COMP0078 Supervised Learning Coursework 2

Group 10

02 Jan 2025

PART I - Rademacher Complexity of Finite Spaces

1.1

We consider a collection $X_1, ..., X_m$, $m \geq 1$, of centered random variables taking values in $[a, b] \subset \mathbb{R}$. Since these random variables are centered, their expectations are equal to:

$$\mathbb{E}[X_i] = 0 \qquad\qquad i = 1, ..., m$$

We define $\bar{X} = \max_i X_i$. We would like to show that for any $\lambda > 0$, we have:

$$\mathbb{E}[\bar{X}] \leq \frac{1}{\lambda} \log \mathbb{E}[e^{\lambda \bar{X}}] \tag{1}$$

Let's start by defining the following function:

$$f(X) = e^{\lambda X} \tag{2}$$

Applying it to the maximum of our collection of random variables, we get:

$$f(\bar{X}) = e^{\lambda \bar{X}} \tag{3}$$

Applying **Jensen's Inequality** to this function leads to:

$$f\left(\mathbb{E}[\bar{X}]\right) \leq \mathbb{E}\left[f(\bar{X})\right]$$

$$e^{\lambda \mathbb{E}[\bar{X}]} \leq \mathbb{E}\left[e^{\lambda \bar{X}}\right] \tag{4}$$

Thus, applying the log function, which is continuous and strictly increasing, the inequation becomes:

$$\log\left(e^{\lambda \mathbb{E}[\bar{X}]}\right) \leq \log\left(\mathbb{E}\left[e^{\lambda \bar{X}}\right]\right) \tag{5}$$

equivalent to:

$$\lambda \mathbb{E}[\bar{X}] \leq \log\left(\mathbb{E}\left[e^{\lambda \bar{X}}\right]\right)$$

From which the constraint $\lambda > 0$ arises, hence leading to the final expression we wanted to prove:

$$\mathbb{E}[\bar{X}] \leq \frac{1}{\lambda} \log\left(\mathbb{E}\left[e^{\lambda \bar{X}}\right]\right) \qquad\qquad \forall \lambda > 0 \tag{6}$$

1.2

We would like to show the following expression:

$$\frac{1}{\lambda} \log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \frac{1}{\lambda} \log(m) + \lambda \frac{(b-a)^2}{8} \tag{7}$$

Since the exponential function is strictly positive and increasing, we can state the following:

$$\mathbb{E}[e^{\lambda \bar{X}}] \leq \sum_{i=1}^{m} \mathbb{E}[e^{\lambda X_i}] \tag{8}$$

We also remind ourselves the random variables in the collection are all centered, hence:

$$\forall i \in \{1, ..., m\} \qquad E[X_i] = 0$$

Therefore:

$$\sum_{i=1}^{m} \mathbb{E}[e^{\lambda X_i}] = \sum_{i=1}^{m} \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \tag{9}$$

Hoeffding's Lemma Let X be a random variable such that $X - \mathbb{E}[X] \in [a, b]$ with $a, b \in \mathbb{R}$ and for any $\lambda > 0$. Then, we have:

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\lambda^2 (b-a)^2/8} \tag{10}$$

Applying *Hoeffding's Lemma* to the sum in the right part of equation 9, we get:

$$\sum_{i=1}^{m} \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \leq \sum_{i=1}^{m} e^{\lambda^2 (b-a)^2/8} \tag{11}$$

Therefore:

$$\mathbb{E}[e^{\lambda \bar{X}}] \leq \sum_{i=1}^{m} e^{\lambda^2 (b-a)^2/8} \tag{12}$$

And since no element in the sum on the right depend on the variable i, we get:

$$\mathbb{E}[e^{\lambda \bar{X}}] \leq m e^{\lambda^2 (b-a)^2/8} \tag{13}$$

Equivalent to:

$$\log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \log\left(m e^{\lambda^2 (b-a)^2/8}\right) \tag{14}$$

$$\log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \log(m) + \log\left(e^{\lambda^2 (b-a)^2/8}\right)$$

$$\log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \log(m) + \frac{\lambda^2 (b-a)^2}{8}$$

Dividing both sides by λ which is strictly positive, we get the final expression:

$$\frac{1}{\lambda} \log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \frac{1}{\lambda} \log(m) + \lambda \frac{(b-a)^2}{8} \qquad \lambda > 0 \tag{15}$$

1.3

From the previous questions, we proved the following expressions, $\forall \lambda > 0$:

$$\mathbb{E}[\bar{X}] \leq \frac{1}{\lambda} \log\left(\mathbb{E}\left[e^{\lambda \bar{X}}\right]\right)$$

And

$$\frac{1}{\lambda} \log\left(\mathbb{E}[e^{\lambda \bar{X}}]\right) \leq \frac{1}{\lambda} \log(m) + \lambda \frac{(b-a)^2}{8}$$

Therefore, we can conclude the following:

$$\forall \lambda > 0 \qquad \mathbb{E}[\bar{X}] \leq \frac{1}{\lambda} \log(m) + \lambda \frac{(b-a)^2}{8} \tag{16}$$

To find the expression of λ in the right-side of the inequation, we must minimize the upper bound of $\mathbb{E}[\bar{X}]$. However, as the value of λ varies from 0 in $\mathbb{R}$, $\frac{1}{\lambda}$ decreases. Therefore, the upper-bound of $\mathbb{E}[\bar{X}]$ needs to be balanced to find it's lower bound:

$$\frac{1}{\lambda} \log(m) = \lambda \frac{(b-a)^2}{8} \tag{17}$$

Equivalent to:

$$\lambda^2 = \frac{8}{(b-a)^2}\log(m)$$

And since $\lambda > 0$:

$$\lambda = \sqrt{\frac{8}{(b-a)^2}\log(m)}$$

$$\lambda = \frac{2}{(b-a)}\sqrt{2\log(m)} \tag{18}$$

Hence, by replacing in the inequation found before, we get:

$$\mathbb{E}[\bar{X}] \leq \frac{1}{\frac{2}{(b-a)}\sqrt{2\log(m)}}\log(m) + \left(\frac{2}{(b-a)}\sqrt{2\log(m)}\right)\frac{(b-a)^2}{8} \tag{19}$$

$$\mathbb{E}[\bar{X}] \leq \frac{(b-a)}{2\sqrt{2}}\sqrt{\log(m)} + \sqrt{\log(m)}\frac{(b-a)}{2\sqrt{2}}$$

$$\mathbb{E}[\bar{X}] \leq \frac{2(b-a)}{2\sqrt{2}}\sqrt{\log(m)}$$

$$\mathbb{E}[\bar{X}] \leq \frac{(b-a)}{\sqrt{2}}\sqrt{\log(m)}$$

$$\mathbb{E}[\bar{X}] \leq \frac{(b-a)\sqrt{2}}{2}\sqrt{\log(m)}$$

Leading finally to:

$$\mathbb{E}[\bar{X}] \leq \frac{(b-a)}{2}\sqrt{2\log(m)} \tag{20}$$

1.4

Problem Setup We define S as a finite set of points in $\mathbb{R}^n$ with cardinality $|S| = m$. We also define its Rademacher complexity as following, with $\sigma_1, ..., \sigma_n$ as Rademacher variables (independent and uniformly sampled from $\{-1, 1\}$):

$$\mathcal{R}(S) = \mathbb{E}_\sigma\left[\max_{x\in S}\frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j\right] \tag{21}$$

We would like to show that:

$$\mathcal{R}(S) \leq \max_{x\in S}\|x\|_2\frac{\sqrt{2\log(m)}}{n} \tag{22}$$

Since $\sigma_1, ..., \sigma_n$ are Rademacher variables, then:

$$\forall j \in \{1, ..., n\}, \forall x_j \in \mathbb{R} \qquad \sigma_j x_j \in \{-x_j, x_j\} \tag{23}$$

We also notice:

$$\mathbb{E}[\sigma_j x_j] = 0 \tag{24}$$

We start by defining the following:

$$X = \frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j$$

Then:

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j\right]$$

We also define:

$$\bar{X} = \max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j$$

And

$$f(X) = e^{\lambda X}$$

Therefore, using *Jensen's Inequality*, we get:

$$f(\mathbb{E}[\bar{X}]) \leq \mathbb{E}[f(\bar{X})]$$

Equivalent to:

$$\exp\left(\lambda \mathbb{E}\left[\max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda \max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right]$$

where the right part is equal to:

$$\mathbb{E}\left[\exp\left(\lambda \max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right] = \mathbb{E}\left[\max_{x \in S} \exp\left(\frac{\lambda}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right]$$

Using the following upper bound to this expression:

$$\mathbb{E}\left[\max_{x \in S} \exp\left(\frac{\lambda}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right] \leq \sum_{x \in S} \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right] \tag{25}$$

And developing this upper bound into:

$$\sum_{x \in S} \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{j=1}^{n} \sigma_j x_j\right)\right] = \sum_{x \in S} \mathbb{E}\left[\prod_{j=1}^{n} \exp\left(\frac{\lambda}{n} \sigma_j x_j\right)\right] = \sum_{x \in S} \prod_{j=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sigma_j x_j\right)\right] \tag{26}$$

We can therefore apply the *Hoeffding's Lemma* as before, but considering this time λ/n instead of λ, and $\sigma_j x_j$ as our random variable (because $\mathbb{E}(\sigma_j x_j) = 0$):

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sigma_j x_j\right)\right] \leq e^{\left(\frac{\lambda}{n}\right)^2 \frac{(b-a)^2}{8}} \tag{27}$$

