}[\sup_{f \in \mathcal{F}}(f(z_j) - f(z'_j)) + \mathbb{E}_{i \neq j}(f(z'_i) - f(z_i))]$$

Adding $\frac{1}{2}$L.H.S and $\frac{1}{2}$R.H.S:

$$\mathbb{E}_S[\phi(S)] \leq \frac{1}{n}\mathbb{E}_{S,S'}\left[\frac{1}{2}\sup_{f \in \mathcal{F}}(f(z'_j) - f(z_j)) + \sum_{i \neq j}(f(z'_i) - f(z_i))\right.$$
$$\left. + \frac{1}{2}\sup_{f \in \mathcal{F}}(f(z_j) - f(z'_j)) + \sum_{i \neq j}(f(z'_i) - f(z_i))\right]$$

Introducing Rademacher random variables σ_j

$$\mathbb{E}_S[\phi(S)] \leq \frac{1}{n}\mathbb{E}_{S,S',\sigma_j}\left[\sup_{f \in \mathcal{F}} \sigma_j(f(z'_j) - f(z_j)) + \sum_{i \neq j}(f(z'_i) - f(z_i))\right]$$

Repeating the procedure for all j's

$$\mathbb{E}_S[\phi(S)] \leq \frac{1}{n}\mathbb{E}_{S,S',\sigma}\left[\sup_{f \in \mathcal{F}} \sum_i(\sigma_i(f(z'_i) - f(z_i))\right]$$

$$\mathbb{E}_S[\phi(S)] \leq \frac{1}{n}\mathbb{E}_{S,S',\sigma_j}\left[\sup_{f \in \mathcal{F}} \sum_i \sigma_i(f(z'_i) + \sup_{f \in \mathcal{F}} \sum_i -\sigma_i(f(z_i)\right]$$

The above holds as $\sup(A + B) \leq \sup(A) + \sup(B)$.

$$\mathbb{E}_S[\phi(S)] \leq \frac{1}{n}\mathbb{E}_{S',\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i(f(z'_i)\right] + \frac{1}{n}\mathbb{E}_{S,\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i(f(z_i)\right]$$

Since, $-\sigma_i$ has the same distribution as σ_i

$$\mathbb{E}_S[\phi(S)] \leq 2\mathfrak{R}_n(\mathcal{F}) \ .$$

Using the bounded difference property of $\hat{\mathfrak{R}}_S(\mathcal{F})$ and using McDiarmid's inequality, we can show that

$$\mathbb{P}\left[\mathfrak{R}_n(\mathcal{F}) - \hat{\mathfrak{R}}_S(\mathcal{F}) \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n \frac{1}{n^2}}\right) \ . \tag{2}$$

The second inequality follows using (2). □

2.1 Elementary Properties

$\mathcal{F}, \mathcal{G}$: Two hypothesis classes, $a \in \mathbb{R}$, constant $b \in \mathbb{R}$

$$a\mathcal{F} = \{af \mid f \in \mathcal{F}\}$$

$$\mathcal{F} + \mathcal{G} = \{f + g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$$

1. $$\mathcal{F} = \{h\} \implies \hat{\mathfrak{R}}_S(\mathcal{F}) = 0$$
2. $$\mathfrak{R}_S(a\mathcal{F}) = |a|\hat{\mathfrak{R}}_S(\mathcal{F})$$
3. $$\mathcal{F} \subseteq \mathcal{G} \implies \hat{\mathfrak{R}}_S(\mathcal{F}) \leq \hat{\mathfrak{R}}_S(\mathcal{G})$$
4. $$\hat{\mathfrak{R}}_S(\mathcal{F} + \mathcal{G}) = \hat{\mathfrak{R}}_S(\mathcal{F}) + \hat{\mathfrak{R}}_S(\mathcal{G})$$
5. $$\hat{\mathfrak{R}}_S(a\mathcal{F} + b) = |a|\hat{\mathfrak{R}}_S(\mathcal{F})$$

These elementary properties allow us to analyze more complex hypothesis classes built from simpler ones.

