Statistical Learning Theory Notes

Nong Minh Hieu[1]

[1] School of Physical and Mathematical Sciences, Nanyang Technological University (NTU - Singapore)

Abstract

This note is based on the lecture notes for "EECS598 - Statistical Learning Theory" course compiled by Clayton Scott.

Contents

1 Probability settings

1.1 Classification problem

Definition 1.1 (Classifier (h))**.**
*In **classification problems**, we consider pairs (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Where:*

- *$\mathcal{X}$ is the space of **feature vectors**.*
- *$\mathcal{Y}$ is the space of **labels**.*

A classifier is a function $h : \mathcal{X} \to \mathcal{Y}$ which aims to assign correct labels to given feature vectors.

Remark : The key assumptions of classification problems are:

- There exists a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$.
- The pairs (x, y) (observed data) are random samples of the random variables pair (X, Y) which has the distribution P_{XY}.

Definition 1.2 (Decomposition of P_{XY})**.**
We can decompose P_{XY} in either of the following two ways:

$$P_{XY} = P_{X|Y} P_Y$$
$$P_{XY} = P_{Y|X} P_X$$

Which can be understood as two possible ways to generate the pairs (x, y) from the joint distribution P_{XY}.

- *The first way is to generate a random label $y \sim P_Y$. Then, generate the feature vector corresponding to that label $x \sim P_{X|Y=y}$.*
- *The second way is to generate a random vector $x \sim P_X$. Then, generate the label corresponding to that feature vector $y \sim P_{Y|X=x}$.*

Proposition 1.1: Law of total expectation

Given $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The **law of total expectation** states that:

$$\begin{aligned}\mathbb{E}_{XY}\Big[\phi(X, Y)\Big] &= \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X, Y)]\Big] \\ &= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[\phi(X, Y)]\Big]\end{aligned}$$

Similar to how P_{XY} is decomposed, law of total expectation describes two way of taking the average value:

- Loop through the labels and take average over the feature vectors corresponding to each label.
- Loop through the feature vectors and take average over the labels corresponding to each vector.

Proof (Proposition 1.1). *We have:*

$$\begin{aligned}
\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(x,y)P_{XY}(x,y)dydx \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(x,y)P_X(x)P_{Y|X}(y|x)dydx \\
&= \int_{\mathcal{X}} P_X(x)\int_{\mathcal{Y}} \phi(x,y)P_{Y|X}(y|x)dydx \\
&= \int_{\mathcal{X}} P_X(x)\mathbb{E}_{Y|X=x}\Big[\phi(X,Y)\Big]dx \\
&= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}\Big[\phi(X,Y)\Big]\Big]
\end{aligned}$$

Applying the same technique, we have $\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] = \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X,Y)]\Big]$. ▫.

Remark : Usually, the label space is discrete and finite, meaning $\mathcal{Y} = \{0,1,2,\ldots,m\}$ for some $m < \infty$. Hence, the expectations over Y can be written as discrete sums:

$$\begin{aligned}
\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] = \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X,Y)]\Big] &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X|Y=y}[\phi(X,Y)] \\
= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[\phi(X,Y)]\Big] &= \mathbb{E}_X\left[\sum_{y\in\mathcal{Y}} \mathbb{E}_{Y=y|X}[\phi(X,Y)]\right]
\end{aligned}$$

Definition 1.3 (Hypothesis space ($\mathcal{H}$)). *The hypothesis space is a collection (family) of classifiers* $h : \mathcal{X} \to \mathcal{Y}$ *that have some common properties:*

$$\mathcal{H} = \Big\{h : \mathcal{X} \to \mathcal{Y}\Big|\textit{some common properties}\Big\}$$

For example, let $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = (0,1)$. *In logistic regression, we assume the classifiers to be logit functions:*

$$\mathcal{H}_{logit} = \Big\{h : \mathbb{R}^d \to (0,1)\Big|h(x) = logit(\beta x) = \frac{1}{1+e^{-\beta x}}, \beta \in \mathbb{R}^{1\times d}\Big\}$$

Definition 1.4 (Learning algorithm ($\mathcal{L}_n$)). *To learn a classifier* $h : \mathcal{X} \to \mathcal{Y}$, *suppose that we have access to a training dataset of* n *data pairs* $\{(X_k, Y_k)\}_{k=1}^n$ *which are assumed to be* ***i.i.d sampled from*** P_{XY}. *The domain of the training data is then* $(\mathcal{X}\times\mathcal{Y})^n$. *A* ***learning algorithm****, denoted as* $\mathcal{L}_n$ *is a function/procedure that derives a classifier* $\hat{h}_n : \mathcal{X} \to \mathcal{Y}$ *from the training data.*

$$\begin{aligned}
&\mathcal{L}_n : (\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H} \\
&\hat{h}_n = \mathcal{L}_n((X_1,Y_1),\ldots,(X_n,Y_n))
\end{aligned}$$

1.2 Goal of classification

Definition 1.5 (Risk ($R(h)$))**.**
*The **risk** of a classifier is defined as followed:*

$$R(h) = P(h(X) \neq Y) = \mathbb{E}[\mathbf{1}_{\{h(X) \neq Y\}}]$$

Where (X, Y) are independent of the training data.

Definition 1.6 (Bayes Risk (R^*))**.**
*The **Bayes risk** is the infimum of the risk taken over all $h : \mathcal{X} \to \mathcal{Y}$, not just for $h \in \mathcal{H}$:*

$$R^* = \inf_{h:\mathcal{X}\to\mathcal{Y}} R(h)$$

Definition 1.7 (Consistency of learning algorithms)**.**
A learning algorithm $\mathcal{L}_n$ is called:

- ***Weakly consistent** if $R(\hat{h}_n) \xrightarrow{p} R^*$:*

$$\lim_{n\to\infty} P(R(\hat{h}_n) \leqslant r) = P(R^* \leqslant r), \ \forall r \geqslant 0$$

- ***Strongly consistent** if $R(\hat{h}_n) \xrightarrow{a.s} R^*$:*

$$P\Big(\lim_{n\to\infty} \Big|R(\hat{h}_n) - R^*\Big| \geqslant \epsilon\Big) = 0, \ \forall \epsilon > 0$$

- ***Universally weakly/strongly consistent** if $\mathcal{L}_n$ is weakly/strongly consistent for all P_{XY}. Meaning, consistency holds without any assumption about P_{XY}.*

2 Bayes classifier

2.1 Properties of Bayes Risk

Overview : Recall that the Bayes classifier is the one with minimum risk and the corresponding risk is called the Bayes Risk. For $\mathcal{Y} = \{0, 1\}$ and defined:

$$\eta(x) = P(Y = 1|X = x)$$

Define the following classifier:

$$h^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geqslant \frac{1}{2} \\ \\ 0 & \text{otherwise} \end{cases}$$

Theorem 2.1: Properties of Bayes classifier

The following properties hold for the Bayes classifier with $\mathcal{Y} = \{0, 1\}$ (Binary classification):

- (i) $R(h^*) = \inf_{h:\mathcal{X}\to\mathcal{Y}}\{R(h)\} = R^*$.
- (ii) $\underbrace{R(h) - R^*}_{\text{Excess risk}} = 2\mathbb{E}_X\left[\left|\eta(x) - \frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]$.
- (iii) $R^* = \mathbb{E}\Big[\min(\eta(X), 1 - \eta(x))\Big]$.

Proof (Theorem 2.1). *Proving each point:*

(i) $R(h^*) = \inf_{h:\mathcal{X}\to\mathcal{Y}}\{R(h)\} = R^*$.
For all $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$\begin{aligned} R(h) &= \mathbb{E}_{XY}\Big[\mathbf{1}_{\{h(X)\neq Y\}}\Big] \\ &= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\Big[\mathbf{1}_{\{Y\neq h(x)\}}\Big]\right] \\ &= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}} \mathbf{1}_{\{y\neq h(x)\}}\right] \\ &= \mathbb{E}_{x\sim X}\Big[\eta(x)\mathbf{1}_{\{h(x)=0\}} + (1 - \eta(x))\mathbf{1}_{\{h(x)=1\}}\Big] \end{aligned}$$

Since the two events $\{h(x) = 1\}$ and $\{h(x) = 0\}$ are mutually exclusive, $R(h)$ is the smallest when we set $h(x) = 1$ when $\eta(x) \geqslant 1 - \eta(x) \implies \eta(x) \geqslant \frac{1}{2}$. Therefore, we have:

$$h^*(x) = \begin{cases} 1 & \textit{if } \eta(x) \geqslant \frac{1}{2} \\ \\ 0 & \textit{otherwise} \end{cases}$$

(ii) $\underbrace{R(h) - R^*}_{\textit{Excess risk}} = 2\mathbb{E}_X\left[\left|\eta(x) - \frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]$.

We have:

$$
\begin{aligned}
R(h) - R^* &= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\left[\mathbf{1}_{\{Y\neq h(x)\}}\right]\right] - \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\left[\mathbf{1}_{\{Y\neq h^*(x)\}}\right]\right] \\
&= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}} \mathbf{1}_{\{y\neq h(x)\}} P(Y=y|X=x)\right] - \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}} \mathbf{1}_{\{y\neq h^*(x)\}} P(Y=y|X=x)\right] \\
&= \mathbb{E}_{x\sim X}\left[\eta(x)\Big(\mathbf{1}_{\{h(x)=0\}} - \mathbf{1}_{\{h^*(x)=0\}}\Big) + (1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1\}} - \mathbf{1}_{\{h^*(x)=1\}}\Big)\right] \\
&= \mathbb{E}_{x\sim X}\Bigg[\eta(x)\Big(\mathbf{1}_{\{h(x)\neq h^*(x), h(x)=0\}} - \mathbf{1}_{\{h(x)\neq h^*(x), h(x)=1\}}\Big) \\
&\qquad + (1-\eta(x))\Big(\mathbf{1}_{\{h(x)\neq h^*(x), h(x)=1\}} - \mathbf{1}_{\{h(x)\neq h^*(x), h(x)=0\}}\Big)\Bigg] \\
&= \mathbb{E}_{x\sim X}\left[(2\eta(x)-1)\mathbf{1}_{\{h(x)\neq h^*(x), h(x)=0\}} + (1-2\eta(x))\mathbf{1}_{\{h(x)\neq h^*(x), h(x)=1\}}\right] \\
&= \mathbb{E}_{x\sim X}\left[\Big|2\eta(x)-1\Big|\mathbf{1}_{\{h(x)\neq h^*(x)\}}\right] \\
&= 2\mathbb{E}_X\left[\left|\eta(X)-\frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]
\end{aligned}
$$

(iii) $R^* = \mathbb{E}\Big[\min(\eta(X), 1-\eta(x))\Big]$.
From (i) we have:

$$
\begin{aligned}
R(h^*) &= \mathbb{E}_{x\sim X}\Big[\eta(x)\mathbf{1}_{\{h^*(x)=0\}} + (1-\eta(x))\mathbf{1}_{\{h^*(x)=1\}}\Big] \\
&= \mathbb{E}_X\Big[\min(\eta(X), 1-\eta(x))\Big]
\end{aligned}
$$

□.

Theorem 2.2: Properties of Bayes classifier (Multi-class)

For multi-class classification with more than two labels : $\mathcal{Y} = \{1, 2, \ldots, M\}$, the Bayes classifier is defined as followed:

$$h^*(x) = \arg\max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\}$$
$$\text{Where : } \eta_y(x) = P(Y = y | X = x)$$

The following properties hold for the Bayes classifier with $\mathcal{Y} = \{1, 2, \ldots, M\}$ (Multi-class classification):

- (i) **Bayes Risk** R^* :

$$R^* = \mathbb{E}_{x \sim X}\Big[1 - \max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\} \Big] = \mathbb{E}_{x \sim X}\Big[\min_{y \in \mathcal{Y}} \overline{\eta_y}(x) \Big]$$

- (ii) **Excess Risk** $R(h) - R^*$:

$$R(h) - R^* = \mathbb{E}_X\Big[\Big(\eta_{y_x^*}(x) - \eta_{y_x}(x) \Big) \mathbf{1}_{\{h(x) \neq h^*(x)\}} \Big]$$

  Where $y_x = h(x)$ is the prediction made by an arbitrary classifier $h : \mathcal{X} \to \mathcal{Y}$ and $y_x^* = h^*(x)$ is the prediction made by the Bayes classifier.

Proof (Theorem 2.2)**.**
(The proof of this theorem has been included in the solution of Exercise 2.1). □.

2.2 Likelihood Ratio Test

Overview : Define $\pi_1 = P(Y = 1)$ and $\pi_0 = P(Y = 0)$ be the prior probabilities. Let $p_1(x) = P(X = x | Y = 1)$ and $p_0(x) = P(X = x | Y = 0)$ be the class-conditional densities. Note that we have:

$$\begin{aligned}
\eta(x) &= P(Y = 1 | X = x) \\
&= \frac{P(X = x | Y = 1) P(Y = 1)}{P(X = x | Y = 1) P(Y = 1) + P(X = x | Y = 0) P(Y = 0)} \\
&= \frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_0 p_0(x)} \\
&= \frac{1}{1 + \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)}}
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
\eta(x) \geqslant \frac{1}{2} &\iff \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)} \leqslant 1 \\
&\iff \frac{p_1(x)}{p_0(x)} \geqslant \frac{\pi_0}{\pi_1}
\end{aligned}$$

Proposition 2.1: Likelihood ratio test

The Bayes classifier h^* can be re-defined as followed:

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} \geqslant \frac{\pi_0}{\pi_1} \\ 0 & \text{otherwise} \end{cases}$$

The fraction $\frac{p_1(x)}{p_0(x)}$ is called the **likelihood ratio**.

2.3 Plug-in classifier

Definition 2.1 (Plug-in classifier)**.**

*A **plug-in classifier** is based on an estimate of $\eta(x)$. This estimate is then plugged into the definition of the Bayes classifier. Suppose that $\widehat{\eta_n}$ is an estimate of η based on n training samples $\{(X_i, Y_i)\}_{i=1}^n$. We define $\widehat{h_n}$ as:*

$$\widehat{h_n} = \begin{cases} 1 & \textit{if } \widehat{\eta_n}(x) \geqslant \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

Corollary 2.1: Excess risk of plug-in classifier

We have the following upper-bound for the excess risk of the plug-in classifier:

$$R(\widehat{h_n}) - R^* \leqslant 2\mathbb{E}_X\Big[\Big|\eta(X) - \widehat{\eta_n}(X)\Big|\Big]$$

Proof (Corollary 2.1)**.**

From theorem 2.1, we have:

$$R(\widehat{h_n}) - R^* = 2\mathbb{E}_X\left[\left|\eta(X) - \frac{1}{2}\right|\mathbf{1}_{\{\widehat{h_n}(X) \neq h^*(X)\}}\right]$$

The indicator term will be non-zero in the above equality if one of the following cases occurs:

$$\begin{cases} \widehat{h_n}(X) = 1, h^*(X) = 0 \\ \widehat{h_n}(X) = 0, h^*(X) = 1 \end{cases} \implies \begin{cases} \widehat{\eta_n}(X) \geqslant \frac{1}{2}, \eta(X) < \frac{1}{2} \\ \widehat{\eta_n}(X) < \frac{1}{2}, \eta(X) \geqslant \frac{1}{2} \end{cases}$$

Case 1 : $\widehat{\eta_n}(X) \geqslant \frac{1}{2}, \eta(X) < \frac{1}{2}$

We have:

$$\eta(X) - \widehat{\eta_n}(X) \leqslant \eta(X) - \frac{1}{2} \quad (\textit{Both sides negative})$$
$$\implies \left|\eta(X) - \widehat{\eta_n}(X)\right| \geqslant \left|\eta(X) - \frac{1}{2}\right|$$

Case 2 : $\widehat{\eta_n}(X) < \frac{1}{2}, \eta(X) \geqslant \frac{1}{2}$

We have:

$$\widehat{\eta_n}(X) - \eta(X) \geqslant \widehat{\eta_n}(X) - \frac{1}{2} \geqslant \eta(X) - \frac{1}{2} \quad \textit{(All positive)}$$

Therefore, we have:

$$\left|\eta(X) - \widehat{\eta_n}(X)\right| \geqslant \left|\eta(X) - \frac{1}{2}\right|$$

For both cases, we have the same $\left|\eta(X) - \widehat{\eta_n}(X)\right| \geqslant \left|\eta(X) - \frac{1}{2}\right|$ *inequality. Therefore, we have:*

$$R(\widehat{h_n}) - R^* \leqslant 2\mathbb{E}_X\Big[\left|\eta(X) - \widehat{\eta_n}(X)\right|\Big]$$

□.

2.4 End of chapter exercises

Exercise 2.1

Extend theorem 2.1 to the multi-class classification case where $\mathcal{Y} = \{1, 2, \ldots, M\}$. In other words, prove theorem 2.2.

Solution (Exercise 2.1).

We re-define the Bayes classifier h^* as followed:

$$h^*(x) = \arg\max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\},$$
$$\eta_y(x) = P(Y = y | X = x)$$

We have:

$$\sum_{y \in \mathcal{Y}} \eta_y(x) = 1, \ \forall x \in \mathcal{X}$$

(i) ***Calculate Bayes risk*** R^*

For any classifier $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$R(h) = \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x)\right]$$

Letting $\hat{y}_x = h(x)$ being h's prediction for a given feature vector $x \in \mathcal{X}$, we have:

$$R(h) = \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}; y \neq \hat{y}_x} \eta_y(x)\right] = \mathbb{E}_{x \sim X}\left[1 - \eta_{\hat{y}_x}(x)\right]$$

In order to minimize $R(h)$, we need $\eta_{\hat{y}_x}(x)$ to be maxmized for all $x \in \mathcal{X}$. Hence, we have:

$$R^* = \mathbb{E}_{x \sim X}\left[1 - \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\}\right]$$

Therefore, we have $h^*(x) = \arg\max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\}$ is the Bayes classifier and the Bayes risk $R^* = \mathbb{E}_{x \sim X}\left[1 - \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\}\right]$.

(ii) ***Calculate excess risk*** $R(h) - R^*$

For any $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$\begin{aligned} R(h) - R^* &= \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x)\right] - \mathbb{E}_{x \sim X}\left[1 - \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\}\right] \\ &= \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) + \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\} - 1\right] \end{aligned}$$

Denote $h^*(x) = y_x^*$ and $h(x) = y_x$. When $h(x) = h^*(x) = y_x^*$, we have:

$$\begin{aligned} \sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) + \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\} &= \sum_{y \in \mathcal{Y}; y \neq y_x} \eta_y(x) + \eta_{y_x^*}(x) \\ &= \sum_{y \in \mathcal{Y}; y \neq y_x^*} \eta_y(x) + \eta_{y_x^*}(x) \\ &= \sum_{y \in \mathcal{Y}} \eta_y(x) = 1 \end{aligned}$$

$$\implies \sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) + \max_{y \in \mathcal{Y}} \Big\{\eta_y(x)\Big\} - 1 = 0$$

When $h(x) \neq h^*(x)$, we have:

$$\begin{aligned}\sum_{y\in\mathcal{Y}} \mathbf{1}_{\{h(x)\neq y\}}\eta_y(x) + \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\} - 1 &= \sum_{y\in\mathcal{Y};y\neq y_x} \eta_y(x) + \eta_{y_x^*}(x) - 1 \\ &= 2\eta_{y_x^*}(x) - 1 + \sum_{y\in\mathcal{Y}\setminus\{y_x,y_x^*\}} \eta_y(x) \\ &= 2\eta_{y_x^*}(x) - \Big(\eta_{y_x}(x) + \eta_{y_x^*}(x)\Big) \\ &= \eta_{y_x^*}(x) - \eta_{y_x}(x).\end{aligned}$$

Therefore, we can re-write the excess risk by multiplying the entire integrand with the indicator function $\mathbf{1}_{\{h(x)\neq h^*(x)\}}$ as followed:

$$R(h) - R^* = \mathbb{E}_{x\sim X}\left[\Big(\eta_{y_x^*}(x) - \eta_{y_x}(x)\Big)\mathbf{1}_{\{h(x)\neq h^*(x)\}}\right]$$

(iii) ***Simpler form of Bayes risk***
From (i) we have:

$$R^* = \mathbb{E}_X\Big[1 - \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\}\Big] = \mathbb{E}_X\Big[\min_{y\in\mathcal{Y}}\Big\{\overline{\eta_y}(x)\Big\}\Big]$$

Where $\overline{\eta_y}(x) = P(Y \neq y|X = x)$.

□.

Exercise 2.2

Define the α**-cost-sensitive risk** of a classifier $h : \mathcal{X} \to \mathcal{Y}$ as followed:

$$R_\alpha(h) = \mathbb{E}_{XY}\Big[(1-\alpha)\mathbf{1}_{\{Y=1,h(X)=0\}} + \alpha\mathbf{1}_{\{Y=0,h(X)=1\}}\Big]$$

Define the Bayes classifier and prove and analogue of theorem 2.1.

Solution (Exercise 2.2)**.**
Using the law of total expectation, we have:

$$\begin{aligned}R_\alpha(h) &= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}}\Big[(1-\alpha)\mathbf{1}_{\{y=1,h(x)=0\}} + \alpha\mathbf{1}_{\{y=0,h(x)=1\}}\Big]P(Y=y|X=x)\right] \\ &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\mathbf{1}_{\{h(x)=0\}} + \alpha(1-\eta(x))\mathbf{1}_{\{h(x)=1\}}\Big]\end{aligned}$$

Since $\mathbf{1}_{\{h(x)=0\}}$ and $\mathbf{1}_{\{h(x)=1\}}$ are mutually exclusive, in order for $R_\alpha(h)$ to be minimize, we define the following Bayes classifier:

$$h^*(x) = \begin{cases}1 & \text{if } \alpha(1-\eta(x)) \leqslant (1-\alpha)\eta(x) \\ \\ 0 & \text{otherwise}\end{cases} = \begin{cases}1 & \text{if } \eta(x) \geqslant \alpha \\ \\ 0 & \text{otherwise}\end{cases}$$

We can also derive a likelihood-ratio test version of the Bayes classifier, we have:

$$\begin{aligned}\eta(x) \geqslant \alpha &\implies \frac{1}{1 + \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)}} \geqslant \alpha \\ &\implies 1 + \frac{\pi_0 \cdot p_0(x)}{\pi_1 \cdot p_1(x)} \leqslant \frac{1}{\alpha} \\ &\implies \frac{p_1(x)}{p_0(x)} \geqslant \frac{\alpha}{1-\alpha}\cdot\frac{\pi_0}{\pi_1}\end{aligned}$$

Hence, we can rewrite the Bayes classifier as followed:

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} \geqslant \frac{\alpha}{1-\alpha} \cdot \frac{\pi_0}{\pi_1} \\ \\ 0 & \text{otherwise} \end{cases}$$

(i) ***Bayes Risk*** R^*_α
We have:

$$\begin{aligned} R^*_\alpha &= R_\alpha(h^*) \\ &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\mathbf{1}_{\{h^*(x)=0\}} + \alpha(1-\eta(x))\mathbf{1}_{\{h^*(x)=1\}}\Big] \\ &= \mathbb{E}_X\Big[\min(\alpha(1-\eta(X)), (1-\alpha)\eta(X))\Big] \end{aligned}$$

(ii) ***Excess Risk*** $R_\alpha(h) - R^*_\alpha$
For an arbitrary $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$\begin{aligned} R_\alpha(h) - R^*_\alpha &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\Big(\mathbf{1}_{\{h(x)=0\}} - \mathbf{1}_{\{h^*(x)=0\}}\Big) + \alpha(1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1\}} - \mathbf{1}_{\{h^*(x)=1\}}\Big)\Big] \\ &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\Big(\mathbf{1}_{\{h(x)=0,h^*(x)=1\}} - \mathbf{1}_{\{h(x)=1,h^*(x)=0\}}\Big) \\ &\quad + \alpha(1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1,h^*(x)=0\}} - \mathbf{1}_{\{h(x)=0,h^*(x)=1\}}\Big)\Big] \\ &= \mathbb{E}_{x\sim X}\Big[\mathbf{1}_{\{h(x)=0,h^*(x)=1\}}(\eta(x)-\alpha) + \mathbf{1}_{\{h(x)=1,h^*(x)=0\}}(\alpha-\eta(x))\Big] \\ &= \mathbb{E}_X\Big[\Big|\eta(X)-\alpha\Big|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\Big] \end{aligned}$$

□.

3 Hoeffding's inequality

3.1 Markov's Inequality

Proposition 3.1: Markov's Inequality

Let U be a non-negative random variable on $\mathbb{R}$, then for all $t > 0$, we have:

$$P(U \geqslant t) \leqslant \frac{1}{t}\mathbb{E}[U]$$

Proof (Proposition 3.1). *We have:*

$$\begin{aligned} tP(U \geqslant t) &= t\mathbb{E}\Big[\mathbf{1}_{\{U \geqslant t\}}\Big] \\ &= t\int_0^\infty \mathbf{1}_{\{x \geqslant t\}} f_U(x)dx \\ &= t\int_t^\infty f_U(x)dx \\ &\leqslant \int_t^\infty x f_U(x)dx \\ &\leqslant \int_0^\infty x f_U(x)dx = \mathbb{E}[U] \\ \implies P(U \geqslant t) &\leqslant \frac{1}{t}\mathbb{E}[U] \end{aligned}$$

□.

Corollary 3.1: Chebyshev's Inequality

Let Z be a random variable on $\mathbb{R}$ with mean μ and variance σ^2, we have:

$$P\Big(\Big|Z - \mu\Big| \geqslant t\Big) \leqslant \frac{\sigma^2}{t^2}$$

Proof (Corollary 3.1). *Using Markov's inequality, we have:*

$$\begin{aligned} P\Big(\Big|Z - \mu\Big| \geqslant t\Big) &= P\Big(\Big|Z - \mu\Big|^2 \geqslant t^2\Big) \\ &\leqslant \frac{\mathbb{E}\Big[\Big|Z - \mu\Big|^2\Big]}{t^2} = \frac{\sigma^2}{t^2} \end{aligned}$$

□.

Corollary 3.2: Chernoff's bounding method

Let Z be a random variable on $\mathbb{R}$, for any $t > 0$, we have:

$$P(Z \geqslant t) \leqslant \inf_{s>0} e^{-st} M_Z(s)$$

Proof (Corollary 3.2).

We have:

$$\begin{aligned} P(Z \geqslant t) &= P(sZ \geqslant st), \quad (t > 0) \\ &= P(e^{sZ} \geqslant e^{st}) \\ &\leqslant \frac{\mathbb{E}\left[e^{sZ}\right]}{e^{st}} = e^{-st} M_Z(s) \quad (\textit{Markov's inequality}) \end{aligned}$$

Since the above inequality holds for all $s > 0$, we can just take the infimum to obtain the tightest bound. Hence, we have:

$$P(Z \geqslant t) \leqslant \inf_{s>0} e^{-st} M_Z(s)$$

□.

3.2 Hoeffding's Inequality

Before diving into Hoeffding's inequality, we need to go through the following lemma (whose proof will not be included) that will help us prove the Hoeffding's inequality:

Lemma 3.1: Hoeffding's lemma

Let V be a random variable on $\mathbb{R}$ with $\mathbb{E}[V] = 0$ and suppose that $a \leqslant V \leqslant b$ with probability one. We have:

$$\mathbb{E}\left[e^{sV}\right] \leqslant \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Proof (Lemma 3.1)**.**
(The proof for this lemma can be found here Wikipedia 2024b). □.

Theorem 3.1: Hoeffding's Inequality

Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables on $\mathbb{R}$ such that $a_i \leqslant Z_i \leqslant b_i$ with probability one for all $1 \leqslant i \leqslant n$. Let $S_n = \sum_{i=1}^n Z_i$. We have:

$$P\Big(\Big|S_n - \mathbb{E}[S_n]\Big| \geqslant t\Big) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad \forall t > 0$$

Proof (Theorem 3.1)**.**

Using the Chernoff's bounds, we have:

$$\begin{aligned}
P\Big(S_n - \mathbb{E}[S_n] \geqslant t\Big) &\leqslant \inf_{s>0} e^{-st} M_{S_n - \mathbb{E}[S_n]}(s) \\
&= \inf_{s>0} e^{-st} \mathbb{E}\Big[e^{s(S_n - \mathbb{E}[S_n])}\Big] \\
&= \inf_{s>0} e^{-st} \mathbb{E}\left[\exp\left(s \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i])\right)\right] \\
&= \inf_{s>0} e^{-st} \mathbb{E}\left[\prod_{i=1}^{n} \exp\Big(s(Z_i - \mathbb{E}[Z_i])\Big)\right] \\
&= \inf_{s>0} e^{-st} \prod_{i=1}^{n} \mathbb{E}\left[\exp\Big(s(Z_i - \mathbb{E}[Z_i])\Big)\right] \quad (\textit{Since all } Z_i - \mathbb{E}[Z_i] \textit{ are independent}) \\
&\leqslant \inf_{s>0} e^{-st} \prod_{i=1}^{n} \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) \quad (\textit{By Hoeffding's lemma}) \\
&= \inf_{s>0} \exp\left(-st + \sum_{i=1}^{n} \frac{s^2(b_i - a_i)^2}{8}\right)
\end{aligned}$$

In order for the above to be minimized, we differentiate the term inside the exponential and set the derivative to 0 *to find the optimal* $s > 0$*. We have:*

$$-t + s \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{4} = 0 \implies s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

Letting $c = \sum_{i=1}^{n}(b_i - a_i)^2$*, we now can derive the tightest Chernoff's bound as followed:*

$$\begin{aligned}
P\Big(S_n - \mathbb{E}[S_n] \geqslant t\Big) &\leqslant \exp\left(-\frac{4t^2}{c} + \frac{16t^2}{c^2} \cdot \frac{c}{8}\right) = \exp\left(-\frac{2t^2}{c}\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)
\end{aligned}$$

Repeating the same argument, we can also prove that:

$$P\Big(\mathbb{E}[S_n] - S_n \geqslant t\Big) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Combining the two sides of the inequality, we have:

$$P\Big(\Big|S_n - \mathbb{E}[S_n]\Big| \geqslant t\Big) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

□.

3.3 Convergence of Empirical Risk

Definition 3.1 (Empirical Risk ($\widehat{R_n}$))**.** *Suppose we are given training data* $\Big\{(X_i, Y_i)_{i=1}^n\Big\}$ *such that each pair* $(X_i, Y_i) \sim P_{XY}$ *are independently identically distributed. Let* $h : \mathcal{X} \to \mathcal{Y}$ *be a classifier. We define the* ***empirical risk*** *to be:*

$$\widehat{R_n}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{h(X_i) \neq Y_i\}}$$

Note that $\mathbb{E}[\widehat{R}_n(h)] = R(h)$ *and* $n\widehat{R}_n(h) \sim Binomial(n, R(h))$. *In the following corollary of the Hoeffding's inequality, we will answer the question* ***how close the empirical risk is as an estimate of true risk*** *or* ***how fast the empirical risk converges to the true risk.***

Corollary 3.3: Convergence of Empirical Risk

Given training data $\left\{(X_i, Y_i)_{i=1}^n\right\}$ such that each pair $(X_i, Y_i) \sim P_{XY}$ are independently identically distributed. Let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier, we have:

$$P\left(\left|\widehat{R}_n(h) - R(h)\right| \geqslant \epsilon\right) \leqslant 2e^{-2n\epsilon^2}, \quad \epsilon > 0$$

Proof (Corollary 3.3)**.**
For all $1 \leqslant i \leqslant n$, *we have* $\mathbf{1}_{\{h(X_i) \neq Y_i\}} \in \{0, 1\}$. *Hence, with probability one,* $0 \leqslant \mathbf{1}_{\{h(X_i) \neq Y_i\}} \leqslant 1$ *and* $b_i = 1, a_i = 0$ *for all* $1 \leqslant i \leqslant n$.

Using the Hoeffding's inequality, we have:

$$\begin{aligned}
P\left(\left|\widehat{R}_n(h) - R(h)\right| \geqslant \epsilon\right) &= P\left(\left|\widehat{R}_n(h) - \mathbb{E}[\widehat{R}_n(h)]\right| \geqslant \epsilon\right) \\
&= P\left(\left|n\widehat{R}_n(h) - \mathbb{E}[n\widehat{R}_n(h)]\right| \geqslant n\epsilon\right) \\
&\leqslant 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \textit{(Hoeffding's inequality)} \\
&= 2e^{-2n\epsilon^2}
\end{aligned}$$

□.

3.4 KL-divergence & Hypothesis Testing

Set-up (Hypothesis Testing) : Suppose that we have $\mathcal{Y} = \{0, 1\}$ and P_{XY} is a distribution on $\mathcal{X} \times \mathcal{Y}$. Let's assume that:

- The prior probabilities π_y are equal.
- The supports of likelihoods p_0, p_1 are the same.
- $0 < \alpha \leqslant p_y(x) \leqslant \beta < \infty$ for all $x \in \mathcal{X}$ such that $p_y(x) > 0$ and for all $y \in \{0, 1\}$.

Now suppose $X_1, \ldots, X_n \sim p_y$ are independently identically distributed where $y \in \{0, 1\}$ is unknown. Can we guess y and how good our guess would be?

Proposition 3.2: KL-divergence hypothesis testing

From the above settings, the optimal classifier is given by the likelihood ratio test:

$$\widehat{h_n}(x) = \begin{cases} 1 & \text{if } \frac{\prod_{i=1}^n p_1(x_i)}{\prod_{i=1}^n p_0(x_i)} \geqslant \frac{\pi_0}{\pi_1} \quad (=1) \\ \\ 0 & \text{otherwise} \end{cases}$$

Where $x = \Big(x_1, \ldots, x_n\Big)$ is an observation of the random vector $X = \Big(X_1, \ldots, X_n\Big)$. Define the class-specific risk $R_y(h)$ be the risk of misclassification when the true label is $Y = y$:

$$R_y(h) = P(h(X) \neq Y | Y = y)$$

Then, we have:

$$R_0(\widehat{h_n}) \leqslant e^{-2nD(p_0||p_1)^2/c}, \text{ where } c = 4(\log\beta - \log\alpha)^2$$

Where $D(p_0||p_1)$ is the KL-divergence of p_1 from p_0. We can prove a similar exponentially decaying bound for $R_1(\widehat{h_n})$.

Proof.

