Learning theory, graded exercise 1

5.1

- Eq 5.2

$$E_{\{S\sim D^m\}}\left[L_D\left(A'(S)\right)\right] \geq \frac{1}{4}$$

We want to show that

$$P\left[L_D\left(A(S)\right) \geq \frac{1}{8}\right] \geq \frac{1}{7}$$

A'(S) is one of the algorithms that performs "As expected" when it comes to learning D (Or it performs worse)

$$L_D: (X \to \{0,1\}) \to [0,1]$$

Thus, we have an upper and a lower bound of $E_{S\sim D}[L_D\left(A'(S)\right)$, so we are now capable of applying Makrov's inequality.

We can express $L_D\left(A'(S)\right)$ as a random variable θ, thus

$$E_{S\sim D}(\theta) \geq \frac{1}{4}$$

Since we have a lower bound, we can still replace it with an equality, Makrov's inequality for a lower bound holds for the smallest possible expected value, it will also hold for all other possible expected values.

$$\mu = \frac{1}{4}$$

We can plug this directly into Makov's inequality, and we get

$$P\left[z > 1 - \frac{7}{8}\right] \geq \frac{\frac{1}{4} - 1 + \frac{7}{8}}{\frac{7}{8}} \geq \frac{1}{7}$$

Thus, we have proved that if the expected value of the loss function of an expected learner is ¼, the probability of the loss being more than 1/4 , will always be 1/7, regardless of the learnier or the number of samples m.

6.2

How to find the VC-dim:

1) **Find a set of size d, that can be shattered**
    a. **A set of features x, so that all of them can be represented, no matter labels that can be chosen**
2) **Find a rule for the labels so that none of the learners can label the correctly, even though the features x are chosen adverserally**

1)

$$k \leq |\, x \,|$$

$$H^X_{=k}\{h \in \{0,1\}^X : |\{x: h(x) = 1 \,| = k\}$$

If $k = |X|$, the learners will be forced to mark every element in x as 1. Thus, no sets are shattered by k = |X|.

Alternatively, if $k < |K|$, each learning wil have a limited number of elements that it can mark as positive or negative examples.

The number of values 1 that can be marked is k, and the maximum number of lements that can be marked as 0 is. $|X| - k$. Thus, the largest sets that can bbe shattered are the ones where any element can be given an arbitrary label. (If not, the labels can be chosen adverseraly to be only 0 of only 1).

Thus, the VC-dimension is

$$VCdim = \min(k, |x| - k\,)$$

2)

A similar case holds for

$$H^X_{\geq k}\{h \in \{0,1\}^X : |\{x: h(x) = 1 \,| \geq k\}$$

But in this case, there is no limit to the number of zeros that can be assigned. As a result, the representability of a set is only limited by k, and if $|C| > k$, the labels can be chosen adversarialy to be only 1, so that none of the learners will be able to correctly label all the points, no matter the value of the x-es.

$$=> VCdim = k$$

6.5

In order to show that $VCdim\left(H_{con}^{d}\right) = 2d$, we have to fufill both the requirements from the previous task.

Both these proofs fall out rather naturally from the 2D-example in the text if we expand the space into d dimensions.

$$\vec{e}_i : Unit\ vector\ along\ the\ axis\ i$$

1) All points can be positioned at $\vec{e}_i\ or\ -\vec{e}_i\ \forall\ i \in [1, d]$. Since these points do not have a position along any of the other axes, we can simply define a hypercube where the distance from the origin to each side is given by $\frac{3}{2}\ or\ \frac{1}{2}$, depending on if a point is positively or negatively labelled. Since each point only has one component, we say that group all the x-es in pairs depending of the axis they are along, and mark them as $x_{i^+}\ and\ x_{i^-}$

   Since $\vec{x}_i\ \vec{e}_j = 0\ \forall j\ \neq i$, the point will be included in the hypercube, regardless of if the scaling is 2 or ½ .

   As a result, we may choose

   The distances-vectors from the origin to each side as

$$\overrightarrow{s_{i^+}} = \left(\frac{1}{2} + \overrightarrow{x_{i^+}} \cdot \overrightarrow{e_i} \cdot y_{i^+}\right)\overrightarrow{e_i}$$

$$and$$

$$\overrightarrow{s_{i^+}} = -\left(\frac{1}{2} + \overrightarrow{x_{i^+}} \cdot (-\overrightarrow{e_i}) \cdot y_{i^+}\right)\overrightarrow{e_i}$$

   Thus, we can choose each side of the hypercube greedily without affecting the others, and all points can thus be represented arbitrarily.

2) **Disproving representation for $|\boldsymbol{C}| = \boldsymbol{2d} + \boldsymbol{1}$**

   We choose the points that have the largest and smallest scalar products with case of of

$$s_{j^+} = \mathrm{e_j} \cdot \max_{i \in [1, 2d+1]} (x_i \cdot e_j)$$

$$s_{j^-} = e_j \cdot \min_{i \in [1, 2d+1]} \left(x_i \cdot e_j\right)$$

   This gives a bounding box around all the points, and since only have 2d vectors to represent the sides of the vector, there is at least one point that could not be chosen.

   Now, if all the first points arrive at the same time, any hypercube that represents these points must also cover the entirety of the minimum bounding box that we created. The final point also has to be inside of this box, so if this point now is negatively labelled, there is no way to correctly label it without mislabelling any of the points we had previously chosen.

