**CS861: Theoretical Foundations of Machine Learning** Lecture 3 - 09/09/2024
University of Wisconsin–Madison, Fall 2024

# Lecture 03: Rademacher Complexity

*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.*

This lecture will introduce the **Bayes optimal classifier**. Then, we will introduce **approximation and estimation error**. Next, we will introduce **McDiarmid's inequality**, a key tool that will be used to establish the **uniform convergence** in the probability of the empirical risk to the true risk for any hypothesis within a given hypothesis class. Finally, we will cover the concept of **empirical Rademacher complexity**, which will help us explicitly bound the difference between the empirical risk and the true risk with high probability.

## 1 Bayes Optimal Classifier

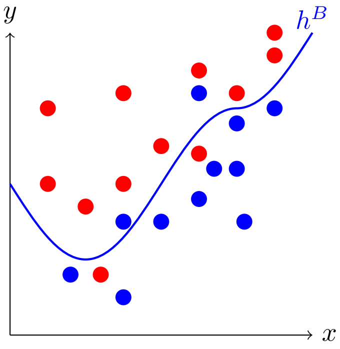

**Figure 1:** An example of Bayes Optimal Classifier

We begin this lecture by introducing the **Bayes optimal classifier**. This classifier always selects the class with the highest probability given the input. For binary classification, the classifier is defined as follows:

$$\begin{aligned} h_B(x) &= \arg\max_{y\in\{0,1\}} \mathbb{P}(Y=y \mid X=x) \\ &= \begin{cases} 1 & \text{if } \mathbb{P}(Y=1 \mid X=x) \geq \frac{1}{2} \\ 0 & \text{if } \mathbb{P}(Y=0 \mid X=x) > \frac{1}{2} \end{cases} \end{aligned}$$

We can show that the Bayes optimal classifier produces the minimum risk among all classifiers. Intuitively, selecting the class with the highest conditional probability minimizes the expected prediction error.

**Theorem:** $\forall h \in h : X \to Y$, $R(h) \geq R(h_B)$

**Proof:** The risk of any classifier $h$ is given by:

$$\begin{aligned}
R(h) &= \mathbb{E}_{XY}[\mathbb{1}(h(X) \neq Y)] \quad \text{(Definition of risk as the expected indicator of misclassification)} \\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbb{1}(h(X) \neq Y) \mid X]] \quad \text{(Applying the law of iterated expectation)} \\
&= \mathbb{E}_X[\mathbb{E}_Y[\mathbb{1}(h(X) \neq Y) \mid Y = 0, X]\,\mathbb{P}(Y = 0 \mid X) + \mathbb{E}_Y[\mathbb{1}(h(X) \neq Y) \mid Y = 1, X]\,\mathbb{P}(Y = 1 \mid X)] \\
&\quad \text{(Separate the cases when } Y = 0 \text{ and } Y = 1) \\
&= \mathbb{E}_X[\mathbb{1}(h(X) \neq 0)\,\mathbb{P}(Y = 0 \mid X) + \mathbb{1}(h(X) \neq 1)\,\mathbb{P}(Y = 1 \mid X)] \\
&= \mathbb{E}_X[\mathbb{1}(h(X) = 1)\,\mathbb{P}(Y = 0 \mid X) + \mathbb{1}(h(X) = 0)\,\mathbb{P}(Y = 1 \mid X)] \\
&= \int_x (\mathbb{1}(h(x) = 1)\,\mathbb{P}(Y = 0 \mid X = x) + \mathbb{1}(h(x) = 0)\,\mathbb{P}(Y = 1 \mid X = x))\,dP_X(x)
\end{aligned}$$

The integrand is minimized at each point if $h(x)$ follows this rule: when $P(Y = 0|X = x) \geq P(Y = 1|X = x)$, we set $h(x) = 0$, ensuring that the smaller of the two terms in the integrand is selected. Similarly, if $P(Y = 1|X = x) \geq P(Y = 0|X = x)$, we set $h(x) = 1$. In other words, by choosing $h(x)$ in this way, we minimize the integrand—and consequently, the risk—for every $x$. Therefore, the risk of $h$ is minimized when $h(x)$ is chosen as follows:

$$h(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{otherwise} \end{cases}$$

This is the same decision rule as the Bayes classifier $h_B$. Therefore, $R(h) \geq R(h_B)$ for all classifiers $h$, proving the result.

