# Theoretical question:

1. We will prove that is PAC learnable in the realize case:

Because we assume realizability, there is s.t .

This means that every positive label samples are inside some ball with radius .

We will propose the next algorithm

Given samples drawn i.i.d from do the following:

1. Find the maximal norm of a positive labeled sample in .
2. Return the hypothesis that corresponds to the maximal norm.

In another words, returns the smallest ball that contains positive samples in .

We know that it exists because we assumed realizability.

We now prove that is PAC using :

Let there be samples drawn i.i.d from with size , , and a reliable distribution .

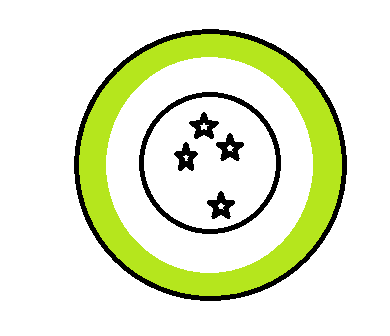
We mark to be the hypothesis returned.

can mistake only by labeling positive samples as negative (due to realizability).

If we denote the d-dimensional ball with radius d, we receive that:

I will assume that (otherwise we have nothing to prove).

We will define a strip ranging from in the following way:



Where is the green strip, the larger ball is , and the smaller ball is .

Because is continuous, we can take s.t .

We notice that if there is s.t , then .

Therefore, in this case .

This means that implies that .

Hence:

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

From the definition of , .

So:

And to achieve that:

This means that the sample complexity is bounded by .

We can see that the sample complexity does not depend over the dimension d.

Let’s assume that is PAC learnable, we will prove that is PAC learnable in expectation.

Because is PAC learnable, there is and an algorithm s.t for any distribution realizable w.r.t , and any sample space with , .

Let there be .

Because is PAC learnable it follows that for sample space with , .

Using the law of total expectation:

We know that from PAC, and it follows that because the expected value of given that is at most .

Thus

We know that , and , because all probabilities are bounded by 1, therefore we conclude that

If we define we will get that is PAC learnable in expectation with the same algorithm .

Let’s assume that is PAC learnable in expectation, we will prove that is PAC learnable.

Because is PAC learnable in expectation, there is an algorithm s.t for PAC in expectation assumption holds.

Let there be .

We need to find out s.t for any distribution realizable w.r.t , and any sample space with , .

Using the Markov inequality (I assume that the loss function range is ):

If we take s.t (because )

we will get that

Hence

So, if we define we will get that is PAC learnable with the same algorithm .

1. We will prove that the VC-dimension of is :

To show that we I will show for all , there is a set of size that shattered by ,

And prove that .

By contradiction, assume that .

So, there is a set of size that shatters.

W.l.o.g we assume that .

For , we define . Let us look on .

Let there be the hypothesis that achieves .

We denote .

We can see that , so by pigeonhole principle there must be s.t for .

From the definition of , it follows that , therefor because follows that . But because , it follows that , and this is contradiction.

To do I will claim the next proposition:

**:** Let there be then the number of maximal chains of 1’s is at most n.

Given 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

**:**

For :

We can see that , and the maximal number of chains of 1’s is indeed (in and and we can see a maximal chain with length 1).

Assuming the induction assumption for , now we will prove for :

Let there be

According to the induction assumption, the maximal number of chains of 1’s in is at most .

If then we can have only one more maximal chain (by

, thus, the number of maximal chains in

is at most .

If then we can create only one more maximal chain (by

, thus, the number of maximal chains in is at most .

Now we will prove that

let there be a set s.t .

Let there be , we will show that

there is s.t .

We will define the parameter of to be

Using the above lemma, we know that the number of maximal chains of 1’s in it at most .

Let us call the set as the set of maximal chains.

We will create interval that fits to a maximal chain:

given :

if : we will create the interval

if : we will create the interval

We can see that the intervals we created are disjoint.

From cardinality concerns, we can find more intervals in s.t no element from is in them and disjoint from the intervals we already created.

Let us define the set of them .

So, if we define we can see that , as required, therefore .

1. We will prove that the VC-dimension of is :

Let there be set of n points.

We will show that shatters

Let there be .

Let us define .