2.2 Ledoux-Talagrand Contraction Lemma

It is often beneficial to separate the analysis of loss functions from the analysis of predictor function classes. For instance, consider the class of linear predictors, $\mathcal{F}$, parameterized by weights w. Linear predictors are frequently applied in both classification and regression tasks, typically paired with distinct loss functions suited to each objective.

Let $h(x) = \langle w, x \rangle$ denote a linear predictor function. This functional form serves as the foundation for algorithms in both classification—where the goal is to assign discrete labels, and regression, where the objective is to predict continuous values.

- Classification loss function :

$$\phi(t) = \min(1, \max(0, 1-t))$$

where

$$t = y\langle w, x\rangle, \quad y \in \{-1, 1\}$$

- Regression loss function :

$$\phi(t) = \min(1, \frac{t^2}{4})$$

where

$$t = y - \langle w, x\rangle$$

To handle more complex loss functions applied to hypothesis outputs, we use contraction inequalities, which show that applying a Lipschitz function ϕ to the elements of $\mathcal{F}$ does not increase the Rademacher complexity by more than a factor determined by the Lipschitz constant.

Lemma 2.3. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a 1-Lipchitz continuous function, i.e.*

$$|\phi(t) - \phi(u)| \leq |t - u|$$

Let $\mathcal{F}$ be the hypothesis class $\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to \mathbb{R}\}$. Define

$$\Phi(\mathcal{F}) = \{\phi(f) \mid f \in \mathcal{F}\}.$$

Then we have

$$\hat{\mathfrak{R}}_S(\phi(\mathcal{F})) \leq \hat{\mathfrak{R}}_S(\mathcal{F}).$$

3 Application: Empirical Rademacher Complexity of Linear Predictors

We now apply the definitions and properties to compute the empirical Rademacher complexity of the linear class $\mathcal{F}$. Using the geometry of linear functions and vector norms, we obtain an upper bound that scales inversely with $\sqrt{n}$, showing how sample size controls complexity.

Lemma 3.1. *Let*

$$z : \{z \mid \|z\| \leq Z\}$$
$$\mathcal{F} = \{z \to \langle w, z\rangle \mid \|w\|_2 \leq W\}$$

Let $S = \{z_1, z_2, \cdots z_n\}$ be a dataset of samples where $z_i \in Z$ then

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \frac{ZW}{\sqrt{n}}.$$

Proof.

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{w:\|w\|\leq W} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle w_i, z_i\rangle \right]$$

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{w:\|w\|\leq W} \frac{1}{n} \sum_{i=1}^{n} \langle w, \sigma_i z_i\rangle \right]$$

Using the Cauchy-Schwartz inequality,

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{w:\|w\|\leq W}\sum_{i=1}^{n}\|w\|_2\|\sigma_i z_i\|\right] \\
&\leq \frac{W}{n}\mathbb{E}_\sigma\left[\|\sum_{i=1}^{n}\sigma_i z_i\|\right] \\
&= \frac{W}{n}\mathbb{E}_\sigma\left[\left(\|\sum_{i=1}^{n}\sigma_i z_i\|_2^2\right)^{\frac{1}{2}}\right] \\
&\leq \frac{W}{n}\left(\mathbb{E}_\sigma\left[\|\sum_{i=1}^{n}\sigma_i z_i\|_2^2\right]\right)^{\frac{1}{2}} \\
&= \frac{W}{n}\left(\mathbb{E}_\sigma\left[\sum_{j=1}^{n}\sum_{i=1}^{n}\sigma_i\sigma_j\langle z_i, z_j\rangle\right]\right)^{\frac{1}{2}} \\
&= \frac{W}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\langle z_i, z_j\rangle\right)^{\frac{1}{2}}
\end{aligned}$$

Since the Rademacher variables are independent with $E[\sigma_i] = 0$, $E[\sigma_j] = 0$ and $E[\sigma_i^2] = 1$ $E[\sigma_j^2] = 1$, the cross terms vanish when $i \neq j$:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &\leq \frac{W}{n}\left(\sum_{i=1}^{n}\|z_i\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{W}{n}\sqrt{n}Z \\
&= \frac{WZ}{\sqrt{n}}
\end{aligned}$$

□