And since before, we stated $\sigma_j x_j \in \{-x_j, x_j\}$, we therefore identify $(b - a) = (x_j - (-x_j)) = 2x_j$. Hence, we get:

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sigma_j x_j\right)\right] \leq e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} \tag{28}$$

Thus:

$$\sum_{x \in S} \prod_{j=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sigma_j x_j\right)\right] \leq \sum_{x \in S} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} \tag{29}$$

Since the cardinality of the set S is $|S| = m$, we can further bound it as following:

$$\sum_{x \in S} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} \leq |S| \max_{x \in S} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} \tag{30}$$

$$= m \max_{x \in S} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} = e^{\log(m)} \max_{x \in S} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}}$$

$$= \max_{x\in S} e^{\log(m)} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} = \max_{x\in S} e^{\log(m)} \prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{x_j^2}{2}} = \max_{x\in S} e^{\log(m) + \left(\frac{\lambda}{n}\right)^2 \frac{\sum_{j=1}^n x_j^2}{2}}$$

And identifying $\|x\| = \sqrt{\sum_{j=1}^n x_j^2}$, we therefore get:

$$\max_{x\in S} e^{\log(m) + \left(\frac{\lambda}{n}\right)^2 \frac{\sum_{j=1}^n x_j^2}{2}} = \max_{x\in S} e^{\log(m) + \left(\frac{\lambda}{n}\right)^2 \frac{\|x\|^2}{2}}$$

Therefore, as a summary, we proved the following:

$$\exp\left(\lambda \mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right]\right) \leq \mathbb{E}\left[\max_{x\in S} \exp\left(\frac{\lambda}{n}\sum_{j=1}^{n}\sigma_j x_j\right)\right] \leq \sum_{x\in S}\prod_{j=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j x_j\right)\right] \leq \sum_{x\in S}\prod_{j=1}^{n} e^{\left(\frac{\lambda}{n}\right)^2 \frac{(2x_j)^2}{8}} \tag{31}$$

$$\leq \max_{x\in S} e^{\log(m) + \left(\frac{\lambda}{n}\right)^2 \frac{\|x\|^2}{2}}$$

Hence, we proved:

$$\exp\left(\lambda \mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right]\right) \leq \max_{x\in S} e^{\log(m) + \left(\frac{\lambda}{n}\right)^2 \frac{\|x\|^2}{2}} \tag{32}$$

And when using the log function on both sides, we get:

$$\lambda \mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right] \leq \max_{x\in S}\left(\log(m) + \frac{\lambda^2 \|x\|^2}{2n^2}\right) \tag{33}$$

$$\mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right] \leq \max_{x\in S}\left(\frac{\log(m)}{\lambda} + \frac{\lambda \|x\|^2}{2n^2}\right) \tag{34}$$

Once again, we need to minimize this upper bound as we did in the previous part, by choosing the right λ. Therefore, we must find the λ solving the equation:

$$\frac{\log(m)}{\lambda} = \frac{\lambda \|x\|^2}{2n^2} \tag{35}$$

Equivalent to:

$$\lambda^2 = \frac{2n^2 \log(m)}{\|x\|^2}$$

And since $\lambda > 0$:

$$\lambda = \sqrt{\frac{2n^2\log(m)}{\|x\|^2}} = \frac{n}{\|x\|}\sqrt{2\log(m)} \tag{36}$$

Replacing in our inequation we had before:

$$\mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right] \leq \max_{x\in S}\left(\frac{\log(m)}{\frac{n}{\|x\|}\sqrt{2\log(m)}} + \frac{n}{\|x\|}\sqrt{2\log(m)}\frac{\|x\|^2}{2n^2}\right) \tag{37}$$

Equivalent to:

$$\mathbb{E}\left[\max_{x\in S} \frac{1}{n}\sum_{j=1}^{n} \sigma_j x_j\right] \leq \max_{x\in S}\left(\frac{\|x\|\sqrt{\log(m)}}{n\sqrt{2}} + \frac{\|x\|\sqrt{2\log(m)}}{2n}\right)$$

$$\mathbb{E}\left[\max_{x\in S}\frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j\right] \leq \max_{x\in S}\left(\frac{\|x\|\sqrt{2\log(m)}}{2n}+\frac{\|x\|\sqrt{2\log(m)}}{2n}\right)$$

$$\mathbb{E}\left[\max_{x\in S}\frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j\right] \leq \max_{x\in S}\left(\frac{\|x\|\sqrt{2\log(m)}}{n}\right) \tag{38}$$

Using the notations of the wording ($\mathbb{E}_\sigma$ instead of $\mathbb{E}$ and $\|x\|_2$ instead of $\|x\|$), we therefore proved:

$$\mathbb{E}_\sigma\left[\max_{x\in S}\frac{1}{n}\sum_{j=1}^{n}\sigma_j x_j\right] \leq \max_{x\in S}\|x\|_2\frac{\sqrt{2\log(m)}}{n} \tag{39}$$

1.5

Problem Setup Let $\mathcal{H}$ be a set of hypotheses $f : \mathcal{X} \to \mathbb{R}$. We assume $\mathcal{H}$ to have a finite cardinality $|\mathcal{H}| < +\infty$. Let $S = (x_i)_{i=1}^n$ be a set of points in $\mathcal{X}$ an input set. We would like to prove an upper bound for empirical Rademacher complexity $\mathcal{R}_S(\mathcal{H})$ where the cardinality of $\mathcal{H}$ appears logarithmically.

The empirical Rademacher complexity of the *set of hypothesis* $\mathcal{H}$, *i.e.* $\mathcal{R}_S(\mathcal{H})$, is given by:

$$\mathcal{R}_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] \tag{40}$$

It's relation to the Rademacher complexity of the *set of hypothesis* $\mathcal{H}$, *i.e.* $\mathcal{R}(\mathcal{H})$, is given by the following expression:

$$\mathcal{R}(\mathcal{H}) = \mathbb{E}_S\left[\mathcal{R}_S(\mathcal{H})\right] \tag{41}$$

Starting again from the *Jensen's Inequality*, given by $f(\mathbb{E}[\bar{X}]) \leq \mathbb{E}[f(\bar{X})]$, with:

$$\bar{X} = \sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)$$

And

$$f(X) = e^{\lambda X}$$

Then:

$$\exp\left(\lambda\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right)\right] \tag{42}$$

And since the exponential function is strictly positive:

$$\mathbb{E}\left[\exp\left(\lambda\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right)\right] \leq \sum_{f\in\mathcal{H}}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] \tag{43}$$

Additionally, σ_j are independent for all $j \in \{1, ..., n\}$. Thus:

$$\sum_{f\in\mathcal{H}}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] = \sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j f(x_j)\right)\right] \tag{44}$$

Once again, since:

$$\forall j \in \{1, ..., n\} \qquad \mathbb{E}[\sigma_j x_j] = 0$$

And since $\forall j \in \{1, ..., n\}, \quad \sigma_j \in \{-1, 1\}$, then:

$$\forall j \in \{1, ..., n\} \qquad \sigma_j f(x_j) \in \{-f(x_j), f(x_j)\}$$

Therefore, we can again apply the *Hoeffding's Lemma* as before, noticing $(b-a) = f(x_j) - (-f(x_j)) = 2f(x_j)$, we therefore have:

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j f(x_j)\right)\right] \leq e^{(\lambda/n)^2 \times ((2f(x_j))^2/8)} \tag{45}$$

Equivalent to:

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j f(x_j)\right)\right] \leq \exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) \tag{46}$$

Therefore, we have:

$$\sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j f(x_j)\right)\right] \leq \sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) \tag{47}$$

Defining the cardinality of the set $\mathcal{H}$ as:

$$|\mathcal{H}| = m < +\infty \tag{48}$$

Then we can further bound:

$$\sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) \leq |\mathcal{H}| \sup_{f\in\mathcal{H}}\prod_{j=1}^{n}\exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) \tag{49}$$

For which the upper bound is equal to:

$$|\mathcal{H}| \sup_{f\in\mathcal{H}}\prod_{j=1}^{n}\exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) = m \sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)$$

Hence, by combining the inequations, we get:

$$\exp\left(\lambda\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right)\right] \leq \sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_j f(x_j)\right)\right] \tag{50}$$

$$\leq \sum_{f\in\mathcal{H}}\prod_{j=1}^{n}\exp\left(\frac{\lambda^2 (f(x_j))^2}{2n^2}\right) \leq |\mathcal{H}| \sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)$$

Leading to finally:

$$\exp\left(\lambda\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right]\right) \leq |\mathcal{H}| \sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right) \tag{51}$$

Applying the log function on both sides:

$$\lambda\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] \leq \log\left(|\mathcal{H}| \sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)\right) \tag{52}$$

The right part of the inequation can be developed into:

$$\log\left(|\mathcal{H}| \sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)\right) = \log(|\mathcal{H}|) + \log\left(\sup_{f\in\mathcal{H}}\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)\right)$$

$$= \log(|\mathcal{H}|) + \sup_{f\in\mathcal{H}}\log\left(\exp\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)\right) = \log(|\mathcal{H}|) + \sup_{f\in\mathcal{H}}\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)$$

Thus:

$$\lambda\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] \leq \log(|\mathcal{H}|) + \sup_{f\in\mathcal{H}}\left(\frac{\lambda^2 \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)$$

Equivalent to:

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(x_j)\right)\right] \leq \frac{\log|\mathcal{H}|}{\lambda} + \sup_{f\in\mathcal{H}}\left(\frac{\lambda \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right)$$

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\log|\mathcal{H}|}{\lambda} + \sup_{f\in\mathcal{H}}\left(\frac{\lambda \sum_{j=1}^{n}(f(x_j))^2}{2n^2}\right) \tag{53}$$

Minimization of the upper bound Now that we found an upper bound for the Empirical Rademacher Complexity, we would like now to minimize it by choosing an appropriate expression of λ. We define:

$$f(\lambda) = \frac{\log |\mathcal{H}|}{\lambda} + \sup_{f \in \mathcal{H}} \left(\frac{\lambda \sum_{j=1}^{n} (f(x_j))^2}{2n^2} \right) \tag{54}$$

Thus:

$$\frac{\partial}{\partial \lambda} f(\lambda) = -\frac{\log |\mathcal{H}|}{\lambda^2} + \sup_{f \in \mathcal{H}} \left(\frac{\sum_{j=1}^{n} (f(x_j))^2}{2n^2} \right) \tag{55}$$

Therefore, solving the following equation to find the minimum of this function:

$$\frac{\partial f}{\partial \lambda} = 0$$

Equivalent to:

$$-\frac{\log |\mathcal{H}|}{\lambda^2} + \sup_{f \in \mathcal{H}} \left(\frac{\sum_{j=1}^{n} (f(x_j))^2}{2n^2} \right) = 0$$

$$\frac{\log |\mathcal{H}|}{\lambda^2} = \sup_{f \in \mathcal{H}} \left(\frac{\sum_{j=1}^{n} (f(x_j))^2}{2n^2} \right) = \frac{1}{2n^2} \sup_{f \in \mathcal{H}} \left(\sum_{j=1}^{n} (f(x_j))^2 \right)$$

$$\lambda^2 = \frac{2n^2 \log |\mathcal{H}|}{\sup_{f \in \mathcal{H}} \left(\sum_{j=1}^{n} (f(x_j))^2 \right)}$$

$$\lambda = \sqrt{\frac{2n^2 \log |\mathcal{H}|}{\sup_{f \in \mathcal{H}} \left(\sum_{j=1}^{n} (f(x_j))^2 \right)}} = \frac{n\sqrt{\log |\mathcal{H}|}}{\sqrt{\sup_{f \in \mathcal{H}} \sum_{j=1}^{n} (f(x_j))^2}}$$

Leading finally to the following minimizer of the function:

$$\lambda = \frac{n\sqrt{2 \log |\mathcal{H}|}}{\sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}} \tag{56}$$

Substituting this expression into the inequation:

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\log |\mathcal{H}|}{\frac{n\sqrt{2 \log |\mathcal{H}|}}{\sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}}} + \left(\frac{n\sqrt{2 \log |\mathcal{H}|}}{\sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}} \right) \times \sup_{f \in \mathcal{H}} \left(\frac{\sum_{j=1}^{n} (f(x_j))^2}{2n^2} \right)$$

Equivalent to:

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\sqrt{\log |\mathcal{H}|} \sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}}{n\sqrt{2}} + \left(\frac{n\sqrt{2 \log |\mathcal{H}|}}{2n^2} \right) \sup_{f \in \mathcal{H}} \left(\frac{\sum_{j=1}^{n} (f(x_j))^2}{\sqrt{\sum_{j=1}^{n} (f(x_j))^2}} \right)$$

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\sqrt{2 \log |\mathcal{H}|} \sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}}{2n} + \left(\frac{\sqrt{2 \log |\mathcal{H}|}}{2n} \right) \sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}$$

$$\mathcal{R}_S(\mathcal{H}) \leq \left(\frac{\sqrt{2 \log |\mathcal{H}|}}{2n} + \frac{\sqrt{2 \log |\mathcal{H}|}}{2n} \right) \sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}$$

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\sqrt{2 \log |\mathcal{H}|}}{n} \sup_{f \in \mathcal{H}} \sqrt{\sum_{j=1}^{n} (f(x_j))^2}$$

Leading to this final expression of the upper bound for the Empirical Rademacher Complexity:

$$\mathcal{R}_S(\mathcal{H}) \leq \frac{\sqrt{2 \log |\mathcal{H}|}}{n} \sqrt{\sup_{f \in \mathcal{H}} \sum_{j=1}^{n} (f(x_j))^2} \tag{57}$$

PART II - Bayes Decision Rule and Surrogate Approach

In this part, we consider the binary classification problem. The classification (or "decision") rule is defined as a binary valued function $c : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \{-1, 1\}$.

The Misclassification Error is defined as following:

$$R(c) = \mathbb{P}_{(x,y)\sim\rho}(c(x) \neq y) \tag{58}$$

in which we assume to sample an input-output pair (x, y) according to a distribution ρ on $\mathcal{X} \times \mathcal{Y}$.

2.1

We define $\mathbf{1}_{y \neq y'}$ as the 0-1 loss such that $\mathbf{1}_{y \neq y'} = 1$ if $y \neq y'$ and 0 otherwise. We would like to show that the Misclassification Error corresponds to the *expected risk* of the 0-1 loss:

$$R(c) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{1}_{c(x)\neq y} d\rho(x, y) \tag{59}$$

We develop the expression of the Misclassification Error given in 58:

$$\begin{aligned}
R(c) &= \mathbb{P}_{(x,y)\sim\rho}(c(x) \neq y) \\
&= \mathbb{P}_{(x,y)\sim\rho}(c(x) = 1, y = (-1), x) + \mathbb{P}_{(x,y)\sim\rho}(c(x) = (-1), y = 1, x) \\
&= \mathbb{P}_{\rho}(x)\mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}c(x) = 1, y = (-1)|x\boldsymbol{)} + \mathbb{P}_{\rho}(x)\mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}c(x) = (-1), y = 1|x\boldsymbol{)} \\
&= \mathbb{P}_{\rho}(x)\left[\mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}c(x) = 1, y = (-1)|x\boldsymbol{)} + \mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}c(x) = (-1), y = 1|x\boldsymbol{)}\right]
\end{aligned}$$

$$R(c) = \mathbb{P}_{\rho}(x)\left[\mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}y = (-1)|x\boldsymbol{)}\mathbf{1}_{c(x)\neq(-1)} + \mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}y = 1|x\boldsymbol{)}\mathbf{1}_{c(x)\neq 1}\right] \tag{60}$$

And therefore, noticing the input-output pair in our expression is sampled from the distribution ρ, we can rewrite:

$$R(c) = \int_{x\in\mathcal{X}} \left[\mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}y = (-1)|x\boldsymbol{)}\mathbf{1}_{c(x)\neq(-1)} + \mathbb{P}_{(x,y)\sim\rho}\boldsymbol{(}y = 1|x\boldsymbol{)}\mathbf{1}_{c(x)\neq 1}\right] d\rho_{\mathcal{X}}(x) \tag{61}$$

$$= \int_{x\in\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} \rho(y|x)\mathbf{1}_{c(x)\neq y}\right) d\rho_{\mathcal{X}}(x)$$

$$= \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{Y}} \mathbf{1}_{c(x)\neq y} d\rho(y|x)\right) d\rho_{\mathcal{X}}(x)$$

$$R(c) = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathbf{1}_{c(x)\neq y} d\rho(x, y) \tag{62}$$

Which is the expression we intended to prove.

2.2

We assume to know ρ. We would like to calculate the closed-form minimizer f_* of the expected risk $\mathcal{E}(f)$ for each of the following losses: the squared loss, the exponential loss, the logistic loss, and the hinge loss.

The *expected risk* in the *surrogate problem* is given by the expression:

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) d\rho(x, y) \tag{63}$$

Which is equivalent to:

$$\mathcal{E}(f) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x) \right) d\rho_{\mathcal{X}}(x) \tag{64}$$

Finding the minimizer of the expected risk $\mathcal{E}(f)$ is equivalent to finding the minimizer of the inner expression that can be solved in a point-wise manner $\forall x \in \mathcal{X}$ since $\mathcal{Y} = \{-1, 1\}$. Such minimizer can be found by differentiating the inner integral with respect to the function f and setting the obtained gradient to zero.

