## Theoretical question:

1. We will prove that $\mathcal{H}_{ball}$ is PAC learnable in the realize case:

   Because we assume realizability, there is $h_{R^*} \in \mathcal{H}_{ball}$ s.t $Y = h_{R^*}(X)$.
   This means that every positive label samples are inside some ball with radius $R^*$.

   We will propose the next algorithm $\mathcal{A}_{ball}$:
   Given samples $S$ drawn i.i.d from $\mathbb{R}^d \times \{0,1\}$, $\mathcal{A}_{ball}$ do the following:
   1. Find the maximal $\ell_2$ norm of a positive labeled sample in $S$.
   2. Return the hypothesis $h_R \in \mathcal{H}_{ball}$ that corresponds to the maximal norm.

   In another words, $\mathcal{A}_{ball}$ returns the smallest ball that contains positive samples in $S$.
   We know that it exists because we assumed realizability.

   We now prove that $\mathcal{H}_{ball}$ is PAC using $\mathcal{A}_{ball}$:

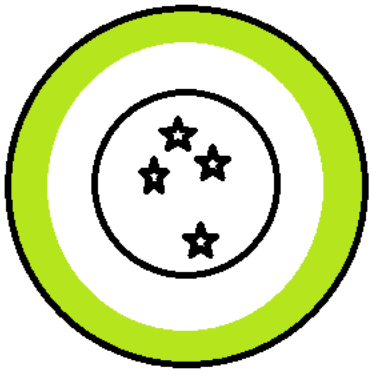
   Let there be samples $S = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d from $\mathbb{R}^d \times \{0,1\}$ with size $n$, $\epsilon, \delta \in (0,1)$, and a reliable distribution $P$.
   We mark $h_R \overset{\text{def}}{=} \mathcal{A}_{ball}(S)$ to be the hypothesis $\mathcal{A}_{ball}$ returned.

   $h_R$ can mistake only by labeling positive samples as negative (due to realizability).
   If we denote $B_R$ the d-dimensional ball with radius d, we receive that:

   $$e_p(h_R) = P[B_{R^*} \backslash B_R]$$

   I will assume that $P[B_{R^*} \backslash B_R] > \epsilon$ (otherwise we have nothing to prove).

   We will define a strip $T$ ranging from $B_{R^*}$ *to* $B_R$ in the following way:

   Where $T$ is the green strip, the larger ball is $B_{R^*}$, and the smaller ball is $B_R$.
   Because $P$ is continuous, we can take $T$ s.t $P[T] = \epsilon$.

   We notice that if there is $x_i$ s.t $x_i \in T$, then $B_{R^*} \backslash B_R \subset T$.
   Therefore, in this case $e_p(h_R) = P[B_{R^*} \backslash B_R] \leq P[T] = \epsilon$.

This means that $e_p(h_R) > \epsilon$ implies that $\forall 1 \leq i \leq n,\ x_i \notin T$.
Hence:

$$P\left[e_p(h_R) > \epsilon\right] \leq P[\forall 1 \leq i \leq n, x_i \notin T] = \prod_{i=1}^{n} P[x_i \notin T] = P[x \notin T]^n$$

Where the last two equalities holds because $S$ is drawn i.i.d.

From the definition of $T$, $P[x \notin T] = 1 - \epsilon$.
So:

$$P\left[e_p(h_R) > \epsilon\right] \leq (1-\epsilon)^n \leq e^{-\epsilon n} < \delta$$

And to achieve that:

$$e^{-\epsilon n} < \delta \Leftrightarrow e^{\epsilon n} > \frac{1}{\delta} \Leftrightarrow n > \frac{ln\frac{1}{\delta}}{\epsilon}$$

This means that the sample complexity $m_h$ is bounded by $\left\lceil \frac{ln\frac{1}{\delta}}{\epsilon} \right\rceil$.
We can see that the sample complexity does not depend over the dimension d.

2. ⇒

Let's assume that $\mathcal{H}$ is PAC learnable, we will prove that $\mathcal{H}$ is PAC learnable in expectation.

Because $\mathcal{H}$ is PAC learnable, there is $m_h$: $(0,1)^2 \to \mathbb{N}$ and an algorithm $\mathcal{A}$ s.t for any distribution $P$ realizable w.r.t $\mathcal{H}$, and any sample space $S$ with $|S| \geq m_h(\epsilon, \delta)$, $P\left[e_p\left(\mathcal{A}(S)\right) > \epsilon\right] < \delta$.

