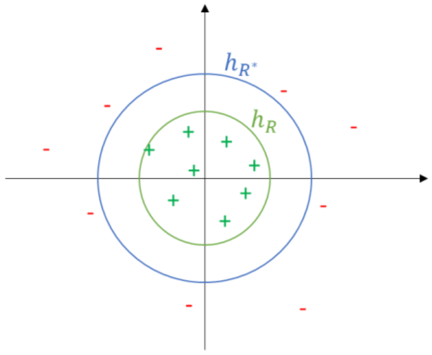
Theoretical Assignment

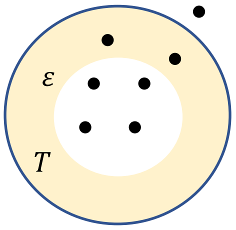
1) **PAC learnability of ℓ_2-balls around the origin.** Given a real number $R \geq 0$ define the hypothesis $h_R : \mathbb{R}^d \to \{0,1\}$ by, $h_R(x) = \begin{cases} 1 & \|x\|_2 \leq R \\ 0 & otherwise \end{cases}$. Consider the hypothesis class $H_{ball} = \{h_R \mid R \geq 0\}$. Prove directly (without using the Fundamental Theorem of PAC Learning) that H_{ball} is PAC learnable in the realizable case (assume for simplicity that the marginal distribution of X is continuous). How does the sample complexity depend on the dimension d? Explain.

Solution:

On input $S = \left\{(x_i, y_i) : i = 1, \dots, n\, ,\ x_i \in \mathcal{X} = \mathbb{R}^d\, ,\ y \in \mathcal{Y} = \{0,1\}\right\}$ the algorithm A return the hypothesis $h_R \in H_{ball}$ where $R = \max\limits_{(x,y)\in S}\{\|x\|_2 : y = 1\}$. We will show that H_{ball} is PAC learnable in the realizable case by using algorithm A.

Denote $h_{R^*} \in H_{ball}$ the real classifier of the space $(\mathcal{X}, \mathcal{Y})$ which is in H_{ball} since we are in the realizable case.





Let be $\varepsilon > 0$, and denote a strip T in a probability of ε inside h_{R^*} on the edge of it. We know that there is such strip from the assumption that the distribution P is continuous. Let R be the ball area that h_R drawn and let R^* be the correspond ball area of the real classifier. We know that:

$$e_P(h_R) = \mathbb{E}[\Delta_{zo}(h_R(X), Y)] = \mathbb{P}(h_R(X) \neq Y) = \mathbb{P}(R^* \setminus R)$$

Assume $\mathbb{P}(R^*) > \varepsilon$. Otherwise, $e_P(h_R) = \mathbb{P}(R^* \setminus R\,) \leq_{R^*\setminus R \subset R^*} \mathbb{P}(R^*) \leq \varepsilon$ and the PAC condition holds with probability 1.

Obviously, for any set of n training points S, there is no negative point inside h_R area. In addition, by the correctness of h_{R^*}, the only points which might be in the area between R and R^* are positive ones. Therefore, if there is a (positive) point $x_i \in T$ then $R^* \setminus R \ \subset T$ because h_R contains this point. Hence, $e_P(h_R) = \mathbb{P}(R^* \setminus R\,) \leq \mathbb{P}(T) = \varepsilon$. Equivalently, if $e_P(h_R) > \varepsilon$ then all the training points of S are not in the strip T. Since $\mathbb{P}(x_i \in T) = \varepsilon$, we get the following conclusion:

$$\mathbb{P}(e_P(h_R) > \varepsilon) \leq \mathbb{P}(\forall (x_i, y_i) \in S : x_i \notin T) =_{IID} (1-\varepsilon)^n \leq_{1-x\leq e^{-x}} e^{-\varepsilon n}$$

Therefore, $\mathbb{P}(e_P(h_R) > \varepsilon) \leq e^{-\varepsilon n} < \delta \Leftrightarrow n > \frac{1}{\varepsilon}\ln\left(\frac{1}{\delta}\right)$.