2.2.a) Squared Loss

The squared loss is given by the following formula:

$$\ell(f(x), y) = (f(x) - y)^2 \tag{65}$$

Thus, $\forall x \in \mathcal{X}$:

$$\frac{\partial}{\partial f} \int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x) = 0 \tag{66}$$

The left part of the equation can be developed as:

$$\frac{\partial}{\partial f} \int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x) = \int_{\mathcal{Y}} \frac{\partial}{\partial f} \ell(f(x), y) d\rho(y|x)$$

$$= \sum_{y\in\mathcal{Y}} \frac{\partial}{\partial f} \ell(f(x), y) \rho(y|x) = \sum_{y\in\mathcal{Y}} \frac{\partial}{\partial f} (f(x) - y)^2 \rho(y|x) = \sum_{y\in\mathcal{Y}} 2(f(x) - y)\rho(y|x)$$

And therefore:

$$\sum_{y\in\mathcal{Y}} 2(f(x) - y)\rho(y|x) = 0 \tag{67}$$

Is equivalent to:

$$2(f(x) - 1)\rho(y = 1|x) + 2(f(x) - (-1))\rho(y = (-1)|x) = 0 \tag{68}$$

$$(f(x) - 1)\rho(y = 1|x) + (f(x) + 1)\rho(y = (-1)|x) = 0$$

$$f(x)\rho(y = 1|x) + f(x)\rho(y = (-1)|x) + \rho(y = (-1)|x) - \rho(y = 1|x) = 0$$

$$f(x) \left(\rho(y = 1|x) + \rho(y = (-1)|x)\right) = \rho(y = 1|x) - \rho(y = (-1)|x)$$

Leading to the equation of f_*:

$$f_*(x) = \frac{\rho(y = 1|x) - \rho(y = (-1)|x)}{\rho(y = (-1)|x) + \rho(y = 1|x)} \tag{69}$$

But since we notice that, $\forall x \in \mathcal{X}$, we have:

$$\rho(y = 1|x) - \rho(y = (-1)|x) = \sum_{y\in\mathcal{Y}} \rho(y|x) = 1$$

Therefore, the final expression of the minimizer f_* for the squared loss case is:

$$f_*(x) = \rho(y = 1|x) - \rho(y = (-1)|x) \tag{70}$$

2.2.b) Exponential loss

The exponential loss is given by:

$$\ell(f(x), y) = \exp\left(-yf(x)\right) = e^{-yf(x)} \tag{71}$$

Using the same process as for the squared loss:

$$\frac{\partial}{\partial f} \int_{\mathcal{Y}} e^{-yf(x)} d\rho(y|x) = 0 \tag{72}$$

Equivalent to:

$$\int_{\mathcal{Y}} \frac{\partial}{\partial f} e^{-yf(x)} d\rho(y|x) = 0$$

$$\sum_{y \in \mathcal{Y}} \frac{\partial}{\partial f} e^{-yf(x)} \rho(y|x) = 0$$

$$\sum_{y \in \mathcal{Y}} -ye^{-yf(x)} \rho(y|x) = 0$$

$$-e^{-f(x)}\rho(y = 1|x) + e^{f(x)}\rho(y = (-1)|x) = 0$$

$$e^{f(x)}\rho(y = (-1)|x) = e^{-f(x)}\rho(y = 1|x)$$

$$e^{2f(x)} = \frac{\rho(y = 1|x)}{\rho(y = (-1)|x)}$$

And using the *logarithmic function* log:

$$2f(x) = \log\left(\frac{\rho(y = 1|x)}{\rho(y = (-1)|x)}\right)$$

Leading to the minimizer equation:

$$f_*(x) = \frac{1}{2} \log\left(\frac{\rho(y = 1|x)}{\rho(y = (-1)|x)}\right) = \frac{\log(\rho(y = 1|x)) - \log(\rho(y = (-1)|x))}{2} \tag{73}$$

2.2.c) Logistic Loss

The logistic loss is given by the expression:

$$\ell(f(x), y) = \log\left(1 + e^{-yf(x)}\right) \tag{74}$$

Replacing it in the equation as before to find the minimizer:

$$\int_{\mathcal{Y}} \frac{\partial}{\partial f} \ell(f(x), y) d\rho(y|x) = \sum_{y \in \mathcal{Y}} \frac{\partial}{\partial f} \log\left(1 + e^{-yf(x)}\right) \rho(y|x) \tag{75}$$

$$= \sum_{y \in \mathcal{Y}} (-ye^{-yf(x)}) \times \frac{1}{1 + e^{-yf(x)}} \rho(y|x)$$

$$= \sum_{y \in \mathcal{Y}} \frac{-ye^{-yf(x)}}{1 + e^{-yf(x)}} \rho(y|x) \tag{76}$$

$$\frac{-e^{-f(x)}\rho(y = 1|x)}{1 + e^{-f(x)}} + \frac{e^{f(x)}\rho(y = (-1)|x)}{1 + e^{f(x)}}$$

And since $e^{-f(x)} = \dfrac{1}{e^{f(x)}}$, then by multiplying the left fraction's numerator and denominator by $e^{f(x)}$:

$$\frac{-e^{-f(x)}\rho(y = 1|x)}{1 + e^{-f(x)}} + \frac{e^{f(x)}\rho(y = (-1)|x)}{1 + e^{f(x)}} = \frac{\frac{1}{e^{-f(x)}}\rho(y = 1|x)}{1 + \frac{1}{e^{-f(x)}}} + \frac{e^{f(x)}\rho(y = (-1)|x)}{1 + e^{f(x)}}$$

$$= \frac{\rho(y=1|x)}{e^{f(x)}+1} + \frac{e^{f(x)}\rho(y=(-1)|x)}{1+e^{f(x)}}$$

And therefore, setting this expression to zero to find the minimizer f_*:

$$\frac{\rho(y=1|x)}{e^{f_*(x)}+1} + \frac{e^{f_*(x)}\rho(y=(-1)|x)}{1+e^{f_*(x)}} = 0 \tag{77}$$

Equivalent to:

$$\frac{\rho(y=1|x)}{e^{f_*(x)}+1} = \frac{e^{f_*(x)}\rho(y=(-1)|x)}{1+e^{f_*(x)}}$$
$$\rho(y=1|x) = e^{f_*(x)}\rho(y=(-1)|x)$$
$$e^{f_*(x)} = \frac{\rho(y=1|x)}{\rho(y=(-1)|x)} \tag{78}$$

By applying the *logarithmic function* on both sides as before, we obtain the minimizer f_*'s expression:

$$f_*(x) = \log\left(\frac{\rho(y=1|x)}{\rho(y=(-1)|x)}\right) = \log(\rho(y=1|x)) - \log(\rho(y=(-1)|x)) \tag{79}$$

2.2.d) Hinge Loss

The Hinge Loss is given by the following expression:

$$\ell(f(x),y) = \max(0, 1-yf(x)) \tag{80}$$

Unlike the other loss functions we studied up until now, it cannot be differentiated. Let's develop the expression of the expected risk with the Hinge Loss:

$$\mathcal{E}(f) = \int_{\mathcal{X}}\left(\int_{\mathcal{Y}} \max(0,1-yf(x))\, d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) = \int_{\mathcal{X}}\left(\sum_{y\in\mathcal{Y}} \max(0,1-yf(x))\,\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \tag{81}$$

Developing the inner integral part of the expression:

$$\sum_{y\in\mathcal{Y}} \max(0,1-yf(x))\,\rho(y|x) = \max(0,1-f(x))\,\rho(y=1|x) + \max(0,1+f(x))\,\rho(y=(-1)|x) \tag{82}$$

We can distinguish three cases as follows:

Case 1. $f(x) < -1$. In such case, we have:

$$\max(0,1-f(x)) = 1-f(x) \qquad \text{and} \qquad \max(0,1+f(x)) = 0 \tag{83}$$

Therefore, our expression becomes:

$$\mathcal{E}(f) = (1-f(x))\,\rho(y=1|x) \qquad f(x) < -1 \tag{84}$$

Case 2. $-1 \leq f(x) \leq 1$. In such case, we have:

$$\max(0,1-f(x)) = 1-f(x) \qquad \text{and} \qquad \max(0,1+f(x)) = 1+f(x) \tag{85}$$

Therefore, our expression becomes:

$$\mathcal{E}(f) = (1-f(x))\,\rho(y=1|x) + (1+f(x))\,\rho(y=(-1)|x) \qquad -1 < f(x) \leq 1 \tag{86}$$

Case 3. $1 < f(x)$. In such case, we have:

$$\max(0,1-f(x)) = 0 \qquad \text{and} \qquad \max(0,1+f(x)) = 1+f(x) \tag{87}$$

Therefore, our expression becomes:

$$\mathcal{E}(f) = (1+f(x))\,\rho(y=(-1)|x) \qquad -1 < f(x) \leq 1 \tag{88}$$

How we should interpret these results - Since both probabilities $\rho(y=1|x)$ and $\rho(y=(-1)|x)$ are greater than zero, the expected risk will have a higher absolute value in Cases 1 and 3, where the model predicts the opposite to the associated probability, compared to that of Case 2. Thus, it can be considered as a penalty for predicting the wrong label. Therefore, the optimal minimizer for the expected risk in the Hinge Loss framework model is comprised within the interval $f(x) \in [-1,1]$. The expected risk expression in Case 2 can be expanded as following:

$$\begin{aligned}
\forall f(x) \in [-1,1], \qquad \mathcal{E}(f) &= (1-f(x))\,\rho(y=1|x) + (1+f(x))\,\rho(y=(-1)|x) \\
&= \rho(y=1|x) - f(x)\rho(y=1|x) + \rho(y=(-1)|x) + f(x)\rho(y=(-1)|x) \\
\mathcal{E}(f) &= 1 + f(x)\left(\rho(y=(-1)|x) - \rho(y=1|x)\right) \qquad -1 \leq f(x) \leq 1 \qquad (89)
\end{aligned}$$

- When $\rho(y=1|x) > \rho(y=(-1)|x)$, the difference $\rho(y=(-1)|x) - \rho(y=1|x)$ will become negative. Therefore, to minimize the expected risk, we must have $f(x) = -1$.
- When $\rho(y=1|x) < \rho(y=(-1)|x)$, the difference $\rho(y=(-1)|x) - \rho(y=1|x)$ will become positive. To minimize the expected risk, we must have $f(x) = +1$.
- When $\rho(y=1|x) \approx \rho(y=(-1)|x)$, the expected risk (and thus the uncertainty) of the model cannot be minimized. It is referred as the *irreducible uncertainty*, because the model's prediction become random due to its modeling limitation.