Let there be $a \in (0,1)$.
Because $\mathcal{H}$ is PAC learnable it follows that for sample space $S$ with $|S| \geq m_h\left(\frac{a}{2}, \frac{a}{2}\right)$, $P\left[e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] < \frac{a}{2}$.

Using the law of total expectation:

$$\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right] = \mathbb{E}\left[e_p\left(\mathcal{A}(S)\right) | e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] P\left[e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] + \mathbb{E}\left[e_p\left(\mathcal{A}(S)\right) | e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}\right] P\left[e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}\right]$$

We know that $P\left[e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] < \frac{a}{2}$ from PAC, and it follows that $\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right) | e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}\right] \leq \frac{a}{2}$ because the expected value of $e_p\left(\mathcal{A}(S)\right)$ given that $e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}$ is at most $\frac{a}{2}$.

Thus

$$\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right] \leq \mathbb{E}\left[e_p\left(\mathcal{A}(S)\right) | e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] \frac{a}{2} + \frac{a}{2} P\left[e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}\right]$$

We know that $P\left[e_p\left(\mathcal{A}(S)\right) \leq \frac{a}{2}\right] \leq 1$, and $\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right) | e_p\left(\mathcal{A}(S)\right) > \frac{a}{2}\right] \leq 1$, because all probabilities are bounded by 1, therefore we conclude that

$$\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right] \leq \frac{a}{2} + \frac{a}{2} = a$$

If we define $N$: $(0,1) \to \mathbb{N}, N(a) = m_h(\frac{a}{2}, \frac{a}{2})$ we will get that $\mathcal{H}$ is PAC learnable in expectation with the same algorithm $\mathcal{A}$.

⇐

Let's assume that $\mathcal{H}$ is PAC learnable in expectation, we will prove that $\mathcal{H}$ is PAC learnable.

Because $\mathcal{H}$ is PAC learnable in expectation, there is an algorithm $\mathcal{A}$ s.t for PAC in expectation assumption holds.
Let there be $\epsilon, \delta \in (0,1)$.
We need to find out $m_h: (0,1)^2 \to \mathbb{N}$ s.t for any distribution $P$ realizable w.r.t $\mathcal{H}$, and any sample space $S$ with $|S| \geq m_h(\epsilon, \delta)$, $P\left[e_p\left(\mathcal{A}(S)\right) > \epsilon\right] < \delta$.

Using the Markov inequality (I assume that the loss function range is $\mathbb{R}^+$):

$$P\left[e_p\left(\mathcal{A}(S)\right) > \epsilon\right] \leq \frac{\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right]}{\epsilon}$$

If we take $S$ s.t $|S| \geq N\left(\frac{\epsilon\delta}{2}\right)$ (because $\epsilon, \delta \in (0,1), \epsilon\delta \in (0,1)$)
we will get that

$$\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right] \leq \frac{\epsilon\delta}{2}$$

Hence $P\left[e_p\left(\mathcal{A}(S)\right) > \epsilon\right] \leq \frac{\mathbb{E}\left[e_p\left(\mathcal{A}(S)\right)\right]}{\epsilon} \leq \frac{\epsilon\delta}{2\epsilon} < \delta$
So, if we define $m_h: (0,1)^2 \to \mathbb{N}, m_h(\epsilon, \delta) = N(\epsilon\delta)$ we will get that $\mathcal{H}$ is PAC learnable with the same algorithm $\mathcal{A}$.

3. We will prove that the VC-dimension of $\mathcal{H}_k$ is $2k$:

   To show that we I will show for all $k$, there is a set of size $2k$ that shattered by $\mathcal{H}_k$, And prove that $VCdim(\mathcal{H}_k) < 2k + 1$.

   ***Lemma***: $VCdim(\mathcal{H}_k) < 2k + 1$

   ***Proof***: By contradiction, assume that $VCdim(\mathcal{H}_k) \geq 2k + 1$.
   So, there is a set $C = \{x_i\}_{i=1}^{2k+1}$ of size $2k + 1$ that $\mathcal{H}_k$ shatters.
   W.l.o.g we assume that $x_1 < x_2 < \cdots < x_{2k+1}$.
   For $1 \leq i \leq 2k + 1$, we define $s_i = \begin{cases} 1 & i \text{ is odd}, \\ 0 & i \text{ is even} \end{cases}$. Let us look on $s \stackrel{\text{def}}{=} \{s_i\}_{i=1}^{2k+1}$.