Proposition 3.2 We can rewrite the optimal classifier as:

$$\widehat{h_n}(X) = \begin{cases} 1 & \textit{if } \widehat{S_n}(X_1, \ldots, X_n) \geqslant 0 \\ \\ 0 & \textit{otherwise} \end{cases}$$

Where we have:

$$\begin{aligned} \widehat{S_n}(X_1, \ldots, X_n) &= \log \frac{\prod_{i=1}^n p_1(X_i)}{\prod_{i=1}^n p_0(X_i)} \\ &= \sum_{i=1}^n \log \frac{p_1(X_i)}{p_0(X_i)} \\ &= \sum_{i=1}^n Z_i \quad \left(\textit{Letting } Z_i = \log \frac{p_1(X_i)}{p_0(X_i)}\right) \end{aligned}$$

Since the likelihoods are bounded, we have:

$$a_i = \log\frac{\alpha}{\beta} \leqslant Z_i \leqslant \log\frac{\beta}{\alpha} = b_i, \quad 1 \leqslant i \leqslant n$$

Now, we have:

$$\begin{aligned} R_0(\widehat{h_n}) &= P(h(X) \neq Y | Y = 0) \\ &= P(\widehat{S_n} \geqslant 0 | Y = 0) \\ &= P(\widehat{S_n} - \mathbb{E}[S_n | Y = 0] \geqslant -\mathbb{E}[S_n | Y = 0] | Y = 0) \end{aligned}$$

To calculate the conditional expectation $\mathbb{E}[S_n | Y = 0]$, we have:

$$\begin{aligned} \mathbb{E}[S_n | Y = 0] &= n\mathbb{E}[Z_1 | Y = 0] \\ &= n \int \log \frac{p_1(x)}{p_0(x)} p_0(x) dx \\ &= -n \int \log \frac{p_0(x)}{p_1(x)} p_0(x) dx = -nD(p_0||p_1) \end{aligned}$$

Therefore, we have:

$$\begin{aligned} R_0(\widehat{h_n}) &= P(\widehat{S_n} - \mathbb{E}[S_n|Y=0] \geqslant nD(p_0||p_1)|Y=0) \\ &\leqslant \exp\left(-\frac{2n^2 D(p_0||p_1)^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \quad (\textit{Hoeffding's inequality}) \end{aligned}$$

For every $1 \leqslant i \leqslant n$*, we have:*

$$\begin{aligned} b_i - a_i &= \log\frac{\beta}{\alpha} - \log\frac{\alpha}{\beta} \\ &= \log\frac{\beta^2}{\alpha^2} = 2\log\frac{\beta}{\alpha} = 2(\log\beta - \log\alpha) \\ \implies \sum_{i=1}^n (b_i - a_i)^2 &= 4n(\log\beta - \log\alpha)^2 \end{aligned}$$

Finally, we have:

$$R_0(\widehat{h_n}) \leqslant \exp\left(-\frac{2nD(p_0||p_1)^2}{4(\log\beta - \log\alpha)^2} \right)$$

Similarly, for $R_1(\widehat{h_n})$*, we have:*

$$R_1(\widehat{h_n}) \leqslant \exp\left(-\frac{2nD(p_1||p_0)^2}{4(\log\beta - \log\alpha)^2} \right)$$

□.

3.5 End of chapter exercises

Exercise 3.1

- (**i**) Apply Chernoff's bounding method to obtain an exponential bound on the tail probability $P(Z \geqslant t)$ for a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$.
- (**ii**) Appealing to the central limit theorem, use part (**i**) to give an approximate bound on the binomial tail. This should not only match the exponential decay given by Hoeffding's inequality, but also reveal the dependence on the variance of the binomial.

Solution (Exercise 3.1)**.**

.

(**i**) ***Chernoff's bounds for*** $Z \sim \mathcal{N}(\mu, \sigma^2)$

Using the Chernoff's bounding method, we have:

$$\begin{aligned} P(Z \geqslant t) &\leqslant \inf_{s>0} e^{-st} M_Z(s) \\ &= \inf_{s>0} \exp\left(-st + \mu s + \frac{1}{2}\sigma^2 s^2 \right) \end{aligned}$$

The above bound is the tightest when the derivative of the term inside the exponential equals zero. Hence, we have:

$$-t + \mu + s\sigma^2 = 0 \implies s = \frac{t-\mu}{\sigma^2}$$

From the above, we have the tightest Chernoff's bound as followed:

$$P(Z \geqslant t) \leqslant \exp\left(-\frac{(t-\mu)^2}{\sigma^2} + \frac{(t-\mu)^2}{2\sigma^2} \right) = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2} \right)$$

(**ii**) ***Binomial tail upper bound***

Let S_n be the binomial random variable such that:

$$S_n = \sum_{i=1}^{n} X_i, \quad X_i \sim Bernoulli(p)$$

For a positive $\epsilon > 0$, we want to know the upper tail bound $P(S_n - \mathbb{E}[S_n] \geqslant \epsilon)$. Letting $\overline{X} = \frac{1}{n} S_n$, we have:

$$\begin{aligned} P(S_n - \mathbb{E}[S_n] \geqslant \epsilon) &= P\left(\overline{X} - \frac{\mathbb{E}[S_n]}{n} \geqslant \frac{\epsilon}{n} \right) \\ &= P\left(\overline{X} - p \geqslant \frac{\epsilon}{n} \right) \\ &= P\left(\frac{\overline{X} - p}{\sqrt{pq}/\sqrt{n}} \geqslant \frac{\epsilon}{\sqrt{npq}} \right), \quad (q = 1-p) \end{aligned}$$

By the Central Limit Theorem, we have:

$$\frac{\overline{X} - p}{\sqrt{pq}/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Hence, as $n \to \infty$, the upper tail bound would be:

$$P(S_n - \mathbb{E}[S_n] \geqslant \epsilon) = P\left(\frac{\overline{X} - p}{\sqrt{pq}/\sqrt{n}} \geqslant \frac{\epsilon}{\sqrt{npq}}\right)$$
$$\leqslant \exp\left(-\frac{\epsilon^2}{2npq}\right) = \exp\left(-\frac{\epsilon^2}{2Var(S_n)}\right)$$

Double-check the bound with Hoeffding's inequality, we have:

$$P(S_n - \mathbb{E}[S_n] \geqslant \epsilon) \leqslant \exp\left(-\frac{2\epsilon^2}{n}\right)$$

□.

Exercise 3.2

Can you remove the assumption in $0 < \alpha \leqslant p_y(x)$? Consider other restrictions on p_y, other concentration inequalities, or other f-divergences.

Solution (Exercise 3.2)**.**
When we remove the assumption that $0 < \alpha \leqslant p_y(x)$, the class-conditional densities are not bounded below. Hence, we have:

$$\exp\left(-\frac{2nD(p_1||p_0)^2}{4(\log\beta - \log\alpha)^2}\right) \to 1 \text{ when } \alpha \to 0$$

In other words, the bound is no longer meaningful. We can instead use the Chernoff bounding method:

$$\begin{aligned} R_0(\widehat{h_n}) &= P(S_n \geqslant 0 | Y = 0) \\ &\leqslant \inf_{s>0} \prod_{i=1}^{n} \mathbb{E}_{q_0}\left[e^{sZ_i}\right] \\ &= \inf_{s>0} \prod_{i=1}^{n} \mathbb{E}_{q_0}\left[\exp\left(s \log \frac{p_1(X_i)}{p_0(X_i)}\right)\right] \\ &= \inf_{s>0} \prod_{i=1}^{n} \mathbb{E}_{q_0}\left[\frac{p_1(X_i)^s}{p_0(X_i)^s}\right] \end{aligned}$$

Taking logarithm from both sides, we have:

$$\begin{aligned} \log R_0(\widehat{h_n}) &\leqslant \inf_{s>0} \sum_{i=1}^{n} \log \mathbb{E}_{q_0}\left[\frac{p_1(X_i)^s}{p_0(X_i)^s}\right] \\ &= \inf_{s>0} \sum_{i=1}^{n} (s-1) R_s(p_1||p_0) \\ &= \inf_{s>0} n(s-1) R_s(p_1||p_0) \end{aligned}$$

Where $R_s(p_1||p_0)$ is the Renyi divergence Wikipedia 2024c. □.

4 Empirical Risk Minimization

4.1 Uniform Deviation Bounds

Definition 4.1 (Empirical Risk Minimization ($\widehat{h_n}$))**.**
Let $\left\{(X_i, Y_i)\right\}_{i=1}^{n}$ *be independently identically distributed random variables sampled from* P_{XY}*. Let* $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ *be a set of classifiers.* ***Empirical Risk Minimization*** *is a learning algorithm such that:*

$$\widehat{h_n} = \arg\min_{h \in \mathcal{H}} \widehat{R_n}(h)$$

Where $\widehat{R_n}$ *is the empirical risk and* $\widehat{h_n}$ *is called the* ***Empirical Risk Minimizer****. An important question is how close* $\widehat{R_n}$ *is to* $R^*_{\mathcal{H}} = \inf_{h \in \mathcal{H}} R(h)$*.*

Overview (Uniform Deviation Bounds) : Previously, we proved the following bound using the Hoeffding's inequality:

$$P\left(\left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon\right) \leqslant \delta$$

Where $\delta = 2e^{-2n\epsilon^2}$. **However, since we do not know $\widehat{h_n}$ (the specific function in $\mathcal{H}$ that minimizes the empirical risk), we look for a bound that is guaranteed to apply for all $h \in \mathcal{H}$. This is called the Uniform Deviation Bound**.

Definition 4.2 (Uniform Deviation Bounds (UDB))**.**
Given a set of classifiers $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$*,* $\epsilon > 0$*, the* ***Uniform Deviation Bounds*** *is the probability that for at least one* $h \in \mathcal{H}$*, the empirical risk deviates away from the true risk by* ϵ *and has the following form:*

$$P\left(\sup_{h \in \mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \leqslant \epsilon\right) \geqslant 1 - \delta$$

$$Or: P\left(\sup_{h \in \mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon\right) \leqslant \delta$$

The above bounds have the following interpretations:

- *The probability that the deviation from the true risk is at most* ϵ *for all functions in* $\mathcal{H}$ *is at least* $1 - \delta$*.*
- *The probability that there exists at least a function in* $\mathcal{H}$ *whose deviation from the true risk is at least* ϵ *is at most* δ*.*

Basically, we want to ***bound the probability that some function deviates too far from the true risk****.*

Theorem 4.1: Uniform Deviation Bounds for finite $\mathcal{H}$

Assume that $|\mathcal{H}| < \infty$. We have:

$$P\left(\sup_{h \in \mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon\right) \leqslant 2|\mathcal{H}|e^{-2n\epsilon^2}$$

Proof (Theorem 4.1). *For $h \in \mathcal{H}$, define the following event:*

$$\Omega_\epsilon(h) = \left\{ \left| \widehat{R_n}(h) - R(h) \right| \geqslant \epsilon \right\}$$

Which is the event that the function h deviates away from the true risk by $\epsilon > 0$. Now, define the following event:

$$\Omega_\epsilon(\mathcal{H}) = \bigcup_{h \in \mathcal{H}} \Omega_\epsilon(h)$$

Which is the event that at least one $h \in \mathcal{H}$ deviates away from the true risk by $\epsilon > 0$. We have:

$$\begin{aligned} P\left(\sup_{h \in \mathcal{H}} \left| \widehat{R_n}(h) - R(h) \right| \geqslant \epsilon \right) &= P(\Omega_\epsilon(\mathcal{H})) \\ &= P\left(\bigcup_{h \in \mathcal{H}} \Omega_\epsilon(h) \right) \\ &\leqslant \sum_{h \in \mathcal{H}} P(\Omega_\epsilon(h)) \\ &\leqslant \sum_{h \in \mathcal{H}} 2e^{-2n\epsilon^2} \quad (\textit{Corollary 3.3}) \\ &= 2|\mathcal{H}|e^{-2n\epsilon^2} \end{aligned}$$

□.

Proposition 4.1: (Probabilistic) Bound on Excess Risk of $\widehat{h_n}$

Suppose that $\mathcal{H}$ satisfies:

$$P\left(\sup_{h \in \mathcal{H}} \left| \widehat{R_n}(h) - R(h) \right| \geqslant \epsilon \right) \leqslant \delta$$

Then, with probability of at least $1 - \delta$, we have the following **upper bound on the Excess Risk of the Empirical Risk Minimizer**:

$$R(\widehat{h_n}) - R^*_{\mathcal{H}} \leqslant 2\epsilon$$

In other words, **with probability $1-\delta$, the empirical risk minimizer deviates from the true risk minimizer by at most 2ϵ.**

Proof (Proposition 4.1). *We have:*

$$P\left(\sup_{h \in \mathcal{H}} \left| \widehat{R_n}(h) - R(h) \right| \geqslant \epsilon \right) \leqslant \delta \implies P\left(\sup_{h \in \mathcal{H}} \left| \widehat{R_n}(h) - R(h) \right| \leqslant \epsilon \right) \geqslant 1 - \delta$$

Hence, with probability $1 - \delta$, for all $h \in \mathcal{H}$, we have:

$$\begin{aligned} \left| \widehat{R_n}(h) - R(h) \right| \leqslant \epsilon &\implies -\epsilon \leqslant \widehat{R_n}(h) - R(h) \leqslant \epsilon \\ &\implies \begin{cases} \widehat{R_n}(h) & \leqslant R(h) + \epsilon \\ R(h) & \leqslant \widehat{R_n}(h) + \epsilon \end{cases} \end{aligned}$$

Therefore:

$$\begin{aligned} R(\widehat{h_n}) &\leqslant \widehat{R}_n(\widehat{h_n}) + \epsilon \\ &\leqslant \widehat{R}_n(h) + \epsilon \quad \textit{(Since } \widehat{h_n} \textit{ minimizes the Empirical Risk)} \\ &\leqslant \Big(R(h) + \epsilon\Big) + \epsilon = R(h) + 2\epsilon \end{aligned}$$

Since $h \in \mathcal{H}$ is an arbitrary choice, we take the infimum over $\mathcal{H}$ to get the tightest bound. We have:

$$\begin{aligned} R(\widehat{h_n}) &\leqslant \inf_{h \in \mathcal{H}} R(h) + 2\epsilon \\ &= R^*_{\mathcal{H}} + 2\epsilon \end{aligned}$$

□.

Remark : We can express the above proposition verbally as "**If the UDB is at most δ, then with probability $1-\delta$, the Excess Risk of the Empirical Risk Minimizer is at most 2ϵ**".
Remark : Note that the above proof assumes that *there exists an empirical risk minimizer*. This is not guaranteed when $|\mathcal{H}|$ is infinite.

Proposition 4.2: (Non-probabilistic) Bound on Excess Risk of $\widehat{h_n}$

We have the following inequality:

$$R(\widehat{h_n}) - R^*_{\mathcal{H}} \leqslant 2 \sup_{h \in \mathcal{H}} \Big| \widehat{R}_n(h) - R(h) \Big|$$

Proof (Proposition 4.2). *Let $h^*_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(h)$. We have:*

$$R(\widehat{h_n}) - R^*_{\mathcal{H}} \leqslant \Big| R(\widehat{h_n}) - \widehat{R}_n(\widehat{h_n}) \Big| + \widehat{R}_n(\widehat{h_n}) - \widehat{R}_n(h^*_{\mathcal{H}}) + \Big| \widehat{R}_n(h^*_{\mathcal{H}}) - R^*_{\mathcal{H}} \Big|$$

Since $\widehat{h_n}$ is the Empirical Risk Minimizer, we have $\widehat{R}_n(\widehat{h_n}) - \widehat{R}_n(h^*_{\mathcal{H}}) \leqslant 0$. Hence:

$$\begin{aligned} R(\widehat{h_n}) - R^*_{\mathcal{H}} &\leqslant \Big| R(\widehat{h_n}) - \widehat{R}_n(\widehat{h_n}) \Big| + \Big| \widehat{R}_n(h^*_{\mathcal{H}}) - R^*_{\mathcal{H}} \Big| \\ &\leqslant 2 \sup_{h \in \mathcal{H}} \Big| \widehat{R}_n(h) - R(h) \Big| \end{aligned}$$

□.

Corollary 4.1: Excess Risk of $\widehat{h_n}$ - $\delta \to \epsilon$ relation

This is a Corollary for both proposition 4.1 and proposition 4.2. If $\mathcal{H}$ is finite, then:

$$P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) \leqslant \underbrace{2|\mathcal{H}|e^{-n\epsilon^2/2}}_{\delta}$$

Equivalently, with probability of at least $1-\delta$, we have:

$$R(\widehat{h_n}) \leqslant R^*_{\mathcal{H}} + \sqrt{\frac{2}{n}\left(\log|\mathcal{H}| - \log\frac{\delta}{2}\right)}$$

Proof (Corollary 4.1)**.**
By proposition 4.2, we have:

$$\begin{aligned} P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) &\leqslant P\left(2 \sup_{h\in\mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon)\right) \\ &= P\left(\sup_{h\in\mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \frac{\epsilon}{2})\right) \\ &\leqslant 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2}\right) \end{aligned}$$

Now, let:

$$\delta = 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2}\right) \implies \epsilon = \sqrt{\frac{2}{n}\left(\log|\mathcal{H}| - \log\frac{\delta}{2}\right)}$$

By proposition 4.1, with at least probability $1-\delta$*, we have:*

$$R(\widehat{h_n}) \leqslant R^*_{\mathcal{H}} + \epsilon = R^*_{\mathcal{H}} + \sqrt{\frac{2}{n}\left(\log|\mathcal{H}| - \log\frac{\delta}{2}\right)}$$

□.

4.2 PAC Learning & Sample Complexity

Definition 4.3 (PAC & Sample Complexity ($N(\epsilon,\delta)$))**.**
We say that $\widehat{h_n}$ *is a* (ϵ,δ)***-learning algorithm*** *for* $\mathcal{H}$ *if there exists a function* $N(\epsilon,\delta)$ *such that:*

$$\forall \epsilon,\delta > 0 : n \geqslant N(\epsilon,\delta) \implies P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) \leqslant \delta$$

Where we have:

- $N(\epsilon,\delta)$ *is called the* ***Sample Complexity****.*
- $\mathcal{H}$ *is called* ***Uniformly Learnable****.*
- $\widehat{h_n}$ *is called* ***Probably Approximately Correct (PAC)****.*

Remark : By corollary 4.1, we have $\delta = 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2}\right)$. Solving for n, we have:

$$N(\epsilon,\delta) = \frac{2}{\epsilon^2}\left(\log|\mathcal{H}| - \log\frac{\delta}{2}\right)$$

4.3 Zero-error case

In the following proposition, we can obtain a tighter bound for the zero empirical risk case. However, it is not particularly useful in many cases.

Proposition 4.3: Zero-error case bound

If $\widehat{R_n}(\widehat{h_n}) = 0$ and $|\mathcal{H}| < \infty$, we have:

$$P\left(\exists h \in \mathcal{H} : \widehat{R_n}(h) = 0, R(h) \geqslant \epsilon\right) \leqslant \underbrace{|\mathcal{H}|e^{-n\epsilon}}_{\delta}$$

Meaning, with probability of at least $1-\delta$, if $\widehat{R_n}(h) = 0$ then $R(h) \leqslant \frac{1}{n}(\log|\mathcal{H}| - \log\delta)$.

Proof (Proposition 4.3). *Let* $\Omega_0(h) = \left\{\widehat{R_n}(h) = 0\right\}$ *and define the event* Ω_ϵ *as:*

$$\Omega_\epsilon = \bigcup_{h\in\mathcal{H};R(h)\geqslant\epsilon} \Omega_0(h) = \left\{\exists h \in \mathcal{H} : \widehat{R_n}(h) = 0, R(h) \geqslant \epsilon\right\}$$

For any $h \in \mathcal{H}$ *such that* $R(h) \geqslant \epsilon$*, we have:*

$$\begin{aligned}
P(\Omega_0(h)) &= P\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(X_i)\neq Y_i\}} = 0\right) \\
&= P\left(\sum_{i=1}^{n} \mathbf{1}_{\{h(X_i)\neq Y_i\}} = 0\right) \\
&= P\left(\bigcup_{i=1}^{n}\left\{h(X_i) = Y_i\right\}\right) \\
&= \prod_{i=1}^{n} P(h(X_i) = Y_i) \quad \textit{(Since all } (X_i, Y_i) \textit{ pairs are independent)}
\end{aligned}$$

Each $\mathbf{1}_{\{h(X_i)\neq Y_i\}}$ *is a Bernoulli variable with hit probability* $p_i = 1 - \mathbb{E}\Big[h(X_i) \neq Y_i\Big] = 1 - R(h)$*. Hence, we have:*

$$\begin{aligned}
P(\Omega_0(h)) &= \prod_{i=1}^{n} P(h(X_i) = Y_i) \\
&= (1 - R(h))^n \\
&\leqslant (1-\epsilon)^n
\end{aligned}$$

Using the inequality $\log(1-\epsilon) \leqslant -\epsilon$*, we have:*

$$\begin{aligned}
P(\Omega_0(h)) &\leqslant (1-\epsilon)^n = e^{n\log(1-\epsilon)} \\
&\leqslant e^{-n\epsilon}
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
P(\Omega_\epsilon) &= P\left(\bigcup_{h\in\mathcal{H};R(h)\geqslant\epsilon} \Omega_0(h)\right) \\
&\leqslant \sum_{h\in\mathcal{H};R(h)\geqslant\epsilon} P(\Omega_0(h)) \\
&\leqslant \sum_{h\in\mathcal{H};R(h)\geqslant\epsilon} e^{-n\epsilon} \\
&\leqslant |\mathcal{H}|e^{-n\epsilon}
\end{aligned}$$

□.

Remark : Note that the bound obtained in proposition 4.3 is NOT the Uniform Deviation Bound (UDB) because we have:

$$\left\{ \sup_{h\in\mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon \right\} = \left\{ \exists h \in \mathcal{H} : \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon \right\}$$

Therefore, we have:

$$\left\{ \exists h \in \mathcal{H} : \widehat{R_n}(h) = 0, R(h) \geqslant \epsilon \right\} \subseteq \left\{ \sup_{h\in\mathcal{H}} \left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon \right\}$$

Remark : This is trivial improvement. However, define the following subset of $\mathcal{H}$:

$$H_\epsilon^+ = \left\{ h \in \mathcal{H} : R(h) \geqslant \epsilon \right\}$$

We can improve the bound in proposition 4.3 as followed:

$$P(\Omega_\epsilon) \leqslant |H_\epsilon^+| e^{-n\epsilon}$$

4.4 End of chapter exercises

Exercise 4.1: Neyman-Pearson Criterion

The probability of error is not the only performance measure for binary classification. Indeed, the probability of error depends on the prior probability of the class label Y, and it may be that the frequency of the classes changes from training to testing data. In such cases, it is desirable to have a performance measure that does not require knowledge of the prior class probability. Let P_y be the class conditional distribution of class $y \in \{0,1\}$. Define $R_y(h) = P_y(h(X) \neq y)$. Also let $\alpha \in (0,1)$. For $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$, define:

$$\begin{aligned} R^*_{\mathcal{H},1} = \inf_{h \in \mathcal{H}} & R_1(h) \\ \text{s.t. } & R_0(h) \leqslant \alpha \end{aligned}$$

In this problem you will investigate a discrimination rule that is probably approximately correct with respect to the above criterion, which is sometimes called the Neyman-Pearson criterion based on connections to the Neyman-Pearson lemma in hypothesis testing.
Suppose we observe $X^y_1, X^y_2, \ldots, X^y_{n_y} \sim P_y$ for $y \in \{0,1\}$. Define the empirical errors:

$$\widehat{R_y}(h) = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{1}_{\{h(X^y_i) \neq y\}}$$

Fix $\epsilon > 0$ and consider the discrimination rule:

$$\begin{aligned} \widehat{h_n} = \arg\min_{h \in \mathcal{H}} & \widehat{R_1}(h) \\ \text{s.t. } & \widehat{R_0}(h) \leqslant \alpha + \frac{\epsilon}{2} \end{aligned}$$

Suppose $\mathcal{H}$ is finite. Show that with high probability:

$$R_0(\widehat{h_n}) \leqslant \alpha + \epsilon \text{ and } R_1(\widehat{h_n}) \leqslant R^*_{\mathcal{H},1} + \epsilon$$

Solution (Exercise 4.1)**.**
We will prove each point one by one:

- (**i**) $R_0(\widehat{h_n}) \leqslant \alpha + \epsilon$ ***with high probability***.

  Claim 1 : $\forall y \in \{0,1\}, \epsilon > 0 : P\left(\sup_{h \in \mathcal{H}} \left| \widehat{R_y}(h) - R_y(h) \right| \geqslant \epsilon \right) \leqslant 2|\mathcal{H}|e^{-2n\epsilon^2}$

  We have that $n\widehat{R_n}(h) \sim Binomial(n, R_y(h))$ *for all* $h \in \mathcal{H}$*. Hence, we have:*

$$\begin{aligned} P\left(\left| \widehat{R_y}(h) - R_y(h) \right| \geqslant \epsilon \right) &= P\left(\left| n\widehat{R_y}(h) - nR_y(h) \right| \geqslant n\epsilon \right) \\ &= P\left(\left| n\widehat{R_y}(h) - \mathbb{E}\left[n\widehat{R_y}(h) \right] \right| \geqslant n\epsilon \right) \\ &\leqslant 2\exp\left(-\frac{2n^2\epsilon^2}{n} \right) = 2e^{-2n\epsilon^2} \quad (\textit{Hoeffding's inequality}) \end{aligned}$$

From the above, we have:

$$\begin{aligned} P\left(\sup_{h\in\mathcal{H}}\left|\widehat{R_y}(h)-R_y(h)\right|\geqslant\epsilon\right) &= P\left(\bigcup_{h\in\mathcal{H}}\left\{\left|\widehat{R_y}(h)-R_y(h)\right|\geqslant\epsilon\right\}\right) \\ &\leqslant \sum_{h\in\mathcal{H}} P\left(\left|\widehat{R_y}(h)-R_y(h)\right|\geqslant\epsilon\right) \\ &\leqslant \sum_{h\in\mathcal{H}} 2e^{-2n\epsilon^2} = 2|\mathcal{H}|e^{-2n\epsilon^2} \end{aligned}$$

From the assumption, we have:

$$\widehat{R_0}(\widehat{h_n})\leqslant\alpha+\frac{\epsilon}{2}$$

Hence, we have:

$$\begin{aligned} R_0(\widehat{h_n}) &= \widehat{R_0}(\widehat{h_n})+R_0(\widehat{h_n})-\widehat{R_0}(\widehat{h_n}) \\ &\leqslant \alpha+\frac{\epsilon}{2}+\left|R_0(\widehat{h_n})-\widehat{R_0}(\widehat{h_n})\right| \\ &\leqslant \alpha+\frac{\epsilon}{2}+\sup_{h\in\mathcal{H}}\left|R_0(\widehat{h_n})-\widehat{R_0}(\widehat{h_n})\right| \end{aligned}$$

From ***Claim 1****, we know that:*

$$\begin{aligned} & P\left(\sup_{h\in\mathcal{H}}\left|R_0(h)-\widehat{R_0}(h)\right|\geqslant\frac{\epsilon}{2}\right)\leqslant 2|\mathcal{H}|e^{-n\epsilon^2/2} \\ \implies & P\left(\sup_{h\in\mathcal{H}}\left|R_0(h)-\widehat{R_0}(h)\right|\leqslant\frac{\epsilon}{2}\right)\geqslant 1-2|\mathcal{H}|e^{-n\epsilon^2/2} \end{aligned}$$

Hence, with probability of at least $1-2|\mathcal{H}|e^{-n\epsilon^2/2}$*, we have:*

$$R_0(\widehat{h_n})\leqslant\alpha+\frac{\epsilon}{2}+\frac{\epsilon}{2}=\alpha+\epsilon$$

- **(ii)** $R_1(\widehat{h_n})\leqslant R^*_{\mathcal{H},1}+\epsilon$ ***with high probability****.*

Claim 2 : $R_1(\widehat{h_n})-R^*_{\mathcal{H},1}\leqslant 2\sup_{h\in\mathcal{H}}\left|\widehat{R_1}(h)-R_1(h)\right|$
Let $h'\in\mathcal{H}$ *be any function such that* $\widehat{R_0}(h')\leqslant\alpha+\frac{\epsilon}{2}$*. We have:*

$$\begin{aligned} R_1(\widehat{h_n})-R^*_{\mathcal{H},1} &= R_1(\widehat{h_n})-\widehat{R_1}(\widehat{h_n})+\widehat{R_1}(\widehat{h_n})-\widehat{R_1}(h')+\widehat{R_1}(h')-R^*_{\mathcal{H},1} \\ &\leqslant \left|R_1(\widehat{h_n})-\widehat{R_1}(\widehat{h_n})\right|+\underbrace{\widehat{R_1}(\widehat{h_n})-\widehat{R_1}(h')}_{\leqslant 0}+\left|\widehat{R_1}(h')-R^*_{\mathcal{H},1}\right| \\ &\leqslant \left|R_1(\widehat{h_n})-\widehat{R_1}(\widehat{h_n})\right|+\left|\widehat{R_1}(h')-R^*_{\mathcal{H},1}\right| \\ &\leqslant 2\sup_{h\in\mathcal{H}}\left|\widehat{R_1}(h)-R_1(h)\right| \end{aligned}$$

From ***Claim 2****, we have:*

$$\begin{aligned} P\Big(R_1(\widehat{h_n})-R^*_{\mathcal{H},1}\geqslant\epsilon\Big) &\leqslant P\Big(2\sup_{h\in\mathcal{H}}\left|\widehat{R_1}(h)-R_1(h)\right|\geqslant\epsilon\Big) \\ &= P\Big(\sup_{h\in\mathcal{H}}\left|\widehat{R_1}(h)-R_1(h)\right|\geqslant\frac{\epsilon}{2}\Big) \\ &\leqslant 2|\mathcal{H}|e^{-n\epsilon^2/2} \quad (\textit{From } \textbf{\textit{Claim 1}}) \\ \implies P\Big(R_1(\widehat{h_n})-R^*_{\mathcal{H},1}\leqslant\epsilon\Big) &\geqslant 1-2|\mathcal{H}|e^{-n\epsilon^2/2} \end{aligned}$$

Hence, with probability of at least $1 - 2|\mathcal{H}|e^{-n\epsilon^2/2}$*, we have that* $R_1(\widehat{h_n}) \leqslant R^*_{\mathcal{H},1} + \epsilon$.

□.

5 Vapnik-Chevronenkis Theory

In the following section, we will review a notion for measuring complexity of function class called **VC dimension**. Later on in this note, we will show that VC dimension is an upper-bound for the **Rademacher Complexity**.

5.1 VC Dimension

Definition 5.1 (Restriction ($N_{\mathcal{H}}$))**.** *Let $\mathcal{H} \in \{0,1\}^{\mathcal{X}}$ be a set of classifiers. The **restriction** of $\mathcal{H}$ to a finite subset $C \subset \mathcal{X}$ where $|C| = n$ is the set of binary vectors defined by:*

$$\boxed{N_{\mathcal{H}}(C) = \Big\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}, x_i \in C\Big\}}$$

Clearly, we have $|N_{\mathcal{H}}(C)| \leqslant 2^n$ (cardinality of powerset of C).

Definition 5.2 (Shattering Coefficient ($S_{\mathcal{H}}$))**.** *The n^{th} **Shattering coefficient** (sometimes called the **Growth function**) is defined as:*

$$\boxed{S_{\mathcal{H}}(n) = \sup_{C \subset \mathcal{X}; |C| = n} \Big|N_{\mathcal{H}}(C)\Big|}$$

Hence, we have:

$$|N_{\mathcal{H}}(C)| \leqslant S_{\mathcal{H}}(n) \leqslant 2^n, \ \forall C \subset \mathcal{X}$$

*Intuitively, the n^{th} shattering coefficient is the size of the largest n-element restriction of $\mathcal{H}$. It measures the **richness** of $\mathcal{H}$.*

*If $S_{\mathcal{H}}(n) = 2^n$. Then $\exists C \subset \mathcal{X}, |C| = n$ such that $\Big|N_{\mathcal{H}}(C)\Big| = 2^n$. We then say that $\mathcal{H}$ **shatters the points in** C.*

Definition 5.3 (VC-dimension ($V_{\mathcal{H}}$))**.** *The **VC dimension** of $\mathcal{H}$ is defined as:*

$$\boxed{V_{\mathcal{H}} = \sup\Big\{n : S_{\mathcal{H}}(n) = 2^n\Big\} = \sup_{C \subset \mathcal{X}} \Big\{|C| : \mathcal{H} \text{ shatters } C\Big\}}$$

If $S_{\mathcal{H}}(n) = 2^n, \forall n \geqslant 1$ then $V_{\mathcal{H}} = \infty$.

Remark : Note that when $|\mathcal{H}| < \infty$, we have:

$$\begin{aligned} |N_{\mathcal{H}}(C)| = \Big|\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}, x_i \in C\}\Big| &\leqslant |\mathcal{H}| \\ \implies S_{\mathcal{H}}(n) &\leqslant |\mathcal{H}| \\ \implies V_{\mathcal{H}} &\leqslant \log_2 |\mathcal{H}| \end{aligned}$$

Remark : To show that $V_{\mathcal{H}} = n$, we must show that there exists at least n points $x_1, \dots, x_n$ shattered by $\mathcal{H}$ and no set of $n+1$ points can be shattered by $\mathcal{H}$.

Remark : From the above definitions, we can understand $N_{\mathcal{H}}, S_{\mathcal{H}}$ and $V_{\mathcal{H}}$ as followed:

- $N_{\mathcal{H}}(C)$: Number of ways to assign labels to $C \subset \mathcal{X}$ of size $n \geqslant 1$.
- $S_{\mathcal{H}}(n)$: Maximum number of ways to assign labels to subsets of size $n \geqslant 1$.
- $V_{\mathcal{H}}$: Maximum subset size $n \geqslant 1$ such that we have 2^n ways to assign labels (fully labelled).

5.2 Sauer's Lemma

Theorem 5.1: Sauer's Lemma

This is a bound on the shatter coefficient. Let $d = V_{\mathcal{H}} \leqslant \infty$. For all $n \geqslant 1$, we have:

$$S_{\mathcal{H}}(n) \leqslant \sum_{k=0}^{d} \binom{n}{k}$$

Proof (Theorem 5.1, cited Garivier 2019). *Given a function class $\mathcal{H}$ and a subset $C \subset \mathcal{X}$. For brevity, denote the restriction of $\mathcal{H}$ to C as:*

$$N_{\mathcal{H}}(C) = \mathcal{H}_C$$

To prove the above theorem, we prove a stronger result: For all subset $C \subset \mathcal{X}$ where $|C| = n$, we have

$$|\mathcal{H}_C| \leqslant \left|\left\{B \subset C : \mathcal{H} \text{ shatters } B\right\}\right| \leqslant \sum_{k=0}^{d} \binom{n}{k}$$

The second inequality holds because no set with size larger than d is shattered by $\mathcal{H}$. To prove that the first inequality holds for subsets of any size $n \geqslant 1$, we prove by induction:

- ***Base case*** *: Let $n = 1$. Hence, we have $C = \{x\}$ for $x \in \mathcal{X}$. Denote that:*

  $$\Phi_C = \left\{B \subset C : \mathcal{H} \text{ shatters } B\right\}$$

  We have:

  $$\begin{cases} \mathcal{H} \text{ shatters } C & \implies \mathcal{H}_C = \{0,1\}, \Phi_C = \{\varnothing, C\} \\ \mathcal{H} \text{ not shatter } C & \implies \mathcal{H}_C = \{0\} \text{ or } \{1\}, \Phi_C = \{\varnothing\} \end{cases}$$

  For both cases, we have $|\mathcal{H}_C| = |\Phi_C|$.

- ***Inductive case*** *: Assume that the first inequality holds for $n = m-1, m \geqslant 2$, We have to prove that it holds for $n = m$. Let $C = \{c_1, \ldots, c_m\}$ and $C' = \{c_2, \ldots, c_m\}$. Define the following label sets:*

  $$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

  $$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

  First, we notice that $Y_0 = \mathcal{H}_{C'}$. Hence, we have:

  $$\begin{aligned} |Y_0| &= |\mathcal{H}_{C'}| \\ &\leqslant \left|\left\{B \subset C' : \mathcal{H} \text{ shatters } B\right\}\right| \\ &= \left|\left\{B \subset C : c_1 \notin B, \mathcal{H} \text{ shatters } B\right\}\right| \end{aligned}$$

Next, we define the following sub-class of $\mathcal{H}$:

$$\mathcal{H}' = \left\{ h \in \mathcal{H} : \exists h' \in \mathcal{H} \ s.t \ h'(c) = \begin{cases} 1 - h(c) & \text{if } c = c_1 \\ h(c) & \text{otherwise} \end{cases} \right\}$$

Note that $Y_1 = \mathcal{H}'_{C'}$ and $\mathcal{H}'$ shatters $B \in C'$ implies $\mathcal{H}'$ shatters $B \cup \{c_1\}$ because for any $h' \in \mathcal{H}'$, there is always another function in $\mathcal{H}'$ that gives the opposite label to c_1. Hence, we have:

$$\begin{aligned} |Y_1| &= |\mathcal{H}'_{C'}| \\ &\leqslant \left| \left\{ B \subset C' : \mathcal{H}' \text{ shatters } B \right\} \right| \\ &= \left| \left\{ B \subset C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\} \right\} \right| \\ &= \left| \left\{ B \subset C : c_1 \in B, \mathcal{H}' \text{ shatters } B \right\} \right| \\ &\leqslant \left| \left\{ B \subset C : c_1 \in B, \mathcal{H} \text{ shatters } B \right\} \right| \end{aligned}$$

From the above, we have:

$$\begin{aligned} |\mathcal{H}_C| &= |Y_0| + |Y_1| \\ &\leqslant \left| \left\{ B \subset C : c_1 \notin B, \mathcal{H} \text{ shatters } B \right\} \right| + \left| \left\{ B \subset C : c_1 \in B, \mathcal{H} \text{ shatters } B \right\} \right| \\ &= \left| \left\{ B \subset C : \mathcal{H} \text{ shatters } B \right\} \right| \\ &\leqslant \sum_{k=0}^{d} \binom{n}{k} \end{aligned}$$

Taking the supremum over $C \subset \mathcal{X}$ for both sides, we have:

$$S_{\mathcal{H}}(n) \leqslant \sum_{k=0}^{d} \binom{n}{k}$$

□.