Hence, by using $N(\varepsilon, \delta) = \left\lfloor \frac{1}{\varepsilon}\ln\left(\frac{1}{\delta}\right)\right\rfloor + 1$, for every $\varepsilon, \delta \in (0,1)$ and for every realizable distribution P over $\mathbb{R}^d$, by using a set of n training examples S which drawn from $IID\ P$, and the algorithm output A, for $n \geq N(\varepsilon, \delta)$:

$$\mathbb{P}\big(e_P(A(S)) > \varepsilon\big) < \delta$$

This means A is a PAC learner for H_{ball}. ■

The dimension d does not appear in the $N(\varepsilon, \delta)$ as a depended parameter, and the whole computation does not involve the dimension d. In other words, the same sample complexity works for every d, and therefore it does not depend on d.

2) PAC in Expectation. Consider learning in the realizable case. We say a hypothesis class H is PAC learnable in expectation using algorithm A if there exists a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P (realizable by H), given a sample set S such that $|S| \geq N(a)$, it holds that, $\mathbb{E}[e_P(A(S))] \leq a$. Show that H is PAC learnable if and only if H is PAC learnable in expectation.

Solution:

For one direction, assume H is PAC learnable, and we will show that H is PAC learnable in expectation.

By the definition, there exists an algorithm A and a function $N(\varepsilon, \delta)$: $(0,1)^2 \to \mathbb{R}$, such that for every $\varepsilon, \delta \in (0,1)$ and for every realizable distribution P, by using a set of n training examples S which drawn from $IID\ P$, the algorithm output $A(S)$, for $n \geq N(\varepsilon, \delta)$ holds that:

$$\mathbb{P}\big(e_P(A(S)) > \varepsilon\big) < \delta$$

Let be $a \in (0,1)$ and using the algorithm A. Then, for training sample S such that $|S| = n$ where $n \geq N\left(\varepsilon = \frac{a}{2}, \delta = \frac{a}{2}\right) = N(a)$ it holds that:

$$\begin{aligned}\mathbb{E}\big[e_P(A(S))\big] &=_{(1)} \mathbb{E}\big[e_P(A(S)) \mid e_P(A(S)) > \varepsilon\big] \cdot \mathbb{P}\big(e_P(A(S)) > \varepsilon\big) \\ &\quad + \mathbb{E}\big[e_P(A(S)) \mid e_P(A(S)) \leq \varepsilon\big] \cdot \mathbb{P}\big(e_P(A(S)) \leq \varepsilon\big) \\ &\leq_{(2)} \mathbb{E}\big[e_P(A(S)) \mid e_P(A(S)) > \varepsilon\big] \cdot \delta + \varepsilon \leq_{(3)} \delta + \varepsilon = a\end{aligned}$$

We used the following claims:

(1) The low of total expectation.

(2) $\mathbb{P}\big(e_P(A(S)) > \varepsilon\big) < \delta$ by the assumption of PAC learnable H.
$\mathbb{E}\big[e_P(A(S)) \mid e_P(A(S)) \leq \varepsilon\big] \leq \varepsilon$ as given $e_P(A(S)) \leq \varepsilon$.
$\mathbb{P}\big(e_P(A(S)) \leq \varepsilon\big) \leq 1$ for any distribution.

(3) $\mathbb{E}\big[e_P(A(S)) \mid e_P(A(S)) > \varepsilon\big] \leq 1$ as $e_P(A(S))$ is conditional variable with the zero-one loss function ($e_P(A(S))$ gets values of 0 and 1).

Therefore, H is learnable in expectation. ■

For the second direction, assume H is PAC learnable in expectation, and we will show that H is PAC learnable. By the definition, there exists an algorithm A and a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P (realizable by H), given a sample set S such that $|S| \geq N(a)$, it holds that, $\mathbb{E}[e_P(A(S))] \leq a$.