   Let there be the hypothesis $h \in \mathcal{H}_k$ that achieves $s$.
   We denote $h = \{[l_1, u_1], \dots, [l_k, u_k]\}, 0 \leq l_1 \leq u_1 \leq l_2 \leq u_2 \leq \dots \leq u_k \leq 1\}$.

   We can see that $|\{\, 1 \leq i \leq 2k + 1 \mid s_i = 1 \,\}| = k + 1$, so by pigeonhole principle there must be $1 \leq i < j \leq 2k + 1$ s.t $x_i, x_j \in [l_m, u_m]$for $m \in \{1, \dots, k\}$.

   From the definition of $s$, it follows that $j > i + 1$, therefor because $h(x_i) = 1$, follows that $h(x_{i+1}) = 0$. But because $x_i < x_{i+1} < x_j$, it follows that $x_{i+1} \in [l_m, u_m]$, and this is contradiction.

   ***Lemma***: $VCdim(\mathcal{H}_k) \geq 2k$

   To do I will claim the next proposition:

   > ***Lemma*:** Let there be $(s_1, \dots, s_{2n}) \in \{0, 1\}^{2n}$, then the number of maximal chains of 1's is at most n.
   >
   > Given $(s_1, \dots, s_n) \in \{0, 1\}^n$, a **chain of 1's** is a sequence of consecutives 1's.
   > E.g., in (1,1,0,1) we have 4 chains of 1's:
   > - (1,1) in positions 1 and 2
   > - (1) in position 1
   > - (1) in position 2
   > - (1) in position 4
   >
   > I will call a chain maximal if it's not part of another chain. In the previous example we have 2 maximal chains:
   > - (1,1) in positions 1 and 2
   > - (1) in position 4

***Proof*:**
For $n = 1$:
We can see that $\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$, and the maximal number of chains of 1's is indeed $n = 1$(in $(0,1)$ and $(1,0)$ and $(1,1)$ we can see a maximal chain with length 1).

Assuming the induction assumption for $n-1$, now we will prove for $n$:
Let there be $s = (s_1, \dots, s_{2n-2}, s_{2n-1}, s_{2n}) \in \{0,1\}^{2n}$.

According to the induction assumption, the maximal number of chains of 1's in $(s_1, \dots, s_{2n-2})$ is at most $n-1$.

If $s_{2n-2} = 0$ then we can have only one more maximal chain (by $(s_{2n-1}, s_{2n}) \in \{(1,0),(0,1),(1,1)\}$), thus, the number of maximal chains in $s$ is at most $n$.

If $s_{2n-2} = 1$ then we can create only one more maximal chain (by $(s_{2n-1}, s_{2n}) = (0,1)$), thus, the number of maximal chains in $s$ is at most $n$.

Now we will prove that $VCdim(\mathcal{H}_k) \geq 2k$

***Proof***: let there be a set $C = \{x_i\}_{i=1}^{2k}$ s.t $x_1 < x_2 < \cdots < x_{2k}$.
Let there be $(s_1, s_2, \dots, s_{2k}) \in 2^{2k}$, we will show that
there is $h \in \mathcal{H}_k$ s.t $\forall\ 1 \leq i \leq 2k, h(x_i) = s_i$.

We will define the parameter of $C$ to be $\Delta(C) \stackrel{\text{def}}{=} \min_{x,y \in C} |x - y|$

Using the above lemma, we know that the number of maximal chains of 1's in $s$ it at most $k$.
Let us call the set $Ch = \{\, c \mid c\ is\ maximal\ chain\ in\ C \,\}$ as the set of maximal chains.
We will create interval that fits to a maximal chain:

given $c = (s_i, \dots, s_j) \in Ch$:
if $|c| = 1$: we will create the interval $I_c = \left[\max\left\{0,\ x_i - \frac{\Delta(C)}{3}\right\}, \min\left\{1, x_i + \frac{\Delta(C)}{3}\right\}\right]$.

if $|c| > 1$: we will create the interval $I_c = \left[\max\left\{0,\ x_i - \frac{\Delta(C)}{3}\right\}, \min\left\{1, x_j + \frac{\Delta(C)}{3}\right\}\right]$.