Therefore, we can deduce the optimal minimizer for the expected risk under the Hinge Loss is the following:

$$f_*(x) = \begin{cases} 1 & \text{when } \rho(y=1|x) < \rho(y=(-1)|x) \\ -1 & \text{when } \rho(y=1|x) > \rho(y=(-1)|x) \end{cases} \qquad (90)$$

2.3

From Question 2.1, we know the Misclassification Error is given by:

$$R(c) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{1}_{c(x)\neq y}\, d\rho(x,y) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \mathbf{1}_{c(x)\neq y}\, d\rho(y|x) \right) d\rho_{\mathcal{X}}(x)$$

Since $\mathcal{Y} = \{-1,1\}$, then the inner integral can be evaluated pointwise $\forall x \in \mathcal{X}$ as we did for the expected risk. Therefore, it becomes:

$$R(c) = \int_{\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} \mathbf{1}_{c(x)\neq y}\, \rho(y|x) \right) d\rho_{\mathcal{X}}(x) \qquad (91)$$

And the inner sum becomes:

$$\sum_{y\in\mathcal{Y}} \mathbf{1}_{c(x)\neq y}\, \rho(y|x) = \mathbf{1}_{c(x)\neq 1}\, \rho(y=1|x) + \mathbf{1}_{c(x)\neq -1}\, \rho(y=(-1)|x) \qquad (92)$$

Thus, for the Misclassification Error to be minimized, we must have:

$$\forall x \in \mathcal{X}, \quad c(x) = y$$

Since we are assuming to know ρ a priori, then $c(x)$ must predict the label with the highest probability. Therefore, the *Bayes Decision rule* is given by:

$$c_*(x) = \begin{cases} 1 & \text{for } \rho(y=1|x) > \rho(y=(-1)|x) \\ -1 & \text{for } \rho(y=1|x) < \rho(y=(-1)|x) \end{cases} \qquad (93)$$

2.4

Regarding the *surrogate frameworks* defined in problem **(2.2)**, they are all Fischer consistent and do respect the Fischer's consistency main property, stating:

$$R(c_*) = \inf_{c:\mathcal{X}\to\{-1,1\}} R(c) \tag{94}$$

We would like to find, for each type of loss studied in the surrogate framework problem **(2.2)**, a mapping $d : \mathbb{R} \to \{-1, 1\}$ such that:

$$R(c_*(x)) = R\left[d(f_*(x))\right]$$

where f_* is the minimizer of the surrogate risk and $R(c_*)$ is defined as stated in *Fischer's consistency* property.

The mapping $d(f_*(x))$ needs to be equal to $c_*(x)$, therefore it needs to minimize the classification error in our problem.

2.4.a) Squared Loss

Using the minimizer's expression found before in 70 for the squared loss:

$$f_*(x) = \rho(y = 1|x) - \rho(y = (-1)|x)$$

Using this expression, we see that:

1. $\rho(y = 1|x) \geq \rho(y = (-1)|x)$ when $f_*(x) \geq 0$, and

2. $\rho(y = 1|x) < \rho(y = (-1)|x)$ when $f_*(x) < 0$

Therefore, we can deduce the optimal mapping which will minimize the Misclassification Error for the squared loss is:

$$d(f_*(x)) = \begin{cases} 1 & \text{if } f_*(x) \geq 0 \\ -1 & \text{if } f_*(x) < 0 \end{cases} \tag{95}$$

2.4.b) Exponential Loss

Using the minimizer's expression found before in 73 for the exponential loss:

$$f_*(x) = \frac{1}{2}\log\left(\frac{\rho(y = 1|x)}{\rho(y = (-1)|x)}\right)$$

Since $\log(1) = 0$, using this expression, we see that:

1. $\rho(y = 1|x) \geq \rho(y = (-1)|x)$ when $f_*(x) \geq 0$, and

2. $\rho(y = 1|x) < \rho(y = (-1)|x)$ when $f_*(x) < 0$

Therefore, we can deduce the optimal mapping which will minimize the Misclassification Error for the exponential loss is the same as the mapping for the squared loss:

$$d(f_*(x)) = \begin{cases} 1 & \text{if } f_*(x) \geq 0 \\ -1 & \text{if } f_*(x) < 0 \end{cases} \tag{96}$$

2.4.c) Logistic Loss

Recalling the minimizer's expression found in equation 79:

$$f_*(x) = \log\left(\frac{\rho(y = 1 \mid \mathbf{x})}{\rho(y = -1 \mid \mathbf{x})}\right)$$

Similarly to the exponential loss case, the minimizer follows the following conditions:

$$\rho(y = 1 \mid x) \geq \rho(y = -1 \mid x) \quad \text{for } f_*(x) \geq 0,$$
$$\rho(y = 1 \mid x) < \rho(y = -1 \mid x) \quad \text{for } f_*(x) < 0.$$

Thus, we deduce the optimal mapping for the logistic loss case to be defined as:

$$d(f_*(x)) = \begin{cases} 1 & \text{for } f_*(x) \geq 0, \\ -1 & \text{for } f_*(x) < 0. \end{cases}$$

2.4.d) Hinge Loss

Since the minimizer's expression in the case of Hinge Loss found in question 2.2.d) is given by:

$$f_*(x) = \begin{cases} 1 & \text{when } \rho(y=1|x) < \rho(y=(-1)|x) \\ -1 & \text{when } \rho(y=1|x) > \rho(y=(-1)|x) \end{cases} \tag{97}$$

Then, the minimizer in such case already does a mapping corresponding to the highest probability calculated for the two labels. We can still associate the minimizer's mapping to the sign function as following:

$$d(f_*(x)) = \text{sign}(f_*(x)) = \begin{cases} 1 & \text{for } f_*(x) \geq 0 \\ -1 & \text{for } f_*(x) < 0 \end{cases} \tag{98}$$

2.5. Comparison Inequality for Least Square Surrogates

We define the function $f_* : \mathcal{X} \to \mathbb{R}$ as the minimizer of the expected risk for the surrogate least squares classification problem.

We also define the sign function as:

$$\text{sign} : \mathbb{R} \to \{-1, 1\} \qquad\qquad \text{sign}(x) = \begin{cases} 1 & \text{for } x \geq 0, \\ -1 & \text{for } x < 0. \end{cases} \tag{99}$$

We would like to prove the comparison inequality:

$$0 \leq R(\text{sign}(f)) - R(\text{sign}(f_*)) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_*)} \tag{100}$$

2.5.1.

We would like to firstly show the intermediate step:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x) \tag{101}$$

where $\mathcal{X}_f = \{x \in \mathcal{X} \quad | \quad \text{sign}(f(x)) \neq \text{sign}(f_*(x))\}$

Starting from the left part of the equation, we firstly need to ascertain its domain of definition:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = |\mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f(x)) \neq y) - \mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f_*(x)) \neq y)|$$

$$= |\mathbb{P}_\rho(x) \left[\mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f(x)) = 1, y = (-1)|x) + \mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f(x)) = (-1), y = 1|x)\right]$$
$$-\mathbb{P}_\rho(x) \left[\mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f_*(x)) = 1, y = (-1)|x) + \mathbb{P}_{(x,y)\sim\rho}(\text{sign}(f_*(x)) = (-1), y = 1|x)\right] |$$

$$= |\mathbb{P}_\rho(x) \times (\mathbb{P}_{(x,y)\sim\rho}(y = (-1)|x)\mathbf{1}_{\text{sign}(f(x))\neq(-1)} + \mathbb{P}_{(x,y)\sim\rho}(y = 1|x)\mathbf{1}_{\text{sign}(f(x))\neq 1}$$
$$-\mathbb{P}_{(x,y)\sim\rho}(y = (-1)|x)\mathbf{1}_{\text{sign}(f_*(x))\neq(-1)} - \mathbb{P}_{(x,y)\sim\rho}(y = 1|x)\mathbf{1}_{\text{sign}(f_*(x))\neq 1})|$$

$$= |\mathbb{P}_\rho(x)| \times |\mathbb{P}_{(x,y)\sim\rho}(y = (-1)|x)\mathbf{1}_{\text{sign}(f(x))\neq(-1)} + \mathbb{P}_{(x,y)\sim\rho}(y = 1|x)\mathbf{1}_{\text{sign}(f(x))\neq 1}$$
$$-\mathbb{P}_{(x,y)\sim\rho}(y = (-1)|x)\mathbf{1}_{\text{sign}(f_*(x))\neq(-1)} - \mathbb{P}_{(x,y)\sim\rho}(y = 1|x)\mathbf{1}_{\text{sign}(f_*(x))\neq 1}|$$

And since $\mathbb{P}_\rho(x) \geq 0$, then $|\mathbb{P}_\rho(x)| = \mathbb{P}_\rho(x)$. Therefore, we notice that:

$$\text{If sign}(f(x)) = \text{sign}(f_*(x)), \quad \text{then } |R(\text{sign}(f)) - R(\text{sign}(f_*))| = 0 \tag{102}$$

Therefore, using the set defined earlier $\mathcal{X}_f$, and thus considering only the cases where $\text{sign}(f(x)) \neq \text{sign }(f_*(x))$, then we can take our expression found in question 2.1 and express it as:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f \times \mathcal{Y}} |\mathbf{1}_{\text{sign}(f(x)) \neq y} - \mathbf{1}_{\text{sign}(f_*(x)) \neq y}| d\rho(x,y) \tag{103}$$

$$= \int_{\mathcal{X}_f} | \sum_{y \in \mathcal{Y}} (\mathbf{1}_{\text{sign}(f(x)) \neq y} - \mathbf{1}_{\text{sign}(f_*(x)) \neq y}) \rho(y|x) | d\rho_{\mathcal{X}}(x)$$

$$= \int_{\mathcal{X}_f} |\rho(y=1|x) \left(\mathbf{1}_{\text{sign}(f(x)) \neq 1} - \mathbf{1}_{\text{sign}(f_*(x)) \neq 1}\right) - \rho(y=(-1)|x) \left(\mathbf{1}_{\text{sign}(f(x)) \neq (-1)} - \mathbf{1}_{\text{sign}(f_*(x)) \neq (-1)}\right) | d\rho_{\mathcal{X}}(x)$$

But since we are in the case of the least square loss (also called squared loss), then:

$$f_*(x) = \rho(y=1|x) - \rho(y=(-1)|x)$$

And since $\text{sign}(f(x)) \neq \text{sign}(f_*(x))$, thus there are two possible cases.