Let be $\varepsilon, \delta \in (0,1)$ and using the same algorithm A. Then, for training sample S such that $|S| = n$ where $n \geq N\left(a = \frac{\delta\varepsilon}{2}\right) = N(\varepsilon, \delta)$, by Markov inequality it holds that:

$$\mathbb{P}\big(e_P(A(S)) > \varepsilon\big) \leq \frac{\mathbb{E}\big[e_P(A(S))\big]}{\varepsilon} \leq \frac{a}{\varepsilon} = \frac{\delta\varepsilon}{2\varepsilon} = \frac{\delta}{2} < \delta$$

Therefore, H is PAC learnable by A. ■

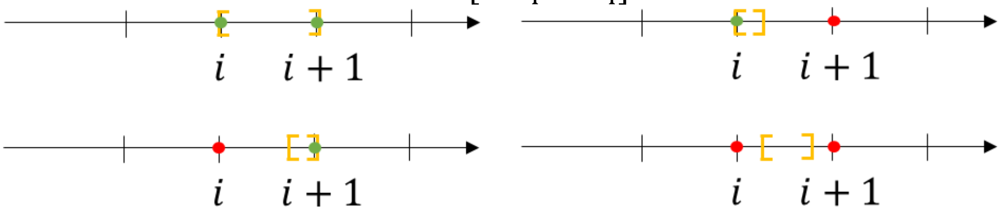
3) **Union of intervals.** Determine the VC-dimension of H_k - the subsets of the real line formed by the union of k intervals. Prove your answer.

Solution:

We prove that $VCdim(H_k) = 2k$.

First of all, we will show that $VCdim(H_k) \geq 2k$. Define a set of unlabeled samples $C = \{1,2,\dots,2k\}$ and let be a dichotomy $S = [s_1,\dots,s_{2k}] \in H_k^C$, an arbitrary classification $s_i \in \{0,1\}$ of the sample C. We will show that the class H_k shatters C by define a classifier $h_I \in H_k$ which realizes the dichotomy S. Define the following $h_I(x) = \begin{cases} 1 & x \in \cup_{i=1}^{k} I_i \\ 0 & otherwise \end{cases}$ where $I = \{I_1,\dots,I_k\} = \{[r_1,l_1],\dots,[r_k,l_k]\}, r_1 \leq l_1 \leq \dots \leq r_k \leq l_k$:

$$\forall i \in \{1,3,\dots,2k-1\}: \quad I_{n_i} = \begin{cases} [i, i+1] & s_i = s_{i+1} = 1 \\ \left[i, i+\frac{1}{4}\right] & s_i = 1, s_{i+1} = 0 \\ \left[i+\frac{3}{4}, i+1\right] & s_i = 0, s_{i+1} = 1 \\ \left[i+\frac{1}{4}, i+\frac{3}{4}\right] & s_i = s_{i+1} = 0 \end{cases}$$



Obviously, I is a set of k closed and disjoint intervals and therefore $h_I \in H_k$. Moreover, by the definition of h_i for every $i \in \{1,\dots,2k\} = C$ it holds that $h_I(i) = s_i$. (If $s_i = 1$ then i is one of the edges of an interval in I and therefore in the union of them; and if $s_i = 0$ then there is no interval in I that contains i).

Secondly, we will show that $VCdim(H_k) \leq 2k$, i.e., for every set of unlabeled samples $C^* = \{x_1,\dots,x_{2k+1}\}$ in size of $2k+1$ where $x_1 < x_2 < \dots < x_{2k+1}$, there exists a dichotomy $S^* = [s_1,\dots,s_{2k+1}]$ where $s_i \in \{0,1\}$ such that for every $h_I \in H_k$ it holds that h_I does not realize S^*.

Define the following dichotomy on C^*: $S^* = [1,0,1,\dots,0,1]$ $i.e., \forall i \in \{1,\dots,2k+1\}$ $s_i = i \bmod 2$. Assume by contradiction that there exists $h_I = \{I_1,\dots,I_k\} \in H_k$ that realizes S^*. Therefore,

(1) $\forall i \in \{1,3,\dots,2k+1\}: s_i = 1 \Rightarrow \exists n_i \in \{1,\dots,k\}\ such\ that\ x_i \in I_{n_i}$

(2) $\forall i \in \{2,4,\dots,2k\}: s_i = 0 \Rightarrow \forall n_i \in \{1,\dots,k\}\ x_i \notin I_{n_i}$

From (1) and (2), and from the order of x_i, we can conclude that every two intervals I_{n_i} and I_{n_j} that are formed in (1) are different. Otherwise, assume i, j are odd and $i < j$. if $x_i, x_j \in I_{n_i}$ then, because there is an even index $i < m < j$, we get that $x_m \in I_{n_i}$ in contrary of (2).