Corollary 5.1: Sauer's lemma - bound on $S_{\mathcal{H}}(n)$ I

If $d = V_{\mathcal{H}} < \infty$, for all $n \geqslant 1$, we have:

$$S_{\mathcal{H}}(n) \leqslant (n+1)^d$$

Proof (Corollary 5.1). *By Binomial theorem, we have:*

$$\begin{aligned} (n+1)^d &= \sum_{k=1}^{d} n^k \binom{d}{k} \\ &= \sum_{k=1}^{d} n^k \frac{d!}{k!(d-k)!} \\ &\geqslant \sum_{k=1}^{d} \frac{n^k}{k!} \geqslant \sum_{k=1}^{d} \frac{n!}{(n-k)!k!} = \sum_{k=1}^{d} \binom{n}{k} \geqslant S_{\mathcal{H}}(n) \end{aligned}$$

□.

Corollary 5.2: Sauer's lemma - bound on $S_{\mathcal{H}}(n)$ II

For all $n \geqslant d = V_{\mathcal{H}}$, we have:

$$S_{\mathcal{H}}(n) \leqslant \left(\frac{ne}{d}\right)^d$$

Proof (Corollary 5.2). *For $\frac{d}{n} < 1$, we have:*

$$\begin{aligned}\left(\frac{d}{n}\right)^d \sum_{k=0}^{d} \binom{n}{k} &\leqslant \sum_{k=0}^{d} \left(\frac{d}{n}\right)^k \binom{n}{k} \\ &\leqslant \sum_{k=0}^{n} \left(\frac{d}{n}\right)^k \binom{n}{k} \\ &= \left(1 + \frac{d}{n}\right)^n \leqslant e^d\end{aligned}$$

Hence, we have:

$$\left(\frac{en}{d}\right)^d \geqslant \sum_{k=0}^{d} \binom{n}{k} \geqslant S_{\mathcal{H}}(n)$$

□.

Corollary 5.3: Sauer's lemma - bound on $S_{\mathcal{H}}(n)$ III

If $V_{\mathcal{H}} = d > 2$, for all $n \geqslant d$, we have:

$$S_{\mathcal{H}}(n) \leqslant n^d$$

Proof (Corollary 5.3). *If $d > 2$ then by corollary 5.2, we have:*

$$\frac{e}{d} < 1 \implies S_{\mathcal{H}}(n) \leqslant \left(\frac{en}{d}\right)^d \leqslant n^d$$

□.

5.3 VC Theorem for classifiers

Theorem 5.2: VC Theorem (for classifiers)

For any $n \geqslant 1$ and $\epsilon > 0$, we have:

$$P\left(\sup_{h \in \mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| \geqslant \epsilon\right) \leqslant 8 S_{\mathcal{H}}(n) e^{-n\epsilon^2/32}$$

Proof (Theorem 5.2). *The proof for this theorem is discussed in further details in Devroye et al. 1996. Specifically, theorem 12.4 (Glivenko-Cantelli Theorem) and theorem 12.5 (VC Theorem). Tighter versions of this inequality is presented in theorem 6.4 and 6.5.* □.

Corollary 5.4: Convergence of Empirical Risk (VC-Theorem)

If $\widehat{h_n}$ is an empirical risk minimizer over $\mathcal{H}$ then:

$$P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) \leqslant 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/128}$$

Proof (Corollary 5.4). *We have:*

$$\begin{aligned} P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) &\leqslant P\left(2\sup_{h\in\mathcal{H}}\left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon\right) \\ &= P\left(\sup_{h\in\mathcal{H}}\left|\widehat{R_n}(h) - R(h)\right| \geqslant \frac{\epsilon}{2}\right) \\ &\leqslant 8S_{\mathcal{H}}(n)e^{-n(\epsilon/2)^2/32} = 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/128} \end{aligned}$$

□.

Corollary 5.5: Excess Risk of $\widehat{h_n}$ - $\delta \to \epsilon$ relation (VC-Theorem)

If $V_{\mathcal{H}} < \infty$ then $\mathcal{H}$ is uniformly learnable by ERM. Specifically, we can define the sample complexity as:

$$N(\epsilon, \delta) = \frac{128}{\epsilon^2}\left(V_{\mathcal{H}}\log(n+1) - \log\frac{\delta}{8}\right)$$

In other words, with probability of at least $1-\delta$, we have:

$$R(\widehat{h_n}) \leqslant R^*_{\mathcal{H}} + \sqrt{\frac{128}{n}\left(V_{\mathcal{H}}\log(n+1) - \log\frac{\delta}{8}\right)}$$

Proof (Corollary 5.5). *Let $\delta = 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/128}$. By corollary 5.4, with probability of at least $1-\delta$, we have:*

$$\begin{aligned} R(\widehat{h_n}) &\leqslant R^*_{\mathcal{H}} + \epsilon \\ &= R^*_{\mathcal{H}} + \sqrt{\frac{128}{n}\left(\log S_{\mathcal{H}}(n) - \log\frac{\delta}{8}\right)} \end{aligned}$$

By Sauer's lemma, we have that $(n+1)^{V_{\mathcal{H}}} \geqslant S_{\mathcal{H}}(n)$ for all $n \geqslant 1$. Hence, we have:

$$\begin{aligned} R(\widehat{h_n}) &\leqslant R^*_{\mathcal{H}} + \sqrt{\frac{128}{n}\left(\log S_{\mathcal{H}}(n) - \log\frac{\delta}{8}\right)} \\ &\leqslant R^*_{\mathcal{H}} + \sqrt{\frac{128}{n}\left(V_{\mathcal{H}}\log(n+1) - \log\frac{\delta}{8}\right)} \end{aligned}$$

Hence, we conclude that $\mathcal{H}$ is PAC-learnable by ERM when $V_{\mathcal{H}} < \infty$ with the following sample complexity:

$$N(\epsilon, \delta) = \frac{128}{\epsilon^2}\left(V_{\mathcal{H}}\log(n+1) - \log\frac{\delta}{8}\right)$$

□.

5.4 VC Classes

Definition 5.4 (VC Class)**.**
*A **VC Class** is a set of classifiers $\mathcal{H}$ such that $V_{\mathcal{H}} < \infty$. In the following section, we will look at some class of classifiers where the VC dimension can be established or bounded.*

Example 1 (Hypercubes) : Consider the set of classifiers

$$\mathcal{H} = \left\{ \mathbf{1}_{\{x \in R\}} : R = \prod_{i=1}^{d} [a_i, b_i],\ a_i < b_i \right\}$$

In this case, for any $d \geqslant 1$, no more than $2d + 1$ points can be shattered by $\mathcal{H}$. Hence, $V_{\mathcal{H}} = 2d$.

Figure 1: Example when $d = 2$. Four points can be shattered by $\mathcal{H}$ but no five points can be shattered by $\mathcal{H}$.

Example 2 (Convex sets in $\mathbb{R}^2$) : Let $\mathcal{X} = \mathbb{R}^2$. Consider the set of classifiers

$$\mathcal{H} = \left\{ \mathbf{1}_{\{x \in C\}} : C \text{ is convex in } \mathbb{R}^2 \right\}$$

In this case, $V_{\mathcal{H}} = \infty$ because for any n points on a circle and for any $1 \leqslant k \leqslant n$, we can draw a polygon that includes k points but not the remaining $n - k$ points for any selection of k in n points (Figure 2).

Example 3 (Finite $|\mathcal{H}|$) : For any function class where $|\mathcal{H}| < \infty$, we have:

$$\begin{aligned} N_{\mathcal{H}}(x_1, \ldots, x_n) &= \left| \left\{ (h(x_1), \ldots, h(x_n)) : h \in \mathcal{H} \right\} \right| \leqslant |\mathcal{H}| \\ &\Longrightarrow S_{\mathcal{H}}(n) \leqslant |\mathcal{H}| \\ &\Longrightarrow V_{\mathcal{H}} \leqslant \log_2 |\mathcal{H}| \end{aligned}$$

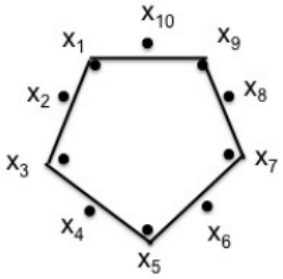

Figure 2: $\mathcal{H}$ can shatter any n points on a circle.

Proposition 5.1: Steele & Dudley bound on $V_{\mathcal{H}}$

Let $\mathcal{F}$ be the set of real-valued functions of the form:

$$\mathcal{F} = \Big\{ f : \mathbb{R}^d \to \mathbb{R} \Big\},\ dim(\mathcal{F}) = m$$

Then, the following set of classifiers:

$$\mathcal{H} = \Big\{ \mathbf{1}_{\{f(x) \geqslant 0\}} : f \in \mathcal{F} \Big\}$$

Has finite VC dimension. Specifically, $V_{\mathcal{H}} \leqslant m$.

Proof (Proposition 5.1). *Suppose that $\mathcal{H}$ shatters $m+1$ points in $C = (x_1, \ldots, x_{m+1})$. Define the linear mapping $L_C : \mathcal{F} \to \mathbb{R}^{m+1}$ such that:*

$$L_C(f) = (f(x_1), \ldots, f(x_{m+1}))^T$$

***Claim :* $L_C(\mathcal{F})$ *is a closed subspace in* $\mathbb{R}^{m+1}$**
Let $\{l_n\}_{n=1}^{\infty} \subset L_C(\mathcal{F})$ be a sequence and let $l_n \to l$ as $n \to \infty$. We can always choose a function $f \in \mathcal{F}$ such that:

$$f(x) = l,\ \forall x \in \mathbb{R}^m$$

Hence, for all $\{l_n\}_{n=1}^{\infty}$ such that $l_n \to l$, the limit $l \in L_C(\mathcal{F})$. Therefore, $L_C(\mathcal{F})$ is closed in $\mathbb{R}^{m+1}$. By the Hilbert Projection Theorem Wikipedia 2024a, we have:

$$\mathbb{R}^{m+1} = L_C(\mathcal{F}) \oplus L_C(\mathcal{F})^{\perp}$$

Since $dim(\mathcal{F}) = m$, we have $dim(L_C(\mathcal{F})) \leqslant m$. Therefore, $dim(L_C(\mathcal{F})^{\perp}) \geqslant 1$ and we have:

$$\forall f \in \mathcal{F}, \exists \gamma \in \mathbb{R}^{m+1} \backslash \{0\} : \gamma^T L(f) = 0$$

Equivalently,

$$\sum_{i=1}^{m+1} \gamma_i f(x_0) = 0 \implies \sum_{i, \gamma_i \geqslant 0} \gamma_i f(x_i) = \sum_{j, \gamma_j < 0} -\gamma_j f(x_j)$$

Since $\mathcal{H}$ shatters $x_1, \ldots, x_{m+1}$. We define a classifier $h \in \mathcal{H}$ such that:

$$h(x_i) = \begin{cases} 1 & \text{if } \gamma_i \geqslant 0 \\ & \implies f(x_i) \geqslant 0 \iff \gamma_i \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

This implies that $\sum_{i:\gamma_i \geqslant 0} \gamma_i f(x_i) \geqslant 0$ *and* $\sum_{j:\gamma_j<0} -\gamma_j f(x_j) < 0$*, which is a contradiction. Therefore, we have* $V_{\mathcal{H}} < \infty$*.* ▫.

Corollary 5.6: Linear classifiers have finite $V_{\mathcal{H}}$

Let $\mathcal{X} = \mathbb{R}^d$ and define a function space $\mathcal{F}$ as:

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f(x) = w^T x + b,\ w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

Then, define the set of linear classifiers as:

$$\mathcal{H} = \left\{ \mathbf{1}_{\{w^T x + b \geqslant 0\}} \middle| w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

We have $V_{\mathcal{H}} \leqslant d + 1$.

Proof (Corollary 5.6). *Since* $dim(\mathcal{F}) = d + 1$*, the above corollary is a direct consequence of proposition 5.1.* ▫.

5.5 VC Theorem for sets

Definition 5.5 (VC Theory for sets). *Let* $\mathcal{G}$ *be a collection of subsets in* $\mathcal{X}$*. Let* $C \subset \mathcal{X}$ *and* $C = \{x_1, \ldots, x_n\}$*. We have the following definitions for restriction of* $\mathcal{G}$ *to* C*, shattering coefficient and VC-dimension of* $\mathcal{G}$*:*

$$\begin{aligned} N_{\mathcal{G}}(C) &= \left| \left\{ G \cap C : G \in \mathcal{G} \right\} \right| \\ S_{\mathcal{G}}(n) &= \sup_{C \subset \mathcal{X}; |C| = n} \left| N_{\mathcal{G}}(C) \right| \\ V_{\mathcal{G}} &= \sup \left\{ n : S_{\mathcal{G}}(n) = 2^n \right\} = \sup_{C \subset \mathcal{X}} \left\{ |C| : \mathcal{G} \text{ shatters } C \right\} \end{aligned}$$

Remark : In analogy to the definitions for classifier, sets and binary classifiers are equivalent via:

$$\begin{aligned} G &\to h_G(x) = \mathbf{1}_{\{x \in G\}} \\ h &\to G_h = \left\{ x : h(x) = 1 \right\} \end{aligned}$$

Theorem 5.3: VC Theorem (for sets)

If $X_1, \ldots, X_n \sim Q$ are identically independently distributed samples. Then for any collection $\mathcal{G}$ of subsets in $\mathcal{X}$, $\epsilon > 0$, we have:

$$P\left(\sup_{G \in \mathcal{G}} \left| \widehat{Q}(G) - Q(G) \right| \geqslant \epsilon \right) \leqslant 8 S_{\mathcal{G}}(n) e^{-n\epsilon^2/32}$$

Where $\widehat{Q}(G)$ is defined (similar to the empirical CDF) as:

$$\widehat{Q}(G) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_1 \in G\}}$$

Proof (Theorem 5.3).
Define the following class of classifiers:

$$\mathcal{H} = \left\{ h_G = \mathbf{1}_{\{x \in G\}} : G \in \mathcal{G} \right\}$$

Define a density function over $\mathcal{X} \times \{0,1\}$ *such that* $\pi_0 = 1$, $P_{X|Y=0} = Q$ *and* $P_{X|Y=1}$ *is arbitrary. For any* $h_G \in \mathcal{H}$, *we have:*

$$\begin{aligned} R(h_G) &= P(h_G(X) \neq Y) \\ &= \pi_0 P_{X|Y=0}(h_G(X) \neq 0) + \pi_1 P_{X|Y=1}(h_G(X) \neq 1) \\ &= \pi_0 P_{X|Y=0}(h_G(X) = 1) + \pi_1 P_{X|Y=1}(h_G(X) = 0) \\ &= P_{X|Y=0}(h_G(X) = 1) \quad (\textit{Since } \pi_0 = 1, \pi_1 = 0) \\ &= Q(G) \end{aligned}$$

Similarly, we have:

$$\widehat{R_n}(h_G) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{h_G(X_i)=1\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in G\}} = \widehat{Q}(G)$$

Therefore:

$$\begin{aligned} P\left(\sup_{G \in \mathcal{G}} \left| \widehat{Q}(G) - Q(G) \right| \geqslant \epsilon \right) &= P\left(\sup_{h_G \in \mathcal{H}} \left| \widehat{R_n}(h_G) - R(h_G) \right| \geqslant \epsilon \right) \\ &\leqslant 8 S_{\mathcal{H}}(n) e^{-n\epsilon^2/32} \quad (\textit{Theorem 5.2}) \\ &= 8 S_{\mathcal{G}}(n) e^{-n\epsilon^2/32} \end{aligned}$$

□.

Corollary 5.7: Dvoretzky-Kiefer-Wolfowitz Inequality

Let $X \sim Q$ be a random variable on the real line $\mathbb{R}$ and denote $G_t = (-\infty, t]$. Then,

$$\begin{aligned} Q(G_t) &= P(X \leqslant t) = F(t) \quad \text{(CDF)} \\ \widehat{Q}(G_t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \leqslant t\}} = \widehat{F_n}(t) \quad \text{(Empirical CDF)} \end{aligned}$$

For all $n \geqslant 1, \epsilon > 0$, we have:

$$P\left(\sup_{t \in \mathbb{R}} \left| \widehat{F_n}(t) - F(t) \right| \geqslant \epsilon \right) \leqslant 8(n+1) e^{-n\epsilon^2/32}$$

Proof (Corollary 5.7).
Let $\mathcal{G} = \left\{ G_t : t \in \mathbb{R} \right\}$, *by theorem 5.3, we have:*

$$\begin{aligned} P\left(\sup_{G \in \mathcal{G}} \left| \widehat{Q}(G) - Q(G) \right| \geqslant \epsilon \right) &= P\left(\sup_{t \in \mathbb{R}} \left| \widehat{Q}(G_t) - Q(G_t) \right| \geqslant \epsilon \right) \\ &= P\left(\sup_{t \in \mathbb{R}} \left| \widehat{F_n}(t) - F(t) \right| \geqslant \epsilon \right) \\ &\leqslant 8 S_{\mathcal{G}}(n) e^{-n\epsilon^2/32} \end{aligned}$$

For any set of n points on the real line, there are $n+1$ ways to label them using half-open intervals. Hence $S_{\mathcal{G}}(n) = n + 1$. Therefore:

$$P\left(\sup_{t \in \mathbb{R}} \left| \widehat{F_n}(t) - F(t) \right| \geqslant \epsilon \right) \leqslant 8(n+1)e^{-n\epsilon^2/32}$$

□.

5.6 End of chapter exercises

Exercise 5.1

Determine the sample complexity $N(\epsilon, \delta)$ for ERM over a class $\mathcal{H}$ with VC dimension $V_{\mathcal{H}} < \infty$.

Solution (Exercise 5.1).

We have:

$$\begin{aligned} P\Big(R(\widehat{h_n}) - R^*_{\mathcal{H}} \geqslant \epsilon\Big) &\leqslant P\left(2 \sup_{h \in \mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| \geqslant \epsilon\right) \\ &= P\left(\sup_{h \in \mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| \geqslant \frac{\epsilon}{2}\right) \\ &\leqslant 8 S_{\mathcal{H}}(n) e^{-n\epsilon^2/128} \quad (\textit{Theorem 5.2}) \\ &\leqslant 8(n+1)^{V_{\mathcal{H}}} e^{-n\epsilon^2/128} \quad (\textit{Corollary 5.1}) \end{aligned}$$

Now let:

$$\begin{aligned} \delta &= 8(n+1)^{V_{\mathcal{H}}} e^{-n\epsilon^2/128} \\ \implies \log\frac{\delta}{8} &= V_{\mathcal{H}} \log(n+1) - \frac{n\epsilon^2}{128} \\ \implies N(\epsilon, \delta) &= \frac{128}{\epsilon^2}\left(V_{\mathcal{H}} \log(n+1) - \log\frac{\delta}{8}\right) \end{aligned}$$

□.

Exercise 5.2

Show that the VC Theorem for sets implies the VC Theorem for classifiers.

Hint : Consider the sets of the form $G' = G \times \{0\} \cup G^c \times \{1\} \subset \mathcal{X} \times \mathcal{Y}$.

Solution (Exercise 5.2).

Given an arbitrary class of classifiers $\mathcal{H}$*. Define the following class of sets:*

$$\mathcal{G} = \Big\{G_h \times \{0\} \cup G_h^c \times \{1\} : h \in \mathcal{H}\Big\}$$

Where for a given $h \in \mathcal{H}$*, we have:*

$$G_h = \Big\{x \in \mathcal{X} : h(x) = 1\Big\}$$

Let P_{XY} *be the density function over* $\mathcal{X} \times \mathcal{Y}$*. For any* $G \in \mathcal{G}$*, we have:*

$$\begin{aligned} P_{XY}(G) &= \pi_0 P_{X|Y=0}(G) + \pi_1 P_{X|Y=1}(G) \\ &= \pi_0 P_{X|Y=0}\Big(G_h \times \{0\} \cup G_h^c \times \{1\}\Big) + \pi_1 P_{X|Y=1}\Big(G_h \times \{0\} \cup G_h^c \times \{1\}\Big) \\ &= \pi_0 P_{X|Y=0}(G_h) + \pi_1 P_{X|Y=1}(G_h^c) \\ &= \pi_0 P_{X|Y=0}(h(X) = 1) + \pi_1 P(X|Y=1)(h(X) = 0) \\ &= P(h(X) \neq Y) \\ &= R(h) \end{aligned}$$

Let $Q = P_{XY}$. We also have:

$$\widehat{Q}(G) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{(X_i,Y_i)\in G_h\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(X_1)\neq Y_i\}} = \widehat{R_n}(h)$$

From the above, we have:

$$\begin{aligned} P\left(\sup_{h\in\mathcal{H}}\left|\widehat{R_n}(h) - R(h)\right| \geqslant \epsilon\right) &= P\left(\sup_{G\in\mathcal{G}}\left|\widehat{Q}(G) - Q(G)\right| \geqslant \epsilon\right) \\ &\leqslant 8S_{\mathcal{G}}(n)e^{-n\epsilon^2/32} \\ &= 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/32} \end{aligned}$$

□.

Exercise 5.3

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ denote two classes of sets:

- (**a**) $\mathcal{G}_1 \cap \mathcal{G}_2 = \Big\{G_1 \cap G_2 : G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\Big\}$.
- (**b**) $\mathcal{G}_1 \cup \mathcal{G}_2 = \Big\{G_1 \cup G_2 : G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\Big\}$.

Show that $S_{\mathcal{G}_1\cap\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$ and $S_{\mathcal{G}_1\cup\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$.

Solution (Exercise 5.3)**.**
Proving each inequality one by one, we have:

Claim : $S_{\mathcal{G}_1\cap\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$
For any $\{x_1,\ldots,x_n\} \subset \mathcal{X}$, denote the following set:

$$\mathcal{F} = \Big\{G_1 \cap \{x_1,\ldots,x_n\} : G_1 \in \mathcal{G}_1\Big\}$$

Then, $\mathcal{F}$ is a collection of subsets of $\{x_1,\ldots,x_n\}$. Furthermore, the cardinality of $\mathcal{F}$ is at most $S_{\mathcal{G}_1}(n)$. Now define the restriction of $\mathcal{G}_1 \cap \mathcal{G}_2$ to $\{x_1,\ldots,x_n\}$:

$$\begin{aligned} \mathcal{G}_1 \cap \mathcal{G}_{2\{x_1,\ldots,x_n\}} &= \Big\{G_1 \cap G_2 \cap \{x_1,\ldots,x_n\} : G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\Big\} \\ &= \bigcup_{F\in\mathcal{F}} \Big\{G_2 \cap F : G_2 \in \mathcal{G}_2\Big\} \end{aligned}$$

For each $F \in \mathcal{F}$, we have $|F| \leqslant n$. Hence, we have:

$$\Big|\Big\{G_2 \cap F : G_2 \in \mathcal{G}_2\Big\}\Big| \leqslant S_{\mathcal{G}_2}(n),\ \forall F \in \mathcal{F}$$

Hence,

$$\begin{aligned} \Big|\mathcal{G}_1 \cap \mathcal{G}_{2\{x_1,\ldots,x_n\}}\Big| &= \left|\bigcup_{F\in\mathcal{F}} \Big\{G_2 \cap F : G_2 \in \mathcal{G}_2\Big\}\right| \\ &\leqslant \sum_{F\in\mathcal{F}} \Big|\Big\{G_2 \cap F : G_2 \in \mathcal{G}_2\Big\}\Big| \\ &\leqslant \sum_{F\in\mathcal{F}} S_{\mathcal{G}_2}(n) = |\mathcal{F}|S_{\mathcal{G}_2}(n) \quad (\textit{Since } |F| \leqslant n, \forall F \in F) \\ &\leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n) \end{aligned}$$

Since the above is a uniform bound, we can take the supremum over all $\{x_1, \dots, x_n\} \subset \mathcal{X}$ and the inequality still holds. Hence,

$$\sup_{\{x_1,\dots,x_n\}\subset\mathcal{X}} \left|\mathcal{G}_1 \cap \mathcal{G}_2\{x_1, \dots, x_n\}\right| = S_{\mathcal{G}_1\cap\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$$

Claim : $S_{\mathcal{G}_1\cup\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$
For any $\{x_1, \dots, x_n\} \subset \mathcal{X}$, we define the following collections of subsets:

$$\mathcal{F}_1 = \left\{G_1 \cap \{x_1, \dots, x_n\} : G_1 \in \mathcal{G}_1\right\}$$
$$\mathcal{F}_2 = \left\{G_2 \cap \{x_2, \dots, x_n\} : G_2 \in \mathcal{G}_2\right\}$$

Then we have:

$$\mathcal{G}_1 \cup \mathcal{G}_2\{x_1,\dots,x_n\} = \bigcup_{F_1\in\mathcal{F}_1}\bigcup_{F_2\in\mathcal{F}_2}\{F_1 \cup F_2\}$$

Since $|\mathcal{F}_1| \leqslant S_{\mathcal{G}_1}(n)$ and $|\mathcal{F}_2| \leqslant S_{\mathcal{G}_2}(n)$, we have:

$$\begin{aligned}\left|\mathcal{G}_1 \cup \mathcal{G}_2\{x_1,\dots,x_n\}\right| &= \left|\bigcup_{F_1\in\mathcal{F}_1}\bigcup_{F_2\in\mathcal{F}_2}\{F_1 \cup F_2\}\right| \\ &\leqslant |\mathcal{F}_1||\mathcal{F}_2| \\ &\leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)\end{aligned}$$

Taking the supremum over $\{x_1, \dots, x_n\} \subset \mathcal{X}$ for both sides, we have:

$$\sup_{\{x_1,\dots,x_n\}\subset\mathcal{X}} \left|\mathcal{G}_1 \cup \mathcal{G}_2\{x_1,\dots,x_n\}\right| = S_{\mathcal{G}_1\cup\mathcal{G}_2}(n) \leqslant S_{\mathcal{G}_1}(n)S_{\mathcal{G}_2}(n)$$

□.

Remark : We can extend the proof for exercise 5.3 to function classes. Hence, we have the following Corollary for the VC-dimension of ensembled classifiers:

Corollary 5.8: VC-dimension bound for ensembled classifiers

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes of classifiers and define:

- **(a)** $\mathcal{H}_1 \cap \mathcal{H}_2 = \left\{\mathbf{1}_{\{h_1(x)=1 \wedge h_2(x)=1\}} : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\right\}$.
- **(b)** $\mathcal{H}_1 \cup \mathcal{H}_2 = \left\{\mathbf{1}_{\{h_1(x)=1 \vee h_2(x)=1\}} : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\right\}$.

We have that $S_{\mathcal{H}_1\cap\mathcal{H}_2}(n) \leqslant S_{\mathcal{H}_1}(n)S_{\mathcal{H}_2}(n)$ and $S_{\mathcal{H}_1\cup\mathcal{H}_2}(n) \leqslant S_{\mathcal{H}_1}(n)S_{\mathcal{H}_2}(n)$. Furthermore, we have:

$$V_{\mathcal{H}_1}, V_{\mathcal{H}_2} < \infty \implies V_{\mathcal{H}_1\cap\mathcal{H}_2} < \infty \text{ and } V_{\mathcal{H}_1\cup\mathcal{H}_2} < \infty$$

Proof (Corollary 5.8). *The translation of the results in exercise 5.3 to function classes is trivial. The main point is to prove the second point of the corollary.*

Assume that the statement does not hold. Hence, $V_{\mathcal{H}_1\cap\mathcal{H}_2} = \infty$ and $S_{\mathcal{H}_1\cap\mathcal{H}_2}(n) = 2^n, \forall n \geqslant 1$. By Sauer's lemma 5.1, since $V_{\mathcal{H}_1}, V_{\mathcal{H}_2} < \infty$, we have:

$$S_{\mathcal{H}_1}(n) \leqslant (n+1)^{V_{\mathcal{H}_1}} \implies \log S_{\mathcal{H}_1}(n) \leqslant V_{\mathcal{H}_1}\log(n+1)$$
$$S_{\mathcal{H}_2}(n) \leqslant (n+1)^{V_{\mathcal{H}_2}} \implies \log S_{\mathcal{H}_2}(n) \leqslant V_{\mathcal{H}_2}\log(n+1)$$

Therefore, we have:

$$\begin{aligned} &\log S_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \leqslant \log S_{\mathcal{H}_1}(n) + \log S_{\mathcal{H}_2}(n) \\ \implies &n \log 2 \leqslant (V_{\mathcal{H}_1} + V_{\mathcal{H}_2}) \log(n+1), \ \forall n \geqslant 1 \end{aligned}$$

However, this is not true since the left-hand-side evolves linearly while the right-hand-side evolves logarithmically. Therefore, we have a contraction $\implies V_{\mathcal{H}_1 \cap \mathcal{H}_2} < \infty$. We can repeat the same argument for $V_{\mathcal{H}_1 \cup \mathcal{H}_2}$. ▫.

Exercise 5.4

Show that the following classes have finite VC dimension by exhibiting an explicit upper-bound on the VC dimension.

- (**a**) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{f(x) \geqslant 0\}} : f \text{ inhomogeneous quadratic polynomial}\}$.
- (**b**) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{x \in C\}} : C \text{ is a closed sphere }\}$.
- (**c**) $\mathcal{X} = \mathbb{R}^2, \mathcal{H} = \{\mathbf{1}_{\{x \in P_k\}} : P_k \text{ is a convex polygon of at most } k \text{ sides}\}$.
- (**d**) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{x \in R_k\}} : R_k \text{ is a union of at most } k \text{ rectangles}\}$.

Solution (Exercise 5.4)**.**
In the following exercise, we will mostly make use of proposition 5.1 to bound the VC dimension of the above function classes:

(**a**) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{f(x) \geqslant 0\}} : f$ ***inhomogeneous quadratic polynomial***$\}$
Define the following function class of d-dimensional inhomogeneous quadratic polynomials:

$$\mathcal{F}_1 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f(\mathbf{x}) = \mathbf{x^T A x} + \mathbf{b^T x} + c, \mathbf{A} \in \mathbb{R}^{d \times d}, \text{ symmetric. } \mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R} \right\}$$

For any $f \in \mathcal{F}_1$, we can write:

$$f(\mathbf{x}) = \underbrace{\sum_{k=1}^{d} a_k x_k^2 + 2 \sum_{i=1}^{d} \sum_{j=i+1}^{d} a_{ij} x_i x_j}_{\mathbf{x^T A x}} + \underbrace{\sum_{l=1}^{d} b_l x_l}_{\mathbf{b^T x}} + c$$

Hence, we have the following basis functions for $\mathcal{F}_1$:

$$\mathcal{B} = \left\{ 1, \underbrace{x_1^2, \ldots, x_d^2}_{d}, \underbrace{x_1 x_2, \ldots, x_{d-1} x_d}_{\frac{d(d-1)}{2}}, \underbrace{x_1, \ldots, x_d}_{d} \right\}$$

Therefore, we have:

$$\dim(\mathcal{F}_1) = |\mathcal{B}| = \frac{d(d-1)}{2} + 2d + 1$$

Therefore, by proposition 5.1, we have:

$$V_{\mathcal{H}} \leqslant \dim(\mathcal{F}_1) = \frac{d(d-1)}{2} + 2d + 1$$

(**b**) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{x \in C\}} : C$ ***is a closed sphere*** $\}$
Denote the following function class:

$$\mathcal{F}_2 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f(\mathbf{x}) = r^2 - \sum_{i=1}^{d} (x_i - c_i)^2, r \in \mathbb{R}, c_i \in \mathbb{R} \right\}$$

Where r denotes the radius of the hypersphere and the vector $(c_1, \ldots, c_d)^T \in \mathbb{R}^d$ is the coordinates of the center. We can rewrite the class of classifiers as followed:

$$\begin{aligned}\mathcal{H} &= \{\mathbf{1}_{\{x \in C\}} : C \text{ is a closed sphere }\} \\ &= \left\{ \mathbf{1}_{\{f(\mathbf{x}) \geqslant 0\}} \middle| f \in \mathcal{F}_2 \right\}\end{aligned}$$

We notice that:

$$\mathcal{F}_2 \subset \mathcal{F}_1 \implies \dim(\mathcal{F}_2) \leqslant \dim(\mathcal{F}_1)$$

Hence, by proposition 5.1, we have:

$$V_{\mathcal{H}} \leqslant \dim(\mathcal{F}_2) \leqslant \dim(\mathcal{F}_1) \leqslant \frac{d(d-1)}{2} + 2d + 1$$

(c) $\mathcal{X} = \mathbb{R}^2, \mathcal{H} = \{\mathbf{1}_{\{x \in P_k\}} : P_k$ ***is a convex polygon of at most*** k ***sides***$\}$
We can think of $\mathbf{1}_{\{x \in P_k\}}$ as an ensemble of k different linear classifiers in $\mathbb{R}^2$. This is illustrated in figure 3.

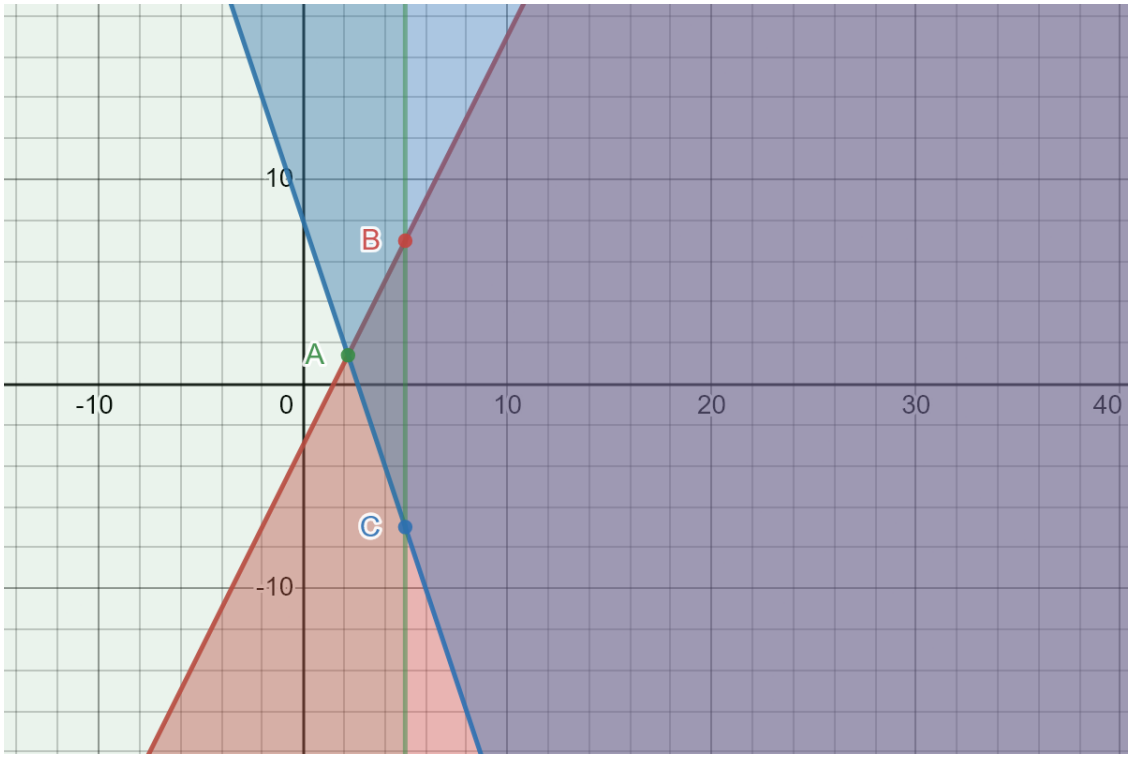

Figure 3: Convex polygon comprises of three linear classifiers of the form $\mathbf{1}_{\{f_i(x,y) \geqslant 0\}}$ where $f_1(x, y) = 2x - y - 3$, $f_2(x, y) = 3x + y - 8$ and $f_3(x, y) = 5 - x$.