We can see that the intervals we created are disjoint.

From cardinality concerns, we can find more $k - |Ch|$ intervals in $[0,1]$ s.t no element from $C$ is in them and disjoint from the intervals we already created.

Let us define the set of them $h'$.

So, if we define $h = \bigcup_{c \in Ch} I_c \cup h'$ we can see that $\forall\ 1 \leq i \leq 2k, h(x_i) = s_i$, as required, therefore $VCdim(\mathcal{H}_k) \geq 2k$.

4. We will prove that the VC-dimension of $\mathcal{H}_{polly}$ is $\infty$:
   Let there be $C = \{(x_1^i, x_2^i)\}_{i=1}^n \subset \mathbb{R}^2$ set of n points.
   We will show that $\mathcal{H}_{polly}$ shatters $C$.

   Let there be $(s_1, s_2, \dots, s_n) \in \{0,1\}^n$.
   Let us define $C' = \{(x_1^i, x_2^i - (1 - s_i))\}_{i=1}^n$.

   If we consider the Lagrange polynomial $L$ for $C'$ we will see that:
   If $s_i = 1$: $L(x_1^i) = x_2^i - (1 - s_i) = x_2^i - (1 - 1) = x_2^i \geq x_2^i$
   Hence $h_L(x_1^i, x_2^i) = 1$

   If $s_i = 0$: $L(x_1^i) = x_2^i - (1 - s_i) = x_2^i - (1 - 0) = x_2^i - 1 < x_2^i$
   Hence $h_L(x_1^i, x_2^i) = 0$

   Therefore, for all $1 \leq i \leq n,\ h_L(x_1^i, x_2^i) = s_i$

   We saw that for every $n,\ \mathcal{H}_{polly}$ shatters a set $C \subset \mathbb{R}^n$ of size $n$, thus $VCdim(\mathcal{H}_{polly}) = \infty$

## Programming Assignment:

a. As we studied in class, in classification problems, the optimal hypothesis (which minimizes the true error) is:

$$h(x) = 1 \Leftrightarrow P[Y = 1 \mid X = x] \geq P[Y = 0 \mid X = x]$$

i.e.,

$$h(x) = 1 \Leftrightarrow P[Y = 1 \mid X = x] \geq \frac{1}{2}$$

We will define $h \in \mathcal{H}_{10}$ s.t applies to the bayes optimal predictor rule:

Thus, if $x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$, we should label it as positive, and if $x \in (0.2, 0.4) \cup (0.6,\ 0.8)$, we should label it as negative.

Hence, we construct $h^* \in \mathcal{H}_{10}$:

$$h = \{[0, 0.1], [0.1, 0.2], [0.4, 0.45], [0.45, 0.5], [0.5, 0.55],$$

$$[0.55, 0.6], [0.8, 0.85], [0.85, 0.9], [0.9, 0.95], [0.95, 1]\}$$

And $h^* \in argmin_{h \in \mathcal{H}_{10}} \mathbb{E}\big[e_p(h)\big]$.

Let's calculate the true error of h:
Let's denote $\Omega = [0, 1] \times \{0,1\}$