Hence, I has $|\{1,3,\dots,2k+1\}| = k+1$ different intervals in contradiction to the fact that $h_I \in H_k$.

All in all, $VCdim(H_k) = 2k$ ∎.

4) **Prediction by polynomials.** Given a polynomial $p: \mathbb{R} \to \mathbb{R}$ define the hypothesis $h_p : \mathbb{R}^2 \to \{0,1\}$ by,

$$h_p(x_1, x_2) = \begin{cases} 1 & p(x_1) \geq x_2 \\ 0 & otherwise \end{cases}$$

Determine the VC-dimension of $H_{poly} = \{h_p \mid p \ is \ a \ polynomial\}$. You can use the fact that given n distinct values $x_1, \dots, x_n \in \mathbb{R}$ and $z_1, \dots, z_n \in \mathbb{R}$ there exists a polynomial p of degree $n-1$ such that $p(x_i) = z_i$ for every $1 \leq i \leq n$.

Solution:

We will show that $VCdim(H_{poly}) = \infty$. Let n be a natural number, and we will show that $VCdim(H_{poly}) \geq n$. Therefore, from the arbitrariness of $n \in \mathbb{N}$ we can conclude that $VCdim(H_{poly}) = \infty$.

Consider a set of unlabeled samples $C = \{(x_i, y_i)\}_{i=1}^{n} = \{(i, 0)\}_{i=1}^{n}$ and let $S = [s_1, \dots, s_n], s_i \in \{0,1\}$ be a dichotomy of each sample.

We use the fact that given n distinct values $x_1, \dots, x_n \in \mathbb{R}$ and $z_1, \dots, z_n \in \mathbb{R}$ there exists a polynomial p of degree $n-1$ such that $p(x_i) = z_i$ for every $1 \leq i \leq n$.

Consider $x_i = i$ and $z_i = \begin{cases} 0 & s_i = 1 \\ -1 & s_i = 0 \end{cases}$ $\forall i \in \{1, \dots, n\}$ and using the above polynomial p in order to define $h_p \in H_{poly}$. We need to show that $\forall i \in \{1, \dots, n\} : h_p\big((i, 0)\big) = s_i$. Indeed,

$$s_i = 1 \Rightarrow p(i) = 0 \Rightarrow p(i) \geq y_i = 0 \Rightarrow h_p\big((i, 0)\big) = 1$$

$$s_i = 0 \Rightarrow p(i) = -1 \Rightarrow p(i) < y_i = 0 \Rightarrow h_p\big((i, 0)\big) = 0 \ \blacksquare$$

Programming Assignment

(1) Union of intervals

(a) Assume that the true distribution $P[x,y] = P[y|x] \cdot P[x]$ is as follows: x is distributed uniformly on the interval $[0,1]$, and denote $U = [0,0.2] \cup [0.4,0.6] \cup [0.8,1]$

$$P[y=1|x] = \begin{cases} 0.8 & x \in U \\ 0.1 & x \in [0,1]/U \end{cases}$$

and $P[y=0|x] = 1 - P[y=1|x]$. Since we know the true distribution P, we can calculate $e_P(h)$ precisely for any hypothesis $h \in H_k$. What is the hypothesis in H_{10} with the smallest error (i.e., $arg \min_{h \in H_{10}} e_P(h)$)?