We define the class of linear classifiers in $\mathbb{R}^2$ as followed:

$$\mathcal{L} = \left\{ \mathbf{1}_{\{f(x,y) \geqslant 0\}} : f(x, y) = ax + by + c \text{ where } a, b, c \in \mathbb{R} \right\}$$

Hence, we can rewrite the classifiers class of convex polygons as:

$$\begin{aligned}\mathcal{H} &= \{\mathbf{1}_{\{x \in P_k\}} : P_k \text{ is a convex polygon of at most } k \text{ sides}\} \\ &= \left\{ \bigwedge_{i=1}^{n} \mathbf{1}_{\{f_i(x,y) \geqslant 0\}} : f_i \in \mathcal{L}, n \leqslant k \right\}\end{aligned}$$

Since $V_{\mathcal{L}} < \infty$, by corollary 5.8, we also have $V_{\mathcal{H}} < \infty$.

(d) $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{\mathbf{1}_{\{x \in R_k\}} : R_k$ ***is a union of at most*** k ***rectangles***$\}$
We can re-define the classifiers class as followed:

$$\begin{aligned}\mathcal{H} &= \{\mathbf{1}_{\{x \in R_k\}} : R_k \text{ is a union of at most } k \text{ rectangles}\} \\ &= \left\{ \bigvee_{i=1}^{n} \mathbf{1}_{\{x \in \mathcal{R}_i\}} : \mathcal{R}_i \text{ is a rectangle }, n \leqslant k \right\}\end{aligned}$$

Since the class of hyper-rectangles have finite VC-dimension of $2d$, by corollary 5.8, we have that $\mathcal{H}$ also has finite VC-dimension. □.

6 Rademacher Complexity

6.1 Bounded Difference Inequality

In the following section, we will discuss another concentration inequality that bounds the difference between functions of random sample and their mean given that the functions satisfy the **bounded difference property**.

Definition 6.1 (Bounded difference property)**.** *Given a real-valued function $\phi : \mathcal{X}^n \to \mathbb{R}$. We say that ϕ satisfies the **bounded difference property** if $\exists c_1, \dots, c_n \in \mathcal{X}$ such that $\forall 1 \leqslant i \leqslant n$:*

$$\sup_{\{x_1,\dots,x_n\}\subset\mathcal{X}, x_i'\in\mathcal{X}} \left|\phi(x_1,\dots,x_i,\dots,x_n) - \phi(x_1,\dots,x_i',\dots,x_n)\right| \leqslant c_i$$

That is, substituting the value at the i^{th} coordinate x_i changes the value of ϕ by at most c_i.

Theorem 6.1: Bounded Difference (McDiarmid's) Inequality

Let $X_1, \dots, X_n$ be independent random variables (not necessarily identically distributed) and $\phi : \mathcal{X}^n \to \mathbb{R}$ be a function satisfying the bounded difference property:

$$\sup_{\{x_1,\dots,x_n\}\subset\mathcal{X}, x_i'\in\mathcal{X}} \left|\phi(x_1,\dots,x_i,\dots,x_n) - \phi(x_1,\dots,x_i',\dots,x_n)\right| \leqslant c_i,\ \forall 1 \leqslant i \leqslant n$$

Then, we have:

$$P\left(\left|\phi(X_1,\dots,X_n) - \mathbb{E}\Big[\phi(X_1,\dots,X_n)\Big]\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right),\ \forall t > 0$$

Remark : Assume that $X_i \in [a_i, b_i]$ and $\phi(X_1, \dots, X_n) = \sum_{i=1}^n X_i$. Then the bounded difference inequality recovers the Hoeffding's inequality 3.1.

Proof (Theorem 6.1)**.** *Define the following random variable:*

$$V_i = \mathbb{E}\Big[\phi\Big|X_1,\dots,X_i\Big] - \mathbb{E}\Big[\phi\Big|X_1,\dots,X_{i-1}\Big]$$

Denote $\phi(X_1,\dots,X_n) = \phi$ and $\mathbb{E}[\phi(X_1,\dots,X_n)] = \mu_\phi$ for brevity, we have:

$$\phi - \mu_\phi = \sum_{i=1}^n V_i$$

Using the Chernoff's bounding method, we have:

$$\begin{aligned} P(\phi - \mu_\phi \geqslant t) &\leqslant \inf_{s>0} e^{-st} M_{\phi-\mu_\pi}(s) \\ &= \inf_{s>0} e^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^n V_i\right)\right] \end{aligned}$$

Claim 1 : For all** $1 \leqslant i \leqslant n$**,** $a_i \leqslant V_i \leqslant b_i$ **and** $b_i - a_i \leqslant c_i$**.
Define the infimum and supremum of V_i as followed:

$$U_i = \sup_{x\in\mathcal{X}}\left\{\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_i=x\Big]-\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]\right\}$$
$$L_i = \inf_{x\in\mathcal{X}}\left\{\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_i=x\Big]-\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]\right\}$$

Clearly, $U_i \geqslant V_i \geqslant L_i$. We have:

$$\begin{aligned}
U_i - L_i &= \sup_{x\in\mathcal{X}}\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1},X_i=x\Big] - \inf_{x\in\mathcal{X}}\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1},X_i=x\Big]\\
&= \sup_{x\in\mathcal{X}}\int \phi(X_1,\ldots,X_{i-1},x,x_{i+1},\ldots,x_n)dP(x_{i+1},\ldots,x_n|X_1,\ldots,X_{i-1},x)\\
&\quad - \inf_{x\in\mathcal{X}}\int \phi(X_1,\ldots,X_{i-1},x,x_{i+1},\ldots,x_n)dP(x_{i+1},\ldots,x_n|X_1,\ldots,X_{i-1},x)\\
&= \sup_{x\in\mathcal{X}}\int \phi(X_1,\ldots,X_{i-1},x,x_{i+1},\ldots,x_n)dP(x_{i+1},\ldots,x_n)\\
&\quad - \inf_{x\in\mathcal{X}}\int \phi(X_1,\ldots,X_{i-1},x,x_{i+1},\ldots,x_n)dP(x_{i+1},\ldots,x_n)\\
&= \sup_{x,y\in\mathcal{X}}\int\Big[\phi(X_1,\ldots,X_{i-1},x,x_{i+1},\ldots,x_n)-\phi(X_1,\ldots,X_{i-1},y,x_{i+1},\ldots,x_n)\Big]dP(x_{i+1},\ldots,x_n)\\
&\leqslant c_i\int dP(x_{i+1},\ldots,x_n) = c_i
\end{aligned}$$

Claim 2 : $\mathbb{E}\Big[V_i\Big|X_1,\ldots,X_{i-1}\Big]=0,\ \forall 1\leqslant i\leqslant n$.
We have:

$$\begin{aligned}
\mathbb{E}\Big[V_i\Big|X_1,\ldots,X_{i-1}\Big] &= \mathbb{E}\left[\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_i\Big]\Big|X_1,\ldots,X_{i-1}\right] - \mathbb{E}\Big[\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]\Big|X_1,\ldots,X_{i-1}\Big]\\
&= \mathbb{E}\left[\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_i\Big]\Big|X_1,\ldots,X_{i-1}\right] - \mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]
\end{aligned}$$

By the tower property, we have:

$$\mathbb{E}\left[\mathbb{E}\Big[\phi\Big|X_1,\ldots,X_i\Big]\Big|X_1,\ldots,X_{i-1}\right] = \mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]$$

Hence,

$$\begin{aligned}
\mathbb{E}\Big[V_i\Big|X_1,\ldots,X_{i-1}\Big] &= \mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big] - \mathbb{E}\Big[\phi\Big|X_1,\ldots,X_{i-1}\Big]\\
&= 0
\end{aligned}$$

From the above claims, we can make use of the Hoeffding's lemma 3.1 to bound the moment

generating functions. We have:

$$\begin{aligned}
P(\phi - \mu_\phi \geqslant t) &= \inf_{s>0} e^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n} V_i\right)\right] \\
&= \inf_{s>0} e^{-st}\mathbb{E}_{X_1,\ldots,X_{n-1}}\mathbb{E}_{X_n|X_1,\ldots,X_{n-1}}\left[\exp\left(s\sum_{i=1}^{n} V_i\right)\right] \\
&= \inf_{s>0} e^{-st}\mathbb{E}_{X_n|X_1,\ldots,X_{n-1}}\left[e^{sV_n}\right]\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(s\sum_{i=1}^{n-1} V_i\right)\right] \\
&\leqslant \inf_{s>0} \exp\left(-st + \frac{s^2c_n^2}{8}\right)\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(s\sum_{i=1}^{n-1} V_i\right)\right] \quad \textit{(Lemma 3.1)} \\
&\vdots \\
&\leqslant \inf_{s>0} \exp\left(-st + s^2\sum_{i=1}^{n}\frac{c_i^2}{8}\right)
\end{aligned}$$

Substituting $s = \frac{4t}{\sum_{i=1}^{n} c_i^2}$ *to minimize the upperbound (just like the proof for Hoeffding's inequality 3.1). We have:*

$$P(\phi - \mu_\phi \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

□.

6.2 Rademacher Complexity

Overview : Rademacher Complexity is a measure for the richness of a class of real-valued functions. In this sense, it is similar to VC dimension. However, unlike VC dimension, the Rademacher Complexity is not restricted to binary functions.

Definition 6.2 (Empirical Rademacher Complexity)**.** *Let* $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$ *be a set of functions* $\mathcal{Z} \to [a,b]$ *where* $a, b \in \mathbb{R}, a < b$*. Let* $Z_1, \ldots, Z_n$ *be an independently identically distributed random sample on* $\mathcal{Z}$ *following some distribution* P*. Denote* $S = (Z_1, \ldots, Z_n)$*, we define the* ***Empirical Rademacher Complexity*** *as:*

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

Where $\sigma = (\sigma_1, \ldots, \sigma_n)^T$*,* $\sigma_i \sim Uniform(-1, 1)$ *are known as* ***Rademacher random variables****. Note that* $\hat{\mathfrak{R}}_S(\mathcal{G})$ *is random due to randomness in* S*.*

Definition 6.3 (Rademacher Complexity)**.** *The* ***Rademacher Complexity*** *of a function class* $\mathcal{G}$ *is defined as:*

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_S\left[\hat{\mathfrak{R}}_S(\mathcal{G})\right]$$

Theorem 6.2: One-sided Rademacher Complexity bound

Let Z be a random variable and $S = (Z_1, \dots, Z_n)$ be an independently identically distributed sample over $\mathcal{Z}$. Consider a class of functions $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$. $\forall \delta > 0, g \in \mathcal{G}$, with at least probability $1-\delta$ with respect to the draw of sample S, we have:

$$\textbf{(i)} \quad \sup_{g\in\mathcal{G}} \left\{ \mathbb{E}\Big[g(Z)\Big] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right\} \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 1/\delta}{2n}}$$

$$\textbf{(ii)} \quad \sup_{g\in\mathcal{G}} \left\{ \mathbb{E}\Big[g(Z)\Big] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right\} \leqslant 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 2/\delta}{2n}}$$

Proof (Theorem 6.2). *For notation brevity, define:*

$$\widehat{\mathbb{E}_S}[g] = \frac{1}{n}\sum_{i=1}^{n} g(Z_i);\ \mathbb{E}\Big[g(Z)\Big] = \mathbb{E}[g]$$

(i) $\mathbb{E}[g] - \widehat{\mathbb{E}_S}[g] \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 1/\delta}{2n}}$
Define the following function:

$$\phi(S) = \sup_{g\in\mathcal{G}} \Big\{ \mathbb{E}[g] - \widehat{\mathbb{E}_S}[g] \Big\}$$

First, we check that ϕ has the bounded difference property. Define $S_i' = (Z_1, \dots, Z_i', \dots, Z_n)$, we have:

$$\begin{aligned}
\Big|\phi(S) - \phi(S_i')\Big| &= \left| \sup_{g\in\mathcal{G}} \Big\{ \mathbb{E}[g] - \widehat{E_S}[g] \Big\} - \sup_{g\in\mathcal{G}} \Big\{ \mathbb{E}[g] - \widehat{\mathbb{E}_{S_i'}}[g] \Big\} \right| \\
&\leqslant \left| \sup_{g\in\mathcal{G}} \Big\{ \widehat{\mathbb{E}_{S_i'}}[g] - \widehat{E_S}[g] \Big\} \right| \quad \Big(\sup A - \sup B \leqslant \sup \Big\{ A - B \Big\} \Big) \\
&= \left| \frac{1}{n} \sup_{g\in\mathcal{G}} \Big(g(Z_i') - g(Z_i) \Big) \right| \\
&\leqslant \frac{b-a}{n}
\end{aligned}$$

By the Bounded Difference Inequality, we have:

$$\begin{aligned}
P\left(\phi(S) - \mathbb{E}[\phi(S)] \leqslant t \right) &\geqslant 1 - \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} (b-a)^2/n^2} \right), \quad t \geqslant 0 \\
&= 1 - \exp\left(- \frac{2nt^2}{(b-a)^2} \right)
\end{aligned}$$

Now, let:

$$\delta = \exp\left(- \frac{2nt^2}{(b-a)^2} \right) \implies t = (b-a)\sqrt{\frac{\log 1/\delta}{2n}}$$

Hence, for all $\delta > 0$, with probability of at least $1-\delta$, we have:

$$\phi(S) \leqslant \mathbb{E}[\phi(S)] + (b-a)\sqrt{\frac{\log 1/\delta}{2n}} \quad (1)$$

Now, to establish (**i**)*, we have to show that* $\mathbb{E}[\phi(S)] \leqslant 2\mathfrak{R}_n(\mathcal{G})$*. Let* $S' = (Z'_1, \ldots, Z'_n)$*. We have:*

$$\begin{aligned}
\mathbb{E}[\phi(S)] &= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}}\Big\{\mathbb{E}[g] - \widehat{E_S}[g]\Big\}\right] \\
&= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}}\Big\{E_{S'}\Big[\widehat{E_{S'}}[g] - \widehat{E_S}[g]\Big]\Big\}\right] \quad \Big(\mathbb{E}[g] = \mathbb{E}_{S'}\widehat{E_{S'}}[g]\Big) \\
&\leqslant \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\Big\{\widehat{E_{S'}}[g] - \widehat{E_S}[g]\Big\}\right] \quad \Big(\textit{Jensen's Inequality :}\ \sup\mathbb{E} \leqslant \mathbb{E}\sup\Big) \\
&= \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\Big(g(Z'_i) - g(Z_i)\Big)\right] \\
&= \mathbb{E}_{\sigma,S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Big(g(Z'_i) - g(Z_i)\Big)\right] \quad (Z_i, Z'_i \textit{ are i.i.d, } \sigma_i \textit{ are symmetric}) \\
&\leqslant E_{\sigma,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z'_i)\right] + E_{\sigma,S}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)g(Z_i)\right] \quad (\sup(f_1+f_2) \leqslant \sup f_1 + \sup f_2) \\
&= E_{\sigma,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z'_i)\right] + E_{\sigma,S}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right] \quad (\textit{Rademacher variables are symmetric}) \\
&= 2\mathfrak{R}_n(\mathcal{G}) \quad (2)
\end{aligned}$$

From (1) *and* (2)*, we have:*

$$\phi(S) \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 1/\delta}{2n}}$$

(**ii**) $\mathbb{E}[g] - \widehat{\mathbb{E}_S}[g] \leqslant 2\hat{\mathfrak{R}}_n(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 2/\delta}{2n}}$

We will first verify that the Empirical Rademacher Complexity satisfies the bounded difference property. Let $S = (X_1, \ldots, X_n)$ *and* $S'_i = (Y_1, \ldots, Y_n)$ *such that* $Y_j = X_j, \forall j \neq i, 1 \leqslant i \leqslant n$*. We have:*

$$\begin{aligned}
\Big|\hat{\mathfrak{R}}_S(\mathcal{G}) - \hat{\mathfrak{R}}_{S'_i}(\mathcal{G})\Big| &\leqslant \left|\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\left(\sum_{j=1}^{n}\sigma_j g(X_j) - \sum_{j=1}^{n}\sigma_j g(Y_j)\right)\right]\right| \\
&= \left|\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{j=1}^{n}\sigma_j(g(X_j) - g(Y_j))\right]\right| \\
&= \left|\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{j=1}^{n}(g(X_j) - g(Y_j))\right]\right| \quad (\sigma_i \textit{ are symmetric}) \\
&= \left|\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\Big(g(X_i) - g(Y_i)\Big)\right]\right| \\
&\leqslant \frac{b-a}{n}
\end{aligned}$$

Therefore, by the bounded difference inequality, with at least $1 - \delta/2$ *probability, we have:*

$$\begin{aligned}
&\hat{\mathfrak{R}}_S(\mathcal{G}) - \mathbb{E}_S\Big[\hat{\mathfrak{R}}_S(\mathcal{G})\Big] \geqslant (a-b)\sqrt{\frac{\log 2/\delta}{2n}} \\
\implies &\hat{\mathfrak{R}}_S(\mathcal{G}) - \mathfrak{R}_n(\mathcal{G}) \geqslant (a-b)\sqrt{\frac{\log 2/\delta}{2n}} \\
\implies &\mathfrak{R}_n(\mathcal{G}) \leqslant \hat{\mathfrak{R}}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}
\end{aligned}$$

From (**i**) *we also have, with a probability of at least* $1-\delta/2$*, we have:*

$$\phi(S) \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}$$

Now denote the following events:

$$A := \left\{ \mathfrak{R}_n(\mathcal{G}) \leqslant \widehat{\mathfrak{R}}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}} \right\}$$
$$B := \left\{ \phi(S) \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}} \right\}$$

We have:

$$\begin{aligned} P\left(\left\{ \phi(S) \leqslant 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 2/\delta}{2n}} \right\} \right) &\geqslant P(A \cap B) \\ &= 1 - P(\overline{A} \cup \overline{B}) \\ &\geqslant 1 - \Big(P(\overline{A}) + P(\overline{B}) \Big) \\ &= 1 - (\delta/2 + \delta/2) = 1 - \delta \end{aligned}$$

□.

Theorem 6.3: Two-sided Rademacher Complexity bound

Consider a set of classifiers $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$. Then, $\forall \delta > 0$, with probability of at least $1-\delta$ with respect to the draw of sample S, we have:

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}\Big[g(Z) \Big] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right| \leqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}$$
$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}\Big[g(Z) \Big] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right| \leqslant 2\widehat{\mathfrak{R}}_n(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 4/\delta}{2n}}$$

Proof (Theorem 6.3). *The proof of this theorem is included in exercise 6.4* □.

6.3 Bounds for binary classification

Overview : Given the following,

- $\mathcal{X}$ is a feature space and $\mathcal{Y} = \{-1, 1\}$ be a label space.
- $\mathcal{H} \subset \{-1,1\}^{\mathcal{X}}$: A set of binary classifiers.
- $\mathcal{G} = \Big\{ g_h : \mathcal{X} \times \mathcal{Y} \to \{0,1\} \Big| g_h(x,y) = \mathbf{1}_{\{h(x) \neq y\}}, h \in \mathcal{H} \Big\}$: Every function $g_h \in \mathcal{G}$ is the risk function of $h \in \mathcal{H}$.
- $S = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ is a sample dataset.

Lemma 6.1: $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_S(\mathcal{H})$

We have the following equality:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_S(\mathcal{H})$$

Proof (Lemma 6.1). *From definition, we have:*

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{G}) &= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{\{h(X_i)\neq Y_i\}}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i \frac{1-Y_i h(X_i)}{2}\right] \\
&= \mathbb{E}_\sigma\left[\frac{1}{2n}\sum_{i=1}^n \sigma_i + \frac{1}{2}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i(-Y_i)h(X_i)\right] \\
&= \frac{1}{2n}\underbrace{\mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i\right]}_{=0} + \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i(-Y_i)h(X_i)\right] \\
&= \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right] \quad (\sigma_i \text{ and } \sigma_i(-Y_i) \text{ has the same distribution}) \\
&= \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H})
\end{aligned}$$

□.

Corollary 6.1: UDB for binary classification using Rademacher Complexity

For all $\delta > 0$, with probability of at least $1-\delta$, we have:

$$\sup_{h\in\mathcal{H}}\left(R(h) - \widehat{R}_n(h)\right) \leqslant \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}}$$
$$\sup_{h\in\mathcal{H}}\left(R(h) - \widehat{R}_n(h)\right) \leqslant \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2n}}$$

We can also derive a two-sided UDB by replacing δ with $\delta/2$:

$$\sup_{h\in\mathcal{H}}\left|R(h) - \widehat{R}_n(h)\right| \leqslant \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log 2/\delta}{2n}}$$
$$\sup_{h\in\mathcal{H}}\left|R(h) - \widehat{R}_n(h)\right| \leqslant \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 4/\delta}{2n}}$$

Proof (Corollary 6.1). *Define the following function space:*

$$\mathcal{G} = \left\{g_h : \mathcal{X}\times\mathcal{Y}\to\{0,1\} \middle| g_h(x,y) = \mathbf{1}_{\{h(x)\neq y\}}, h\in\mathcal{H}\right\}$$

By theorem 6.2, we have:

$$\begin{aligned}
\sup_{g_h\in\mathcal{G}}\left(\mathbb{E}[g_h] - \widehat{E}_S[g_h]\right) &= \sup_{h\in\mathcal{H}}\left(\mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}] - \widehat{E}_S[\mathbf{1}_{\{h(X)\neq Y\}}]\right) \\
&= \sup_{h\in\mathcal{H}}\left(R(h) - \widehat{R}_n(h)\right) \\
&\leqslant 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2n}} \quad (\textit{Theorem 6.2}) \\
&= \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}} \quad \left(\textit{Since } \hat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H})\right)
\end{aligned}$$

Similarly, from theorem 6.2, we can also derive the following inequality:

$$\sup_{h\in\mathcal{H}}\left(R(h)-\widehat{R_n}(h)\right)\leqslant\hat{\mathfrak{R}}_S(\mathcal{H})+3\sqrt{\frac{\log 2/\delta}{2n}}$$

By replacing δ with $\delta/2$, following the same arguments, we obtain the two-sided bounds. □.

6.4 Tighter VC inequalities

In the following section, we will prove tighter VC inequalities compared to theorem 5.2 with the help of Massart's lemma A.2. After that, we will go ahead and look at even tighter versions in the practice exercises.

Theorem 6.4: One-sided VC Inequality

For $0<\delta<1$, with probability of at least $1-\delta$, we have:

$$\sup_{h\in\mathcal{H}}\left(R(h)-\widehat{R_n}(h)\right)\leqslant\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log(1/\delta))}{n}}$$

Equivalently, for all $\epsilon>0$, we have:

$$P\left(\sup_{h\in\mathcal{H}}\left(R(h)-\widehat{R_n}(h)\right)\geqslant\epsilon\right)\leqslant S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}$$

Proof (Theorem 6.4). *Let $\mathcal{H}\subset\{-1,1\}^{\mathcal{X}}$ and $S=(X_1,\ldots,X_n)$ be a random sample. Denote the restriction of $\mathcal{H}$ to S as:*

$$\mathcal{H}_S=\left\{(h(X_1),\ldots,h(X_n)):h\in\mathcal{H}\right\}$$

Hence, for any $u\in\mathcal{H}_S$, we have $\|u\|_2=\sqrt{n}$. By Massart's lemma A.2, we have:

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &= \mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{u\in\mathcal{H}_S}\frac{1}{n}\sum_{i=1}^{n}\sigma_i u_i\right]\\
&\leqslant \mathbb{E}_S\left[\frac{\sup_{u\in\mathcal{H}}\|u\|_2\sqrt{2\log|\mathcal{H}_S|}}{n}\right] \quad (\textit{Massart's lemma A.2})\\
&= \mathbb{E}_S\left[\frac{\sqrt{n}\sqrt{2\log|\mathcal{H}_S|}}{n}\right]=\mathbb{E}_S\left[\sqrt{\frac{2\log|\mathcal{H}_S|}{n}}\right]\\
&\leqslant \sqrt{\frac{2\log\mathbb{E}[|\mathcal{H}_S|]}{n}} \quad (\textit{Jensen's inequality})\\
&\leqslant \sqrt{\frac{2\log S_{\mathcal{H}}(n)}{n}} \quad (|\mathcal{H}_S|\leqslant S_{\mathcal{H}}(n))
\end{aligned}$$

Combine the above inequality with corollary 6.1, with probability of at least $1-\delta$, we have:

$$\begin{aligned}
\sup_{h\in\mathcal{H}}\left(R(h)-\widehat{R}_n(h)\right) &\leqslant \mathfrak{R}_n(\mathcal{H})+\sqrt{\frac{\log 1/\delta}{2n}} \\
&\leqslant \sqrt{\frac{2\log S_{\mathcal{H}}(n)}{n}}+\sqrt{\frac{\log 1/\delta}{2n}} \\
&\leqslant 2\sqrt{\frac{2\log S_{\mathcal{H}}(n)}{n}+\frac{\log 1/\delta}{2n}} \quad (\sqrt{a}+\sqrt{b}\leqslant 2\sqrt{a+b}) \\
&= \sqrt{\frac{8(\log S_{\mathcal{H}}(n)+(\log 1/\delta)/4)}{n}} \qquad (*) \\
&\leqslant \sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 1/\delta)}{n}}
\end{aligned}$$

Now we set:

$$\epsilon=\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 1/\delta)}{n}} \implies \delta = S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}$$

Hence, for all $\epsilon>0$, we have:

$$P\left(\sup_{h\in\mathcal{H}}\left(R(h)-\widehat{R}_n(h)\right)\geqslant\epsilon\right)\leqslant S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}$$

□.

Theorem 6.5: Two-sided VC Inequality

For $0<\delta<1$, with probability of at least $1-\delta$, we have:

$$\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_n(h)\right|\leqslant\sqrt{\frac{8(\ln S_{\mathcal{H}}(n)+\ln(2/\delta))}{n}}$$

Equivalently, for any $\epsilon>0$,

$$P\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_n(h)\right|\geqslant\epsilon\right)\leqslant 2S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}$$

Proof (Theorem 6.5)**.**
The proof of this theorem is included in exercise 6.5. A tighter bound is presented in exercise 6.6.
□.

6.5 End of chapter exercises

Exercise 6.1

Can you improve the constants in the empirical Rademacher complexity bound:

$$\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\Big[g(Z)\Big]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right\}\leqslant 2\widehat{\mathfrak{R}}_S(\mathcal{G})+3(b-a)\sqrt{\frac{\log 2/\delta}{2n}}$$

through a single, direct application of the bounded difference inequality?

Solution (Exercise 6.1 - Wrong, to be fixed later). *Given the sample $S=\{Z_1,\ldots,Z_i,\ldots,Z_n\}$ and define $S_i=\{Z_1,\ldots,Z_i',\ldots,Z_n\}$ for $1\leqslant i\leqslant n$. Define the following function:*

$$\begin{aligned}\phi(S)&=\widehat{\mathbb{E}_S}[g]-\widehat{\mathfrak{R}}_S(\mathcal{G})\\&=\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right]\end{aligned}$$

Note that we have:

$$\mathbb{E}[\phi(S)]=\mathbb{E}\Big[\widehat{\mathbb{E}_S}[g]-\widehat{\mathfrak{R}}_S(\mathcal{G})\Big]=\mathbb{E}[g]-\mathfrak{R}_n(\mathcal{G})$$

We now check the bounded difference property of $\phi(S)$, for $1\leqslant i\leqslant n$, we have:

$$\begin{aligned}|\phi(S)-\phi(S_i)|&=\left|\frac{1}{n}\Big(g(Z_i)-g(Z_i')\Big)-\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sigma_i\Big(g(Z_i)-g(Z_i')\Big)\right]\right|\\&=\left|\frac{1}{n}\Big(g(Z_i)-g(Z_i')\Big)-\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\Big(g(Z_i)-g(Z_i')\Big)\right]\right|\quad(\textit{Since }\sigma_i\in\{-1,1\})\\&\leqslant 2\cdot\frac{b-a}{n}\end{aligned}$$

Hence, by the bounded difference inequality, let $\epsilon>0$, we have:

$$\begin{aligned}P\Big(\phi(S)-\mathbb{E}[\phi(S)]\geqslant\epsilon\Big)&\leqslant\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}4(b-a)^2/n^2}\right)\\&=\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right)\end{aligned}$$

Letting $\delta=\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right)$, we have:

$$\epsilon=(b-a)\sqrt{\frac{2\log 1/\delta}{n}}$$

Therefore, with probability of at least $1-\delta$ $(\delta>0)$, we have:

$$\begin{aligned}&\phi(S)-\mathbb{E}[\phi(S)]\leqslant(b-a)\sqrt{\frac{2\log 1/\delta}{n}}\\\Longrightarrow\ &\widehat{\mathbb{E}_S}[g]-\mathbb{E}[g]\leqslant(\widehat{\mathfrak{R}}_S(\mathcal{G})-\mathfrak{R}_n(\mathcal{G}))+(b-a)\sqrt{\frac{2\log 1/\delta}{n}}\end{aligned}$$

□.

Exercise 6.2

Definition (Partition classifier) : Let $\Pi = \Big\{A_1, \ldots, A_k\Big\}$ be a fixed partition of $\mathcal{X}$. The partition classifiers assign labels that corresponds to the partitions. Specifically, we define the following set of classifiers:

$$\mathcal{H} = \left\{ h : \mathcal{X} \to \{a_1, \ldots, a_k\} \middle| h(x) = \sum_{j=1}^{k} a_j \mathbf{1}_{\{x \in A_j\}},\ a_j \in \{-1, 1\} \right\}$$

Then we have $|\mathcal{H}| = 2^k$. Derive the exact empirical Rademacher complexity of $\mathcal{H}$.

Solution (Exercise 6.2). *We have:*

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma\left[\sup_{a_1, \ldots, a_k} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{k} a_j \mathbf{1}_{\{X_i \in A_j\}} \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma\left[\sup_{a_1, \ldots, a_k} \sum_{j=1}^{k} \sum_{i=1}^{n} \sigma_i a_j \mathbf{1}_{\{X_i \in A_j\}} \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma\left[\sup_{a_1, \ldots, a_k} \sum_{j=1}^{k} \sum_{i \in [n]; X_i \in A_j} \sigma_i a_j \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma\left[\sum_{j=1}^{k} \sup_{a_j \in \{-1,1\}} \sum_{i \in [n]; X_i \in A_j} \sigma_i a_j \right] \\
&= \frac{1}{n} \sum_{j=1}^{k} \mathbb{E}_\sigma\left[\sup_{a_j \in \{-1,1\}} a_j \sum_{i \in [n]; X_i \in A_j} \sigma_i \right]
\end{aligned}$$

Given a sample of Rademacher variables, to maximize the sum of those variables, the strategy is to multiply the entire sample with the dominant sign. For example,

$$\begin{aligned}
\{-1, -1, 1\} &\to \{1, 1, -1\} \qquad \Sigma = 1 \\
\{1, 1, -1\} &\to \{1, 1, -1\} \qquad \Sigma = 1
\end{aligned}$$

This is equivalent to taking the absolute value of the sum. Hence, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{n} \sum_{j=1}^{k} \mathbb{E}_\sigma\left[\left| \sum_{i \in [n]; X_i \in A_j} \sigma_i \right| \right] \\
&= \frac{1}{n} \sum_{j=1}^{k} \mathbb{E}_\sigma\left[\left| \sum_{i \in \mathcal{N}_j} \sigma_i \right| \right], \quad \mathcal{N}_j = \Big\{ i : X_i \in A_j \Big\}
\end{aligned}$$

We can model $\left| \sum_{i \in \mathcal{N}_j} \sigma_i \right|$ *as the absolute difference between two binomial variables with distribution* $Binomial(n_j, p = \frac{1}{2})$ *where* $n_j = |\mathcal{N}_j|$. *For each* $j \in \{1, \ldots, k\}$, *denote that:*

$$\left| \sum_{i \in \mathcal{N}_j} \sigma_i \right| = \Big| X_j - Y_j \Big| \text{ where } X_j, Y_j \sim Binomial\left(n_j, p = \frac{1}{2} \right)$$

Compute $P\Big(\Big|X_j - Y_j\Big| = z\Big)$ ***for*** $z \neq 0$
For the case when $X_j \neq Y_j$*, we have:*

$$P\Big(\Big|X_j - Y_j\Big| = z\Big) = P\Big(X_j = z + Y_j\Big) + P\Big(Y_j = z + X_j\Big)$$

Since the two events are independent and has the same probability (due to the fact that X_j *and* Y_j *are identically distributed). By the law of total probability, we have:*

$$\begin{aligned}
P\Big(X_j = z + Y_j\Big) &= \sum_{y=0}^{n_j - z} P(X_j = z + Y_j | Y_j = y) P(Y_j = y) \\
&= \sum_{y=0}^{n_j - z} \left[\binom{n_j}{z+y} \frac{1}{2^{n_j}} \right] \left[\binom{n_j}{y} \frac{1}{2^{n_j}} \right] \\
&= \frac{1}{2^{2n_j}} \sum_{y=0}^{n_j - z} \binom{n_j}{y} \binom{n_j}{z+y} \\
&= \frac{1}{2^{2n_j}} \sum_{y=0}^{n_j - z} \binom{n_j}{y} \binom{n_j}{n_j - z - y} \\
&= \frac{1}{2^{2n_j}} \binom{2n_j}{n_j - z} \quad \textit{(Vandermonde's Identity)}
\end{aligned}$$

Hence, for $X_j \neq Y_j$*, we have:*

$$\begin{aligned}
P\Big(\Big|X_j - Y_j\Big| = z\Big) &= 2 \times \frac{1}{2^{2n_j}} \binom{2n_j}{n_j - z} \\
&= \frac{1}{2^{2n_j - 1}} \binom{2n_j}{n_j - z}
\end{aligned}$$

Compute $\mathbb{E}\Big[\Big|X_j - Y_j\Big|\Big]$
Now that we have the PMF for $\Big|X_j - Y_j\Big|$*, we have:*

$$\begin{aligned}
\mathbb{E}\Big[\Big|X_j - Y_j\Big|\Big] &= \sum_{z=1}^{n_j} z \times P\Big(\Big|X_j - Y_j\Big| = z\Big) \\
&= \frac{1}{2^{2n_j - 1}} \sum_{z=1}^{n_j} z \times \binom{2n_j}{n_j - z}
\end{aligned}$$

Now, plug the above into the formula of the empirical Rademacher Complexity, we have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{n} \sum_{j=1}^{k} \mathbb{E}_\sigma \left[\left| \sum_{i \in \mathcal{N}_j} \sigma_i \right| \right] \\
&= \frac{1}{n} \sum_{j=1}^{k} \mathbb{E}\Big[\Big|X_j - Y_j\Big|\Big] \\
&= \frac{1}{n} \sum_{j=1}^{k} \frac{1}{2^{2n_j - 1}} \sum_{z=1}^{n_j} z \times \binom{2n_j}{n_j - z}
\end{aligned}$$

□.