Case 1 - $\text{sign}(f(x)) = y$ and $\text{sign}(f_*(x)) \neq y$

In this case:

$$\mathbf{1}_{\text{sign}(f(x)) \neq 1} - \mathbf{1}_{\text{sign}(f_*(x)) \neq 1} = \mathbf{1}_{\text{sign}(f(x)) \neq -1} - \mathbf{1}_{\text{sign}(f_*(x)) \neq -1} = -1 \tag{104}$$

And therefore:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} | - \rho(y=1|x) + \rho(y=(-1)|x) | d\rho_{\mathcal{X}}(x)$$

$$= \int_{\mathcal{X}_f} | - f_*(x) | d\rho_{\mathcal{X}}(x)$$

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x) \tag{105}$$

Case 2 - $\text{sign}(f(x)) \neq y$ and $\text{sign}(f_*(x)) = y$

In this case:

$$\mathbf{1}_{\text{sign}(f(x)) \neq 1} - \mathbf{1}_{\text{sign}(f_*(x)) \neq 1} = \mathbf{1}_{\text{sign}(f(x)) \neq -1} - \mathbf{1}_{\text{sign}(f_*(x)) \neq -1} = 1 \tag{106}$$

And therefore:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |\rho(y=1|x) - \rho(y=(-1)|x)| d\rho_{\mathcal{X}}(x)$$

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x) \tag{107}$$

To conclude, we can see that in both cases, the final expression we wanted to prove is valid:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x) \tag{108}$$

2.5.2

We would like to prove the following:

$$\int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x) \leq \int_{\mathcal{X}_f} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \leq \sqrt{\mathbb{E}[|f(x) - f_*(x)|^2]} \tag{109}$$

Where $\mathbb{E}$ denotes the expectation with respect to $\rho_{\mathcal{X}}$.

Firstly, recalling the definition of $\mathcal{X}_f$ as:

$$\mathcal{X}_f = \{x \in \mathcal{X} \quad | \quad \text{sign}(f(x)) \neq \text{sign}(f_*(x))\}$$

Therefore we can treat the following two cases:

Case 1 : $f(x) > 0$ and $f_*(x) < 0$

Then:

$$f_*(x) - f(x) < f_*(x) < 0$$

which implies:

$$|f_*(x) - f(x)| > |f_*(x)| > 0 \tag{110}$$

Case 2 : $f(x) < 0$ and $f_*(x) > 0$

Then:

$$f_*(x) - f(x) > f_*(x) > 0$$

which implies:

$$|f_*(x) - f(x)| > |f_*(x)| > 0 \tag{111}$$

Therefore, we can state that we have:

$$\forall x \in \mathcal{X}_f \qquad |f_*(x) - f(x)| > |f_*(x)| > 0 \tag{112}$$

And hence:

$$\int_{x \in \mathcal{X}_{\{}} |f_*(x)| d\rho_{\mathcal{X}}(x) \leq \int_{x \in \mathcal{X}_{\{}} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \tag{113}$$

From there, using **Cauchy-Schwartz inequality**, which states the following:

$$\left(\int f(x) g(x) \right)^2 \leq \int f^2(x) dx \int g^2(x) dx \tag{114}$$

Then in our case, since:

$$\int_{x \in \mathcal{X}_{\{}} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) = \sqrt{\left(\int_{x \in \mathcal{X}_{\{}} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \right)^2}$$

we therefore have (reminding that the square-root function is continuous and strictly increasing):

$$\sqrt{\left(\int_{x \in \mathcal{X}_{\{}} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \right)^2} \leq \sqrt{\int_{\mathcal{X}_f} |f_*(x) - f(x)|^2 d\rho_{\mathcal{X}}(x) \int_{\mathcal{X}_f} d\rho_{\mathcal{X}}(x)} \tag{115}$$

$$= \sqrt{\int_{\mathcal{X}_f} |f_*(x) - f(x)|^2 d\rho_{\mathcal{X}}(x)}$$

$$= \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}} \left(|f_*(x) - f(x)|^2 \right)}$$

And since the signs of $f(x)$ and $f_*(x)$ are opposite, then:

$$|f_*(x) - f(x)| = |f(x) - f_*(x)|$$

Therefore, our expression becomes:

$$\sqrt{\mathbb{E}_{\rho_{\mathcal{X}}} \left(|f(x) - f_*(x)|^2 \right)} \tag{116}$$

Therefore, we proved the following:

$$\int_{x \in \mathcal{X}_f} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \leq \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}} \left(|f(x) - f_*(x)|^2 \right)} \tag{117}$$

And thus, we proved overall the following expression we intended to prove:

$$\int_{x \in \mathcal{X}_{\{}} |f_*(x)| d\rho_{\mathcal{X}}(x) \leq \int_{x \in \mathcal{X}_f} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \leq \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}} \left(|f(x) - f_*(x)|^2 \right)} \tag{118}$$

2.5.3

We would like to prove the following:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}}\left(|f(x) - f_*(x)|^2\right)} \tag{119}$$

Let's start from the left side of the equation, using the expression for the expected risk over $\mathcal{X}_f$. Firstly, this expression $\forall x \in \mathcal{X}$ is:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(f_*(x), y) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x)$$

Thus, $\forall x \in \mathcal{X}_f$:

$$\begin{aligned} \mathcal{E}(f) - \mathcal{E}(f_*) &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) - \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \ell(f_*(x), y) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \qquad (120) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \ell(f(x), y) d\rho(y|x) - \int_{\mathcal{Y}} \ell(f_*(x), y) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \end{aligned}$$

And since we find ourselves in the case of the *expected risk* for the *surrogate least square problem*, then:

$$\begin{cases} \ell(f(x), y) &= (f(x) - y)^2 \\ \ell(f_*(x), y) &= (f_*(x) - y)^2 \end{cases} \tag{121}$$

Therefore, our equation becomes:

$$\begin{aligned} \mathcal{E}(f) - \mathcal{E}(f_*) &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left(\ell(f(x), y) - \ell(f_*(x), y)\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left((f(x) - y)^2 - (f_*(x) - y)^2\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left((f(x) - y)^2 - (f_*(x) - y)^2\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left(f^2(x) - 2yf(x) + y^2 - f_*^2(x) + 2yf_*(x) - y^2\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left(f^2(x) - 2yf(x) + y^2 - f_*^2(x) + 2yf_*(x) - y^2\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} \left(\int_{\mathcal{Y}} \left(f^2(x) - f_*^2(x) - 2y(f(x) - f_*(x))\right) d\rho(y|x)\right) d\rho_{\mathcal{X}}(x) \end{aligned}$$

Defining the expectation with respect to $\rho(y|x)$ in general:

$$\mathbb{E}_{\rho(y|x)}[X] = \int_{\mathcal{Y}} X d\rho(y|x) \tag{122}$$

And noticing that for y, we have:

$$\mathbb{E}_{\rho(y|x)}[y] = \sum_{y \in \mathcal{Y}} y\rho(y|x) = \rho(y = 1|x) - \rho(y = (-1)|x) = f_*(x) \tag{123}$$

Since we found this expression in question 2.2.a).

Then our expression becomes:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \int_{\mathcal{X}_f} \mathbb{E}_{\rho(y|x)}\left[f^2(x) - (\mathbb{E}_{\rho(y|x)}[y])^2 - 2y(f(x) - \mathbb{E}_{\rho(y|x)}[y])\right] d\rho_{\mathcal{X}}(x) \tag{124}$$

And since $f(x)$ does not depend on y and $\mathbb{E}_{\rho(y|x)}[\mathbb{E}_{\rho(y|x)}[y]] = \mathbb{E}_{\rho(y|x)}[y]$, we get:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \int_{\mathcal{X}_f} f^2(x) - (\mathbb{E}_{\rho(y|x)}[y])^2 - 2\mathbb{E}_{\rho(y|x)}[y]f(x) + (\mathbb{E}_{\rho(y|x)}[y])^2 d\rho_{\mathcal{X}}(x)$$

$$= \int_{\mathcal{X}_f} f^2(x) - (\mathbb{E}_{\rho(y|x)}[y])^2 - 2\mathbb{E}_{\rho(y|x)}[y]f(x) + 2(\mathbb{E}_{\rho(y|x)}[y])^2 d\rho_{\mathcal{X}}(x)$$

$$= \int_{\mathcal{X}_f} \left(f^2(x) - 2\mathbb{E}_{\rho(y|x)}[y]f(x) + (\mathbb{E}_{\rho(y|x)}[y])^2\right) d\rho_{\mathcal{X}}(x)$$

$$= \int_{\mathcal{X}_f} \left(f(x) - \mathbb{E}_{\rho(y|x)}[y]\right)^2 d\rho_{\mathcal{X}}(x) \tag{125}$$

$$= \int_{\mathcal{X}_f} \left(f(x) - f_*(x)\right)^2 d\rho_{\mathcal{X}}(x) \tag{126}$$

$$= \int_{\mathcal{X}_f} |f(x) - f_*(x)|^2 d\rho_{\mathcal{X}}(x) \tag{127}$$

$$= \mathbb{E}\left[|f(x) - f_*(x)|^2\right] \tag{128}$$

Therefore, we proved the expression we were supposed to prove:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \mathbb{E}\left[|f(x) - f_*(x)|^2\right] \tag{129}$$