Solution:

For any hypothesis $h_I \in H_k$, consider $\mathcal{X} = [0,1]$ and Y the classifications of X. Denote $I = \{I_1, \dots, I_k\}$ and $U_I = \bigcup_{i=1}^{k} I_i$. Define for $A \subset [0,1]$ the weight of A as $w(A) = \mathbb{P}(A)$ in the uniform distribution on $[0,1]$. So, we can express $e_P(h_I)$:

$$e_P(h_I) = \mathbb{E}[\Delta_{zo}(h_I(X),Y)] = \sum_{x \in \mathcal{X}} \Delta_{zo}(h_I(x),0) \cdot \mathbb{P}(X=x, Y=0) + \Delta_{zo}(h_I(x),1) \cdot \mathbb{P}(X=x, Y=1)$$

$$= \sum_{x \in \mathcal{X}} \Delta_{zo}(h_I(x),0) \cdot \mathbb{P}(Y=0 \mid X=x) \cdot \mathbb{P}(X=x) + \Delta_{zo}(h_I(x),1) \cdot \mathbb{P}(Y=1 \mid X=x) \cdot \mathbb{P}(X=x)$$

$$= \sum_{\substack{x \in \mathcal{X} \\ h_I(x)=1}} \mathbb{P}(Y=0 \mid X=x) \cdot \mathbb{P}(X=x) + \sum_{\substack{x \in \mathcal{X} \\ h_I(x)=0}} \mathbb{P}(Y=1 \mid X=x) \cdot \mathbb{P}(X=x)$$

$$= \sum_{\substack{x \in U \\ h_I(x)=1}} 0.2 \cdot \mathbb{P}(X=x) + \sum_{\substack{x \in [0,1]/U \\ h_I(x)=1}} 0.9 \cdot \mathbb{P}(X=x) + \sum_{\substack{x \in U \\ h_I(x)=0}} 0.8 \cdot \mathbb{P}(X=x)$$

$$+ \sum_{\substack{x \in [0,1]/U \\ h_I(x)=0}} 0.1 \cdot \mathbb{P}(X=x)$$

$$= \sum_{x \in U \cap U_I} 0.2 \cdot \mathbb{P}(X=x) + \sum_{x \in ([0,1]/U) \cap U_I} 0.9 \cdot \mathbb{P}(X=x) + \sum_{x \in U \cap ([0,1]/U_I)} 0.8 \cdot \mathbb{P}(X=x)$$

$$+ \sum_{x \in [0,1]/(U \cup U_I)} 0.1 \cdot \mathbb{P}(X=x)$$

$$= 0.2 \cdot w(U \cap U_I) + 0.9 \cdot w(([0,1]/U) \cap U_I) + 0.8 \cdot w(U \cap ([0,1]/U_I)) + 0.1 \cdot w([0,1]/(U \cup U_I))$$

Let $h_I \in H_{10}$ and denote $I = \{[r_1, l_1], [r_2, l_2], \dots, [r_{10}, l_{10}]\}$ where $0 \le r_1 \le l_1 \le \dots \le r_{10} \le l_{10} \le 1$. We want to minimize $e_P(h_I)$. Denote, $s_1 = w(U \cap U_I)$, $s_2 = w([0,1]/(U \cup U_I))$. Then, we can present $e_P(h_I)$ as function of $s_1 \in [0,0.6]$ and $s_2 \in [0,0.4]$:

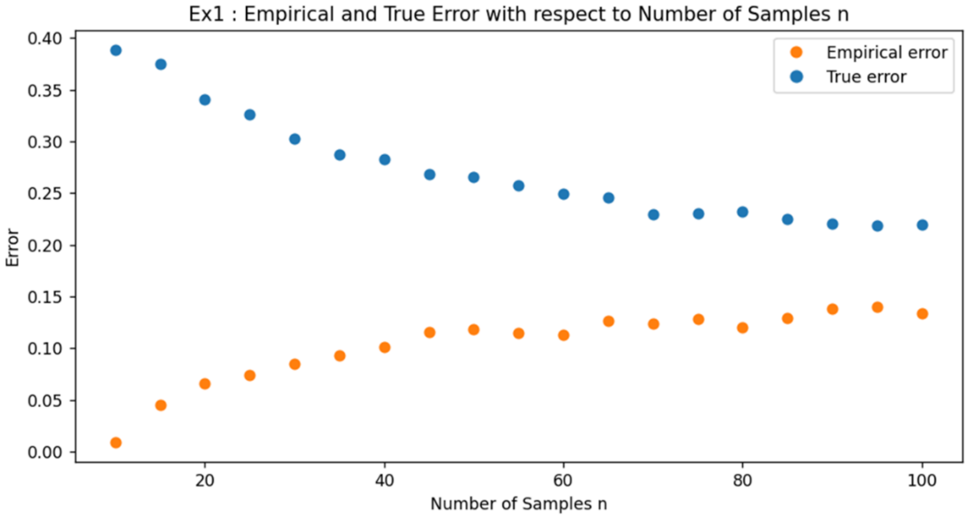
$$\begin{aligned} e_P(h_I) &= f(s_1, s_2) \\ &= 0.2 \cdot s_1 + 0.9 \cdot (1 - w(U) - s_2) + 0.8 \cdot (w(U) - s_1) + 0.1 \cdot s_2 \\ &= 0.2 \cdot s_1 + 0.9 \cdot (0.4 - s_2) + 0.8 \cdot (0.6 - s_1) + 0.1 \cdot s_2 \\ &= 0.84 - 0.6 s_1 - 0.8 s_2 \end{aligned}$$