## 2 Estimation and Approximation Errors

Suppose an algorithm (such as Empirical Risk Minimization (ERM)) chooses a hypothesis $\hat{h} \in \mathcal{H}$ from the hypothesis class $\mathcal{H}$. Then the *excess risk*, i.e the difference between the risk of the selected hypothesis $\hat{h}$ and the Bayes optimal hypothesis $h_B$ can be decomposed as follows:

$$\underbrace{R(\hat{h}) - R(h^B)}_{\text{excess risk}} = \underbrace{R(\hat{h}) - R(h^*)}_{\text{Estimation Error}} + \underbrace{R(h^*) - R(h^B)}_{\text{Approximation Error}}$$

Here, $h^*$ is the best hypothesis within the class $\mathcal{H}$ that minimizes the true risk:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h).$$

The two terms in this decomposition represent the following:

- **Estimation Error:** $R(\hat{h}) - R(h^*)$
  This is the error that results from selecting $\hat{h}$ based on the empirical data. It represents the difference between the empirical minimizer $\hat{h}$ and the best possible hypothesis $h^*$ in the class $\mathcal{H}$.

- **Approximation Error:** $R(h^*) - R(h^B)$
  This is the error caused by the fact that the best hypothesis $h^*$ in $\mathcal{H}$ may not be as good as the Bayes optimal classifier $h_B$. The approximation error arises from the limitations of the hypothesis class $\mathcal{H}$.

Our focus so far, and going forward, is in bounding the **estimation error**. Controlling the approximation error typically requires additional assumptions about the problem.

To reduce estimation error, it is necessary to have enough data and choose the best hypothesis based on that data. We show an example in Figure 2, where we can see that expanding from $\mathcal{H}$ to $\widetilde{\mathcal{H}}$ would reduce the approximation error, but would also then potentially increase the estimation error by having a larger class of hypotheses to "test" on the given data.

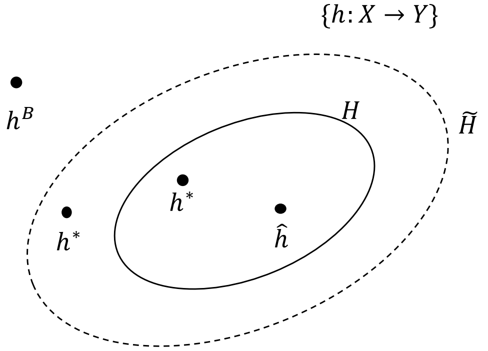

**Figure 2:** A visual representation of the Approximation vs. Estimation error trade-off

## 3 McDiarmid's Inequality

Next, we introduce McDiarmid's inequality, a concentration inequality that provides a bound on the difference between a function's value when evaluated on a sample and its expected value. We will apply McDiarmid's inequality in the context of Rademacher complexity. To begin, we need to define the bounded difference property, which is a requirement for applying McDiarmid's inequality.

**Definition 1** (Bounded Difference Property)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. The function $f$ satisfies the bounded difference property if there exist constants $c_1, \ldots, c_n \in \mathbb{R}$ such that for all $k \in \{1, \ldots, n\}$:*

$$\sup_{z_1,\ldots,z_k,\ldots,z_n,\tilde{z}_k} |f(z_1, \ldots, z_k, \ldots, z_n) - f(z_1, \ldots, \tilde{z}_k, \ldots, z_n)| \le c_k$$

Intuitively, the bounded difference property means that changing any one input to the function $f$ results in a controlled, finite change in the output. Now, we formally state McDiarmid's inequality.

