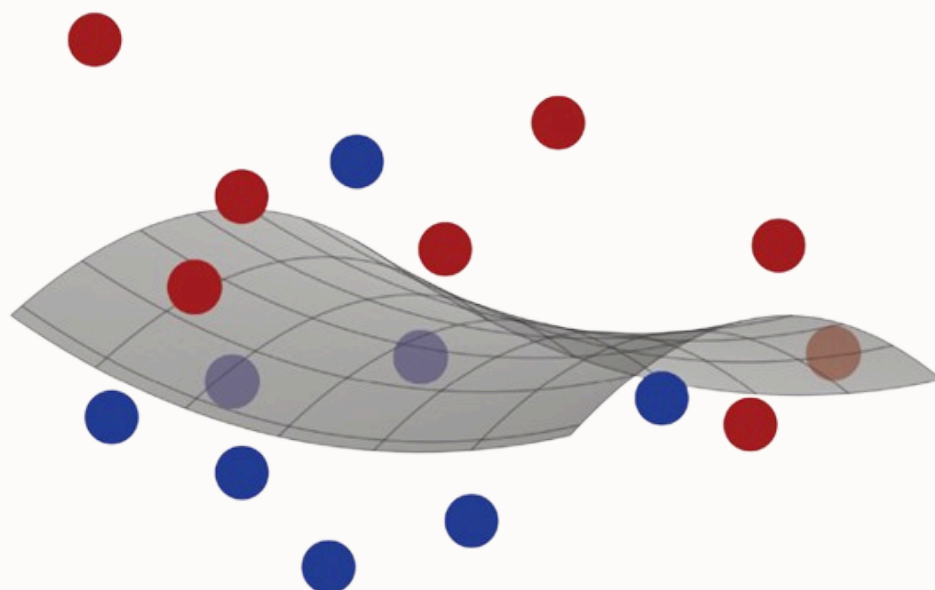


# Foundations of Machine Learning

**DAY - 12**

**Rademacher Complexity**




# Rademacher Complexity and VC-Dimension - Overview

Created By: **Birva Dave**

 [medium.com/@birva1809](https://medium.com/@birva1809)

 [github.com/Birva1809](https://github.com/Birva1809)

 [linkedin.com/in/birva-dave](https://linkedin.com/in/birva-dave)

 [birvadave1809@gmail.com](mailto:birvadave1809@gmail.com)


- 
- In machine learning, hypothesis sets are often infinite. This presents challenges for learning guarantees, as traditional sample complexity bounds may not provide meaningful insights in such cases. Despite this, it is possible to efficiently learn from finite samples even when dealing with infinite hypothesis sets, as shown in specific cases like axis-aligned rectangles. The goal is to generalize such results and derive broad learning guarantees for infinite hypothesis sets.
  - A common strategy is to reduce the problem to analyzing finite subsets using specific complexity measures. Three such measures are:
    - Rademacher complexity
    - Growth function
    - VC-dimension
  - These help in deriving learning guarantees, especially by transitioning from data-independent to data-dependent bounds.


# Rademacher Complexity

Created By: **Birva Dave**

 [medium.com/@birva1809](https://medium.com/@birva1809)

 [github.com/Birva1809](https://github.com/Birva1809)

 [linkedin.com/in/birva-dave](https://linkedin.com/in/birva-dave)

 [birvadave1809@gmail.com](mailto:birvadave1809@gmail.com)

## Purpose

- Rademacher complexity quantifies the richness of a hypothesis class by measuring its ability to fit random noise.

## Empirical Rademacher Complexity

- Let:
- $G$  be a set of functions  $g:Z \rightarrow [a,b]$
- $S=(z_1,...,z_m)$  be a fixed sample from  $Z$
- $\sigma_i$  be i.i.d. Rademacher variables taking values in  $\{-1,+1\}$
- Then the empirical Rademacher complexity is:

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

- This measures the average correlation between the outputs of functions in  $G$  and random noise over the sample  $S$ .