Quick analysis shows that $\nabla f = \begin{pmatrix} -0.6 \\ -0.4 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, therefore, f gets its minimum at the edges where $s_1 = 0,0.6$ or $s_2 = 0,0.4$. It is easy to see that we get the minimum of f where $s_1 = 0.6$ and $s_2 = 0.4$: $f(0.6,0.4) = \mathbf{0.16}$.

This scenario describes a set $\boldsymbol{U_I = U}$. Therefore, $h_I \in H_{10}$ that minimizes $e_P(h_I)$ could be the hypothesis with the following intervals:

$$I = \left\{ \begin{matrix} [0,0.05], [0.05,0.1], [0.1,0.15], [0.15,0.2], \\ [0.4,0.45], [0.45,0.5], [0.5,0.6], \\ [0.8,0.85], [0.85,0.9], [0.9,1] \end{matrix} \right\}$$

(b) Results:



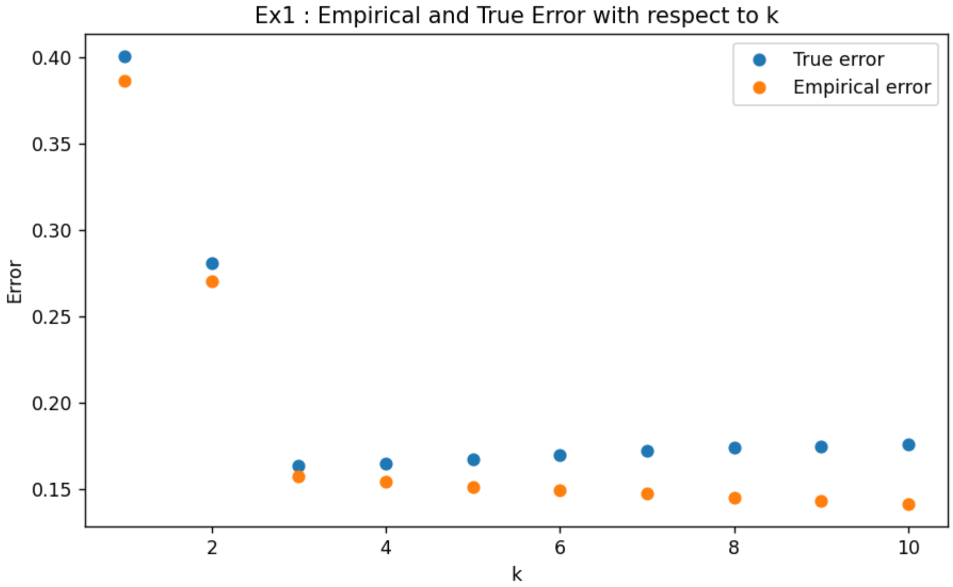
Discussion:

It can be seen in the graph that the empirical error is increasing as the number of samples, n, are getting large; while the true error is decreasing at the same context. These results can be explained by understanding the case we examined: As n increases, we have more and more x points at the real line that are chosen uniformly. Therefore, the ERM algorithm can get closer to the best classifier in $h_I \in H_3$, where $I = \{[0,0.2],[0.4,0.6],[0.8,1]\}$. Recall that the true distribution P does not form exactly this classifier, so this experiment is under the agnostic frame. However, as I explained in section (a), this h_I is the best classifier of P. Hence, by using more labeled points, the ERM algorithm can expose the interval I and can get closer to better classifier with smaller true error.