Exercise 6.3

Let $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2 \subset [a,b]^{\mathcal{Z}}$. Let $c, d \in \mathbb{R}$. Prove the following properties of the empirical Rademacher Complexity:

- $\widehat{\mathfrak{R}}_S(c\mathcal{G}+d) = |c|\widehat{\mathfrak{R}}_S(\mathcal{G})$ where $c\mathcal{G}+d = \Big\{g'(z) = c\cdot g(z)+d \Big| g \in \mathcal{G}\Big\}$.
- $\widehat{\mathfrak{R}}_S(conv(\mathcal{G})) = \widehat{\mathfrak{R}}_S(\mathcal{G})$ where $conv(\mathcal{G}) = \Big\{ \sum_{i=1}^n \alpha_i g_i \Big| \alpha_i \geqslant 0, \sum_{i=1}^n \alpha_i = 1, g_i \in \mathcal{G}\Big\}$.
- $\widehat{\mathfrak{R}}_S(\mathcal{G}_1+\mathcal{G}_2) = \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + \widehat{\mathfrak{R}}_S(\mathcal{G}_2)$ where $\mathcal{G}_1+\mathcal{G}_2 = \Big\{g_1+g_2 \Big| g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\Big\}$.

As a result, the above properties also apply to the Rademacher Complexity.

Solution (Exercise 6.3)**.**

Proving each property one by one, we have:

(i) ***Prove that*** $\widehat{\mathfrak{R}}_S(c\mathcal{G}+d) = |c|\widehat{\mathfrak{R}}_S(\mathcal{G})$
where $c\mathcal{G}+d = \Big\{g'(z) = c\cdot g(z)+d \Big| g \in \mathcal{G}\Big\}$**.**
Given a sample $S = \{Z_1, \dots, Z_n\}$*. Denote* $\mathcal{G}' = c\mathcal{G}+d$*, we have:*

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}') &= \mathbb{E}_\sigma\left[\sup_{g'\in\mathcal{G}'} \frac{1}{n}\sum_{i=1}^n \sigma_i g'(Z_i)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n \sigma_i\Big(c\cdot g(Z_i)+d\Big)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n \sigma_i\Big(c\cdot g(Z_i)\Big)\right] + \mathbb{E}_\sigma\left[d\cdot\sum_{i=1}^n \sigma_i\right] \\
&= |c|\cdot\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n \sigma_i g(Z_i)\right] + |d|\cdot\underbrace{\mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i\right]}_{=0} \qquad \textit{(Since } \sigma_i\textit{'s are symmetric)} \\
&= |c|\cdot\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n \sigma_i g(Z_i)\right] = |c|\cdot\widehat{\mathfrak{R}}_S(\mathcal{G})
\end{aligned}$$

(ii) ***Prove that*** $\widehat{\mathfrak{R}}_S(\mathcal{G}_1+\mathcal{G}_2) = \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + \widehat{\mathfrak{R}}_S(\mathcal{G}_2)$
where $\mathcal{G}_1+\mathcal{G}_2 = \Big\{g_1+g_2 \Big| g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\Big\}$**.**

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}_1+\mathcal{G}_2) &= \mathbb{E}_\sigma\left[\sup_{g_1\in\mathcal{G}_1, g_2\in\mathcal{G}_2} \frac{1}{n}\sum_{i=1}^n \sigma_i\cdot(g_1(Z_1)+g_2(Z_2))\right] \\
&= \mathbb{E}_\sigma\left[\sup_{g_1\in\mathcal{G}_1} \frac{1}{n}\sum_{i=1}^n \sigma_i g_1(Z_i)\right] + \mathbb{E}_\sigma\left[\sup_{g_2\in\mathcal{G}_2} \frac{1}{n}\sum_{i=1}^n \sigma_i g_2(Z_i)\right] \\
&= \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + \widehat{\mathfrak{R}}_S(\mathcal{G}_2)
\end{aligned}$$

(iii) ***Prove that*** $\widehat{\mathfrak{R}}_S(conv(\mathcal{G})) = \widehat{\mathfrak{R}}_S(\mathcal{G})$
where $conv(\mathcal{G}) = \Big\{ \sum_{i=1}^n \alpha_i g_i \Big| \alpha_i \geqslant 0, \sum_{i=1}^n \alpha_i = 1, g_i \in \mathcal{G}\Big\}$**.**
We can rewrite $conv(\mathcal{G})$ *as:*

$$conv(\mathcal{G}) = \bigcup_{\{\alpha_1,\dots,\alpha_n\}\subset\mathbb{R}_+, \sum_{i=1}^n \alpha_i = 1} \left\{\sum_{i=1}^n \alpha_i \mathcal{G}\right\}$$

Hence, for any set of $\{\alpha_1, \dots, \alpha_n\} \subset \mathbb{R}_+$ *such that* $\sum_{i=1}^n \alpha_i = 1$*, we have:*

$$\begin{aligned}\hat{\mathfrak{R}}_S(conv(\mathcal{G})) &= \sum_{i=1}^n |\alpha_i| \hat{\mathfrak{R}}_S(\mathcal{G}) \\ &= \hat{\mathfrak{R}}_S(\mathcal{G}) \sum_{i=1}^n \alpha_i \\ &= \hat{\mathfrak{R}}_S(\mathcal{G})\end{aligned}$$

□.

Exercise 6.4

Prove theorem 6.3 (Two-sided Rademacher Complexity bounds).

Solution (Exercise 6.4)**.**

Applying theorem 6.2 for $-\mathcal{G}$*, for any* $0 < \delta < 1$*, we have:*

$$\begin{aligned}&P\left(\sup_{g' \in -\mathcal{G}} \left\{\mathbb{E}[g'(Z)] - \frac{1}{n}\sum_{i=1}^n g'(Z_i)\right\} \leqslant 2\mathfrak{R}_n(-\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right) \geqslant 1 - \delta/2 \\ \implies &P\left(\sup_{g \in \mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right\} \leqslant 2\mathfrak{R}_n(-\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right) \geqslant 1 - \delta/2 \\ \implies &P\left(\sup_{g \in \mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right\} \geqslant 2\mathfrak{R}_n(-\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right) \leqslant \delta/2\end{aligned}$$

Since $\mathfrak{R}_n(-\mathcal{G}) = |-1| \cdot \mathfrak{R}_n(\mathcal{G}) = \mathfrak{R}_n(\mathcal{G})$*, we have:*

$$P\left(\sup_{g \in \mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right\} \geqslant 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right) \leqslant \delta/2$$

Now, we have:

$$\begin{aligned}P(\sup|A| \geqslant \epsilon) &= P\Big(\Big\{\sup\{A\} \geqslant \epsilon\Big\} \cup \Big\{\sup\{-A\} \geqslant \epsilon\Big\}\Big) \\ &\leqslant P\Big(\Big\{\sup\{A\} \geqslant \epsilon\Big\}\Big) + P\Big(\Big\{\sup\{-A\} \geqslant \epsilon\Big\}\Big)\end{aligned}$$

Letting $\epsilon = 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}$*, we have:*

$$\begin{aligned}P\left(\sup_{g \in \mathcal{G}} \left|\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right| \geqslant \epsilon\right) &\leqslant P\left(\sup_{g \in \mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right\} \geqslant \epsilon\right) \\ &\quad + \underbrace{P\left(\sup_{g \in \mathcal{G}} \left\{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right\} \geqslant \epsilon\right)}_{\textit{Original one-sided bound}} \\ &\leqslant \delta/2 + \delta/2 = \delta\end{aligned}$$

Applying theorem 6.2 and the same reasoning for $-\mathcal{G}$ *for the empirical Rademacher Complexity bound, we also have:*

$$P\left(\sup_{g \in \mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right| \geqslant 2\hat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log 4/\delta}{2n}}\right) \leqslant \delta/2 + \delta/2 = \delta$$

□.

Exercise 6.5

Prove theorem 6.5 (Two-sided VC inequality).

Solution (Exercise 6.5)**.**

Applying theorem 6.4 for $-\mathcal{H}$, with probability at $1-\delta/2$, we have:

$$\begin{aligned}
&\sup_{h'\in -\mathcal{H}}\left\{R(h')-\widehat{R}_n(h')\right\}\leqslant\sqrt{\frac{8(\log S_{-\mathcal{H}}(n)+\log 2/\delta)}{n}}\\
\implies&\sup_{h\in\mathcal{H}}\left\{R(-h)-\widehat{R}_n(-h)\right\}\leqslant\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 2/\delta)}{n}}\quad(S_{\mathcal{H}}(n)=S_{-\mathcal{H}}(n))\\
\implies&\sup_{h\in\mathcal{H}}\left\{\left[1-R(h)\right]-\left[1-\widehat{R}_n(h)\right]\right\}\leqslant\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 2/\delta)}{n}}\\
\implies&\sup_{h\in\mathcal{H}}\left\{\widehat{R}_n(h)-R(h)\right\}\leqslant\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 2/\delta)}{n}}
\end{aligned}$$

Now we set:

$$\epsilon=\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 2/\delta)}{n}}\implies\delta=2S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}$$

Therefore, for $\epsilon>0$, we have:

$$\begin{aligned}
P\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_n(h)\right|\geqslant\epsilon\right)&\leqslant P\left(\sup_{h\in\mathcal{H}}\left\{R(h)-\widehat{R}_n(h)\right\}\geqslant\epsilon\right)+P\left(\sup_{h\in\mathcal{H}}\left\{\widehat{R}_n(h)-R(h)\right\}\geqslant\epsilon\right)\\
&\leqslant\delta/2+\delta/2=\delta\\
&=2S_{\mathcal{H}}(n)e^{-n\epsilon^2/8}
\end{aligned}$$

In other words, with probability of at least $1-\delta$, we have:

$$\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_n(h)\right|\leqslant\sqrt{\frac{8(\log S_{\mathcal{H}}(n)+\log 2/\delta)}{n}}$$

□.

Exercise 6.6

Improve theorem 6.5 by improving the exponential's power and compromising the constant in front of the exponential.

Solution (Exercise 6.6)**.**

By corollary 6.1, with probability of at least $1-\delta$, we have:

$$\begin{aligned}
\sup_{h\in\mathcal{H}}\left\{R(h)-\widehat{R}_n(h)\right\}&\leqslant\mathfrak{R}_n(\mathcal{H})+\sqrt{\frac{\log 1/\delta}{2n}}\\
&\leqslant\sqrt{\frac{2\log S_{\mathcal{H}}(n)}{n}}+\sqrt{\frac{\log 1/\delta}{2n}}\quad\textit{(Massart's Lemma)}\\
&\leqslant\sqrt{2\left(\frac{2\log S_{\mathcal{H}}(n)}{n}+\frac{\log 1/\delta}{2n}\right)}\quad(\sqrt{a}+\sqrt{b}\leqslant\sqrt{2(a+b)})\quad(**)\\
&=\sqrt{\frac{4\log S_{\mathcal{H}}(n)+\log 1/\delta}{n}}
\end{aligned}$$

Now we set:

$$\epsilon = \sqrt{\frac{4\log S_{\mathcal{H}}(n) + \log 1/\delta}{n}} \implies \delta = S_{\mathcal{H}}(n)^4 e^{-n\epsilon^2}$$

Hence, for $\epsilon > 0$, we have:

$$P\left(\sup_{h\in\mathcal{H}} \left\{ R(h) - \widehat{R}_n(h) \right\} \geqslant \epsilon \right) \leqslant S_{\mathcal{H}}(n)^4 e^{-n\epsilon^2}$$

Consequently, we have the two-sided bound:

$$P\left(\sup_{h\in\mathcal{H}} \left| R(h) - \widehat{R}_n(h) \right| \geqslant \epsilon \right) \leqslant 2S_{\mathcal{H}}(n)^4 e^{-n\epsilon^2}$$

***Remark** : Notice that in $(**)$ we used a tighter inequality $\sqrt{a} + \sqrt{b} \leqslant \sqrt{2(a+b)}$ instead of $\sqrt{a} + \sqrt{b} \leqslant 2\sqrt{a+b}$ like Clayton Scott's proof in theorem 6.4.* ▫.

7 Kernels and Hilbert Spaces

7.1 Pre-Hilbert Spaces

Definition 7.1 (Pre-Hilbert Spaces)**.**

A real inner product space (IPS) is a pair $\left(V, \langle ., . \rangle\right)$ *where* V *is a real vector space equipped with and inner product* $\langle ., . \rangle : V \times V \to \mathbb{R}$ *that satisfies:*

- ***Positive semi-definiteness*** *:* $\langle u, u \rangle \geqslant 0, \ \forall u \in V$ *and* $\langle u, u \rangle = 0 \iff u = 0$.
- ***Symmetry*** *:* $\langle u, v \rangle = \langle v, u \rangle, \ \forall u, v \in V$.
- ***Linearity*** *:* $\langle a_1 u_1 + a_2 u_2, v \rangle = a_1 \langle u_1, v \rangle + a_2 \langle u_2, v \rangle, \ \forall u_1, u_2, v \in V$ *and* $a_1, a_2 \in \mathbb{R}$.

We can also call $\left(V, \langle ., . \rangle\right)$ *a* ***Pre-Hilbert Space****.*

Proposition 7.1: Normed induced by Pre-Hilbert Spaces

If V is a Pre-Hilbert space with an inner product $\langle ., . \rangle$. Then,

$$\|u\| = \sqrt{\langle u, u \rangle}, \ \forall u \in V$$

Is a norm on V.

Proof (Proposition 7.1)**.**

To prove that $\|.\|$ *is a norm on* V*, we have to prove that it satisfies the following properties: absolute homogeneity, positive semi-definiteness and triangle inequality.*

- ***Absolute homogeneity*** *: Let* $\alpha \in \mathbb{R}$ *and* $u \in V$*, we have*

$$\begin{aligned} \|\alpha u\| &= \langle \alpha u, \alpha u \rangle^{\frac{1}{2}} \\ &= \left(\alpha^2 \langle u, u \rangle\right)^{\frac{1}{2}} \quad \textit{(Linearity of inner product)} \\ &= |\alpha| \cdot \langle u, u \rangle^{\frac{1}{2}} \\ &= |\alpha| \cdot \|u\| \end{aligned}$$

- ***Positive semi-definiteness*** *: This property is inferred directly from the positive semi-definiteness of inner product.*
- ***Triangle inequality*** *: Let* $u, v \in V$*, we have*

$$\begin{aligned} \|u + v\| &= \langle u + v, u + v \rangle^{\frac{1}{2}} \\ &= \left(\langle u, u \rangle + 2\langle u, v \rangle + \langle v, v \rangle\right)^{\frac{1}{2}} \\ &= \left(\|u\|^2 + 2\langle u, v \rangle + \|v\|^2\right)^{\frac{1}{2}} \\ &\leqslant \left(\|u\|^2 + 2\|u\| \cdot \|v\| + \|v\|^2\right)^{\frac{1}{2}} \quad \textit{(Cauchy-Schwarz Inequality)} \\ &= \|u\| + \|v\| \end{aligned}$$

Note that the proof for triangle inequality makes use of the Cauchy-Schwarz inequality, which we will prove shortly. □.

Proposition 7.2: Cauchy-Schwarz Inequality

Let $\Big(V, \big\langle ., . \big\rangle\Big)$ be a Pre-Hilbert space, we have:

$$\Big|\big\langle u, v \big\rangle\Big| \leqslant \|u\| \cdot \|v\|, \ \forall u, v \in V$$

Proof (Proposition 7.2). *Consider the following matrix:*

$$G(u,v) = \begin{pmatrix} \big\langle u, u \big\rangle & \big\langle u, v \big\rangle \\ \big\langle v, u \big\rangle & \big\langle v, v \big\rangle \end{pmatrix} = \begin{pmatrix} \|u\|^2 & \big\langle u, v \big\rangle \\ \big\langle v, u \big\rangle & \|v\|^2 \end{pmatrix}$$

Notice that all elements of $G(u,v)$ are non-negative. Hence, $G(u,v)$ is positive semi-definite and we have:

$$\begin{aligned} \det\Big(G(u,v)\Big) \geqslant 0 &\implies \|u\|^2 \cdot \|v\|^2 - \big\langle u, v \big\rangle \cdot \big\langle v, u \big\rangle \geqslant 0 \\ &\implies \|u\|^2 \cdot \|v\|^2 \geqslant \big\langle u, v \big\rangle \cdot \big\langle v, u \big\rangle \\ &\implies \|u\| \cdot \|v\| \geqslant \Big|\big\langle u, v \big\rangle\Big| \end{aligned}$$

□.

7.2 Hilbert Spaces

In this section we will discuss what a metric space is and the criteria for a Pre-Hilbert space to be a Hilbert space.

Definition 7.2 (Cauchy Sequence). *Let M be a set and $\{x_n\}_{n=1}^{\infty}$ be a sequence in M. We say that $\{x_n\}_{n=1}^{\infty}$ is a **Cauchy sequence** if the elements get closer and closer as $n \to \infty$. Formally, for all $\epsilon > 0$, we have:*

$$\exists N \in \mathbb{N} \text{ such that: } \forall i, j \geqslant N, \ d(x_i, x_j) < \epsilon$$

Where $d : M \times M \to \mathbb{R}$ is a metric, which is explained in the definition below.

Definition 7.3 (Metric Space). *A metric space is a pair $\Big(M, d\Big)$ where M is a set and $d : M \times M \to \mathbb{R}$ satisfies:*

- ***Positive semi-definiteness** : $d(x,y) \geqslant 0, \ \forall x, y \in M$ and $d(x,y) = 0 \iff x = y$.*
- ***Symmetry** : $d(x,y) = d(y,x), \ \forall x, y \in M$.*
- ***Triangle inequality** : $d(x,y) \leqslant d(x,z) + d(y,z), \ \forall x, y, z \in M$.*

*We say that $\Big(M, d\Big)$ is a **complete metric space** if and only if **every Cauchy sequence converges to an element in** M.*

Definition 7.4 (Hilbert Space)**.**

A Pre-Hilbert Space $\left(V, \langle ., . \rangle\right)$ *is a* ***Hilbert Space*** *if* $\left(V, d_{\|.\|}\right)$ *is a complete metric space where* $d_{\|.\|}$ *is the metric induced by the inner product:*

$$d_{\|.\|}(x, y) = \|x - y\|, \ x, y \in V$$

Where $\|u\| = \sqrt{\langle u, u \rangle}$ *for* $u \in V$*.*

7.3 Important theorems in Hilbert Spaces

In this section, we will discuss two important theorems, namely the **Hilbert Projection Theorem** and the **Riesz Representation Theorem**.

7.3.1 Projection Theorem

The **Hilbert Projection Theorem** talks about the condition when any vector in a Hilbert Space will have a unique projection onto a subspace. Before proving the theorem, we look at the Polarization Identity.

Note : From here, we denote Hilbert Spaces as H.

Lemma 7.1: Polarization Identity

Let H be a Hilbert Space. For all $f, g \in H$, we have:

$$\|f + g\|^2 + \|f - g\|^2 = 2\left(\|f\|^2 + \|g\|^2\right)$$

Proof (Lemma 7.1)**.**

We have:

$$\begin{aligned} \|f + g\|^2 &= \langle f + g, f + g \rangle \\ &= \|f\|^2 + 2\langle f, g \rangle + \|g\|^2 \quad (1) \\ \|f - g\|^2 &= \langle f - g, f - g \rangle \\ &= \|f\|^2 - 2\langle f, g \rangle + \|g\|^2 \quad (2) \end{aligned}$$

Combining (1) *and* (2) *yields the Polarization Identity.* □.

Theorem 7.1: Hilbert Projection Theorem

Let H be a Hilbert Space and $V \subseteq H$ be a closed subspace. For any $f \in H$, there exists a unique $v \in V$ such that:

$$\|f - v\| = \inf_{v^* \in V} \|f - v^*\|$$

In other words, **there exists a unique minimizer inside the subspace if and only if the subspace is closed**.

Proof (Theorem 7.1)**.**

In this proof, without loss of generality, we work with the minimizer for the norm squared rather than the norm.

Claim : *V* ***is closed*** $\implies$ ***unique minimizer exists***
Suppose that V is a closed subspace of H. Since H is a complete metric space, V is also complete. Now let:

$$\alpha^2 = \inf_{v^* \in V} \|f - v^*\|^2$$

Since for all $v \in V$, we have $\|f - v\|^2 \geqslant \alpha^2$. Choose a sequence $\{v_n\}_{n=1}^{\infty} \subset V$ such that:

$$0 \leqslant \|f - v_n\|^2 - \alpha^2 \leqslant \frac{1}{n} \quad (1)$$

For any $m, n \in \mathbb{N}$, we have:

$$\|f - v_n\|^2 + \|f - v_m\|^2 - 2\alpha^2 \leqslant \frac{1}{m} + \frac{1}{n}$$

By Polarization Identity 7.1, we have $\|f - v_n\|^2 + \|f - v_m\|^2 = \frac{1}{2}\Big(\|v_n - v_m\|^2 + \|2f - (v_n + v_m)\|^2\Big)$. Hence, we have:

$$\|v_n - v_m\|^2 + \|2f - (v_n + v_m)\|^2 - 4\alpha^2 \leqslant \frac{2}{m} + \frac{2}{n}$$
$$\implies \|v_n - v_m\|^2 + 4\left(\left\|f - \frac{v_n + v_m}{2}\right\|^2 - \alpha^2\right) \leqslant \frac{2}{m} + \frac{2}{n} \quad (2)$$

Since $\frac{v_n + v_m}{2} \in V$, we have:

$$\left\|f - \frac{v_n + v_m}{2}\right\|^2 \geqslant \alpha \implies \|v_n - v_m\|^2 + 4\left(\left\|f - \frac{v_n + v_m}{2}\right\|^2 - \alpha^2\right) \geqslant \|v_n - v_m\|^2 \quad (3)$$

From (2) and (3), we have:

$$\|v_n - v_m\|^2 \leqslant \frac{2}{m} + \frac{2}{n}$$

Taking $m, n \to \infty$, we have $\|v_n - v_m\|^2 \to 0$. Therefore, $\{v_n\}_{n=1}^{\infty}$ is a Cauchy sequence. Since V is a complete metric space, $v_n \to v \in V$. From (1) we have:

$$\|f - v\|^2 = \lim_{n \to \infty} \|f - v_n\|^2 = \alpha^2$$

Hence, We have proven that a minimizer indeed exists in V. Now, we have to prove that the minimizer is actually unique. Suppose that we have another $\hat{v} \in V$ such that:

$$\hat{v} = \arg\min_{v^* \in V} \|f - v^*\|^2$$

Then, by the Polarization Identity 7.1, we have:

$$\begin{aligned}
2\alpha^2 &= \|f - v\|^2 + \|f - \hat{v}\|^2 \\
&= \frac{1}{2}\Big(\|v - \hat{v}\|^2 + \|2f - (v + \hat{v})\|^2\Big) \quad \textit{(Polarization Identity)} \\
&= \frac{1}{2}\|v - \hat{v}\|^2 + 2\left\|f - \frac{v + \hat{v}}{2}\right\|^2 \\
&\geqslant \frac{1}{2}\|v - \hat{v}\|^2 + 2\alpha^2 \quad \left(\textit{Since } \frac{v + \hat{v}}{2} \in V\right)
\end{aligned}$$

From the above, we have:

$$0 \geqslant \frac{1}{2}\|v - \hat{v}\|^2 \geqslant 0 \implies \|v - \hat{v}\|^2 = 0$$

Therefor, $v = \hat{v}$ and the minimizer is unique.

Claim : unique minimizer exists $\implies$ V is closed
Suppose that V is not closed, meaning it does not contain all its limit points. Hence, there exists a sequence $\{f_n\}_{n=1}^{\infty} \subset V$ such that $f_n \to f \in H \backslash V$. By the assumption, there exists $v \in V$ such that:

$$\|f - v\|^2 = \inf_{v^* \in V} \|f - v^*\|^2$$

But then, we have:

$$0 \leqslant \|f - v\|^2 \leqslant \|f - f_n\|^2$$

Therefore,

$$0 \leqslant \|f - v\|^2 \leqslant \lim_{n \to \infty} \|f - f_n\|^2 = 0 \implies f = v$$

This means that $f \in V \implies$ contradiction. Hence, we can conclude that V is a closed subspace of H.

□.

Corollary 7.1: Hilbert Spaces as direct sum

Let H be a Hilbert Space and $V \subseteq H$ is a closed subspace. Then, we can write:

$$H = V \oplus V^{\perp}$$

In other words, we can define:

$$H = \left\{ x + y \middle| x \in V, y \in V^{\perp} \right\}$$

Proof (Corollary 7.1)**.**
By the Projection Theorem, there exists a unique $x \in V$ such that:

$$\|f - x\|^2 \leqslant \|f - v\|^2, \ \forall v \in V$$

Let $v = x + \lambda z$ where $\lambda \in \mathbb{R}$ is a scalar and $z \in V$. Then, we have:

$$\begin{aligned} \|f - x\|^2 &\leqslant \|f - (x + \lambda z)\|^2, \ \forall z \in V \\ &= \|(f - x) - \lambda z\|^2 \\ &= \|f - x\|^2 + |\lambda|^2 \|z\|^2 - 2\lambda \langle f - x, z \rangle \end{aligned}$$

The above inequality holds if and only if:

$$|\lambda|^2 \|z\|^2 - 2\lambda \langle f - x, z \rangle \geqslant 0$$

Plugging $\lambda = \frac{\langle f - x, z \rangle}{\|z\|^2}$ into the above inequality, we have:

$$\frac{\left| \langle f - x, z \rangle \right|^2}{\|z\|^2} \leqslant 0 \implies \langle f - x, z \rangle = 0, \ \forall z \in V$$

From the above, we conclude that for all $f \in H$, we have:

$$f = x + y \text{ where } \begin{cases} x &= \arg\min_{v^* \in V} \|f - v^*\|^2 \\ y &= f - x, \ y \in V^{\perp} \end{cases}$$

In other words, we have:

$$H = V \oplus V^{\perp}$$

□.

Corollary 7.2: $H = \ker(\Phi) \oplus \ker(\Phi)^{\perp}$

Let H be a Hilbert Space and $\Phi : H \to \mathbb{R}$ (or $\mathbb{C}$) be a bounded linear functional (whose definition is stated below). We define the **kernel** of Φ as the set:

$$\ker(\Phi) = \left\{ x \in H \;\middle|\; \Phi(x) = 0 \right\}$$

Then, we can write H as the direct sum:

$$H = \ker(\Phi) \oplus \ker(\Phi)^{\perp}$$

Proof (Corollary 7.2). *Since Φ is a bounded linear functional in a complete metric space, it is continuous. Hence, the pre-image of any closed subset of $\mathbb{R}$ is closed. Hence, we have:*

$$\ker(\Phi) = \Phi^{-1}(\{0\}) \text{ is closed}$$

Therefore, by corollary 7.1, we have:

$$H = \ker(\Phi) \oplus \ker(\Phi)^{\perp}$$

□.

7.3.2 Representation Theorem

Definition 7.5 (Bounded linear functional). *Let V be a complete metric space (or Banach space). A mapping $\Phi : V \to \mathbb{R}$ (or $\mathbb{C}$) is called a **bounded linear functional** if it satisfies:*

- ***Linearity*** *: $\Phi(\alpha f + \beta g) = \alpha\Phi(f) + \beta\Phi(g)$ for $\alpha, \beta \in \mathbb{R}$ and $f, g \in V$.*
- ***Boundedness*** *: $\exists C > 0 : |\Phi(f)| \leqslant C \cdot \|f\|, \ \forall f \in V$.*

Definition 7.6 (Dual space). *Let V be a complete metric space (or Banach space). The dual space of V, denoted as V^*, is the space of all bounded linear functionals from V to $\mathbb{R}$ (or $\mathbb{C}$):*

$$V^* = \left\{ \Phi : V \to \mathbb{R} \middle| \Phi \text{ linear and bounded} \right\}$$

For any $\Phi \in V^*$, we define the operator norm as:

$$\|\Phi\|_{V*} = \sup_{v \in V, \|v\|=1} \left| \Phi(z) \right|$$

Theorem 7.2: Riesz Representation Theorem

Let H be a Hilbert Space and $\Phi : H \to \mathbb{R}$ (or $\mathbb{C}$) be a bounded linear functional. Then,

- $\exists g \in H$ such that $\Phi(f) = \langle f, g \rangle, \ \forall f \in H$.
- $\|\Phi\|_{H*} = \|g\|_H$.

We call $g \in H$ the **Riesz representation** of Φ.

Proof (Theorem 7.2)**.** _____

□.

A Related topics

A.1 Neyman-Pearson Lemma

A.1.1 Type I & Type II errors

Overview : In a hypothesis test, we are interested in testing a given null hypothesis H_0 against some alternative hypothesis H_1. Hence, we define some rejection region $\mathcal{R} \subset \mathbb{R}$ such that:

$$x \in \mathcal{R} \implies \text{reject } H_0$$

Equivalently, denote that $\overline{\mathcal{R}}$ is the acceptance region. We define the following conditional probability densities:

- $f_1(x)$: Density given that H_1 is true.
- $f_0(x)$: Density given that H_0 is true.

Definition A.1 (Type I & Type II errors)**.** *In hypothesis testing, we define the type I error as the probability that we falsely reject the null hypothesis given that the null hypothesis is true. On the other hands, type II error is the probability that we falsely accept the null hypothesis given that the hypothesis is not true:*

$$\alpha = P_{H_0}(\mathcal{R}) = \int_{\mathcal{R}} f_0(x)dx$$
$$\beta = P_{H_1}(\overline{\mathcal{R}}) = 1 - \int_{\mathcal{R}} f_1(x)dx$$

There is a trade-off between type I and type II errors as illustrated in the figure below:

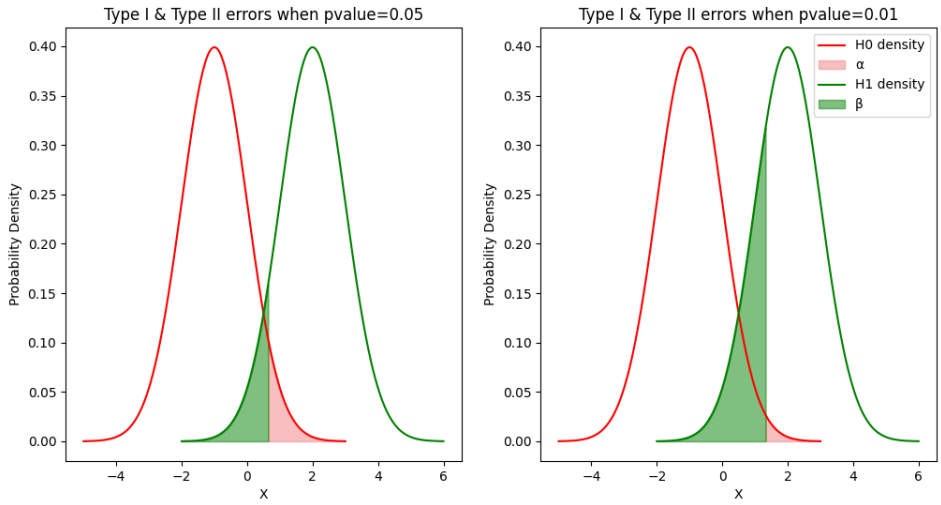

Figure 4: Trade-off between type I and type II errors

Definition A.2 (Power of hypothesis test)**.**
Given a hypothesis test used to test a null hypothesis H_0 against an alternative hypothesis H_1. The probability:

$$P_{H_1}(\mathcal{R}) = \int_{\mathcal{R}} f_1(x)dx = 1 - \beta$$

*Which denotes the probability that we correctly reject the null hypothesis given that H_1 is true is called the **Power** of the hypothesis test. Later on we will see that using **Neyman-Pearson Lemma**, we can prove any hypothesis test has the power of at most the likelihood ratio test's power.*

A.1.2 Neyman-Pearson Lemma

Overview : The Neyman-Pearson Lemma is concerned with maximizing the power of hypothesis test subjected to a certain degree of type I error. Formally, we are trying to solve the following constrained optimization problem:

$$\begin{aligned}\text{maximize} : P_{H_1}(\mathcal{R}) &= \int_{\mathcal{R}} f_1(x)dx \\ \text{subjected to} : P_{H_0}(\mathcal{R}) &= \int_{\mathcal{R}} f_0(x)dx \leqslant \alpha\end{aligned}$$

Theorem A.1: Neyman-Pearson Lemma

Let H_0 and H_1 be simple hypotheses. For a constant $c > 0$, suppose the likelihood ratio test rejects H_0 when $L(X) > c$ has significance level $\alpha \in (0, 1)$. **Then for any other test of H_0 with significance level of at most α, its power against H_1 is at most the power of the likelihood ratio test.**

$$\mathcal{R} = \left\{ x \in \mathbb{R} : L(x) = \frac{f_1(x)}{f_0(x)} > c \right\}$$

Proof (Theorem A.1)**.**
Note that the rejection region $\mathcal{R} = \left\{ x \in \mathbb{R} : L(x) = \frac{f_1(x)}{f_0(x)} > c \right\}$ maximizes the quantity:

$$\int_{\mathcal{R}} (f_1(x) - cf_0(x))dx$$

Because $f_1(x) - cf_0(x) < 0$ for all $x \notin \mathcal{R}$. Therefore, for any other test with rejection region $\mathcal{R}'$ with significance level of at most α, we have:

$$\begin{aligned}\int_{\mathcal{R}} (f_1(x) - cf_0(x))dx &\geqslant \int_{\mathcal{R}'} (f_1(x) - cf_0(x))dx \\ \implies P_{H_1}(\mathcal{R}) - P_{H_1}(\mathcal{R}') &\geqslant c\left(\int_{\mathcal{R}} f_0(x)dx - \int_{\mathcal{R}'} f_0(x)dx \right) \\ &= c\left(\alpha - \int_{\mathcal{R}'} f_0(x)dx \right)\end{aligned}$$

Since $\int_{\mathcal{R}'} f_0(x)dx \leqslant \alpha$, *we have:*

$$P_{H_1}(\mathcal{R}) - P_{H_1}(\mathcal{R}') \geqslant 0 \implies P_{H_1}(\mathcal{R}') \leqslant P_{H_1}(\mathcal{R})$$

Hence, for any test with significance level of at most α, *the power is at most the power of the likelihood ratio test* $P_{H_1}(\mathcal{R})$. □.

A.2 Rademacher Complexity bound for linear function classes

In the following section, we will explore a simple exercise for bounding the Rademacher Complexity for linear function classes. We will first look into the important theorems and lemmas. Then, we will tackle the problem at the end of this section.

A.2.1 Problem Statement

Problem : Let $\mathcal{F}$ be a linear function class defined as followed:

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f(x) = wx, \|w\|_2 \leqslant R \right\}$$

Our objective is to prove the following Rademacher Complexity bound:

$$\mathfrak{R}_n(\mathcal{F}) \leqslant \tilde{O}\left(\frac{R}{\sqrt{n}}\right)$$

Before solving the above problem, we have to get familiar with the definition of **covering number** and some related lemmas.