**Theorem 1** (McDiarmid's Inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that satisfies the bounded difference property with bounds $c_1, \ldots, c_n$. Let $Z_1, \ldots, Z_n$ be independent random variables. Then for any $\epsilon > 0$, the following holds:*

$$\mathbb{P}(f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \ge \epsilon) \le \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^n c_k^2}\right)$$

*Similarly,*

$$\mathbb{P}(\mathbb{E}[f(Z_1, \ldots, Z_n)] - f(Z_1, \ldots, Z_n) \ge \epsilon) \le \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^n c_k^2}\right)$$

To see McDiarmid's inequality in action, consider the following example:

**Example 2.** We will apply McDiarmid's inequality to show that:

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

First, we will show that $\widehat{R}(h)$ satisfies the bounded difference property. Let $X_1, \dots, X_n \in \mathbb{R}^d$ and $Y \in \{0,1\}$ be random variables. Define the random variable $Z_i = \mathbb{1}(h(X_i) \neq Y_i)$, and recall that:

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i) = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Thus, $\widehat{R}(h)$ can be written as a function of $Z_1, \dots, Z_n$. Let $k \in \{1, \dots, n\}$. The maximum change in $\widehat{R}(h)$ due to a change in $Z_k$ is bounded by $\frac{1}{n}$.

$$\begin{aligned}
&\sup_{Z_1,\dots,Z_k,\dots,Z_n,\tilde{Z}_k} \left| \frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\left(\sum_{i=1}^{k-1} Z_i + \tilde{Z}_k + \sum_{i=k+1}^{n} Z_i\right)\right| \quad \text{(expand sum and substitute } \tilde{Z}_k \text{ for } Z_k\text{)} \\
&= \sup_{Z_k,\tilde{Z}_k} \left| \frac{1}{n} Z_k - \frac{1}{n}\tilde{Z}_k \right| = \frac{1}{n} \quad \text{(maximum change is bounded by } \frac{1}{n}\text{)}
\end{aligned}$$

Thus, $\widehat{R}(h)$ satisfies the bounded difference property, and the maximum difference for any input change is $\frac{1}{n}$. Now, we can apply McDiarmid's inequality.

$$\begin{aligned}
\mathbb{P}(\widehat{R}(h) - R(h) > \epsilon) &= \mathbb{P}\left(\widehat{R}(h) - \mathbb{E}[\widehat{R}(h)] > \epsilon\right) \quad \text{(by definition of } R(h)\text{)} \\
&\leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^{n}\left(\frac{1}{n}\right)^2}\right) = \exp(-2n\epsilon^2) \quad \text{(apply McDiarmid's inequality and simplify the sum)}
\end{aligned}$$

By applying the same reasoning, we can also bound the probability that the true risk $R(h)$ exceeds the empirical risk $\widehat{R}(h)$ by more than $\epsilon$:

$$\mathbb{P}(R(h) - \widehat{R}(h) > \epsilon) \leq \exp(-2n\epsilon^2) \quad \text{(applying McDiarmid's inequality similarly for } R(h) - \widehat{R}(h)\text{)}$$

Next, using the union bound, we can combine the two bounds. The union bound states that the probability of either event occurring (i.e., $\widehat{R}(h) > R(h) + \epsilon$ or $R(h) > \widehat{R}(h) + \epsilon$) is at most the sum of their probabilities:

$$\begin{aligned}
\mathbb{P}(|R(h) - \widehat{R}(h)| > \epsilon) &= \mathbb{P}(R(h) - \widehat{R}(h) > \epsilon) + \mathbb{P}(\widehat{R}(h) - R(h) > \epsilon) \\
&\leq \exp(-2n\epsilon^2) + \exp(-2n\epsilon^2) \quad \text{(summing the two bounds)} \\
&= 2\exp(-2n\epsilon^2) \quad \text{(final result combining both terms)}
\end{aligned}$$

Thus, we have the desired result:

$$\mathbb{P}(|R(h) - \widehat{R}(h)| > \epsilon) \leq 2\exp(-2n\epsilon^2)$$

# 4 Uniform convergence

We aim to show that for any small $\varepsilon > 0$, the following holds:

$$\mathbb{P}(\forall h \in \mathcal{H}, |\widehat{R}(h) - R(h)| \leq \varepsilon) \geq \gamma$$

where $\gamma \in [0,1]$ is a large value (i.e., close to 1).