In summary we proved in **subsection 2.5.1** that:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| = \int_{\mathcal{X}_f} |f_*(x)| d\rho_{\mathcal{X}}(x)$$

But since in **subsection 2.5.2**, we demonstrated the following:

$$\int_{x \in \mathcal{X}_{\{}} |f_*(x)| d\rho_{\mathcal{X}}(x) \leq \int_{x \in \mathcal{X}_f} |f_*(x) - f(x)| d\rho_{\mathcal{X}}(x) \leq \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}}\left(|f(x) - f_*(x)|^2\right)}$$

Therefore, we can assert:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| \leq \sqrt{\mathbb{E}_{\rho_{\mathcal{X}}}\left(|f(x) - f_*(x)|^2\right)} \tag{130}$$

And since, in **subsection 2.5.3**, we showed:

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \mathbb{E}\left[|f(x) - f_*(x)|^2\right]$$

Therefore, we can conclude:

$$|R(\text{sign}(f)) - R(\text{sign}(f_*))| \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_*)} \tag{131}$$

And since the left part represents the norm of $R(\text{sign}(f)) - R(\text{sign}(f_*))$ that will always be positive, we can therefore add:

$$0 \leq |R(\text{sign}(f)) - R(\text{sign}(f_*))|$$

Therefore, we proved throughout *section 2.5* the **comparison inequality** for the **surrogate least squares classification problem** obtained in *section 2.2*:

$$0 \leq |R(\text{sign}(f)) - R(\text{sign}(f_*))| \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_*)} \tag{132}$$

PART III

1.

The *One-versus-Rest* (OvR) method is a technique to generalize a two-class classifier for k classes by training k separate binary classifiers. Each classifier is trained to distinguish one class as positive $(+1)$ while treating all other $k-1$ classes as negative (-1). During prediction, all k classifiers evaluate the input, producing confidence scores or decision values. The final class prediction corresponds to the classifier with the highest confidence score. This method is popular due to its simplicity and effectiveness. However, it requires training k classifiers, which can be computationally demanding for large k, and the confidence scores of individual classifiers may not be calibrated to represent probabilities accurately.

2.

To generalize Kernel Perceptron to k-class classification, we can follow the steps. First, train k separate Kernel Perceptron classifiers, one for each class. Each classifier is responsible for distinguishing one class (+1) from all other classes (-1). Then during prediction, evaluate all k classifiers for a given input and assign the class corresponding to the classifier with the highest confidence score.

Representation The weight $w(\cdot)$ is not explicitly stored as a vector. Instead, it is represented as a sum of the coefficients α_i (stored for each training sample x_i) multiplied by the kernel evaluations $K(x_i, \cdot)$. The kernel function $K(x_i, x)$ computes the similarity between the stored training sample x_i and the input x.

Evaluation: To evaluate $w(x)$ for a given input x: Loop through all stored training samples (x_i, α_i). Then compute the kernel function $K(x_i, x)$ for each sample and multiply the result by α_i and sum the results:

$$w(x) = \sum_{i=0}^{m} \alpha_i K(x_i, x).$$

Adding new terms to the sum during training When a sample (x_t, y_t) is misclassified, the following steps occur

- Compute the kernel $K(x_t, \cdot)$ for the current sample x_t.
- Update the coefficient $\alpha_t = y_t$ (positive or negative depending on the true label).
- Add the term $\alpha_t K(x_t, \cdot)$ to the representation of $w(\cdot)$.

Number of Epochs The number of epochs refers to how many times the model iterates over the training dataset. Too few epochs may lead to underfitting, where the model fails to learn the decision boundary, while too many epochs can result in overfitting, where the model adapts too closely to the training data and fails to generalize. For this task, start with a moderate number of epochs (e.g., 5) and monitor the training error. If the error stabilizes (no updates to α_t), additional epochs are unnecessary. If convergence is not reached, incrementally increase the number of epochs.

3.

The results in The table 1 demonstrate the performance of polynomial kernel perceptrons across degrees $d = 1, \dots, 7$ in terms of training and testing error rates. For $d = 1$, the test error is relatively high at $9.46\% \pm 0.70\%$, reflecting underfitting due to low model complexity. As d increases, the test error decreases significantly, reaching a minimum of $2.63\% \pm 0.05\%$ at $d = 5$. This indicates that the model captures the data patterns most effectively at this degree. Beyond $d = 5$, the test error begins to stabilize or slightly increase (e.g., $3.06\% \pm 0.22\%$ at $d = 7$), which suggests overfitting as the model becomes overly complex and loses its ability to generalize. The small standard deviations across all degrees indicate consistent performance over multiple runs.

Degree	**Train Error (Mean ± Std) (%)**	**Test Error (Mean ± Std) (%)**
1	8.43 ± 0.97	9.46 ± 0.70
2	0.84 ± 0.18	3.49 ± 0.70
3	0.11 ± 0.02	2.72 ± 0.30
4	0.08 ± 0.04	3.09 ± 0.30
5	0.03 ± 0.00	2.63 ± 0.05
6	0.03 ± 0.01	3.01 ± 0.11
7	0.03 ± 0.03	3.06 ± 0.22

Table 1: Train and Test Error Rates (Mean ± Std) for Polynomial Kernel Perceptron

4.

The results in table 2 of the 20 cross-validation runs show a mean train error of $0.05\% \pm 0.02\%$ and a mean test error of $2.84\% \pm 0.38\%$, with the best polynomial degree averaging at 5.3 ± 0.78. The low train error indicates that the model fits the training data very well, while the slightly higher test error reflects generalization to unseen data. The optimal degree of $d^* = 5.3$ suggests that moderate complexity in the polynomial kernel captures the patterns in the data effectively. Increasing the degree beyond this point may lead to overfitting, as the model becomes too tailored to the training set.

Metric	**Mean ± Std (%)**
Best Degree	5.30 ± 0.78
Train Error	0.05 ± 0.02
Test Error	2.84 ± 0.38

Table 2: Summary Results of Cross-Validation over 20 Runs

5.

The confusion matrix in Table 3 reveals that the model achieves high classification accuracy, with most confusion rates being below 1%. Errors are more prevalent between visually similar digits, such as *3* and *5*, *3* and *7* or *4* and *3*, which share structural similarities. Standard deviations for confusion rates are consistently low, indicating stable performance across the 20 runs. Overall, the off-diagonal elements show sparse errors, demonstrating the model's robustness in distinguishing distinct digit classes while highlighting challenges in differentiating similar ones.

	0	1	2	3	4	5	6	7	8	9
0	0.00 ± 0.00	0.07 ± 0.13	0.06 ± 0.13	0.11 ± 0.18	0.05 ± 0.12	0.12 ± 0.18	0.25 ± 0.21	0.05 ± 0.12	0.11 ± 0.15	0.10 ± 0.15
1	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.23	0.04 ± 0.17	0.31 ± 0.29	0.02 ± 0.08	0.17 ± 0.33	0.04 ± 0.17	0.06 ± 0.14	0.04 ± 0.14
2	0.40 ± 0.38	0.27 ± 0.41	0.00 ± 0.00	0.51 ± 0.64	0.64 ± 0.50	0.16 ± 0.25	0.16 ± 0.30	0.62 ± 0.41	0.18 ± 0.35	0.62 ± 0.35
3	0.32 ± 0.29	0.30 ± 0.30	0.59 ± 0.53	0.00 ± 0.00	1.64 ± 1.07	0.03 ± 0.14	0.58 ± 0.61	0.33 ± 0.40	0.12 ± 0.30	0.24 ± 0.33
4	0.26 ± 0.44	0.54 ± 0.41	0.52 ± 0.79	0.03 ± 0.12	0.00 ± 0.00	0.80 ± 0.65	0.93 ± 0.68	0.45 ± 0.64	0.50 ± 0.68	0.24 ± 0.33
5	0.82 ± 0.74	0.14 ± 0.28	0.36 ± 0.55	1.12 ± 0.81	0.71 ± 0.91	0.00 ± 0.00	0.22 ± 0.30	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
6	0.83 ± 0.80	0.22 ± 0.30	0.24 ± 0.38	0.00 ± 0.00	0.00 ± 0.00	0.13 ± 0.25	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
7	0.10 ± 0.24	0.33 ± 0.44	0.29 ± 0.51	1.57 ± 1.14	0.53 ± 0.49	0.21 ± 0.39	0.21 ± 0.39	0.00 ± 0.00	0.24 ± 0.33	0.00 ± 0.00
8	0.60 ± 0.40	0.46 ± 0.55	0.50 ± 0.52	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.32	0.00 ± 0.00	0.63 ± 0.71	0.00 ± 0.00	0.00 ± 0.00
9	0.28 ± 0.46	0.09 ± 0.23	0.15 ± 0.29	0.21 ± 0.30	0.22 ± 0.30	0.94 ± 0.91	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 3: Averaged Confusion Matrix over 20 Runs (Values in percentages %)

6.