$$e_P(h) = \mathbb{E}[\Delta_{zo}(h(X), Y)] = P[h(X) \neq Y]$$

$$= P[\{\, (x, y) \in \Omega \mid h(x) \neq y \,\}]$$

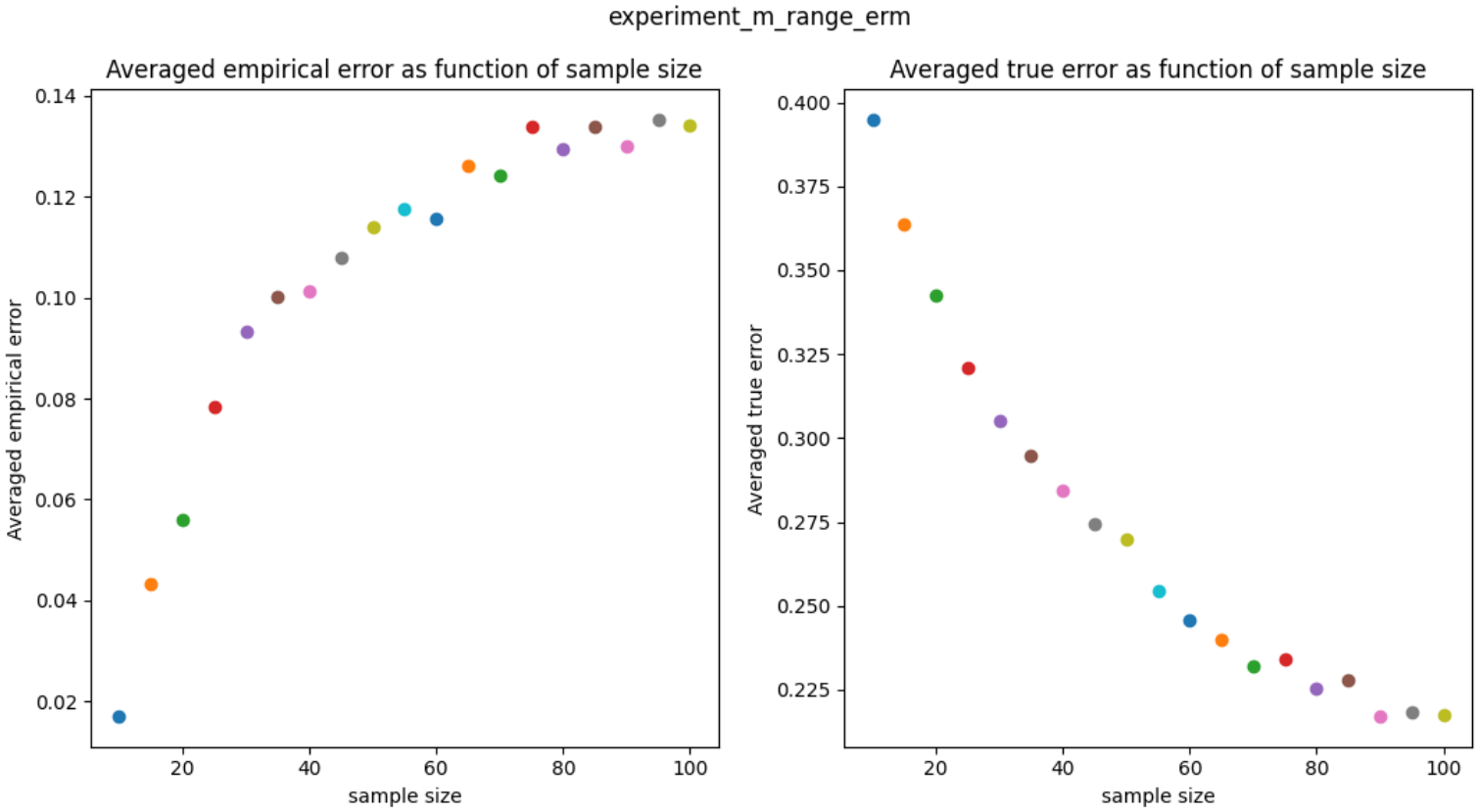
$$= P[\{\, (x, 0) \in \Omega \mid h(x) = 1 \,\}] + P[\{\, (x, 1) \in \Omega \mid h(x) = 0 \,\}]$$

$$= P[\{\, (x, 0) \in \Omega \mid x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \,\}] + P[\{\, (x, 1) \in \Omega \mid x \notin [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \,\}]$$

$$= P[Y = 0 | X \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] P\big[X \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]\big] + P[Y = 1 | X \in (0.2, 0.4) \cup (0.6, 0.8)] P[X \in (0.2, 0.4) \cup (0.6, 0.8)]$$

$$= 0.2 * 0.6 + 0.1 * 0.4 = 0.16$$

b.