In a previous lecture, we analyzed the following probability for an arbitrary $h \in \mathcal{H}$:

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq \delta$$

where $\delta \in [0,1]$ is a small value (i.e., close to 0), derived using bounds such as Hoeffding's inequality.

We then applied the union bound and obtained:

$$\mathbb{P}(\exists h \in \mathcal{H}, |\widehat{R}(h) - R(h)| > \varepsilon) \leq \sum_{h \in \mathcal{H}} \mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq |\mathcal{H}| \cdot \delta$$

However, the above result becomes meaningless if $|\mathcal{H}| \cdot \delta \geq 1$, which is guaranteed to happen if $|\mathcal{H}| = \infty$, regardless of how small $\delta > 0$ is. Therefore, we need a bound that is still useful when the hypothesis class $\mathcal{H}$ is potentially infinite.

To address this, we consider the following quantity:

$$f(S) := \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h))$$

where

$$S := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

with $X_i \in \mathcal{X}$ and $Y_i \in \{0,1\}$, and $\widehat{R}_S(h) := \frac{1}{n} \sum_{(X_i,Y_i) \in S} \mathbb{1}_{\{h(X_i) \neq Y_i\}}$.

Next, to apply McDiarmid's inequality, define

$$\tilde{S} := \{(X_1, Y_1), \ldots, (\tilde{X}_k, \tilde{Y}_k), \ldots, (X_n, Y_n)\}.$$

Then, we can analyze the difference between $f(S)$ and $f(\tilde{S})$ as follows:

$$\begin{aligned}
\sup_{S \cup \tilde{S}} |f(S) - f(\tilde{S})| &= \sup_{S \cup \tilde{S}} \left| \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h)) - \sup_{h \in \mathcal{H}} (\widehat{R}_{\tilde{S}}(h) - R(h)) \right| \\
&\leq \sup_{S \cup \tilde{S}} \sup_{h \in \mathcal{H}} \left| (\widehat{R}_S(h) - R(h)) - (\widehat{R}_{\tilde{S}}(h) - R(h)) \right| \quad \text{(rearranging terms)} \\
&= \sup_{S \cup \tilde{S}} \sup_{h \in \mathcal{H}} \left| \widehat{R}_S(h) - \widehat{R}_{\tilde{S}}(h) \right| \quad \text{(canceling out the } R(h) \text{ terms)} \\
&= \sup_{S \cup \tilde{S}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \left( \mathbb{1}_{\{h(X_k) \neq Y_k\}} - \mathbb{1}_{\{h(\tilde{X}_k) \neq \tilde{Y}_k\}} \right) \right| \\
&\leq \frac{1}{n} \quad \text{(since the indicator function values can differ by at most 1)}
\end{aligned}$$

Thus, the bounded difference property holds with $c_k = \frac{1}{n}$ for all $k$, allowing us to apply McDiarmid's inequality. Applying it, we get:

$$\begin{aligned}
\mathbb{P}_{S \sim P_{X,Y}} (f(S) - \mathbb{E}[f(S)] > \varepsilon) &= \mathbb{P}_{S \sim P_{X,Y}} \left( \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h)) - \mathbb{E}[\sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h))] > \varepsilon \right) \\
&\leq \exp(-2n\varepsilon^2) \quad \text{(applying McDiarmid's inequality)}
\end{aligned}$$