A.2.2 Covering Number

Definition A.3 (ϵ-Cover)**.**
Let Q *be a set. A subset* $\mathcal{C} \subset Q$ *is called an* ϵ***-cover*** *of* Q *with respect to a metric* ρ *if:*

$$\forall v \in Q, \exists v' \in \mathcal{C} : \rho(v, v') \leqslant \epsilon$$

Basically, $\mathcal{C}$ *can be thought of as a collection of centers of* ϵ***-balls overlapping*** Q *(Figure 5).*

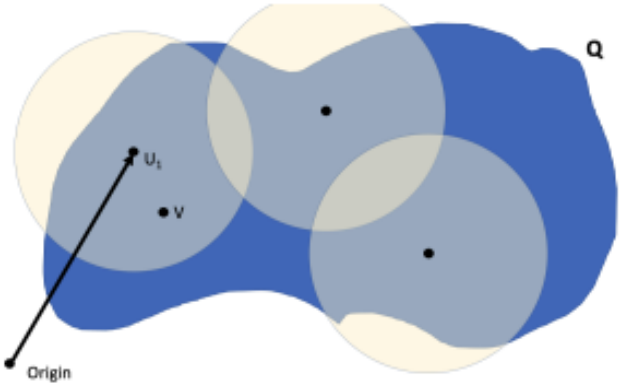

Figure 5: Examples of ϵ-balls. The centers of the balls would form an ϵ-cover if they completely contain Q.

Definition A.4 (Covering Number of sets ($\mathcal{N}(Q, \epsilon, \rho)$)**.**
*The **covering number** of a set Q is defined as the size of the smallest ϵ-cover needed to completely contain Q. In other words, it is the minimum number of ϵ-balls needed to completely contain Q:*

$$\mathcal{N}(Q, \epsilon, \rho) = \textit{minimum size of } \epsilon\textit{-cover of } Q \textit{ w.r.t } \rho$$

Visual illustration of covering number is included in figure 6.

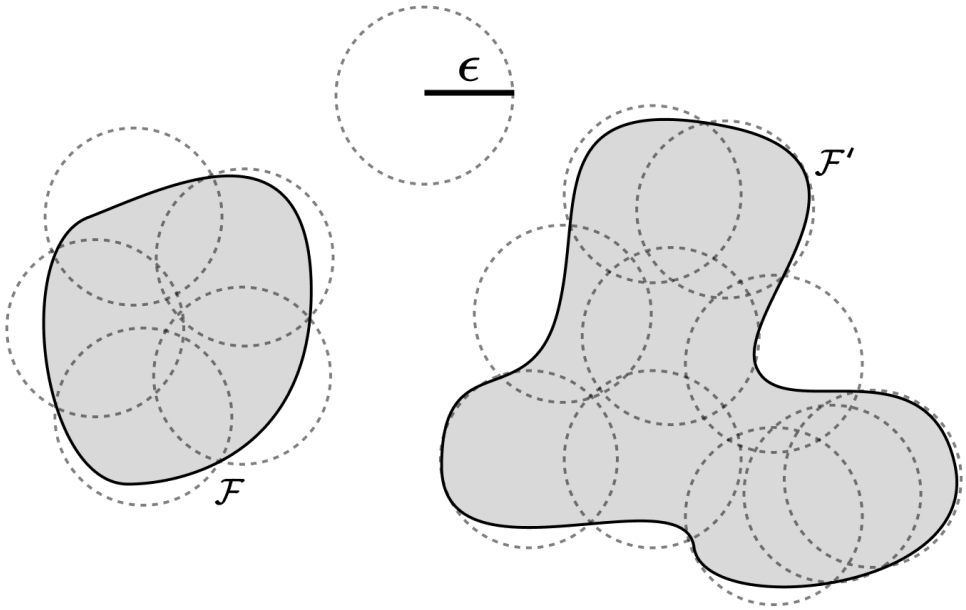

Figure 6: Example of covering number. For $\mathcal{F}$, the covering number is 5. For $\mathcal{F}'$, the covering number is 10.

Definition A.5 (Covering number of function class ($\mathcal{N}(\mathcal{F}, \epsilon, \rho)$))**.**
Let $\mathcal{F}$ be a function class. Define the restriction of $\mathcal{F}$ to the observations $\{x_1, \ldots, x_n\}$ as:

$$\mathcal{F}_{x_1,\ldots,x_n} = \Big\{ (f(x_1), \ldots, f(x_n)) : f \in \mathcal{F} \Big\}$$

*Then, we can define the **Covering Number** for $\mathcal{F}$ as followed:*

$$\mathcal{N}(\mathcal{F}, \epsilon, \rho) = \sup_{x_1,\ldots,x_n} \mathcal{N}(\mathcal{F}_{x_1,\ldots,x_n}, \epsilon, \rho)$$

*We can call $\mathcal{N}(\mathcal{F}_{x_1,\ldots,x_n}, \epsilon, \rho)$ the **Empirical Covering Number**.*

A.2.3 Massart's Lemma

Theorem A.2: Massart's Lemma

Let $A \subseteq \mathbb{R}^n$ and $|A| < \infty$. Let $r = \sup_{u \in A} \|u\|_2$ ($\|.\|_2$ is the l^2 norm). Then,

$$\mathbb{E}_\sigma \left[\frac{1}{n} \sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right] \leqslant \frac{r\sqrt{2 \ln |A|}}{n}$$

Where $u = \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix}^T$.

Proof (Theorem A.2). *For all $t \geqslant 0$, we have:*

$$\begin{aligned}
\exp\left(t\mathbb{E}_\sigma \left[\sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right] \right) &= \exp\left(\mathbb{E}_\sigma \left[t \sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right] \right) \\
&\leqslant \mathbb{E}_\sigma \left[\exp\left(t \sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right) \right] \quad (\textit{Jensen's Inequality}) \\
&= \mathbb{E}_\sigma \left[\sup_{u \in A} \exp\left(t \sum_{i=1}^{n} \sigma_i u_i \right) \right] \quad (\textit{Exponential is strictly increasing}) \\
&\leqslant \sum_{u \in A} \mathbb{E}_\sigma \left[\exp\left(t \sum_{i=1}^{n} \sigma_i u_i \right) \right]
\end{aligned}$$

Since all Rademacher variables σ_i are independent, we have:

$$\begin{aligned}
\exp\left(t\mathbb{E}_\sigma \left[\sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right] \right) &\leqslant \sum_{u \in A} \prod_{i=1}^{n} \mathbb{E}_{\sigma_i} \Big[\exp(t\sigma_i u_i) \Big] \\
&= \sum_{u \in A} \prod_{i=1}^{n} M_{\sigma_i}(tu_i)
\end{aligned}$$

We know that $\mathbb{E}[\sigma_i] = 0$ for every $1 \leqslant i \leqslant n$ and $-1 \leqslant \sigma_i \leqslant 1$ with probability 1. Hence, by Hoeffding's lemma 3.1, we have:

$$\begin{aligned}
\exp\left(t\mathbb{E}_\sigma \left[\sup_{u \in A} \sum_{i=1}^{n} \sigma_i u_i \right] \right) &\leqslant \sum_{u \in A} \prod_{i=1}^{n} M_{\sigma_i}(tu_i) \\
&\leqslant \sum_{u \in A} \prod_{i=1}^{n} \exp\left(\frac{4t^2 u_i^2}{8} \right) = \sum_{u \in A} \prod_{i=1}^{n} \exp\left(\frac{t^2 u_i^2}{2} \right) \\
&= \sum_{u \in A} \exp\left(\frac{t^2}{2} \sum_{i=1}^{n} u_i^2 \right) \\
&= \sum_{u \in A} \exp\left(\frac{t^2 \|u\|_2^2}{2} \right) \\
&\leqslant \sum_{u \in A} \exp\left(\frac{t^2 r^2}{2} \right) = |A| \exp\left(\frac{t^2 r^2}{2} \right)
\end{aligned}$$

Taking the log from both sides, we have:

$$t\mathbb{E}_\sigma\left[\sup_{u\in A}\sum_{i=1}^{n}\sigma_i u_i\right] \leqslant \log|A| + \frac{t^2r^2}{2}$$
$$\implies \mathbb{E}_\sigma\left[\sup_{u\in A}\sum_{i=1}^{n}\sigma_i u_i\right] \leqslant \inf_{t>0}\left(\frac{\log|A|}{t} + \frac{tr^2}{2}\right)$$

Setting the right hand side's derivative with respect to $t \geqslant 0$ to 0, we find that $t = r^{-1}\sqrt{2\log|A|}$. Plugging this value into the above bound, we have:

$$\mathbb{E}_\sigma\left[\sup_{u\in A}\sum_{i=1}^{n}\sigma_i u_i\right] \leqslant r\sqrt{2\log|A|}$$

□.

Corollary A.1: Massart's lemma bound on $\widehat{\mathfrak{R}}_S(\mathcal{F})$

Let $\mathcal{F}$ be a function class and let Q be its output space. Given a dataset $S = \{x_1, \ldots, x_n\}$, define the following norm for every $f \in \mathcal{F}$:

$$\|f\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(x_1)^2}$$

Then, we have the following bound on the Empirical Rademacher Complexity:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leqslant \sup_{f\in\mathcal{F}}\|f\|_2\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Proof (Corollary A.1)**.**
Let $r = \sup_{q\in Q}\|q\|_2$ where the norm applied on the output space is the l^2 norm (without the scaling factor $\frac{1}{\sqrt{n}}$). We have:

$$\begin{aligned}\sup_{f\in\mathcal{F}}\|f\|_2 &= \sup_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2}\\ &= \frac{1}{\sqrt{n}}\sup_{q\in Q}\sqrt{\sum_{i=1}^{n} q_i^2}\\ &= \frac{r}{\sqrt{n}}\\ \implies r &= \sqrt{n}\sup_{f\in\mathcal{F}}\|f\|_2\end{aligned}$$

We have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{q\in Q}\frac{1}{n}\sum_{i=1}^{n}\sigma_i q_i\right] \\
&\leqslant \frac{r\sqrt{2\log|Q|}}{n} \quad \textit{(Massart's lemma A.2)} \\
&= \sup_{f\in\mathcal{F}}\|f\|_2\frac{\sqrt{2n\log|Q|}}{n} \\
&= \sup_{f\in\mathcal{F}}\|f\|_2\sqrt{\frac{2\log|\mathcal{F}|}{n}}
\end{aligned}$$

□.

A.2.4 Dudley's Theorem

Theorem A.3: Dudley's Theorem

If $\mathcal{F}$ is a function class from $\mathcal{Z}\to\mathbb{R}$ (where $\mathcal{Z}$ is a vector space and the norm of $f\in\mathcal{F}$ is not necessarily bounded), then:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leqslant 12\int_0^{\sup_{f\in\mathcal{F}}\|f\|_2}\sqrt{\frac{\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2)}{n}}d\epsilon$$

Where for any $f\in\mathcal{F}$, given a sample $x_1,\dots,x_n$, we define the $\|.\|_2$ norm as followed:

$$\|f\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}f(x_i)^2}$$

Proof (Theorem A.3)**.**
The main idea of the proof is via chaining. Define the following sequence $\{\epsilon_j\}_{j=0}^n$:

$$\begin{cases}\epsilon_0 &= \sup_{f\in\mathcal{F}}\|f\|_2\\ \epsilon_j &= 2^{-j}\epsilon_0\end{cases}$$

Next, we define the sequence of ϵ-covers $\mathcal{N}_{\epsilon_j}$ corresponding to each ϵ_j. By definition, we have:

$$\forall f\in\mathcal{F},\exists g_j\in\mathcal{N}_{\epsilon_j}:\|f-g_j\|_2\leqslant\epsilon_j$$

We can write any $f\in\mathcal{F}$ as the following telescoping sum:

$$f = f-g_n+\sum_{j=1}^{n}\Big(g_j-g_{j-1}\Big)$$

Now, define the Empirical Rademacher Complexity as followed:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right] = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\langle\sigma,f\rangle\middle|x_1,\dots,x_n\right]$$

We have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left\langle\sigma, f\right\rangle\middle| x_1,\ldots,x_n\right] \\
&= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left\langle\sigma, f-g_n+\sum_{j=1}^{n}\left(g_j-g_{j-1}\right)\right\rangle\middle| x_1,\ldots,x_n\right] \\
&\leqslant \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left\langle\sigma, f-g_n\right\rangle+\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\left\langle\sigma, g_j-g_{j-1}\right\rangle\middle| x_1,\ldots,x_n\right] \quad (\sup\Sigma\leqslant\Sigma\sup) \\
&\leqslant \|\sigma\|_2\cdot\|f-g_n\|_2+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\left\langle\sigma, g_j-g_{j-1}\right\rangle\middle| x_1,\ldots,x_n\right] \quad (\textit{Cauchy-Schwarz}) \\
&\leqslant \epsilon_n+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\left\langle\sigma, g_j-g_{j-1}\right\rangle\middle| x_1,\ldots,x_n\right] \\
&\leqslant \epsilon_n+\sum_{j=1}^{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left\langle\sigma, g_j-g_{j-1}\right\rangle\middle| x_1,\ldots,x_n\right]
\end{aligned}$$

From here, note that:

$$\|g_j-g_{j-1}\|_2\leqslant\|g_j-f\|_2+\|g_{j-1}-f\|_2\leqslant\epsilon_j+\epsilon_{j-1}=3\epsilon_j$$

Also, define the following classes of functions:

$$\mathcal{H}_j=\left\{g_j-g_{j-1}\middle| g_j\in\mathcal{N}_{\epsilon_j}, g_{j-1}\in\mathcal{N}_{\epsilon_{j-1}}\right\}$$

Continuing the above, we have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leqslant \epsilon_n+\sum_{j=1}^{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}_j}\left\langle\sigma, h\right\rangle\middle| x_1,\ldots,x_n\right] \\
&= \epsilon_n+\sum_{j=1}^{n}\widehat{\mathfrak{R}}_S(\mathcal{H}_j) \\
&\leqslant \epsilon_n+\sum_{j=1}^{n}\sup_{h\in\mathcal{H}_j}\|h\|_2\cdot\sqrt{\frac{2\log|\mathcal{H}_j|}{n}} \quad (\textit{Massart's lemma}) \\
&\leqslant \epsilon_n+\sum_{j=1}^{n}6(\epsilon_j-\epsilon_{j-1})\cdot\sqrt{\frac{2\log|\mathcal{H}_j|}{n}} \quad \left(\sup_{h\in\mathcal{H}_j}\|h\|_2\leqslant 3\epsilon_j\leqslant 6(\epsilon_j-\epsilon_{j+1})\right)
\end{aligned}$$

Now we have:

$$|\mathcal{H}_j|\leqslant|\mathcal{N}_{\epsilon_j}|\cdot|\mathcal{N}_{\epsilon_{j-1}}|\leqslant|\mathcal{N}_{\epsilon_j}|^2$$

Hence, we have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leqslant \epsilon_n + 6\sum_{j=1}^{n}(\epsilon_j - \epsilon_{j-1}) \cdot \sqrt{\frac{2\log|\mathcal{N}_{\epsilon_j}|^2}{n}} \\
&= \epsilon_n + 12\sum_{j=1}^{n}(\epsilon_j - \epsilon_{j-1}) \cdot \sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}} \\
&\leqslant \epsilon_n + 12\sum_{j=1}^{n}\int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}}dt \\
&\leqslant \epsilon_n + 12\sum_{j=1}^{n}\int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt \\
&\leqslant \epsilon_n + 12\int_{\epsilon_{n+1}}^{\epsilon_0} \sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt
\end{aligned}$$

Now take $n \to \infty$, we notice that $\epsilon_n \to 0$ and the above inequality holds for every ϵ-cover for $0 < \epsilon < \epsilon_0 = \sup_{f\in\mathcal{F}}\|f\|_2$. Therefore,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leqslant 12\int_0^{\sup_{f\in\mathcal{F}}\|f\|_2} \sqrt{\frac{\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2)}{n}}d\epsilon$$

□.

Remark : Using theorem A.3, we can translate the covering number to the Empirical Rademacher Complexity given that the covering number has some special formulation and the function class $\mathcal{F}$ is bounded in $[-1, 1]$. For example:

- (**i**) $\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2) \approx (1/\epsilon)^R$.
  Then we have $\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2) \approx R\log(1/\epsilon)$. Therefore,
  $$\int_0^1 \sqrt{\frac{\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2)}{n}}d\epsilon = \int_0^1 \sqrt{\frac{R\log(1/\epsilon)}{n}}d\epsilon \approx \sqrt{\frac{R}{n}}$$
- (**ii**) $\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2) \approx a^{R/\epsilon}$.
  Then we have $\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2) \approx \frac{R}{\epsilon}\log a$. Therefore,
  $$\begin{aligned}
  \int_0^1 \sqrt{\frac{\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|_2)}{n}}d\epsilon &\approx \int_0^1 \sqrt{\frac{R}{n\epsilon}\log a}d\epsilon \\
  &= \sqrt{\frac{R}{n}\log a}\int_0^1 \sqrt{\frac{1}{\epsilon}}d\epsilon \\
  &= \tilde{O}\left(\sqrt{\frac{R}{n}}\right)
  \end{aligned}$$
- (**iii**) $\mathcal{N}(\mathcal{F},\epsilon,\|.\|^2) \approx a^{R/\epsilon^2}$.
  Then, we have $\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|^2) \approx \frac{R}{\epsilon^2}\log a$. Therefore,
  $$\int_0^1 \sqrt{\frac{\log\mathcal{N}(\mathcal{F},\epsilon,\|.\|^2)}{n}}d\epsilon \approx \sqrt{\frac{R}{n}\log a}\int_0^1 \frac{1}{\epsilon}d\epsilon = \infty$$

A.2.5 Bound on covering number of linear function class

Theorem A.4: Maurey's Sparsification Lemma

In a Hilbert space $\mathcal{H}$ equiped with a norm $\|.\|$, let $f \in \mathcal{H}$ where $f = \sum_{j=1}^{d} w_j g_j, \quad g_j \in \mathcal{H}$ such that $\|g_j\| \leqslant b, w_j \geqslant 0$ and $\alpha = \sum_{j=1}^{d} w_j \leqslant 1$. Then, for any $n \geqslant 1$, there exists non-negative integers $k_1, \ldots, k_d \geqslant 0$ such that $\sum_{j=1}^{d} k_j \leqslant n$ and:

$$\left\| f - \sum_{j=1}^{d} \frac{k_j}{n} g_j \right\|^2 \leqslant \frac{\alpha b^2 - \|f\|^2}{n}$$

Remark : The intuition of theorem A.4 is that if we have a function from a Hilbert space which can be written as linear combination of bounded functions. Then, we can estimate the coefficients w_j using integer solutions k_j. The higher n gets, the more accurately we can estimate the coefficients.

Proof (Theorem A.4). *The proof of this theorem is too technical for this note. However, a rigorous proof is provided in the following paper Bartlett et al. 2017 (Lemma A.6).* □.

Lemma A.1: Bound on $\mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)$ for linear functions

Let $\mathcal{F}$ be a linear function class defined as followed:

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f(x) = wx, \|w\|_q \leqslant a, \|x\|_p \leqslant b \right\}$$

Where p, q are Holder conjugates and $2 \leqslant p \leqslant \infty$. Then, we have:

$$\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2) \leqslant \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \log(2d + 1)$$

The proof of this lemma is discussed in Theorem 3 in Zhang 2002.

Proof (Lemma A.1). *For $f \in \mathcal{F}$, we can write:*

$$f(x) = \sum_{j=1}^{d} x_j \cdot w_j$$

$$\|x\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p} \leqslant b$$

$$\|w\|_q = \left(\sum_{j=1}^{d} |w_j|^q \right)^{1/q} \leqslant a$$

We define the following:

$$g_j(x) = \frac{n^{1/p} ab}{|x_j|} x_j, \quad w'_j = \frac{|x_j|}{n^{1/p} ab} w_j, \quad (n \geqslant 1)$$

By Holder's inequality, we can check that:

$$\begin{aligned}\sum_{j=1}^{d}|w'_j| &= \sum_{j=1}^{d}\frac{|x_j|}{n^{1/p}ab}|w_j| \\ &\leqslant \frac{1}{n^{1/p}ab}\left(\sum_{j=1}^{d}|x_j|^p\right)^{1/p}\left(\sum_{j=1}^{d}|w_j|^q\right)^{1/q} \\ &= \frac{\|x\|_p\|w\|_q}{n^{1/p}ab} \\ &\leqslant \frac{1}{n^{1/p}} \\ &\leqslant 1\end{aligned}$$

By the definition of g_j, we know that:

$$\forall x \in \mathbb{R}^d : g_j(x) \leqslant n^{1/p}ab \implies \|g_j\|_2 \leqslant n^{1/2}ab$$

Furthermore, for all $f \in \mathcal{F}$, we can write:

$$f(x) = \sum_{j=1}^{d} w_j x_j \implies f := \sum_{j=1}^{d}|w'_j| \cdot [sgn(w'_j)g_j]$$

Hence, if we choose $\lceil (ab/\epsilon)^2 \rceil \geqslant K \geqslant (ab/\epsilon)^2$, by theorem A.4, we can always find integers $k_1, \ldots, k_d$ such that $\sum_{j=1}^{d}|k_j| \leqslant K$ and:

$$\begin{aligned}\left\|f - \sum_{j=1}^{d}\frac{k_j}{K}g_j\right\|_2^2 &\leqslant \frac{na^2b^2\sum_{j=1}^{d}|w'_j| - \|f\|_2^2}{K} \\ &\leqslant \frac{na^2b^2 - \|f\|_2^2}{K} \\ &\leqslant \frac{na^2b^2}{K} \\ &\leqslant n\epsilon^2\end{aligned}$$

Remark *: Note that the above norm refers to the standard l^2 norm. However, when we switch to the norm used in covering number and Rademacher Complexity (with the scaling factor $\frac{1}{\sqrt{n}}$), the n on the right-hand-side vanishes.*

From the above, the covering number of $\mathcal{F}$ will not exceed the number of integer solutions for $\sum_{j=1}^{d}|k_j| \leqslant K$, which is $(2d+1)^K$. Hence, we have:

$$\begin{aligned}\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2) &\leqslant K\log(2d+1) \\ &\leqslant \left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil \log(2d+1)\end{aligned}$$

□.

A.2.6 Rademacher Complexity bound for linear functions class

In this section, we finally solve the problem stated in section A.2.1. First, consider the following lemma, then prove the proposition A.1:

Theorem A.5: Dudley's Entropy Integral bound

Let $\mathcal{F}$ be a real-valued function class and assume that $\mathbf{0} \in \mathcal{F}$, $S = \{x_1, \dots, x_n\}$ is a dataset. Then,

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leqslant \inf_{\alpha > 0} \left(4\alpha + 12 \int_{\alpha}^{\sup_{f \in \mathcal{F}} \|f\|_2} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} d\epsilon \right)$$

Where we define the $\|.\|_2$ norm as followed:

$$\|f\|_2 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} f(x_i)^2}, \ f \in \mathcal{F}$$

Proof (Theorem A.5). *This proof extend the proof for the more general Dudley's theorem A.3. From the proof, we have:*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leqslant \epsilon_n + 12 \int_{\epsilon_{n+1}}^{\epsilon_0} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt$$

Where $\epsilon_0 = \sup_{f \in \mathcal{F}} \|f\|_2$ and for every $j \geqslant 1$ we have $\epsilon_j = 2^{-j} \epsilon_0$. $\mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)$ is the size of the minimum ϵ-cover.

Let $\alpha \geqslant 0$, pick the sample size n such that:

$$n = \sup_{j \geqslant 1} \left\{ j : \epsilon_j \geqslant 2\alpha \right\}$$

Hence, for any choice of sample size of at least $n + 1$, we have:

$$\epsilon_{n+1} \leqslant 2\alpha \implies \epsilon_n = 2\epsilon_{n+1} \leqslant 4\alpha$$

We also have:

$$\epsilon_{n+1} = \frac{\epsilon_n}{2} \geqslant \frac{2\alpha}{2} = \alpha$$

Therefore, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &\leqslant \epsilon_n + 12 \int_{\epsilon_{n+1}}^{\epsilon_0} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt \\
&\leqslant 4\alpha + 12 \int_{\epsilon_{n+1}}^{\epsilon_0} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt \quad (\epsilon_n \leqslant 4\alpha) \\
&\leqslant 4\alpha + 12 \int_{\alpha}^{\epsilon_0} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt \quad (\epsilon_{n+1} \geqslant \alpha) \\
&= 4\alpha + 12 \int_{\alpha}^{\sup_{f \in \mathcal{F}} \|f\|_2} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt
\end{aligned}$$

Since the left-hand-side does not depend on α, we can take the infimum over $\alpha > 0$ to obtain the tightest bound:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leqslant \inf_{\alpha > 0} \left(4\alpha + 12 \int_{\alpha}^{\sup_{f \in \mathcal{F}} \|f\|_2} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)}{n}} dt \right)$$

□.

Proposition A.1: $\mathfrak{R}_n(\mathcal{F})$ bound for linear function class (covering number)

Given the following function class $\mathcal{F}$ whose range is bounded within $[-1, 1]$:

$$\mathcal{F} = \Big\{ f : \mathbb{R}^d \to \mathbb{R} \Big| f(x) = wx, \|w\|_2 \leqslant a, \|x\|_2 \leqslant b \Big\}$$

Then, we can obtain the following bound for the Rademacher Complexity using Covering number:

$$\mathfrak{R}_n(\mathcal{F}) \leqslant \tilde{O}\left(\frac{R}{\sqrt{n}}\right), \quad R = ab$$

Proof (Proposition A.1)**.**

From lemma A.1, we have the following bound on the covering number of $\mathcal{F}$:

$$\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2) \leqslant \left\lceil \frac{R^2}{\epsilon^2} \right\rceil \log(2d+1) < 2\frac{R^2}{\epsilon^2} \log(2d+1) = \frac{R^2}{\epsilon^2} \log(4d^2+4d+1)$$

The second inequality holds under the assumption that $R^2 > \epsilon^2$. Let $D = 4d^2 + 4d + 1$, we have:

$$\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2) < \frac{R^2}{\epsilon^2} \log D$$

Since $\mathcal{F}$ is bounded in the $[-1, 1]$ range, we have that $\sup_{f \in \mathcal{F}} \|f\|_2 \leqslant 1$. Hence, we have:

$$\begin{aligned}
\int_{\alpha}^{\sup_{f\in\mathcal{F}} \|f\|_2} \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)} d\epsilon &\leqslant \int_{\alpha}^{1} \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)} d\epsilon \\
&< R\sqrt{\log D} \int_{\alpha}^{1} \frac{1}{\epsilon} d\epsilon \\
&= R\sqrt{\log D}\Big(- \log \alpha \Big)
\end{aligned}$$

Using theorem A.5, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &\leqslant \inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{\sup_{f\in\mathcal{F}} \|f\|_2} \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)} d\epsilon \right) \\
&< \inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{n}} R\sqrt{\log D}\Big(- \log \alpha \Big) \right)
\end{aligned}$$

Letting $\alpha = \frac{3R}{\sqrt{n}}$, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &< \frac{12R}{\sqrt{n}} + \frac{12R}{\sqrt{n}} \sqrt{\log D}\left(- \log \frac{3R}{\sqrt{n}} \right) \\
&= \frac{12R}{\sqrt{n}} \left[1 - \sqrt{\log D} \log \frac{3R}{\sqrt{n}} \right] \\
&= \tilde{O}\left(\frac{R}{\sqrt{n}}\right)
\end{aligned}$$

Since the right-hand-side does not depend on the sample S, we can just take expectation over the samples for both sides and we have:

$$\mathfrak{R}_n(\mathcal{F}) < \tilde{O}\left(\frac{R}{\sqrt{n}}\right)$$

Remark *: In fact, we do not have to specify the bound range* $[-1,1]$ *of* $\mathcal{F}$ *and make the above result more general:*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) < \frac{12R}{\sqrt{n}}\left[1 + \log \sup_{f\in\mathcal{F}} \|f\|_2 - \sqrt{\log D}\log\frac{3R}{\sqrt{n}}\right] = \tilde{O}\left(\frac{R}{\sqrt{n}}\right)$$

□.

Proposition A.2: $\mathfrak{R}_n(\mathcal{F})$ bound for linear function class (Depending on d)

Given the following function class $\mathcal{F}$ whose range is bounded within $[-1,1]$:

$$\mathcal{F} = \left\{f : \mathbb{R}^d \to \mathbb{R} \middle| f(x) = wx, \|w\|_2 \leqslant a, \|x\|_2 \leqslant b\right\}$$

Then, we can obtain the following bound for the Rademacher Complexity:

$$\mathfrak{R}_n(\mathcal{F}) \leqslant \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

Proof (Proposition A.2)**.**
By lemma A.8 of Long et al. 2020, a ball in $\mathbb{R}^d$ *with radius* $\kappa > 0$*, denoted* $\mathcal{B}_\kappa$*, with respect to the* l^2 *norm has the following covering number bound:*

$$\mathcal{N}(\mathcal{B}_\kappa, \epsilon, \|.\|_2) \leqslant \left\lceil\left(\frac{3\kappa}{\epsilon}\right)^d\right\rceil \leqslant \left(1 + \frac{3\kappa}{\epsilon}\right)^d$$

Let W *be the set of vectors in* $\mathbb{R}^d$ *whose* l^2 *norm is bounded by* $a \in \mathbb{R}$*. By the above lemma, the covering number of* W *has the following bound:*

$$\mathcal{N}(W, \epsilon, \|.\|_2) \leqslant \left(1 + \frac{3a}{\epsilon}\right)^d$$

Bounding the covering number of $\mathcal{F}$
Given a sample $S = \{x_1, \ldots, x_n\}$ *in* $\mathbb{R}^d$*, we define the norm of* $f \in \mathcal{F}$ *as followed:*

$$\|f\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2}$$

For all $w \in W$*, we have a corresponding function in* $\mathcal{F}$*, denote it as* f_w*. Denote* $\mathcal{C}_\epsilon(W)$ *as the* ϵ*-cover of* W *and let* $\bar{w} \in \mathcal{C}_\epsilon(W)$ *be the member of the* ϵ*-cover that is closest to* $w \in W$*. We have:*

$$\begin{aligned}
\|f_w - f_{\bar{w}}\|_2 &= \sqrt{\frac{1}{n}\sum_{i=1}^n (f_w - f_{\bar{w}})(x_i)^2} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^n \langle w - \bar{w}, x_i\rangle^2} \\
&\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^n \|w - \bar{w}\|_2^2 \cdot \|x_i\|_2^2} \quad \textit{(Cauchy-Schwarz)} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^n \epsilon^2 b^2} = \epsilon b
\end{aligned}$$

From the above, we have:

$$\forall f_w \in \mathcal{F} : \exists \bar{w} \in \mathcal{C}_{\epsilon/b}(W) \text{ such that } \|f_w - f_{\bar{w}}\|_2 \leqslant \epsilon$$

Therefore,

$$\mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2) \leqslant \left(1 + \frac{3ab}{\epsilon}\right)^d$$

By Dudley's Entropy Integral bound A.5, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &\leqslant \inf_{\alpha>0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{\sup_{f\in\mathcal{F}} \|f\|_2} \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|.\|_2)} d\epsilon \right\} \\
&\leqslant \inf_{\alpha>0} \left\{ 4\alpha + 12\sqrt{\frac{d}{n}} \int_\alpha^1 \sqrt{\log\left(1 + \frac{3ab}{\epsilon}\right)} d\epsilon \right\} \quad \left(\sup_{f\in\mathcal{F}} \|f\|_2 \leqslant 1\right) \\
&\leqslant \inf_{\alpha>0} \left\{ 4\alpha + 12\sqrt{\frac{d}{n}}(1-\alpha) \sqrt{\log\left(1 + \frac{3ab}{\alpha}\right)} \right\}
\end{aligned}$$

Setting a small value for $\alpha > 0$, we have:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leqslant \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

□.

Proposition A.3: $\mathfrak{R}_n(\mathcal{F})$ bound for linear function class (no covering number)

The following proposition presents an upper bound for the Rademacher Complexity without dependence on the dimensionality d. Let $\mathcal{F}$ be a function class (not necessarily linear) and Q is its output space such that:

$$R = \sup_{q\in Q} \|q\|_2$$

Then, we have the following bound on the Rademacher Complexity:

$$R_n(\mathcal{F}) \leqslant \frac{R}{\sqrt{n}}$$

Proof (Proposition A.3)**.**

Let S be samples with n elements, we have:

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{q\in Q}\langle\sigma, q\rangle\right] \\
&\leqslant \frac{1}{n}\mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{q\in Q}\|\sigma\|_2\cdot\|q\|_2\right] \quad (\textit{Cauchy-Schwarz}) \\
&= \frac{1}{n}\mathbb{E}_S\mathbb{E}_\sigma\left[\|\sigma\|_2\cdot\sup_{q\in Q}\|q\|_2\right] \\
&= \frac{R}{n}\mathbb{E}_S\mathbb{E}_\sigma\Big[\|\sigma\|_2\Big] \\
&= \frac{R}{n}\mathbb{E}_\sigma\Big[\|\sigma\|_2\Big] \\
&= \frac{R}{n}\mathbb{E}_\sigma\left[\sqrt{\sum_{i=1}^{n}\sigma_i^2}\right] \\
&\leqslant \frac{R}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}_\sigma[\sigma_i^2]} \quad (\textit{Jenen's Inequality}) \\
&= \frac{R}{\sqrt{n}}
\end{aligned}
$$

□.