## Expected Rademacher Complexity

- Let  $D$  be the data distribution. The Rademacher complexity of  $G$  is:

$$\mathcal{R}_m(G) = \mathbb{E}_S[\hat{R}_S(G)]$$

where  $S \sim D_m$ .

## Generalization Bounds via Rademacher Complexity

For functions  $g \in G$ , where  $G$  maps  $Z \rightarrow [0, 1]$ , with probability at least  $1-\delta$  over a sample  $S$  of size  $m$ , the following bounds hold:

- Bound with expected Rademacher complexity:

# Rademacher Complexity

## Purpose

- Rademacher complexity quantifies the richness of a hypothesis class by measuring its ability to fit random noise.

## Empirical Rademacher Complexity

- Let:
- $G$  be a set of functions  $g:Z \rightarrow [a,b]$
- $S=(z_1,...,z_m)$  be a fixed sample from  $Z$
- $\sigma_i$  be i.i.d. Rademacher variables taking values in  $\{-1,+1\}$
- Then the empirical Rademacher complexity is:

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

- This measures the average correlation between the outputs of functions in  $G$  and random noise over the sample  $S$ .

## Expected Rademacher Complexity

- Let  $D$  be the data distribution. The Rademacher complexity of  $G$  is:

$$\mathcal{R}_m(G) = \mathbb{E}_S[\hat{R}_S(G)]$$

where  $S \sim D_m$ .

## Generalization Bounds via Rademacher Complexity

- For functions  $g \in G$ , where  $G$  maps  $Z \rightarrow [0, 1]$ , with probability at least  $1-\delta$  over a sample  $S$  of size  $m$ , the following bounds hold:

# Rademacher Complexity

- Bound with expected Rademacher complexity:

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

- Bound with empirical Rademacher complexity:

$$\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^m g(z_i) + 2\hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

- These bounds are derived using McDiarmid's inequality and show how generalization performance relates to the complexity of the hypothesis set.

## Rademacher Complexity and Binary Classification

- Let:
  - $H$  be a hypothesis class of functions  $h: X \rightarrow \{-1, +1\}$
  - $G$  be the corresponding loss class under the 0-1 loss:
  - $G = \{(x, y) \mapsto 1 [h(x) \neq y]: h \in H\}$

For a sample  $S = \{(x_i, y_i)\}_{i=1}^m$ , let  $S_X = (x_1, \dots, x_m)$ . Then:

$$\hat{\mathcal{R}}_S(G) = \frac{1}{2} \hat{\mathcal{R}}_{S_X}(H)$$


- This relation simplifies analysis for binary classification, allowing the use of Rademacher complexity bounds directly on the hypothesis set  $H$ .


# Rademacher Complexity

Created By: **Birva Dave**

 [medium.com/@birva1809](https://medium.com/@birva1809)

 [github.com/Birva1809](https://github.com/Birva1809)

 [linkedin.com/in/birva-dave](https://linkedin.com/in/birva-dave)

 [birvadave1809@gmail.com](mailto:birvadave1809@gmail.com)

---

## Binary Classification Generalization Bound

- With probability at least  $1-\delta$ , for all  $h \in H$ :
  - Risk bound using expected Rademacher complexity:

$$\mathbb{E}_{(x,y) \sim D}[\mathbf{1}[h(x) \neq y]] \leq \frac{1}{m} \sum_{i=1}^m \mathbf{1}[h(x_i) \neq y_i] + \mathcal{R}_m(H) + \sqrt{\frac{\log(1/\delta)}{2m}}$$

- Risk bound using empirical Rademacher complexity:

$$\mathbb{E}_{(x,y) \sim D}[\mathbf{1}[h(x) \neq y]] \leq \frac{1}{m} \sum_{i=1}^m \mathbf{1}[h(x_i) \neq y_i] + \hat{R}_{S_X}(H) + \sqrt{\frac{\log(1/\delta)}{2m}}$$

## Why This Matters

- Rademacher complexity provides data-dependent, computable bounds for generalization.
- It is especially useful when traditional VC-dimension-based bounds are too loose or hard to calculate.
- Though computing empirical Rademacher complexity is NP-hard for some classes, it remains a powerful theoretical tool.