The probability above can be interpreted as:

$$\mathbb{P} \left( \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h)) > \mathbb{E}[\sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h))] + \varepsilon \right),$$

which is equivalent to:

$$\mathbb{P} \left( \exists h \in \mathcal{H}, (\widehat{R}_S(h) - R(h)) > \mathbb{E} \left[ \sup_{h \in \mathcal{H}} (\widehat{R}_S(h) - R(h)) \right] + \varepsilon \right).$$

Therefore, by McDiarmid's inequality, we can assert that, for all $h \in \mathcal{H}$, with probability at least $1 - \exp(-2n\varepsilon^2)$, we have:

$$\widehat{R}_S(h) - R(h) \leq \mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))] + \varepsilon.$$

It would be helpful to quantify or at least bound the term $\mathbb{E}[\sup_{h\in\mathcal{H}}(\widehat{R}_S(h) - R(h))]$, which is generally not equal to $\sup_{h\in\mathcal{H}} \mathbb{E}[(\widehat{R}_S(h) - R(h))] = 0$, given the non-linearity of the supremum operator.

Next, we will introduce the Rademacher complexity to help us bound this term.

# 5 Rademacher Complexity

**Definition 2.** *A Rademacher random variable $\sigma \in \{-1, 1\}$ is defined such that $\mathbb{P}(\sigma = -1) = \mathbb{P}(\sigma = 1) = \frac{1}{2}$.*

**Definition 3** (Empirical Rademacher Complexity)**.** *Let $S := \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an observed sample of n points, and let $\sigma := (\sigma_1, \dots, \sigma_n) \in \{-1, 1\}^n$ be n independent Rademacher random variables. The Empirical Rademacher Complexity is defined as:*

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

*where $\ell(\cdot)$ is a loss function, such as $\ell(h(x_i), y_i) = \mathbb{1}_{\{h(x_i)\neq y_i\}}$ in the case of a classification problem.*

The intuition behind this definition is that it measures the capacity of a hypothesis class $\mathcal{H}$ to align with random noise (represented by the Rademacher variables $\sigma_i$). In other words, it quantifies how well the hypothesis class $\mathcal{H}$ can fit random patterns in the data.

For better clarity, let us rewrite the definition as follows: let $\tilde{\ell} := (\ell(h(x_1), y_1), \dots, \ell(h(x_n), y_n))$. Then, we can express the empirical Rademacher complexity as:

$$\widehat{Rad}(S, \mathcal{H}) := \mathbb{E}_\sigma \left[ \sup_{h\in\mathcal{H}} \frac{1}{n} \sigma \cdot \tilde{\ell} \right],$$

where $\sigma \cdot \tilde{\ell}$ represents the dot product between the vector of Rademacher random variables $\sigma$ and the vector of loss values $\tilde{\ell}$.

This expression measures how well the hypothesis class $\mathcal{H}$ can correlate with the random vector of Rademacher variables. Intuitively, more complex or flexible hypothesis classes can align more closely with random noise, leading to a higher Rademacher complexity.

*(Note: Recall that the dot product between two vectors a and b is equal to the cosine of the angle between them, scaled by the product of their norms. When these vectors have zero mean, the cosine of the angle is equivalent to their correlation coefficient.)*

## 5.1 Example of computing Rademacher Complexity

Consider the sample $S = \{(x_1, y_1 = 0), (x_2, y_2 = 1)\}$, where $x_1 < x_2$. We will examine two hypothesis classes:

- $\mathcal{H}_1 = \{h_a(x) = \mathbb{1}_{\{x\geq a; a\in\mathbb{R}\}}\}$
- $\mathcal{H}_2 = \{h_a(x) = \mathbb{1}_{\{x\geq a; a\in\mathbb{R}\}}\} \cup \{h_a(x) = \mathbb{1}_{\{x\leq a; a\in\mathbb{R}\}}\}$

In this example, we compute the empirical Rademacher complexity for both hypothesis classes by considering different combinations of Rademacher variables, $\sigma = (\sigma_1, \sigma_2)$.