If we consider the Lagrange polynomial for we will see that:

If :

Hence

If :

Hence

Therefore, for all

We saw that for every , shatters a set of size , thus

# Programming Assignment:

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

i.e.,

We will define s.t applies to the bayes optimal predictor rule:

Thus, if , we should label it as positive, and if

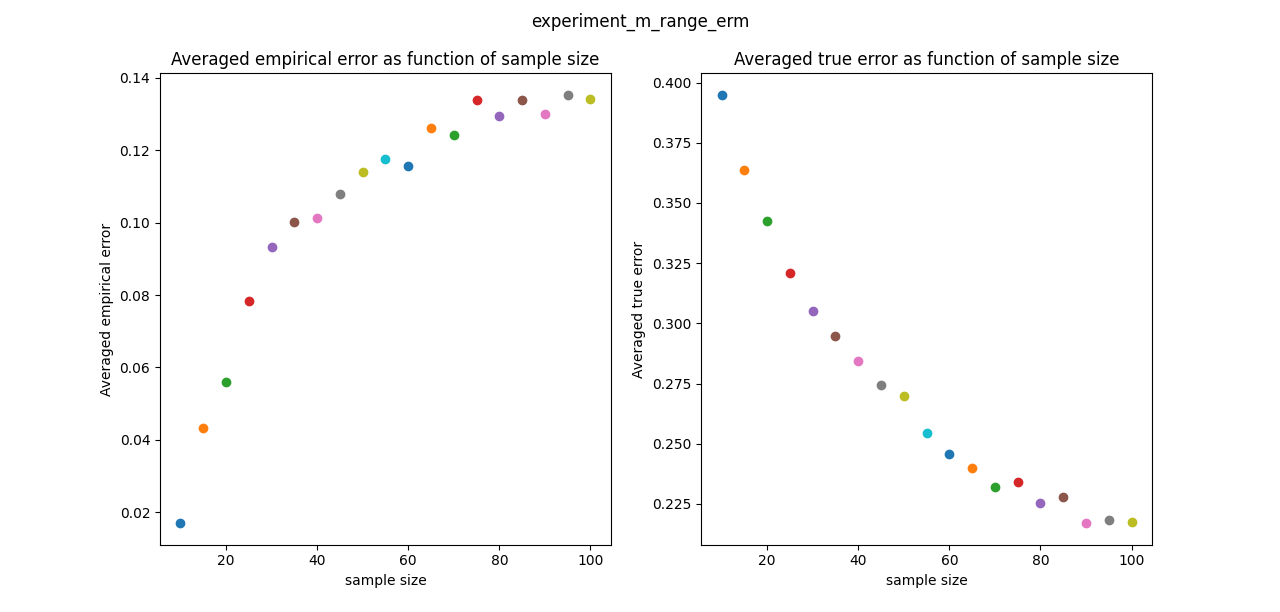
, we should label it as negative.

Hence, we construct :

And .

Let’s calculate the true error of h:

Let’s denote

1. 

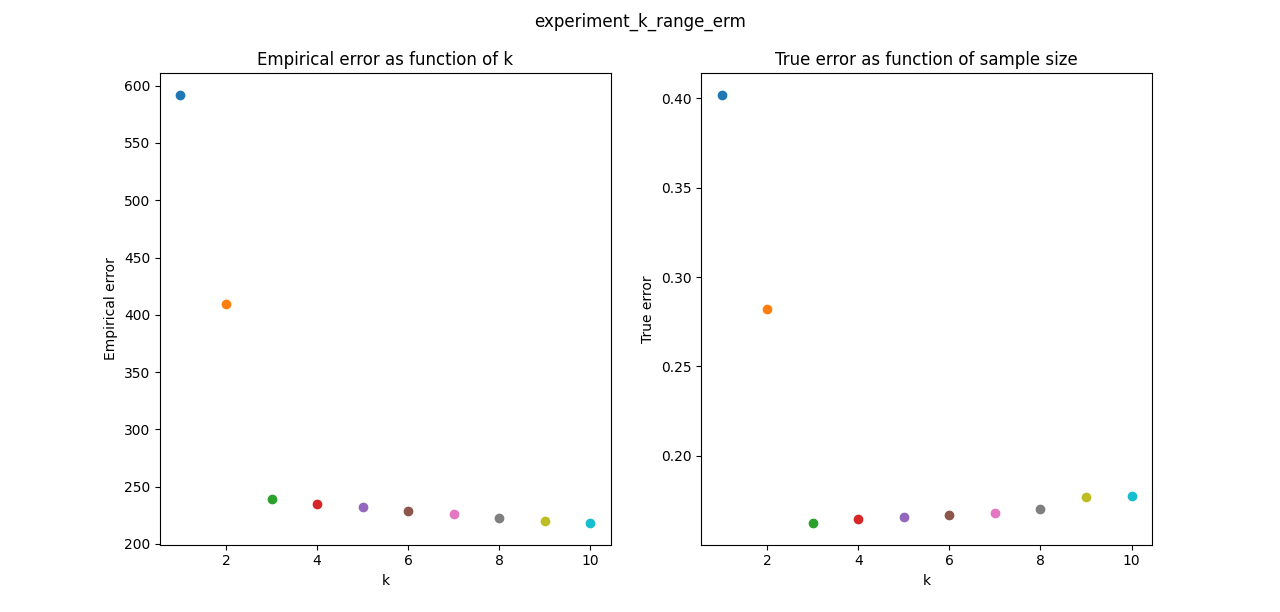
We can see 2 results from our experiment:

* 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 of is finite, and this means that 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 .

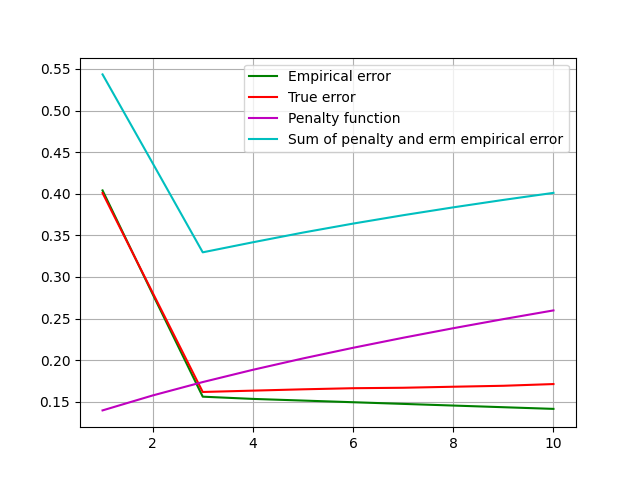
The increasing of the averaged empirical error as the sample size also increasing also makes sense. Because we are limited to only 3 intervals, as the sample size increase, we can’t have perfect fit for our samples.

1. 

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

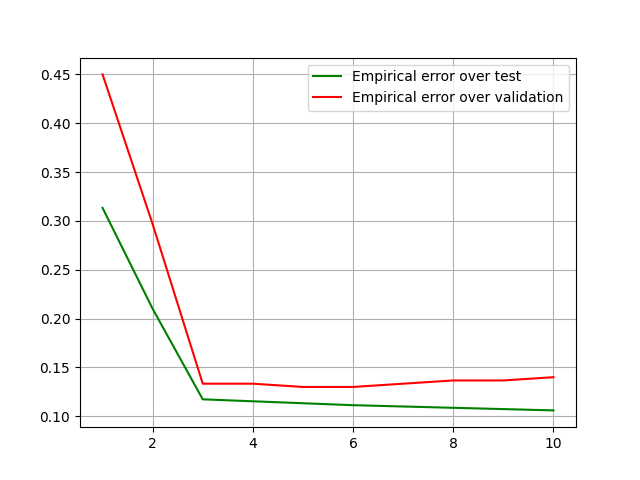
The k that minimizes the empirical error is , but has large true error compared to another k’s.

1. 

The k that minimizes the sum of the penalty function and the empirical error is

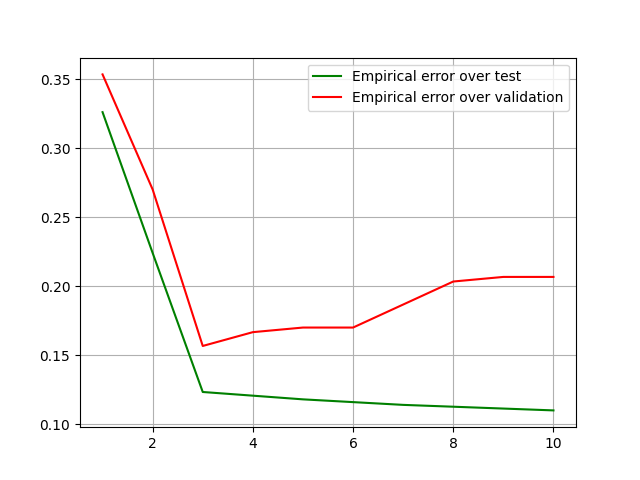
, and we can see that, indeed is the smallest one, which is better from the value of k we choose in the last article.

The SRM gave a tighter bound to the true error.

1.   
     
   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

In another session of learning, the chosen hypothesis was with ,

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.