A.3 Rademacher Complexity of the ramp loss

A.3.1 Problem statement

Problem : Consider the multi-class classification problem with K labels ($K \geqslant 2$). Given the following function class

$$W = \left\{ \mathbf{w} \in \mathbb{R}^{K \times d} \middle| \|\mathbf{w}\|_F \leqslant R \right\}$$
$$\mathcal{F} = \left\{ f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}^K \middle| f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}\mathbf{x}; \mathbf{w} \in W, \|x\|_2 \leqslant 1 \right\}$$

and consider the following loss function:

$$l_r(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} \geqslant r \\ 1 - \mathbf{z}/r & \text{if } \mathbf{z} \in (0, r) \\ 1 & \text{if } \mathbf{z} \leqslant 0 \end{cases}$$

Derive the bound for the Rademacher complexity of the loss function class $\mathfrak{R}_n(\mathcal{L}_r)$ where:

$$\mathcal{L}_r = \left\{ (x, y) \mapsto l_r\Big(f_{\mathbf{w}}(x)_y - \max_{k \neq y} f_{\mathbf{w}}(x)_k\Big) \middle| f_{\mathbf{w}} \in \mathcal{F} \right\}$$

A.3.2 Approach 1 : Using covering number

Overview : Let $L_{\mathbf{w}} \in \mathcal{L}_r$ where

$$L_{\mathbf{w}}(x, y) = l_r\Big(f_{\mathbf{w}}(x)_y - \max_{k \neq y} f_{\mathbf{w}}(x)_k\Big)$$

and $S = \Big\{(x_i, y_i)\Big\}_{i=1}^n$ be a sample dataset. Define the following norm:

$$\|L_{\mathbf{w}}\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^n L_{\mathbf{w}}(x_i, y_i)^2}$$

We have to find an ϵ-covering with respect to $\|.\|_2$ norm, denoted as $\mathcal{C}_\epsilon(\mathcal{L}_{r|_S}, \|.\|_2)$. Meaning, $\forall L_{\mathbf{w}} \in \mathcal{L}_r$, $\exists L_{\bar{\mathbf{w}}} \in \mathcal{C}_\epsilon(\mathcal{L}_{r|_S}, \|.\|_2)$ such that:

$$\|L_{\mathbf{w}} - L_{\bar{\mathbf{w}}}\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^n \Big(L_{\mathbf{w}}(x_i, y_i) - L_{\bar{\mathbf{w}}}(x_i, y_i)\Big)^2} \leqslant \epsilon$$

1. Bounding the covering number of $\mathcal{L}_r$: Let $\mathbf{w}, \bar{\mathbf{w}} \in W$. For any pair $(x_i, y_i) \in S$, we have:

$$\begin{aligned}
\Big|L_{\mathbf{w}}(x_i, y_i) - L_{\bar{\mathbf{w}}}(x_i, y_i)\Big| &= \Big|l_r\Big(f_{\mathbf{w}}(x_i)_{y_i} - \max_{k \neq y_i} f_{\mathbf{w}}(x_i)_k\Big) - l_r\Big(f_{\bar{\mathbf{w}}}(x_i)_{y_i} - \max_{k \neq y_i} f_{\bar{\mathbf{w}}}(x_i)_k\Big)\Big| \\
&\leqslant \frac{1}{r}\Big|\Big(f_{\mathbf{w}}(x_i)_{y_i} - f_{\bar{\mathbf{w}}}(x_i)_{y_i}\Big) + \Big(\max_{k \neq y_i} f_{\bar{\mathbf{w}}}(x_i)_k - \max_{k \neq y_i} f_{\mathbf{w}}(x_i)_k\Big)\Big| \quad (l_r \text{ is } 1/r - \text{Lipchitz}) \\
&\leqslant \frac{1}{r}\Big|\Big(f_{\mathbf{w}}(x_i)_{y_i} - f_{\bar{\mathbf{w}}}(x_i)_{y_i}\Big) + \Big(\max_{k \neq y_i}\Big\{f_{\bar{\mathbf{w}}}(x_i)_k - f_{\mathbf{w}}(x_i)_k\Big\}\Big)\Big| \\
&\leqslant \frac{1}{r}\Big|f_{\mathbf{w}}(x_i)_{y_i} - f_{\bar{\mathbf{w}}}(x_i)_{y_i}\Big| + \frac{1}{r}\max_{k \neq y_i}\Big|f_{\bar{\mathbf{w}}}(x_i)_k - f_{\mathbf{w}}(x_i)_k\Big| \\
&\leqslant \frac{2}{r}\sup_{\substack{x_i \in S \\ j \in \{1,\ldots,K\}}}\Big|f^{(j)}_{\mathbf{w}}(x_i) - f^{(j)}_{\bar{\mathbf{w}}}(x_i)\Big| \\
&= \frac{2}{r}\sup_{j \in \{1,\ldots,K\}}\|f^{(j)}_{\mathbf{w}} - f^{(j)}_{\bar{\mathbf{w}}}\|_\infty \\
&= \frac{2}{r}\max\Big\{\Big|(\mathbf{w}_j - \bar{\mathbf{w}}_j)x_i\Big|\Big\}_{\substack{i \in \{1,\ldots,n\}, \\ j \in \{1,\ldots,K\}}} \quad (1)
\end{aligned}$$

Now, define the following class of functions:

$$\begin{aligned}
\mathcal{H} &= \Big\{h : \mathbb{R}^{Kd} \to \mathbb{R} : h(\mathbf{x}) = \beta\mathbf{x}, \beta \in \mathcal{B}\Big\} \\
\mathcal{B} &= \Big\{\beta \in \mathbb{R}^{1 \times Kd} : \beta = \begin{pmatrix}\mathbf{w}_1 & \ldots & \mathbf{w}_K\end{pmatrix},\ \mathbf{w} \in W\Big\}
\end{aligned}$$

Basically, we construct the set of parameters $\beta \in \mathbb{R}^{1 \times Kd}$ by concatenating rows of vectors $\mathbf{w} \in W$ horizontally. By the definition of W, we have that $\|\beta\|_2 \leqslant R,\ \forall \beta \in \mathcal{B}$.

For any sample $S = \{x_i\}_{i=1}^n$ of vectors $x_i \in \mathbb{R}^d$, construct a new sample S' of $\mathbb{R}^{Kd}$ vectors by creating K vectors from each $x_i \in S$:

$$x_i \longrightarrow \begin{cases}
x_{i,1} &= \begin{pmatrix} x_i & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}\end{pmatrix}^T \\
x_{i,2} &= \begin{pmatrix} \mathbf{0} & x_i & \mathbf{0} & \ldots & \mathbf{0}\end{pmatrix}^T \\
&\vdots \\
x_{i,K} &= \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & x_i\end{pmatrix}^T
\end{cases}$$

For all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, K\}$, we have $\|x_{i,j}\|_2 \leqslant 1$. Hence, applying Zhang's theorem 4 Zhang 2002 on $\mathcal{H}_{|S'}$, we have:

$$\log_2 \mathcal{N}\Big(\mathcal{H}_{|S'}, \epsilon, \|.\|_\infty\Big) \leqslant \frac{36R^2}{\epsilon^2}\log_2\left(\frac{8RnK}{\epsilon} + 6nK + 1\right) \leqslant \frac{36R^2}{\epsilon^2}\log_2\left(\left(\frac{8R}{\epsilon} + 7\right)nK\right)$$

In other words, if we denote $\mathcal{C}_\epsilon(\mathcal{H}_{|S'}, \|.\|_\infty)$ as the minimum ϵ-covering for $\mathcal{H}_{|S'}$ with respect to the $\|.\|_\infty$ norm, we have:

$$\Big|\mathcal{C}_\epsilon(\mathcal{H}_{|S'}, \|.\|_\infty)\Big| \leqslant \frac{36R^2}{\epsilon^2}\log_2\left(\left(\frac{8R}{\epsilon} + 7\right)nK\right)$$

And for any $h \in \mathcal{H}$ parameterized by $\beta = \begin{pmatrix}\mathbf{w_1} & \ldots & \mathbf{w_K}\end{pmatrix}$, there exists $\bar{h} \in \mathcal{C}_\epsilon(\mathcal{H}_{|S'}, \|.\|_\infty)$,

parameterized by $\bar{\beta} = \begin{pmatrix} \bar{\mathbf{w}}_1 & \dots & \bar{\mathbf{w}}_K \end{pmatrix}$, such that $\|h - \bar{h}\|_\infty \leqslant \epsilon$. Hence, we have:

$$\begin{aligned} \|h - \bar{h}\|_\infty &= \sup_{x_{i,j} \in S'} \Big| h(x_{i,j}) - \bar{h}(x_{i,j}) \Big| \\ &= \sup_{x_{i,j} \in S'} \Big| (\beta - \bar{\beta}) x_{i,j} \Big| \\ &= \max_{\substack{x_i \in S \\ j \in \{1,\dots,K\}}} \Big| (\mathbf{w}_j - \bar{\mathbf{w}}_j) x_i \Big| < \epsilon \quad (2) \end{aligned}$$

From (1) and (2), we have:

$$\log_2 \mathcal{N}\Big(\mathcal{L}_{r|S}, \epsilon, \|.\|_2\Big) \leqslant \log_2 \mathcal{N}\Big(\mathcal{H}_{|S'}, r\epsilon/2, \|.\|_\infty\Big) \leqslant \frac{144R^2}{r^2\epsilon^2} \log_2 \left(\left(\frac{16R}{r\epsilon} + 7 \right) nK \right)$$

2. Dudley's Entropy Integral : Using theorem A.5, we have:

$$\hat{\mathfrak{R}}_S(\mathcal{L}_r) \leqslant 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}\Big(\mathcal{L}_r, \epsilon, \|.\|_2\Big)} d\epsilon, \ \alpha > 0$$

Note that we made use of the fact that $\sup_{\mathbf{z} \in \mathbb{R}} l_r(\mathbf{z}) = 1$. Hence, $\sup_{L_\mathbf{w} \in \mathcal{L}_r} \|L_\mathbf{w}\|_2 = 1$. Hence, the upper limit of the integral is 1. From the above covering number bound, we have:

$$\begin{aligned} \hat{\mathfrak{R}}_S(\mathcal{L}_r) &\leqslant 4\alpha + \frac{144R}{r\sqrt{n}} \int_\alpha^1 \frac{1}{\epsilon} \sqrt{\log \left(\left(\frac{16R}{r\epsilon} + 7 \right) nK \right)} d\epsilon \\ &< 4\alpha + \frac{144R}{r\sqrt{n}} \cdot \sqrt{\log \left(\left(\frac{16R}{r\alpha} + 7 \right) nK \right)} \int_\alpha^1 \frac{1}{\epsilon} d\epsilon \\ &= 4\alpha + \frac{144R}{r\sqrt{n}} \cdot \sqrt{\log \left(\left(\frac{16R}{r\alpha} + 7 \right) nK \right)} \Big(- \log \alpha \Big) \end{aligned}$$

Setting $\alpha = \frac{36R}{r\sqrt{n}}$, we have the following upperbound:

$$\hat{\mathfrak{R}}_S(\mathcal{L}_r) \leqslant \frac{144R}{r\sqrt{n}} \left[1 + \sqrt{\log \left(\left(\frac{4}{9}\sqrt{n} + 7 \right) nK \right)} \log \left(\frac{r\sqrt{n}}{36R} \right) \right]$$

Therefore, we have:

$$\mathfrak{R}_n(\mathcal{L}_r) \leqslant \tilde{O}\left(\frac{R}{r\sqrt{n}} \right)$$

A.3.3 Approach 2 : Using contraction inequality

1. Find Rademacher complexity of each output component : First, we bound the covering number of each component of the output then work out the Rademacher complexity.
For any $f_\mathbf{w} \in \mathcal{F}$, we have:

$$f_\mathbf{w} = \begin{pmatrix} f_\mathbf{w}^{(1)}(x) \\ \vdots \\ f_\mathbf{w}^{(K)}(x) \end{pmatrix} = \begin{pmatrix} \mathbf{w}_1 x \\ \vdots \\ \mathbf{w}_K x \end{pmatrix}$$

Where $\mathbf{w}_j$ is the j^{th} row of $\mathbf{w}$ for $j = \{1, \ldots, K\}$. Define the following classes:

$$\mathcal{F}_j = \left\{ f_{\mathbf{w}}^{(j)} : f_{\mathbf{w}} \in \mathcal{F} \right\} = \left\{ \mathbf{x} \mapsto \mathbf{w}_j \mathbf{x} : \mathbf{w} \in W \right\}, \quad j \in \{1, \ldots, K\}$$

Since for any $\mathbf{w} \in W$, we have $\|\mathbf{w}\|_F \leqslant R$. Therefore, for all $j \in \{1, \ldots, K\}$, we have $\|\mathbf{w}_j\|_2 \leqslant R$. Hence, by theorem 3 in Zhang 2002, we have:

$$\log_2 \mathcal{N}\left(\mathcal{F}_j, \epsilon, \|.\|_2\right) \leqslant \left\lceil \frac{R^2}{\epsilon^2} \right\rceil \log_2(2d+1) \leqslant 2\frac{R^2}{\epsilon^2} \log_2(2d+1)$$

Assuming that $R \geqslant \epsilon$. Hence, by Dudley's A.5, we have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}_j) &\leqslant 4\alpha + \frac{12R}{\sqrt{n}} \sqrt{\log D} \int_{\alpha}^{R} \frac{1}{\epsilon} d\epsilon, \quad D = (2d+1)^2 \\
&= 4\alpha + \frac{12R}{\sqrt{n}} \sqrt{\log D} \log \frac{R}{\alpha} \\
&= \frac{12R}{\sqrt{n}} \left(1 + \sqrt{\log D} \log \frac{\sqrt{n}}{3} \right), \quad \text{Setting } \alpha = \frac{3R}{\sqrt{n}} \\
\implies \mathfrak{R}_n(\mathcal{F}_j) &\leqslant \frac{12R}{\sqrt{n}} \left(1 + \sqrt{\log D} \log \frac{\sqrt{n}}{3} \right) = \tilde{O}\left(\frac{R}{\sqrt{n}} \right)
\end{aligned}$$

2. Using l_∞ contraction inequality : By theorem 1 in Foster et al. 2019, if we have $\phi_1, \ldots, \phi_n$, where $\phi_i : \mathbb{R}^K \to \mathbb{R}$, being L-Lipchitz with respect to the l_∞ norm, meaning $|\phi_j(x) - \phi_j(y)| \leqslant L \cdot \|x - y\|_\infty$. We have:

$$\frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f_{\mathbf{w}} \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \phi_i(f_{\mathbf{w}}(x_i)) \right] \leqslant \tilde{O}\left(L\sqrt{K} \right) \cdot \max_{j \in \{1, \ldots, K\}} \mathfrak{R}_n(\mathcal{F}_j)$$

For any $z \in \mathbb{R}^K$, define the function $\psi_j(z) = z_j - \max_{k \neq j} z_k$ and for any $(x_i, y_i) \in S$, let $\phi_i(f_{\mathbf{w}}(x_i)) = (l_r \circ \psi_{y_i})(f_{\mathbf{w}}(x_i))$. We have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{L}_r) &= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f_{\mathbf{w}} \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i l_r \left(f_{\mathbf{w}}(x_i)_{y_i} - \max_{k \neq y_i} f_{\mathbf{w}}(x_i)_k \right) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f_{\mathbf{w}} \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \phi_i(f_{\mathbf{w}}(x_i)) \right]
\end{aligned}$$

Following similar arguments as (1), we know that $l_r \circ \psi_j, \quad j \in \{1, \ldots, K\}$ is $2/r$-Lipchitz continuous with respect to the l_∞ norm. Hence, we have:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{L}_r) &\leqslant \tilde{O}\left(\frac{2\sqrt{K}}{r} \right) \cdot \max_{j \in \{1, \ldots, K\}} \mathfrak{R}_n(\mathcal{F}_j) \\
&= \tilde{O}\left(\frac{R\sqrt{K}}{r\sqrt{n}} \right)
\end{aligned}$$

A.3.4 Approach 3 : Stacking covering numbers

1. Stacking covering numbers : By Zhang's theorem 4 Zhang 2002, we have:

$$\log_2 \mathcal{N}\left(\mathcal{F}_{j|S}, \epsilon, \|.\|_\infty\right) \leqslant \frac{36R^2}{\epsilon^2} \log_2 \left(\frac{8Rn}{\epsilon} + 6n + 1 \right), \quad j \in \{1, \ldots, K\}$$

Denote $C_\epsilon^{(j)}$ as the minimum (internal) ϵ-cover of $\mathcal{F}_j = \Big\{\mathbf{x} \mapsto \mathbf{w}_j\mathbf{x} : \mathbf{w} \in W\Big\}$ with respect to the $\|.\|_\infty$ norm. Meaning, for any $f_{\mathbf{w}} \in \mathcal{F}$:

$$\exists f_{\bar{\mathbf{w}}}^{(j)} \in C_\epsilon^{(j)} : \|f_{\mathbf{w}}^{(j)} - f_{\bar{\mathbf{w}}}^{(j)}\|_\infty = \sup_{i\in\{1,\dots,n\}} \Big|f_{\mathbf{w}}^{(j)}(x_i) - f_{\bar{\mathbf{w}}}^{(j)}(x_i)\Big| < \epsilon, \quad \bar{\mathbf{w}} \in W$$

Therefore, for any $f_{\mathbf{w}} \in \mathcal{F}$:

$$\exists f_{\bar{\mathbf{w}}} \in C_\epsilon^{(1)} \times \cdots \times C_\epsilon^{(K)} : \max_{j\in\{1,\dots,K\}} \|f_{\mathbf{w}}^{(j)} - f_{\bar{\mathbf{w}}}^{(j)}\|_\infty = \max_{\substack{j\in\{1,\dots,K\}\\ i\in\{1,\dots,n\}}} \Big|f_{\mathbf{w}}^{(j)}(x_i) - f_{\bar{\mathbf{w}}}^{(j)}(x_i)\Big| < \epsilon \quad (3)$$

However, note that even though $f_{\bar{\mathbf{w}}} \in C_\epsilon^{(1)} \times \cdots \times C_\epsilon^{(K)}$, it does not necessarily mean $f_{\bar{\mathbf{w}}} \in \mathcal{F}$. Because in the worst case senario, we have:

$$\|\bar{\mathbf{w}}\|_F = \sqrt{\sum_{j=1}^K \|\bar{\mathbf{w}}_j\|_2^2} \leqslant \sqrt{\sum_{j=1}^K R^2} = R\sqrt{K}$$

Hence, $C_\epsilon^{(1)} \times \cdots \times C_\epsilon^{(K)}$ is an external ϵ-cover. From (1), (3) and lemma A.2, we have:

$$\begin{aligned}
\mathcal{N}\Big(\mathcal{L}_{r|S}, \epsilon, \|.\|_2\Big) &\leqslant \mathcal{N}^{ext}\Big(\mathcal{L}_{r|S}, \epsilon/2, \|.\|_2\Big) \\
&\leqslant \Big|C_{r\epsilon/4}^{(1)} \times \cdots \times C_{r\epsilon/4}^{(K)}\Big| \\
&\leqslant \prod_{j=1}^K \Big|C_{r\epsilon/4}^{(j)}\Big| \\
\implies \log\mathcal{N}\Big(\mathcal{L}_{r|S}, \epsilon, \|.\|_2\Big) &\leqslant \sum_{j=1}^K \log\Big|C_{r\epsilon/4}^{(j)}\Big| \\
&\leqslant \frac{576R^2K}{r^2\epsilon^2}\log\left(\frac{32Rn}{r\epsilon} + 6n + 1\right) \\
&\leqslant \frac{576R^2K}{r^2\epsilon^2}\log\left(\left(\frac{32R}{r\epsilon} + 7\right)n\right)
\end{aligned}$$

2. Dudley's Entropy Integral : By Dudley's A.5, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{L}_r) &\leqslant 4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log\mathcal{N}\Big(\mathcal{L}_{r|S}, \epsilon, \|.\|_2\Big)}d\epsilon, \quad \alpha > 0 \\
&= 4\alpha + \frac{288R\sqrt{K}}{r\sqrt{n}}\int_\alpha^1 \frac{1}{\epsilon}\sqrt{\log\left(\left(\frac{32R}{r\epsilon} + 7\right)n\right)}d\epsilon \\
&< 4\alpha + \frac{288R\sqrt{K}}{r\sqrt{n}} \cdot \sqrt{\log\left(\left(\frac{32R}{r\alpha} + 7\right)n\right)}\int_\alpha^1 \frac{1}{\epsilon}d\epsilon \\
&= 4\alpha + \frac{288R\sqrt{K}}{r\sqrt{n}} \cdot \sqrt{\log\left(\left(\frac{32R}{r\alpha} + 7\right)n\right)}\Big(-\log\alpha\Big)
\end{aligned}$$

Setting $\alpha = \frac{72R\sqrt{K}}{r\sqrt{n}}$, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{L}_r) &< \frac{288R\sqrt{K}}{r\sqrt{n}}\left[1 + \sqrt{\log\left(\left(\frac{4\sqrt{n}}{9\sqrt{K}} + 7\right)n\right)}\log\left(\frac{r\sqrt{n}}{72R\sqrt{K}}\right)\right] \\
&= \tilde{O}\left(\frac{R\sqrt{K}}{r\sqrt{n}}\right)
\end{aligned}$$

A.4 Important lemmas and theorems for A.3

A.4.1 l_∞ Contraction Inequality

Theorem A.6: l_∞ Contraction Inequality

Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$ and $\phi_1, \dots, \phi_n$ be L-Lipchitz with respect to the l_∞ norm. We have:

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \phi_i\Big(f(x_i)\Big)\right] \leqslant \tilde{O}\Big(L\sqrt{K}\Big) \cdot \max_{j\in\{1,\dots,K\}} \mathfrak{R}_n(\mathcal{F}_j)$$

Proof (Theorem A.6). *The proof included in theorem 1 of Foster et al. 2019 made use of the fat-shattering coefficients, which is not introduced in this note. Hence, it will not be rewritten here.* □.

A.4.2 External-internal ϵ-covers

Lemma A.2: External-internal covering numbers

Let $(X, \|.\|)$ be a normed space and consider a subspace $V \subset X$. Let $\epsilon > 0$ and denote $\mathcal{N}(V, \epsilon, \|.\|)$ as the (internal) covering number of V, $\mathcal{N}^{ext}(V, \epsilon, \|.\|)$ as the external covering number of V. We have:

$$\mathcal{N}\Big(V, \epsilon, \|.\|\Big) \leqslant \mathcal{N}^{ext}\Big(V, \epsilon/2, \|.\|\Big)$$

Proof (Lemma A.2). *Let $\mathcal{C}^{ext}_{\epsilon/2}\Big(V, \|.\|\Big) = \Big\{v_1^{ext}, \dots, v_{N_0}^{ext}\Big\}$ be the minimal $\epsilon/2$-cover of V with respect to the norm $\|.\|$ where $N_0 = \Big|\mathcal{C}^{ext}_{\epsilon/2}\Big(V, \|.\|\Big)\Big| = \mathcal{N}^{ext}\Big(V, \epsilon/2, \|.\|\Big)$. We have:*

$$V \subseteq \bigcup_{i=1}^{N_0} \mathcal{B}_{\epsilon/2}(v_i^{ext})$$

Where for $\epsilon > 0$, $\mathcal{B}_\epsilon(x)$ is the ϵ-ball centered around x. For every $v_i^{ext} \in \mathcal{C}^{ext}_{\epsilon/2}\Big(V, \|.\|\Big)$, we have:

$$V \cap \mathcal{B}_{\epsilon/2}(v_i^{ext}) \neq \varnothing$$

Otherwise, v_i^{ext} is redundant which contradicts the fact that $\mathcal{C}^{ext}_{\epsilon/2}\Big(V, \|.\|\Big)$ is a minimum external $\epsilon/2$-cover of V. Hence, for all v_i^{ext}, we have:

$$\exists v_i^{in} \in V \cap \mathcal{B}_{\epsilon/2}(v_i^{ext}) : \mathcal{B}_{\epsilon/2}(v_i^{ext}) \subset \mathcal{B}_\epsilon(v_i^{in})$$

Therefore, we have:

$$V \subseteq \bigcup_{i=1}^{N_0} \mathcal{B}_{\epsilon/2}(v_i^{ext}) \subseteq \bigcup_{i=1}^{N_0} \mathcal{B}_\epsilon(v_i^{in})$$

Notice that from the above, it is possible to cover V with fewer than N_0 ϵ-balls $\mathcal{B}_\epsilon(v_i^{in})$. Hence, we have:

$$\mathcal{N}\Big(V, \epsilon, \|.\|\Big) \leqslant N_0 = \mathcal{N}^{ext}\Big(V, \epsilon/2, \|.\|\Big)$$

□.

A.5 Rademacher Complexity of ramp loss - two layers case

A.5.1 Problem Statement

Problem : Consider the multi-class classification problem with $K \geqslant 2$ labels. Given a tuple of matrices $(M^{(1)}, M^{(2)})$ (that represents initialized weight matrices). Define the following class of tuples of weight matrices:

$$\mathcal{A} = \left\{ \left(A^{(1)}, A^{(2)}\right) : \|A^{(i)}\|_\sigma \leqslant s_i,\ \|(A^{(i)} - M^{(i)})^T\|_{2,1} \leqslant a_i,\ \|A^{(2)} - M^{(2)}\|_F \leqslant a_* \right\}$$

Where $a_i, s_i, R > 0$, $x \in \mathbb{R}^{d_0}, A^{(1)} \in \mathbb{R}^{d_1 \times d_0}, A^{(2)} \in \mathbb{R}^{d_2 \times d_1}$ $(d_2 = K)$. $\|.\|_\sigma$ denotes the spectral norm and $\|.\|_{p,q}$ denotes the matrix (p,q) norm defined as:

$$\|A\|_{p,q} = \left(\sum_j \left(\sum_i |A_{ij}|^p \right)^{q/p} \right)^{1/q}$$

Define the following class of two-layer neural networks:

$$\mathcal{F}_\mathcal{A} = \left\{ x \mapsto A^{(2)}\sigma(A^{(1)}x) : \left(A^{(1)}, A^{(2)}\right) \in \mathcal{A},\ \sigma \text{ is } \rho\text{-Lipchitz w.r.t } l^2\text{-norm},\ \|x\|_2 \leqslant 1 \right\}$$

Derive the Rademacher Complexity bound for the class of loss functions:

$$\mathcal{L}_r = \left\{ (x,y) \mapsto l_r\Big(F_\mathbf{A}(x)_y - \max_{k \neq y} F_\mathbf{A}(x)_k\Big) \,\middle|\, F_\mathbf{A} \in \mathcal{F}_\mathcal{A} \right\}$$

Where l_r is the ramp loss with margin $r \in (0,1)$.

From this point on, we will refer to $\|.\|_p$ as the l^p norm and the $\|.\|_p^S$ as the norm of a function f defined over a sample S. Specifically:

$$\|f\|_p^S = \left(\frac{1}{|S|} \sum_{x_i \in S} f(x_i)^p \right)^{1/p}$$

We also define the ∞-norm over the sample of a function f as followed:

$$\|f\|_\infty^S = \sup_{x_i \in S} |f(x_i)|$$

A.5.2 Solution

1. Bounding the covering number for $\mathcal{L}_r$: Let $L_\mathbf{A} \in \mathcal{L}_r$ parameterized by $\mathbf{A} \in \mathcal{A}$ be defined as:

$$L_\mathbf{A}(x,y) = l_r\Big(F_\mathbf{A}(x)_y - \max_{k \neq y} F_\mathbf{A}(x)_k\Big)$$

Then let $S = \Big\{(x_i, y_i)\Big\}_{i=1}^{n}$ be a sample dataset, for $L_{\mathbf{A}}, L_{\bar{\mathbf{A}}} \in \mathcal{L}_r$ and $(x_i, y_i) \in S$, we have:

$$
\begin{aligned}
\Big|L_{\mathbf{A}}(x_i, y_i) - L_{\bar{\mathbf{A}}}(x_i, y_i)\Big| &= \Big|l_r\Big(F_{\mathbf{A}}(x_i)_y - \max_{k\neq y} F_{\mathbf{A}}(x_i)_k\Big) - l_r\Big(F_{\bar{\mathbf{A}}}(x_i)_y - \max_{k\neq y} F_{\bar{\mathbf{A}}}(x_i)_k\Big)\Big| \\
&\leqslant \frac{1}{r}\Big|\Big(F_{\mathbf{A}}(x_i)_y - F_{\bar{\mathbf{A}}}(x_i)_y\Big) + \Big(\max_{k\neq y} F_{\bar{\mathbf{A}}}(x_i)_k - \max_{k\neq y} F_{\mathbf{A}}(x_i)_k\Big)\Big| \quad (l_r \text{ is } 1/r\text{-Lipchitz}) \\
&\leqslant \frac{1}{r}\Big|F_{\mathbf{A}}(x_i)_y - F_{\bar{\mathbf{A}}}(x_i)_y\Big| + \frac{1}{r}\max_{k\neq y}\Big|F_{\bar{\mathbf{A}}}(x_i)_k - F_{\mathbf{A}}(x_i)_k\Big| \\
&\leqslant \frac{2}{r}\max_{j\in\{1,\dots,K\}}\Big|F_{\mathbf{A}}(x_i)_j - F_{\bar{\mathbf{A}}}(x_i)_j\Big| \\
&= \frac{2}{r}\max_{j\in\{1,\dots,K\}}\Big|A_j^{(2)}\sigma\Big(A^{(1)}x_i\Big) - \bar{A}_j^{(2)}\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big| \\
&= \frac{2}{r}\max_{j\in\{1,\dots,K\}}\Big|A_j^{(2)}\Big(\sigma\Big(A^{(1)}x_i\Big) - \sigma\Big(\bar{A}^{(1)}x_i\Big)\Big) + \Big(A_j^{(2)} - \bar{A}_j^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big| \\
&\leqslant \frac{2}{r}\max_{j\in\{1,\dots,K\}}\Big|A_j^{(2)}\Big(\sigma\Big(A^{(1)}x_i\Big) - \sigma\Big(\bar{A}^{(1)}x_i\Big)\Big)\Big| + \frac{2}{r}\max_{j\in\{1,\dots,K\}}\Big|\Big(A_j^{(2)} - \bar{A}_j^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big| \\
&= \frac{2}{r}\Big\|A^{(2)}\Big(\sigma\Big(A^{(1)}x_i\Big) - \sigma\Big(\bar{A}^{(1)}x_i\Big)\Big)\Big\|_\infty + \frac{2}{r}\Big\|\Big(A^{(2)} - \bar{A}^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big\|_\infty \\
&\leqslant \frac{2}{r}\|A^{(2)}\|_{2,\infty}\Big\|\sigma\Big(A^{(1)}x_i\Big) - \sigma\Big(\bar{A}^{(1)}x_i\Big)\Big\|_2 + \frac{2}{r}\Big\|\Big(A^{(2)} - \bar{A}^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big\|_\infty \quad (*) \\
&\leqslant \frac{2a_*}{r}\Big\|\sigma\Big(A^{(1)}x_i - \bar{A}^{(1)}x_i\Big)\Big\|_2 + \frac{2}{r}\Big\|\Big(A^{(2)} - \bar{A}^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_i\Big)\Big\|_\infty \\
&\leqslant \frac{2a_*\rho}{r}\Big\|A^{(1)}x_i - \bar{A}^{(1)}x_i\Big\|_2 + \frac{2}{r}\max_{x_*\in S}\Big\|\Big(A^{(2)} - \bar{A}^{(2)}\Big)\sigma\Big(\bar{A}^{(1)}x_*\Big)\Big\|_\infty
\end{aligned}
$$

The $\|.\|_{2,\infty}$ norm used in $(*)$ does not refer to the (p,q)-norm as defined above. Given a matrix $A \in \mathbb{R}^{m\times n}$, the $\|.\|_{2,\infty}$ norm is defined as $\|A\|_{2,\infty} = \max_{1\leqslant i\leqslant m}\|A_i\|_2$, which is the maximum l^2 norm of row vectors. Hence, $\|A^{(2)}\|_{2,\infty} \leqslant \|A^{(2)}\|_F \leqslant a_*$.

Use the notation $\mathcal{C}_\epsilon(X, \|.\|)$ to denote a minimal ϵ-cover of a set X with respect to the norm $\|.\|$. We will proceed to cover $\mathcal{L}_r$ with the following strategy:

- Derive the covering number with granularity $\epsilon_1 > 0$ with respect to the $\|.\|_2^S$ norm over the original sample dataset S for the following class:

  $$\mathcal{G}_1 = \Big\{x \mapsto A^{(1)}x : \|A^{(1)}\|_\sigma \leqslant s_1, \|(A^{(1)} - M^{(1)})^T\|_{2,1} \leqslant a_1\Big\}$$

  Denote the desired cover as $\mathcal{C}_1 = \mathcal{C}_{\epsilon_1}(\mathcal{G}_1, \|.\|_2^S)$. We can easily derive the cardinality of $\mathcal{C}_1$ using lemma 3.2 in Bartlett et al. 2017.

- With granularity $\epsilon_2 > 0$, define the cover $\mathcal{C}_2$ as followed:

  $$\mathcal{C}_2 = \bigcup_{\bar{A}^{(1)}\in\mathcal{C}_1} \mathcal{C}_{\epsilon_2}\Big(\mathcal{G}_2, \|.\|_*^{Z(\bar{A}^{(1)})}\Big)$$

  $$\text{Where } \begin{cases} \mathcal{G}_2 &= \Big\{z \mapsto A^{(2)}z : \|A^{(2)}\|_\sigma \leqslant s_2,\ \|A^{(2)} - M^{(2)}\|_F \leqslant a_*\Big\} \\ Z(\bar{A}^{(1)}) &= \Big\{\sigma\Big(\bar{A}^{(1)}x_i\Big) : x_i \in S\Big\}, \text{ for every } \bar{A}^{(1)} \in \mathcal{C}_1 \end{cases}$$

  We denote the $\|.\|_*^Z$ norm over a sample dataset Z of a function $g \in \mathcal{G}_2$ as followed:

  $$\|g\|_*^Z = \max_{z_*\in Z}\max_{j\in\{1,\dots,d_2\}}|g_j(z_*)| = \max_{z_*\in Z}\|g(z_*)\|_\infty$$

We have:

$$|\mathcal{C}_2| \leqslant \sum_{\bar{A}^{(1)} \in \mathcal{C}_1} \left| \mathcal{C}_{\epsilon_2}\left(\mathcal{G}_2, \|.\|_*^{Z(\bar{A}^{(1)})}\right)\right| \leqslant |\mathcal{C}_1| \cdot \sup_{\bar{A}^{(1)} \in \mathcal{C}_1} \mathcal{N}\left(\mathcal{G}_2, \epsilon_2, \|.\|_*^{Z(\bar{A}^{(1)})}\right)$$

- For all $L_{\mathbf{A}} \in \mathcal{L}_r,\ \mathbf{A} = \left(A^{(1)}, A^{(2)}\right)$, we can choose $\bar{\mathbf{A}} = \left(\bar{A}^{(1)}, \bar{A}^{(2)}\right)$ such that $\bar{A}^{(1)} \in \mathcal{C}_1$ and $\bar{A}^{(2)} \in \mathcal{C}_{\epsilon_2}\left(\mathcal{G}_2, \|.\|_*^{Z(\bar{A}^{(1)})}\right) \subset \mathcal{C}_2$ (for the sake of brevity, we do not write elements of $\mathcal{C}_1, \mathcal{C}_2$ as functions but parameters instead) and:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|A^{(1)}x_i - \bar{A}^{(1)}x_i\right\|_2^2} \leqslant \epsilon_1, \quad \max_{z^* \in Z(\bar{A}^{(1)})}\left\|\left(A^{(2)} - \bar{A}^{(2)}\right)z_*\right\|_\infty \leqslant \epsilon_2$$

We have:

$$\begin{aligned}
&\|L_{\mathbf{A}} - L_{\bar{\mathbf{A}}}\|_2^S \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|L_{\mathbf{A}}(x_i, y_i) - L_{\bar{\mathbf{A}}}(x_i, y_i)\right|^2} \\
&\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2a_*\rho}{r}\left\|A^{(1)}x_i - \bar{A}^{(1)}x_i\right\|_2 + \frac{2}{r}\left\|\left(A^{(2)} - \bar{A}^{(2)}\right)\sigma\left(\bar{A}^{(1)}x_i\right)\right\|_\infty\right]^2} \\
&\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2a_*\rho}{r}\left\|A^{(1)}x_i - \bar{A}^{(1)}x_i\right\|_2 + \frac{2}{r}\epsilon_2\right]^2} \\
&\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2a_*\rho}{r}\left\|A^{(1)}x_i - \bar{A}^{(1)}x_i\right\|_2\right]^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2\epsilon_2}{r}\right]^2} \quad \text{(Minkowski's Inequality)} \\
&\leqslant \frac{2a_*\rho\epsilon_1}{r} + \frac{2\epsilon_2}{r}
\end{aligned}$$

For an arbitrary granularity $\epsilon > 0$ that we want to impose on the cover of $\mathcal{L}_r$, in order to make $\left|L_{\mathbf{A}}(x_i, y_i) - L_{\bar{\mathbf{A}}}(x_i, y_i)\right| \leqslant \epsilon$, set:

$$\epsilon_1 = \frac{r\epsilon}{8a_*\rho}, \quad \epsilon_2 = \frac{3r\epsilon}{8} \implies \|L_{\mathbf{A}} - L_{\bar{\mathbf{A}}}\|_2^S \leqslant \frac{\epsilon}{4} + \frac{3\epsilon}{4} = \epsilon$$

- Find the cardinality of $\mathcal{C}_1, \mathcal{C}_2$ with granularities ϵ_1, ϵ_2 defined above. Then, cover $\mathcal{L}_r$ using the fact that:

$$\begin{aligned}
\log \mathcal{N}\left(\mathcal{L}_r, \epsilon, \|.\|_2^S\right) &\leqslant \log|\mathcal{C}_1| + \log|\mathcal{C}_2| \\
&\leqslant 2\log|\mathcal{C}_1| + \log\left(\sup_{\bar{A}^{(1)} \in \mathcal{C}_1} \mathcal{N}\left(\mathcal{G}_2, \epsilon_2, \|.\|_*^{Z(\bar{A}^{(1)})}\right)\right) \quad (1)
\end{aligned}$$