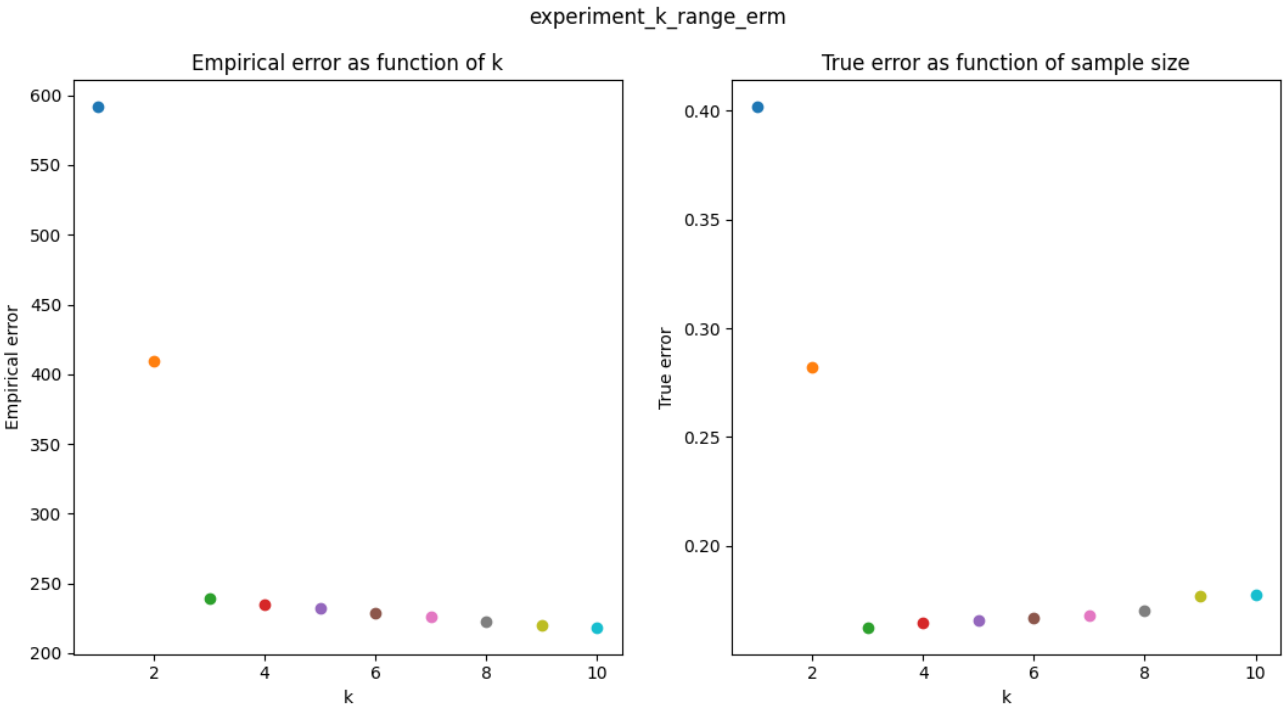

We can see 2 results from our experiment:

- The averaged true error decreasing as the sample size increasing
- The averaged empirical error increasing as the sample size also increasing

The fact that the true error decreasing is not surprising, because we proved in the theoretical question that the $VCDim$ of $\mathcal{H}_k$ is finite, and this means that $\mathcal{H}_3$ is Pac learnable in the agnostic case.
Therefore, the true error of the hypothesis returned by ERM should converge to the best true error, which is $0.16$.

The increasing of the averaged empirical error as the sample size also increasing also makes sense. Because we are limited to only 3 intervals, as the sample size increase, we can't have perfect fit for our samples.

c.

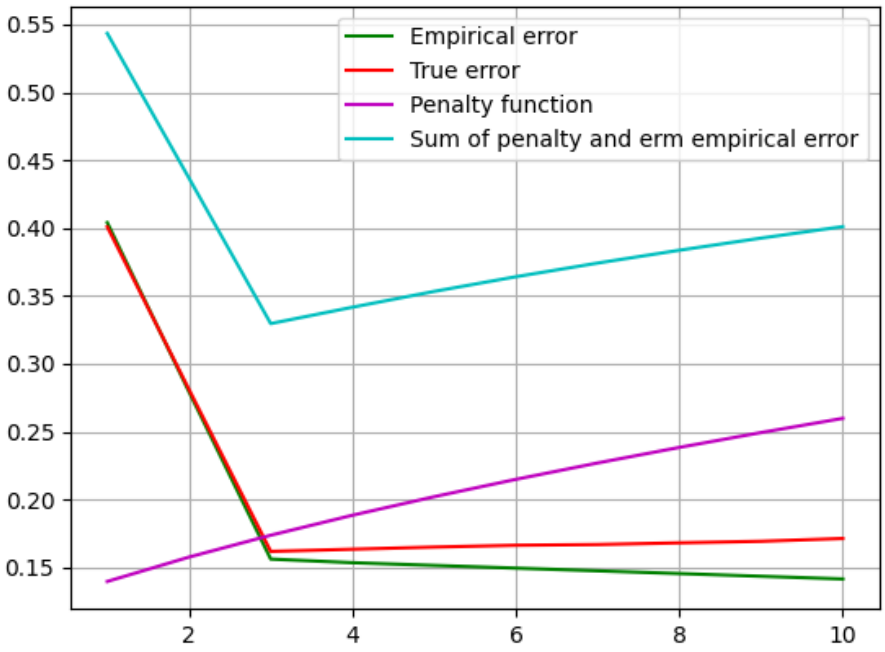



We can see 2 results from our experiment:

- The true error as a function of k is relatively small when k grows
- The empirical error decreasing as a function of k

The k that minimizes the empirical error is $k = 10$, but $h_{ERM_{10}}$ has large true error compared to another k's.

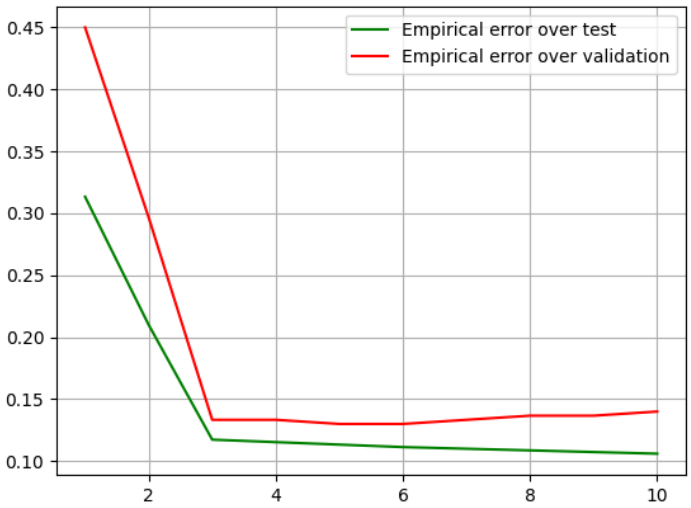
d.



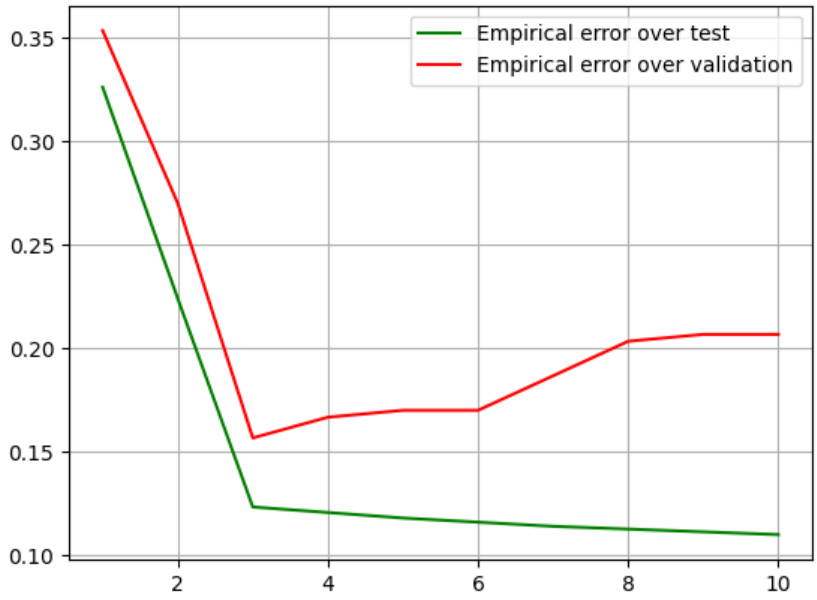

The k that minimizes the sum of the penalty function and the empirical error is $k = 3$, and we can see that, $e_p(h_{ERM_3})$ indeed is the smallest one, which is better from the value of k we choose in the last article.
The SRM gave a tighter bound to the true error.

e.

Running the holdout-validation process a few times gave me different results occasionally. For example, in one session of learning, the chosen hypothesis was with $k = 5$.



In another session of learning, the chosen hypothesis was with $k = 3$,
Which is the hypothesis with best true error.




We can see that indeed the holdout-validation process indeed managed to give a better hypothesis, without knowing strict bounds over our hypothesis classes.