The difficulty in predicting these specific images arises from several factors inherent to the dataset and the nature of pixelated image classification. First, many of the misclassified digits appear visually ambiguous due to distortion, pixelation, or missing details. For example, the digit '6" predicted as '1" shows a lack of connectivity in its lower loop, which makes it resemble a vertical stroke. Similarly, the digit '9" predicted as '7" suffers from an incomplete or faint curvature, leading the model to confuse it with a straight-lined digit. Second, some digits naturally share similar shapes, especially in low-resolution images. For instance, '5" predicted as '8" demonstrates how overlapping features such as curved tops and closed loops can mislead the classifier. The digit '4" misclassified as '9" has a vertical stroke that can resemble the curvature of '9" in pixelated form. Lastly, the pixelation limits the model's ability to capture finer details and distinguish between subtle variations in shape. While the model performs well on clearer samples, it struggles with these edge cases, which reflect the inherent challenge of generalizing to ambiguous data.

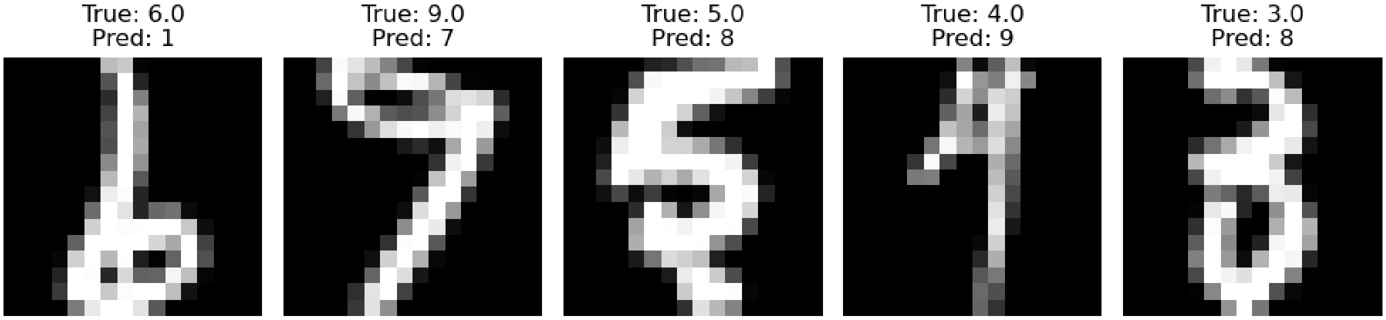

Figure 1: The visualization of the five digits along with their labels

7.

(a) The chosen set of c values, $[0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10]$, is designed to explore a broad range of kernel widths for the Gaussian kernel. The parameter c determines the 'width" of the kernel, influencing how the model weighs training points based on their distance. Small c values, such as 0.01 and 0.05, correspond to broader kernels, where training points have a wider influence, capturing more global patterns in the data. Moderate c values, such as $0.1, 0.5,$ and 1, provide a balance between global and local influences. Larger c values, such as $2, 5,$ and 10, result in narrower kernels, focusing on more localized patterns and emphasizing nearby training points. This range ensures that the model is evaluated across different levels of locality, from capturing broader patterns to focusing on finer details.

(b) The Table 4 displays the mean train and test errors for various values of the Gaussian kernel parameter c. The train error decreases and reaches zero as c increases, demonstrating that the model becomes increasingly overfitted to the training data for larger c values. The test error, however, exhibits a minimum at $c = 0.01$, achieving the best performance at $2.83\% \pm 0.38\%$. For $c > 0.01$, the test error begins to rise, which indicates overfitting as the model prioritizes minimizing the training error at the expense of generalization. Notably, at $c = 10$, the test error increases significantly to $19.52\% \pm 0.71\%$, further indicating the detrimental effects of overfitting.

c	**Train Error (%)**	**Test Error (%)**
0.01	0.0565 ± 0.0193	2.8306 ± 0.3848
0.05	0.0202 ± 0.0246	3.9489 ± 0.3676
0.1	0.0175 ± 0.0217	5.3360 ± 0.6223
0.5	0.0000 ± 0.0000	6.8118 ± 0.5600
1	0.0000 ± 0.0000	6.8011 ± 0.5394
2	0.0000 ± 0.0000	6.8414 ± 0.5173
5	0.0067 ± 0.0293	6.8522 ± 0.6331
10	0.0000 ± 0.0000	19.5215 ± 0.7161

Table 4: Train and Test Errors with Mean ± Standard Deviation for Gaussian Kernel

(c) The results in Table 5 from the cross-validation procedure using the Gaussian kernel indicate that the optimal value for the parameter c is consistently selected as $c^* = 0.01 \pm 0.00$ across all 20 runs. This value corresponds to the smallest tested c, which leads to a minimal mean train error of $0.06\% \pm 0.02\%$ and a mean test error of $2.83\% \pm 0.38\%$. The small train error reflects that the model is effectively fitting the training data, while the low test error suggests good generalization performance.

Metric	**Value (mean ± std)**
Mean selected c^*	0.01 ± 0.00
Mean Train Error	$0.06\% \pm 0.02\%$
Mean Test Error	$2.83\% \pm 0.38\%$

Table 5: Summary of Gaussian Kernel Results for Cross-Validation

(d) The Gaussian kernel consistently achieves lower test errors compared to the polynomial kernel, with a mean test error of $2.83\% \pm 0.38\%$ for the Gaussian kernel versus a minimum test error of $2.63\% \pm 0.05\%$ at $d = 5$ for the polynomial kernel. The Gaussian kernel's flexibility in adapting to data geometry due to its localized influence offers superior performance in this dataset. In contrast, the polynomial kernel's performance is more sensitive to the choice of degree, and overfitting is observed for higher degrees. Overall, the Gaussian kernel demonstrates better generalization, particularly for datasets with complex patterns, while the polynomial kernel requires careful tuning to achieve comparable performance.

8.

(a) The **One-versus-One (OvO)** method is a multiclass classification approach that breaks down the problem into multiple binary classification tasks. For k classes, it trains $\frac{k(k-1)}{2}$ classifiers, each focusing on distinguishing between two specific classes while ignoring the rest. During prediction, every classifier votes for one of its two classes, and the class with the most votes across all classifiers is chosen as the final output. OvO is computationally efficient for binary classifiers like SVMs, as each classifier only works with a subset of the data. However, it can become expensive for large k due to the quadratic growth in the number of classifiers.

(b) The results from the One-versus-One (OvO) method with the polynomial kernel for degrees $d = 1, \ldots, 7$ has shown in Table 6. It shows consistent improvements in both train and test error rates as the degree increases. For lower degrees (e.g., $d = 1$), the test error is relatively high, indicating underfitting. As the degree increases to $d = 4$ or $d = 5$, the test error reaches its minimum, suggesting that the model achieves a good balance between complexity and generalization. Beyond $d = 5$, the train error continues to decrease slightly, while the test error remains stable, indicating potential overfitting is minimal.

Degree (d)	**Train Error (% Mean ± Std)**	**Test Error (% Mean ± Std)**
1	4.31 ± 1.00	6.85 ± 1.04
2	0.65 ± 0.26	3.84 ± 0.53
3	0.28 ± 0.13	3.44 ± 0.47
4	0.08 ± 0.04	3.29 ± 0.49
5	0.07 ± 0.06	3.25 ± 0.51
6	0.05 ± 0.03	3.36 ± 0.42
7	0.03 ± 0.02	3.42 ± 0.40

Table 6: Train and Test Errors (%) with Mean ± Standard Deviation for the One-versus-One Method using the Polynomial Kernel.

(c) The results shown in Table 7 demonstrate that the One-Versus-One classification method with the polynomial kernel achieves excellent performance. The mean train error is extremely low at 0.08% ± 0.06%, indicating effective model training and overfitting prevention. Similarly, the mean test error is 3.27% ± 0.50%, suggesting robust generalization across datasets. The best degree (d^*) is 4.75 ± 0.94, which aligns with the observation that mid-range polynomial degrees generally balance model complexity and overfitting. Overall, this approach provides a reliable and accurate classification framework.

Metric	**Mean (%)**	**Std Dev (%)**
Train Error	0.0800	0.0600
Test Error	3.2700	0.5000
Best Degree (d^*)	4.7500	0.9421

Table 7: Mean and standard deviation of train and test errors with the best degree (d^*) using One-Versus-One polynomial kernel cross-validation.

(d) Comparing the One-versus-One (OvO) and One-versus-Rest (OvR) methods using the polynomial kernel highlights their distinct strengths. The OvR method achieves the lowest test error at $d = 5$, with $2.63\% \pm 0.05\%$, slightly outperforming the OvO method, which has a test error of $3.25\% \pm 0.51\%$ at the same degree. However, the OvR method exhibits consistently lower training errors. The OvO method, while achieving marginally higher test errors, provides more consistent results across degrees. Additionally, OvO handles pairwise class separation, which may lead to improved robustness in multiclass classification but at the cost of increased computational complexity. OvO requires $\binom{k}{2}$ classifiers, making it more expensive than the k classifiers required for OvR. In conclusion, OvR offers slightly better test performance at optimal degrees and is computationally efficient, whereas OvO provides stability and robustness in scenarios where computational resources are not a limiting factor.