1.1. Bounding $|\mathcal{C}_1|$: By lemma 3.2 in Bartlett et al. 2017, we have:

$$\log|\mathcal{C}_1| = \log \mathcal{N}\left(\mathcal{G}_1, \epsilon_1, \|.\|_2^S\right) \leqslant \left\lceil \frac{a_1^2}{\epsilon_1^2} \right\rceil \log(2d_0d_1)$$

Setting $\epsilon_1 = \frac{r\epsilon}{8a_*\rho}$ and let $W = \max\{d_0, d_1, d_2\}$, we have:

$$\log|\mathcal{C}_1| \leqslant \left\lceil \frac{64a_*^2 a_1^2 \rho^2}{r^2\epsilon^2} \right\rceil \log(2d_0 d_1) \leqslant \left\lceil \frac{64a_*^2 a_1^2 \rho^2}{r^2\epsilon^2} \right\rceil \log(2W^2)$$

Assuming $64a_*^2 a_1^2 \rho^2 \geqslant r^2\epsilon^2$, we have:

$$\log|\mathcal{C}_1| \leqslant \left\lceil \frac{64a_*^2 a_1^2 \rho^2}{r^2\epsilon^2} \right\rceil \log(2W^2) \leqslant \frac{128a_*^2 a_1^2 \rho^2}{r^2\epsilon^2} \log(2W^2) \quad (2)$$

1.2. Bounding $|\mathcal{C}_2|$: for any member $\bar{A}^{(1)} \in \mathcal{C}_1$, we have:

$$Z(\bar{A}^{(1)}) = \{z_1, \ldots, z_n\}, \text{ where } z_i = \sigma\Big(\bar{A}^{(1)} x_i\Big),\ x_i \in S$$

For all $z_i \in Z(\bar{A}^{(1)})$, we have:

$$\|z_i\|_2 = \left\|\sigma\Big(\bar{A}^{(1)} x_i\Big)\right\|_2 \leqslant \rho\left\|\bar{A}^{(1)} x_i\right\|_2 \leqslant \rho s_1$$

To cover $\mathcal{G}_2$ with $\|.\|_*^{Z(\bar{A}^{(1)})}$ norm, define a new sample $Z'(\bar{A}^{(1)})$ derived from $Z(\bar{A}^{(1)})$ as followed: For every $z_i \in Z(\bar{A}^{(1)})$, create d_2 new $\mathbb{R}^{d_2 \cdot d_1}$ vectors $\{z_{i,j}\}_{j=1}^{d_2}$ such that the $((j-1)\times d_1 + 1)$ to $j \times d_1$ entries of $z_{i,j}$ are the same as z_i, the remaining entries are 0.

$$z_i \mapsto \begin{cases} z_{i,1} &= \begin{pmatrix} z_i & \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \ldots & \vec{\mathbf{0}}_{d_1} \end{pmatrix}^T \\ z_{i,2} &= \begin{pmatrix} \vec{\mathbf{0}}_{d_1} & z_i & \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \ldots & \vec{\mathbf{0}}_{d_1} \end{pmatrix}^T \\ &\vdots \\ z_{i,d_2} &= \begin{pmatrix} \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \vec{\mathbf{0}}_{d_1} & \ldots & z_i \end{pmatrix}^T \end{cases}$$

Where each $\vec{\mathbf{0}}_{d_1}$ represents a 0-vector with d_1 entries. Also, define the following new function class:

$$\mathcal{G}_2' = \left\{ z' \mapsto \tilde{A}^{(2)} z' : \tilde{A}^{(2)} = \begin{pmatrix} A_1^{(2)} & \ldots & A_{d_2}^{(2)} \end{pmatrix},\ A^{(2)} \in \mathcal{G}_2,\ z' \in Z'(\bar{A}^{(1)}) \right\}$$

Specifically, each parameter $\tilde{A}^{(2)}$ of $\mathcal{G}_2'$ is created by concatenating rows of matrices $A^{(2)} \in \mathcal{G}_2$ to form $\mathbb{R}^{1\times d_2 \cdot d_1}$ row vectors. We realize that to construct a cover for $\mathcal{G}_2$ with respect to $\|.\|_*^{Z(\bar{A}^{(1)})}$ is equivalent to constructing a cover for $\mathcal{G}_2'$ with respect to $\|.\|_\infty^{Z'(\bar{A}^{(1)})}$. Hence, we have:

$$\mathcal{N}\Big(\mathcal{G}_2, \epsilon_2, \|.\|_*^{Z(\bar{A}^{(1)})}\Big) = \mathcal{N}\Big(\mathcal{G}_2', \epsilon_2, \|.\|_\infty^{Z'(\bar{A}^{(1)})}\Big)$$

Since for all $A^{(2)} \in \mathcal{G}_2$, $\|A^{(2)}\|_F \leqslant a_*$, we have $\|\tilde{A}^{(2)}\|_2 \leqslant a_*,\ \forall \tilde{A}^{(2)} \in \mathcal{G}_2'$. Also, we have $\|z'\|_2 \leqslant \rho s_1,\ \forall z' \in Z'(\bar{A}^{(1)})$. Therefore, by theorem 4 in Zhang 2002, we have:

$$\begin{aligned} \log\mathcal{N}\Big(\mathcal{G}_2, \epsilon_2, \|.\|_*^{Z(\bar{A}^{(1)})}\Big) &= \log\mathcal{N}\Big(\mathcal{G}_2', \epsilon_2, \|.\|_\infty^{Z'(\bar{A}^{(1)})}\Big) \\ &\leqslant \frac{36a_*^2 s_1^2 \rho^2}{\epsilon_2^2} \log\left(\left(\frac{8a_* s_1 \rho}{\epsilon_2} + 7\right) n d_2\right) \\ &\leqslant \frac{36a_*^2 s_1^2 \rho^2}{\epsilon_2^2} \log\left(\left(\frac{8a_* s_1 \rho}{\epsilon_2} + 7\right) n W\right) \end{aligned}$$

Setting $\epsilon_2 = \frac{3r\epsilon}{8}$, we have:

$$\log\left(\sup_{\bar{A}^{(1)}\in\mathcal{C}_1}\mathcal{N}\Big(\mathcal{G}_2,\epsilon_2,\|.\|_*^{Z(\bar{A}^{(1)})}\Big)\right) \leqslant \frac{36a_*^2s_1^2\rho^2}{\epsilon_2^2}\log\left(\left(\frac{8a_*s_1\rho}{\epsilon_2}+7\right)nW\right)$$
$$= \frac{256a_*^2s_1^2\rho^2}{r^2\epsilon^2}\log\left(\left(\frac{64a_*s_1\rho}{3r\epsilon}+7\right)nW\right) \quad (3)$$

1.3. Combine $\mathcal{C}_1, \mathcal{C}_2$: We can cover $\mathcal{L}_r$ with respect to the $\|.\|_2^S$ norm using the Cartesian product $\mathcal{C}_1 \times \mathcal{C}_2$. Hence, $\mathcal{N}\Big(\mathcal{L}_r,\epsilon,\|.\|_2^S\Big) \leqslant \Big|\mathcal{C}_1\times\mathcal{C}_2\Big| \leqslant |\mathcal{C}_1|\cdot|\mathcal{C}_2|$. Hence, From (1), (2) and (3), we have:

$$\log\mathcal{N}\Big(\mathcal{L}_r,\epsilon,\|.\|_2^S\Big) \leqslant 2\log|\mathcal{C}_1| + \log\left(\sup_{\bar{A}^{(1)}\in\mathcal{C}_1}\mathcal{N}\Big(\mathcal{G}_2,\epsilon_2,\|.\|_*^{Z(\bar{A}^{(1)})}\Big)\right)$$
$$\leqslant \frac{256a_*^2\rho^2}{r^2\epsilon^2}\left[a_1^2\log(2W^2) + s_1^2\log\left(\left(\frac{64a_*s_1\rho}{3r\epsilon}+7\right)nW\right)\right]$$
$$\leqslant \frac{256a_*^2\rho^2}{r^2\epsilon^2}\left[2a_1^2\log(2W) + s_1^2\log\left(\left(\frac{64a_*s_1\rho}{3r\epsilon}+7\right)nW\right)\right]$$
$$\leqslant \frac{256a_*^2\rho^2}{r^2\epsilon^2}\max\{a_1\sqrt{2},s_1\}^2\log\left(\left(\frac{64a_*s_1\rho}{3r\epsilon}+7\right)2nW^2\right)$$
$$\leqslant \frac{256a_*^2\rho^2}{r^2\epsilon^2}\max\{a_1\sqrt{2},s_1\}^2\log\left(\left(\frac{64a_*\rho}{3r\epsilon}\max\{a_1\sqrt{2},s_1\}+7\right)2nW^2\right)$$

Letting $R = \frac{16a_*\rho}{r}\max\{a_1\sqrt{2},s_1\}$, we have:

$$\log\mathcal{N}\Big(\mathcal{L}_r,\epsilon,\|.\|_2^S\Big) \leqslant \frac{R^2}{\epsilon^2}\log\left(\left(\frac{4R}{3\epsilon}+7\right)2nW^2\right)$$

Note that for all $\bar{A}^{(1)} \in \mathcal{C}_1, \bar{A}^{(2)} \in \mathcal{C}_2$, we have $(\bar{A}^{(1)}, \bar{A}^{(2)}) \in \mathcal{A}$. Therefore, $\mathcal{C}_1 \times \mathcal{C}_2$ is an internal cover.

2. Dudley's entropy integral : For $\alpha > 0$, we have:

$$\widehat{\mathfrak{R}}_S(\mathcal{L}_r) \leqslant 4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1\sqrt{\mathcal{N}\Big(\mathcal{L}_r,\epsilon,\|.\|_2^S\Big)}d\epsilon$$
$$\leqslant 4\alpha + \frac{12R}{\sqrt{n}}\int_\alpha^1\frac{1}{\epsilon}\sqrt{\log\left(\left(\frac{4R}{3\epsilon}+7\right)2nW^2\right)}d\epsilon$$
$$\leqslant 4\alpha + \frac{12R}{\sqrt{n}}\sqrt{\log\left(\left(\frac{4R}{3\alpha}+7\right)2nW^2\right)}(-\log\alpha)$$

Setting $\alpha = \frac{3R}{\sqrt{n}}$, we have:

$$\widehat{\mathfrak{R}}_S(\mathcal{L}_r) \leqslant \frac{12R}{\sqrt{n}}\left[1+\log\frac{\sqrt{n}}{3R}\sqrt{\log\left(\left(\frac{4\sqrt{n}}{9}+7\right)2nW^2\right)}\right]$$
$$= \tilde{O}\left(\frac{R}{\sqrt{n}}\right)$$

A.6 Rademacher Complexity of ramp loss for neural networks

A.6.1 Problem Statement

Problem : Given a tuple $\begin{pmatrix} M^{(1)} & \dots & M^{(L)} \end{pmatrix}$ of $L \geqslant 1$ reference matrices (that represents initial weights in a neural network). Let $\mathcal{A}$ be the class of tuples of L weight matrices defined as followed:

$$\mathcal{A} = \left\{ \begin{pmatrix} A^{(1)} & \dots & A^{(L)} \end{pmatrix} : \|A^{(i)}\|_\sigma \leqslant s_i, \|(A^{(i)} - M^{(i)})^T\|_{2,1} \leqslant a_i, \|A^{(L)} - M^{(L)}\|_F \leqslant a_* \right\}$$

Where $\|.\|_\sigma$ denotes the spectral norm and $\|.\|_{p,q}$ denotes the matrix (p,q) norm defined as $\|A\|_{p,q} = \left(\sum_j \left(\sum_i |A_{ij}|^p \right)^{q/p} \right)^{1/q}$. Therefore, the $\|.\|_{2,1}$ is defined as:

$$\|A\|_{2,1} = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} |A_{ij}|^2 \right)^{1/2} = \sum_{j=1}^{m} \|A_{.,j}\|_2, \quad A \in \mathbb{R}^{m \times n}$$

Define the following class of neural networks:

$$\mathcal{F}_\mathcal{A} = \left\{ x \mapsto A^{(L)} \left(\bigcirc_{k=1}^{L-1} \sigma_k \circ A^{(k)} \right)(x) : \sigma_k \text{ is } \rho_k\text{-Lipchitz w.r.t } l_2 \text{ norm}, \begin{pmatrix} A^{(1)} & \dots & A^{(L)} \end{pmatrix} \in \mathcal{A} \right\}$$

We denote $d_0, d_1, \dots, d_L$ as the widths of each layer of the neural networks where d_0 denotes the dimensionality of the input and $d_L = K$ be the size of the output.
Derive the bound for the Rademacher Complexity of the loss function class:

$$\mathcal{L}_r = \left\{ (x,y) \mapsto l_r\Big(F_\mathbf{A}(x)_y - \max_{k \neq y} F_\mathbf{A}(x)_k \Big) \,\middle|\, F_\mathbf{A} \in \mathcal{F}_\mathcal{A} \right\}$$

Where l_r is the ramp loss function with margin $r \in (0,1)$.

A.6.2 Neural networks covering bounds with general norm

Theorem A.7: Neural networks covering bound with general norm

Let $L \geqslant 1$ be a natural number and $\epsilon_1, \dots, \epsilon_L > 0$ be given as covering number granularities. Given the following:

- A sequence of vector spaces $\mathcal{V}_0, \dots, \mathcal{V}_L$ endowed with norms $|.|_0, \dots, |.|_L$.
- A sequence of vector spaces $\mathcal{W}_1, \dots, \mathcal{W}_L$ endowed with norms $\|.\|_1, \dots, \|.\|_L$.
- A sequence of real positive numbers $c_1, \dots, c_L$ and linear operators $A^{(i)} : \mathcal{V}_i \to \mathcal{W}_{i+1}$ associated with the operator norm:

$$\|A^{(i)}\|_{op} = \sup_{|Z|_i \leqslant 1} \|A_i Z\|_{i+1} \leqslant c_i, \ \forall i \in \{1, \dots, L\}$$

- A sequence of real positive numbers $\rho_1, \dots, \rho_L$ and activation functions $\sigma_i : \mathcal{W}_i \to \mathcal{V}_i$ such that σ_i are ρ_i-Lipchitz:

$$|\sigma_i(z_1) - \sigma_i(z_2)|_i \leqslant \rho_i \|z_1 - z_2\|_i$$

- Let $\mathcal{A} \subseteq \mathcal{B}_1 \times \cdots \times \mathcal{B}_L$ be a class of tuples of matrices $(A^{(1)}, \dots, A^{(L)})$ such that each $A^{(i)} \in \mathcal{B}_i$ satisfies $\|A^{(i)}\|_{op} \leqslant c_i$.
- Define the class of neural networks $\mathcal{F}_{\mathcal{A}}$ as followed:

$$\mathcal{F}_{\mathcal{A}} = \left\{ x \mapsto A^{(L)} \left(\bigcirc_{k=1}^{L-1} \sigma_k \circ A^{(k)} \right)(x) : (A^{(1)}, \dots, A^{(L)}) \in \mathcal{A} \right\}$$

- Let τ be the aggregated granularity defined as $\tau = \sum_{j=1}^{L} \epsilon_j \left(\prod_{l=j+1}^{L} c_l \rho_{l-1} \right)$ and $Z \subset \mathcal{V}_0$ be a sample dataset. We have:

$$\mathcal{N}\left(\mathcal{F}_{\mathcal{A}|Z}, \tau, \|.\|_L\right) \leqslant |\mathcal{C}_1| \cdot \prod_{i=2}^{L} \sup_{\substack{(A^{(1)}, \dots, A^{(i-1)}) \\ A^{(j)} \in \mathcal{B}_j, j \leqslant i-1}} \mathcal{N}\left(\left\{ A^{(i)} F_{\mathbf{A}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i \right\}, \epsilon_i, \|.\|_i \right)$$

Proof (Theorem A.7). *We will prove the above theorem inductively.*

Base case *: Suppose we have a sample dataset $Z \subset \mathcal{V}_0$. Construct a minimum ϵ_1-cover for the weight matrices in the 1^{st} layer called $\mathcal{C}_1$ such that for all $A^{(1)} \in \mathcal{B}_i$, there exists $\bar{A}^{(1)} \in \mathcal{C}_1$ such that:*

$$\left\| A^{(1)} Z - \bar{A}^{(1)} Z \right\|_1 \leqslant \epsilon_1$$

Inductive step *: For $i, j \geqslant 2, j \geqslant i$, define $\mathbf{A}^{(\mathbf{i,j})} \in \mathcal{B}_i \times \cdots \times \mathcal{B}_j$ as the extraction of layers i to j for every $\mathbf{A} \in \mathcal{A}$. Hence, define the function $F_{\mathbf{A}^{(\mathbf{i,j})}}$ as followed:*

$$F_{\mathbf{A}^{(\mathbf{i,j})}}(Z) = \sigma_j \left(A^{(j)} \sigma_{j-1} \left(A^{(j-1)} \dots \sigma_i \left(A^{(i)} Z \right) \dots \right) \right)$$

We need to construct a cover $\mathcal{C}_i$ with respect to norm $\|.\|_i$ with granularity $\tau_i > 0$ such that for all $\mathbf{A}^{(\mathbf{1,i})} \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_i$, there exists $\bar{\mathbf{A}}^{(\mathbf{1,i})} \in \mathcal{C}_i$ such that:

$$\left\| A^{(i)} F_{\mathbf{A}^{(\mathbf{1,i-1})}}(Z) - \bar{A}^{(i)} F_{\bar{\mathbf{A}}^{(\mathbf{1,i-1})}}(Z) \right\|_i \leqslant \tau_i$$

Suppose that we have constructed covers $\mathcal{C}_{i-1}, \mathcal{C}_{i-2}, \ldots, \mathcal{C}_1$ with respect to norms $\|.\|_{i-1}, \|.\|_{i-2}, \ldots, \|.\|_1$ with granularities $\tau_{i-1}, \tau_{i-2}, \ldots, \tau_1$ (where $\mathcal{C}_1$ is the cover constructed in the base case with granularity $\tau_1 = \epsilon_1$). We have:

$$\begin{aligned}
&\left\|A^{(i)}F_{\mathbf{A}^{(1,i-1)}}(Z) - \bar{A}^{(i)}F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i \\
&\quad\leqslant \left\|\left(A^{(i)} - \bar{A}^{(i)}\right)F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i + \left\|A^{(i)}\left(F_{\mathbf{A}^{(1,i-1)}}(Z) - F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right)\right\|_i \\
&\quad\leqslant \left\|\left(A^{(i)} - \bar{A}^{(i)}\right)F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i + \left\|A^{(i)}\right\|_{op}\cdot\left|F_{\mathbf{A}^{(1,i-1)}}(Z) - F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right|_{i-1} \\
&\quad\leqslant \left\|\left(A^{(i)} - \bar{A}^{(i)}\right)F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i + c_i\left|F_{\mathbf{A}^{(1,i-1)}}(Z) - F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right|_{i-1} \\
&\quad\leqslant \left\|\left(A^{(i)} - \bar{A}^{(i)}\right)F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i + c_i\rho_{i-1}\left\|A^{(i-1)}F_{\mathbf{A}^{(1,i-2)}}(Z) - \bar{A}^{(i-1)}F_{\bar{\mathbf{A}}^{(1,i-2)}}(Z)\right\|_{i-1}
\end{aligned}$$

For each $\bar{\mathbf{A}}^{(1,i-1)} \in \mathcal{C}_{i-1}$, construct the minimum ϵ_i-cover $\mathcal{C}_i(\bar{\mathbf{A}}^{(1,i-1)})$ for the class $\left\{A^{(i)}F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i\right\}$. Then, we have:

$$\begin{aligned}
\left|\mathcal{C}_i(\bar{\mathbf{A}}^{(1,i-1)})\right| &\leqslant \sup_{\bar{\mathbf{A}}^{(1,i-1)}\in\mathcal{C}_{i-1}} \mathcal{N}\left(\left\{A^{(i)}F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i\right\}, \epsilon_i, \|.\|_i\right) \\
&\leqslant \sup_{\substack{(A^{(1)},\ldots,A^{(i-1)}) \\ A^{(j)}\in\mathcal{B}_j, j\leqslant i-1}} \mathcal{N}\left(\left\{A^{(i)}F_{\mathbf{A}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i\right\}, \epsilon_i, \|.\|_i\right)
\end{aligned}$$

Now, construct $\mathcal{C}_i$ as followed:

$$\begin{aligned}
\mathcal{C}_i &= \bigcup_{\bar{\mathbf{A}}^{(1,i-1)}\in\mathcal{C}_{i-1}} \mathcal{C}_i(\bar{\mathbf{A}}^{(1,i-1)}) \\
\implies |\mathcal{C}_i| &\leqslant \sum_{\bar{\mathbf{A}}^{(1,i-1)}\in\mathcal{C}_{i-1}} \left|\mathcal{C}_i(\bar{\mathbf{A}}^{(1,i-1)})\right| \\
&\leqslant |\mathcal{C}_{i-1}|\cdot \sup_{\substack{(A^{(1)},\ldots,A^{(i-1)}) \\ A^{(j)}\in\mathcal{B}_j, j\leqslant i-1}} \mathcal{N}\left(\left\{A^{(i)}F_{\mathbf{A}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i\right\}, \epsilon_i, \|.\|_i\right) \\
&\leqslant |\mathcal{C}_1|\cdot \prod_{k=2}^{i} \sup_{\substack{(A^{(1)},\ldots,A^{(k-1)}) \\ A^{(j)}\in\mathcal{B}_j, j\leqslant k-1}} \mathcal{N}\left(\left\{A^{(k)}F_{\mathbf{A}^{(1,k-1)}}(Z) : A^{(k)} \in \mathcal{B}_k\right\}, \epsilon_k, \|.\|_k\right)
\end{aligned}$$

The last inequality was achieved by expanding the expression inductively. From the construction of $\mathcal{C}_i$, for all $\mathbf{A}^{(1,i)} \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_i$, we can choose $\bar{\mathbf{A}}^{(1,i)} \in \mathcal{C}_i$ such that:

$$\begin{aligned}
&\left\|A^{(i)}F_{\mathbf{A}^{(1,i-1)}}(Z) - \bar{A}^{(i)}F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i \\
&\quad\leqslant \left\|\left(A^{(i)} - \bar{A}^{(i)}\right)F_{\bar{\mathbf{A}}^{(1,i-1)}}(Z)\right\|_i + c_i\rho_{i-1}\left\|A^{(i-1)}F_{\mathbf{A}^{(1,i-2)}}(Z) - \bar{A}^{(i-1)}F_{\bar{\mathbf{A}}^{(1,i-2)}}(Z)\right\|_{i-1} \\
&\quad\leqslant \epsilon_i + c_i\rho_{i-1}\left(\epsilon_{i-1} + c_{i-1}\rho_{i-2}\left\|A^{(i-2)}F_{\mathbf{A}^{(1,i-3)}}(Z) - \bar{A}^{(i-2)}F_{\bar{\mathbf{A}}^{(1,i-3)}}(Z)\right\|_{i-2}\right) \\
&\quad\vdots \\
&\quad\leqslant \sum_{j=1}^{i}\epsilon_j\left(\prod_{l=j+1}^{i} c_l\rho_{l-1}\right)
\end{aligned}$$

By induction we have the size of the cover $\mathcal{C}_L$ satisfies the following upper bound:

$$|\mathcal{C}_L| \leqslant |\mathcal{C}_1|\cdot \prod_{i=2}^{L} \sup_{\substack{(A^{(1)},\ldots,A^{(i-1)}) \\ A^{(j)}\in\mathcal{B}_j, j\leqslant i-1}} \mathcal{N}\left(\left\{A^{(i)}F_{\mathbf{A}^{(1,i-1)}}(Z) : A^{(i)} \in \mathcal{B}_i\right\}, \epsilon_i, \|.\|_i\right)$$

With granularity $\tau = \sum_{j=1}^{L} \epsilon_j \left(\prod_{l=j+1}^{L} c_l \rho_{l-1} \right)$. ▫.

A.6.3 Solution to A.6.1 - without applying theorem A.7

In this section, we will solve the problem in A.6.1 using the same proof technique that was used for theorem A.7 without applying it directly.

1. Construct cover for $\mathcal{L}_r$ inductively: Given a sample dataset $S = \left\{(x_i, y_i)\right\}_{i=1}^{n}$ and $\epsilon > 0$, we need to construct an ϵ-cover for $\mathcal{L}_r$ with respect to the $\|.\|_2^S$ defined as:

$$\forall L_{\mathbf{A}} \in \mathcal{L}_r : \|L_{\mathbf{A}}\|_2^S = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |L_{\mathbf{A}}(x_i, y_i)|^2}$$

Where we use the notation $L_{\mathbf{A}}$ to specify that $L_{\mathbf{A}}$ is parameterized by the tuple of weight matrices $\mathbf{A} = (A^{(1)}, \dots, A^{(L)}) \in \mathcal{A}$. Let $L_{\mathbf{A}}, L_{\bar{\mathbf{A}}} \in \mathcal{L}_r$ be the loss functions for $F_{\mathbf{A}}, F_{\bar{\mathbf{A}}} \in \mathcal{F}_{\mathcal{A}}$, we have:

$$|L_{\mathbf{A}}(x_i, y_i) - L_{\bar{\mathbf{A}}}(x_i, y_i)| \leqslant \frac{2}{r} \max_{j \in \{1,\dots,K\}} \left| F_{\mathbf{A}}(x_i)_j - F_{\bar{\mathbf{A}}}(x_i)_j \right|$$

For each $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$, we define $F_{\mathbf{A}}$ as the entire network and $F_{\mathbf{A}^{(1,k)}}$ as the extraction of the first $k \geqslant 1$ layers with activation in the k^{th} layer. Specifically:

$$F_{\mathbf{A}}(x) = A^{(L)}\sigma_{L-1}\Big(A^{(L-1)}\sigma_{L-2}\Big(\dots A^{(2)}\sigma_1(A^{(1)}x)\dots\Big)\Big)$$
$$F_{\mathbf{A}^{(1,k)}}(x) = \sigma_k\Big(A^{(k)}\sigma_{k-1}\Big(A^{(k-1)}\sigma_{k-2}\Big(\dots A^{(2)}\sigma_1(A^{(1)}x)\dots\Big)\Big)\Big)$$

Expanding $\max_{j\in\{1,\dots,K\}} \left| F_{\mathbf{A}}(x_i)_j - F_{\bar{\mathbf{A}}}(x_i)_j \right|$, we have:

$$\begin{aligned}
&\max_{j\in\{1,\dots,K\}} \left| F_{\mathbf{A}}(x_i)_j - F_{\bar{\mathbf{A}}}(x_i)_j \right| \\
&\quad= \left\| A^{(L)} F_{\mathbf{A}^{(1,L-1)}}(x_i) - \bar{A}^{(L)} F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) \right\|_\infty \\
&\quad= \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) + A^{(L)}\left(F_{\mathbf{A}^{(1,L-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i)\right) \right\|_\infty \\
&\quad\leqslant \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) \right\|_\infty + \left\| A^{(L)}\left(F_{\mathbf{A}^{(1,L-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i)\right) \right\|_\infty \\
&\quad\leqslant \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) \right\|_\infty + \left\| A^{(L)}\left(F_{\mathbf{A}^{(1,L-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i)\right) \right\|_2 \\
&\quad\leqslant \max_{x_*\in S} \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_*) \right\|_\infty + \|A^{(L)}\|_\sigma \left\| F_{\mathbf{A}^{(1,L-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) \right\|_2 \\
&\quad\leqslant \max_{x_*\in S} \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_*) \right\|_\infty + s_L \left\| F_{\mathbf{A}^{(1,L-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_i) \right\|_2 \\
&\quad\leqslant \max_{x_*\in S} \left\| \left(A^{(L)} - \bar{A}^{(L)}\right) F_{\bar{\mathbf{A}}^{(1,L-1)}}(x_*) \right\|_\infty + s_L\rho_{L-1} \left\| A^{(L-1)} F_{\mathbf{A}^{(1,L-2)}}(x_i) - \bar{A}^{(L-1)} F_{\bar{\mathbf{A}}^{(1,L-2)}}(x_i) \right\|_2
\end{aligned}$$

For $1 \leqslant k \leqslant L-1$, construct the cover $\mathcal{C}_k$ with respect to the $\|.\|_2^S$ for the following class:

$$\mathcal{F}_k = \left\{ x \mapsto A^{(k)} F_{\mathbf{A}^{(1,k-1)}}(x) : \|A^{(k)}\|_\sigma \leqslant s_k,\ \|(A^{(k)} - M^{(k)})^T\|_{2,1} \leqslant a_i \right\}$$

For every $x_i \in S$, we have:

$$\begin{aligned}
&\left\|A^{(k)} F_{\mathbf{A^{(1,k-1)}}}(x_i) - \bar{A}^{(k)} F_{\bar{\mathbf{A}}^{(1,k-1)}}(x_i)\right\|_2 \\
&\qquad = \left\|\left(A^{(k)} - \bar{A}^{(k)}\right) F_{\bar{\mathbf{A}}^{(1,k-1)}}(x_i)\right\|_2 + \left\|A^{(k)}\left(F_{\mathbf{A}^{(1,k-1)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,k-1)}}(x_i)\right)\right\|_2 \\
&\qquad = \left\|\left(A^{(k)} - \bar{A}^{(k)}\right) F_{\bar{\mathbf{A}}^{(1,k-1)}}(x_i)\right\|_2 + s_k \rho_{k-1} \left\|A^{(k-1)}\left(F_{\mathbf{A}^{(1,k-2)}}(x_i) - F_{\bar{\mathbf{A}}^{(1,k-2)}}(x_i)\right)\right\|_2
\end{aligned}$$

Expand the above inductively in a similar manner as theorem A.7 and for each layer index $1 \leqslant j \leqslant k$, denote $\mathcal{B}_l = \left\{A^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l} : \|A^{(l)}\|_\sigma \leqslant s_l, \|(A^{(l)} - M^{(l)})^T\|_{2,1} \leqslant a_l\right\}$ and $\mathcal{C}_1$ is the cover of the first layer, we have:

$$|\mathcal{C}_k| \leqslant |\mathcal{C}_1| \cdot \prod_{j=2}^{k} \sup_{\substack{(A^{(1)},\ldots,A^{(j-1)}) \\ \forall l \leqslant j-1 : A^{(l)} \in \mathcal{B}_l}} \mathcal{N}\left(\left\{x \mapsto A^{(j)} F_{\mathbf{A}^{(1,j-1)}}(x) : A^{(j)} \in \mathcal{B}_j\right\}, \epsilon_j, \|.\|_2^S\right)$$

Where $\epsilon_1, \ldots, \epsilon_k$ are known positive constants. Denote $F_{\mathbf{A}^{(1,j-1)}}(x) = x$ for $j = 1$ and also denote $W = \max_{0 \leqslant l \leqslant L-1} d_l$. By lemma 3.2 from Bartlett et al. 2017, we have:

$$\begin{aligned}
\log|\mathcal{C}_k| &\leqslant \sum_{j=1}^{k} \sup_{\substack{(A^{(1)},\ldots,A^{(j-1)}) \\ \forall l \leqslant j-1 : A^{(l)} \in \mathcal{B}_l}} \log \mathcal{N}\left(\left\{x \mapsto A^{(j)} F_{\mathbf{A}^{(1,j-1)}}(x) : A^{(j)} \in \mathcal{B}_j\right\}, \epsilon_j, \|.\|_2^S\right) \\
&\leqslant \sum_{j=1}^{k} \sup_{\substack{(A^{(1)},\ldots,A^{(j-1)}) \\ \forall l \leqslant j-1 : A^{(l)} \in \mathcal{B}_l, \|x\|_2 \leqslant 1}} \frac{a_j^2 \|F_{\mathbf{A}^{(1,j-1)}}(x)\|_2^2}{\epsilon_j^2} \log(2 d_{j-1} d_j) \\
&\leqslant \sum_{j=1}^{k} \sup_{\substack{(A^{(1)},\ldots,A^{(j-1)}) \\ \forall l \leqslant j-1 : A^{(l)} \in \mathcal{B}_l, \|x\|_2 \leqslant 1}} \frac{a_j^2 \|F_{\mathbf{A}^{(1,j-1)}}(x)\|_2^2}{\epsilon_j^2} \log(2W^2)
\end{aligned}$$

To bound the output at the $(j-1)^{th}$ layer, we have:

$$\begin{aligned}
\left\|F_{\mathbf{A}^{(1,j-1)}}(x)\right\|_2 &= \left\|\sigma_{j-1}\left(A^{(j-1)} \sigma_{j-2}\left(\ldots A^{(2)} \sigma_1(A^{(1)} x)\ldots\right)\right)\right\|_2 \\
&\leqslant \rho_{j-1} s_{j-1} \left\|\sigma_{j-2}\left(A^{(j-2)} \sigma_{j-3}\left(\ldots A^{(2)} \sigma_1(A^{(1)} x)\ldots\right)\right)\right\| \\
&\leqslant \prod_{l=1}^{j-1} \rho_l s_l
\end{aligned}$$

Hence, we have:

$$\log|\mathcal{C}_k| \leqslant \sum_{j=1}^{k} \frac{a_j^2 \prod_{1 \leqslant l < j} \rho_l^2 s_l^2}{\epsilon_j^2} \log(2W^2)$$

B List of Definitions

C Important Theorems

D Important Corollaries

E Important Propositions

F Important Lemmas

G References

References

Bartlett, Peter, Dylan J. Foster, and Matus Telgarsky (2017). *Spectrally-normalized margin bounds for neural networks*. arXiv: 1706.08498 [cs.LG].

Devroye, Luc, László Györfi, and Gábor Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer New York. ISBN: 9781461207115. DOI: 10.1007/978-1-4612-0711-5. URL: http://dx.doi.org/10.1007/978-1-4612-0711-5.

Durrett, Rick (2010). *Probability: Theory and Examples*. 4th. USA: Cambridge University Press. ISBN: 0521765390.

Fan, Zhou (2016). *Statistics 200: Introduction to Statistical Inference - Lecture 6*. https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture06.pdf. [Accessed 07-01-2024].

Foster, Dylan J. and Alexander Rakhlin (2019). *ℓ_∞ Vector Contraction for Rademacher Complexity*. arXiv: 1911.06468 [cs.LG].

Garivier, Aurelien (2019). *Lecture notes in Machine Learning Theory*. URL: https://www.math.univ-toulouse.fr/~agarivie/sites/default/files/5_VC.pdf.

Hoogland, Jesse (n.d.). *Generalization, from thermodynamics to statistical physics - ai alignment forum*. URL: https://www.alignmentforum.org/posts/uG7oJkyLBHEw3MYpT/generalization-from-thermodynamics-to-statistical-physics.

Ledent, Antoine et al. (2021). *Norm-based generalisation bounds for multi-class convolutional neural networks*. arXiv: 1905.12430 [cs.LG].

Long, Philip M. and Hanie Sedghi (2020). *Generalization bounds for deep convolutional neural networks*. arXiv: 1905.12600 [cs.LG].

undefinedinlar, Erhan (2011). *Probability and Stochastics*. Springer New York. ISBN: 9780387878591. DOI: 10.1007/978-0-387-87859-1. URL: http://dx.doi.org/10.1007/978-0-387-87859-1.

Wikipedia (2023). *Vitali set — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Vitali%20set&oldid=1187241923. [Online; accessed 24-December-2023].

— (2024a). *Hilbert projection theorem — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Hilbert%20projection%20theorem&oldid=1172787172. [Online; accessed 11-January-2024].

— (2024b). *Hoeffding's lemma — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Hoeffding's%20lemma&oldid=1114715065. [Online; accessed 04-January-2024].

— (2024c). *Rényi entropy — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=R%C3%A9nyi%20entropy&oldid=1190869396. [Online; accessed 05-January-2024].

Zhang, Tong (2002). "Covering Number Bounds of Certain Regularized Linear Function Classes". In: *Journal of Machine Learning Research* 2. URL: https://www.jmlr.org/papers/volume2/zhang02b/zhang02b.pdf.