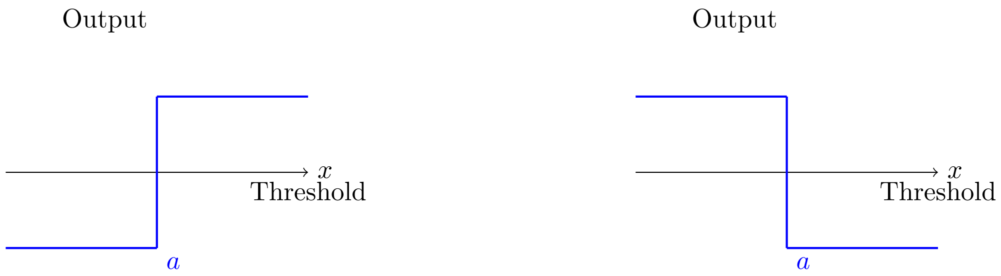

**Figure 3:** Two example threshold functions, where the hypothesis is either $h(x) = \mathbb{1}(x \geq a)$ or $h(x) = \mathbb{1}(x \leq a)$

### 5.1.1 For $\mathcal{H}_1$:

Let's consider the Rademacher variables $\sigma_1, \sigma_2 \in \{-1, +1\}$ and the empirical loss $\ell(h(x_i), y_i) = \mathbb{1}_{\{h(x_i) \neq y_i\}}$. We compute $\widehat{Rad}(S, \mathcal{H}_1)$ based on different values of $\sigma$.

- $\sigma = (+1, +1)$:

$$\begin{aligned}\widehat{Rad}(S, \mathcal{H}_1) &= \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}_1} \frac{1}{2} (\sigma_1 \ell(h(x_1), y_1) + \sigma_2 \ell(h(x_2), y_2)) \right] \\ &= \frac{1}{2} (\ell_1 + \ell_2) \quad \text{(the supremum depends on the choice of threshold } a\text{)}\end{aligned}$$

- $\sigma = (-1, +1)$:

$$\widehat{Rad}(S, \mathcal{H}_1) = \frac{1}{2} (-\ell_1 + \ell_2)$$

- $\sigma = (+1, -1)$:

$$\widehat{Rad}(S, \mathcal{H}_1) = \frac{1}{2} (\ell_1 - \ell_2)$$

- $\sigma = (-1, -1)$:

$$\widehat{Rad}(S, \mathcal{H}_1) = \frac{1}{2} (-\ell_1 - \ell_2)$$

For $\mathcal{H}_1$, the overall empirical Rademacher complexity can be computed by averaging these values, yielding $\widehat{Rad}(S, \mathcal{H}_1) = \frac{3}{8}$.

### 5.1.2 For $\mathcal{H}_2$:

Now, let's consider the extended hypothesis class $\mathcal{H}_2$. The presence of hypotheses of the form $h_a(x) = \mathbb{1}_{\{x \leq a; a \in \mathbb{R}\}}$ allows for more flexibility, which increases the empirical Rademacher complexity.

- $\sigma = (+1, +1)$:

$$\widehat{Rad}(S, \mathcal{H}_2) = \frac{1}{2} (\ell_1 + \ell_2) \quad \text{(since } \mathcal{H}_2 \text{ provides more flexibility)}$$

- $\sigma = (-1, +1)$:

$$\widehat{Rad}(S, \mathcal{H}_2) = \frac{1}{2} (-\ell_1 + \ell_2)$$

- $\sigma = (+1, -1)$:

$$\widehat{Rad}(S, \mathcal{H}_2) = \frac{1}{2}\left(\ell_1 - \ell_2\right)$$

- $\sigma = (-1, -1)$:

$$\widehat{Rad}(S, \mathcal{H}_2) = \frac{1}{2}\left(-\ell_1 - \ell_2\right)$$

For $\mathcal{H}_2$, the empirical Rademacher complexity is $\widehat{Rad}(S, \mathcal{H}_2) = \frac{1}{2}$, which is higher than that of $\mathcal{H}_1$ due to the additional flexibility of the hypothesis class.

## Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Justin Kiefel and Albert Dorador.