On the other hand, as n increases, the classifier of ERM algorithm yields worst empirical error. The main reason is that ERM algorithm tries to fit more and more labeled points with a classifier $h \in H_3$ which is not the true classifier of these points. When n is small, the algorithm can handle classifying these points with 3-interval-classifier from H_3 and therefore can reach small empirical error. However, when n is large, any 3-interval-classifier exposes more and more pointes that can't be fit to the output classifier.

Moreover, as I analyzed in section (a), both errors are getting more closer to the minimal possible true error: 0.16, and the rate of the increasing and decreasing are getting smaller, as n increases. We can guess that the empirical test converges to the optimal true error e_P^*.

(c) Results:



Empirical (k) = [0.386, 0.27, 0.1573, 0.154, 0.1513, 0.1493, 0.1473, 0.1453, 0.1433, 0.1413]

True (k) = [0.40039, 0.28059, 0.1633, 0.16489, 0.1674, 0.16958, 0.17202, 0.1737, 0.17479, 0.1758]

The best ERM hypothesis is where k=3.

Discussion:

The true error gets its best value (minimum error) where k=3 with error of 0.1633. The true error decreases dramatically until k=3 and then increases moderately. These results can be explained by the fact that H_3 is the best hypothesis class for describing the real, since P is "almost" a classifier that uses 3 intervals. Moreover, in section (a) we saw that the value of the best true error that $h \in H_k$ can get is 0.16, which is very close to the best true error of ERM hypothesis in the best case where $k = 3$. Where $k = 1,2$ we got quite bad hypotheses, because there is no a good classifier from H_1 or H_2 with 1 or 2 intervals that fit to "almost" 3-interval-classifier as P behaves (the intervals of P are strictly separated). However, where $k \geq 3$ we find a good classifier when we take consecutive intervals for describing one big interval of P (as I did in section (a) for the case of $k = 10$).

The empirical error decreases while k increases. The most dramatic change occurred until $k = 3$ and afterwards there are small changes in the values of the empirical errors. As k increases, the classifier class H_k gets more and more expressive, and therefore the ERM hypothesis is able to fit more points from the 1500 given labeled points. That's why the empirical error decreasing. In the same manner of the former explanation, the point $k = 3$ is a turning point with respect to the decreasing rate.

So, the hypothesis with the smallest empirical error for ERM is where $k^* = 10$. However, this is not a good choice for hypothesis, since the true error that we got is not the optimal one. This is a case of overfitting – out class is too much expressive. The ERM algorithm is tring to be as much accurate on the data set as it can, but the return hypothesis is getting away from the optimal one.

(d) Results & Discussion:

Test error (k) = [0.4266. 0.27, 0.1533, 0.1533, 0.1533, 0.1533, 0.1533, 0.1533, 0.1533, 0.1533]

The best classifiers (with the smallest test error) are where $k = 3,4,\dots,10$. All those classifiers are very close to each other. For example, the intervals set of $k = 3$ is:

$$[(0.00094, 0.1953), (0.3991, 0.5952), (0.8006, 0.99995)]$$

which is very close to the optimal classifier as mentioned in section (a). Similarly, for $k = 10$ we got the following classifier intervals:

[(0.00094, 0.02803), (0.02834, 0.08702), (0.08931, 0.1071), (0.1118, 0.1382), (0.1395, 0.1818), (0.1827, 0.1953), (0.3991, 0.5952), (0.8006, 0.8170), (0.8202, 0.8686), (0.8733, 0.99995)]

As we can see, even where $k = 10$, the intervals cover almost all the 1's areas of the intervals that belongs to $k = 3$ classifier. The probability to get points in the test samples that are in these tiny gaps between the two intervals sets, is small enough, so the test error does not change on these two hypotheses.

Taking the hypothesis with the minimal test error from this method does not necessarily give us the hypothesis with the optimal true error. However, we can choose the first hypothesis with small enough error, which the next hypotheses are close enough to this hypothesis.