   Thus, no set of dimension 2d +1 can be shattered

The VD-dimension of a d-dimensional hypercube is 2d

6.8

I wasted quite a few hours on trying to find a method of classifying points by multiplying some kind of base frequency, so that each new point in a sequence qould be made into a top or abottom point without disturbing any of the others. After a few hours I had to give up. So, instead, I will use the frequency-based solution that will correctly classify the points, but they are not guaranted to be at the top or bottom of the sine-curve.

We are trying to classify a (possibly infinite) set of points, on a numerical line between (0,1). In order for us to find an algorithm that lets us do that, we have two requirements.

1) We need to be able to do some semi-local choice about ω_i, so that the choice will not have consequences for the later choices that we can not handle when we get there.
2) All choices must result in some ω_i, the last of which will be the ω in our learner.

$\sin(x)$ can sometimes be a bit difficult to work with, but we do have some useful properties.

1) $\sin(x + n \cdot 2\pi) = \sin x \ , \forall\, n \in N$
2) $\sin(-x) = -\sin(x)$

So, now let us use these properties to start making an algorithm.

1) If each point $x_i$has a distance that is $4\ x_{i-1}$, the sine-curve we have chosen up until now will not be able to affect it.
    a. If we have $\sin\left(\frac{\pi}{2}x\right)$ to label the point $x = 1$ as positive, we can freely subtract any whole multiple of 2π from $\frac{\pi}{2} \cdot 4$ in order to try to cancel it, and if n =1, that is the case. Thus, we can expand this to get a rule where the previous points will not affect the classification of the later points if $x_{i+1} = 4\ x_i$
2) We also need to guarantee that the later choices will not be able to change the sign of the sine-function at any previous points
    a. In practice, this will mean that the sum of all later contributions must be smaller than $\frac{\pi}{2}$ in either direction.
        i. Since we know that the values x_i quadruples at each iteration, and that ω_i and at each timestep so that the x and ω_i normalizes each other and we get $\sin\left(\pm\frac{\pi}{2}\right)$.Because of this, we simply have to show that: $\sin((\omega_0 + \sum_{i=1}^{\infty}\omega_i)x_i)$ still has the same sign, since all other cases can be reduced to cases on the same form, due to the fact that all previous parts will amount to $n2\pi$
        ii. We know that $|\omega_i| = \frac{1}{4}|\omega_{i-1}|$, so we can simply take the worst-case where all following ω_i have the same sign, which gives us:

$$\sum_{i=1}^{\infty}\frac{1}{4^i} = \frac{1}{4}\cdot\frac{1}{1-\frac{1}{4}} = \frac{1}{3} < \frac{1}{2}$$

        iii. Thus, the sign does not change, and our algorithm will be able to label an infinite number of x_i on this form, regardless of y_i

The VC-dimension of $\boldsymbol{sin(\theta x)}$ is infinite if $\boldsymbol{\theta \in R}$

6.9

This time, we have the learner

$$H = \left\{h_{a,b,s} : a \leq b, s \in \{-1,1\}\right\}$$

$$h_{a,b,s}(x) = \ s \ if \ x \in [a,b], and - s \ otherwise$$

This example is somewhat similar to the one with the hypercube, but the addition of the extra sign gives us some extra flexibility. For instance, three points on a line can be classified without any problems.

$$---X----O----X—$$
$$-----|---|----$$

$$---O----O----X—$$
$$--|---------|---$$

$$---O----O----O—$$
$$--|----------|-$$

The cases for where we surround x instead can found instead, simply by flipping the line, or changing the value of s, which is equivalent to inverting all labels.

But in order to show the VC-dimesjon, we will employ a similar example as in the task with the hypercube, except for the fact that we will "trap" it instead.

1) Imagine a line with three points

   $$---X----O----X—$$

   In order to correctly label these three points, the entire box has to be between the first and the third point.
   Now, if the forth point is to the right of all other points and has the label O, the s with the same sign is forced to be between the x-es, while the one one with the opposite one will find itself in the same situation, except for the fact that the impossible point is the one furthest to the left

   $$---X---O---X---O----$$

This same example can be generalized to a case, where we only talk about the first, secondt, third and forth point from the left. Thus, we can say that we did not lose any generality by the simplifications we did, and thus:

The VC-dimension of H is 3

7.3 (1)

$w(h)$ specifies our preference for certain hypotheses, but we do require $\sum w(h) \leq 1$in order for the sum to be finite, so that we can reason about it in a good way.

In an actual learning case, we'd most likely choose each w(h) based on a priori knowledge, as well as the size of smallest H_n that each h belongs to.

Since of H is an union of infinitely many sets H_n, it will not necessarily make sense to talk about |H|, and ivied all classes based on that.

Instead, we define the subsets H_n that H is a union of to follow the structure

$$|H_n| \geq |H_{n-1}|$$

Now, we can give the larger ypothesis a larger weight, even though it might not reflect the exact size of the set. But since the sets are finite, we can still distribute the weight of the sets evenly among the members.

Now, we have to choose a set of weights that will converge, even though it will be an infinite sum.

We modify the basic integral trick of decreasing sequences in order to get an upper bound of the sum.

$$\sum_{i}^{\infty} \frac{1}{n^2} \leq \frac{1}{1^2} + \int_1^{\infty} \frac{1}{x^2} dx = 1 + 1 = 2$$

(This is because each element in the sum can be seen as an integral over a column, but if we integrate over a decreasing function over the same interval, and the final sum of the function is the same, the integral of the continuous function must be larger, and we have an upper bound. The value of 1 is simply to move the integral so that it fulfils our requirements.

Thus, the weight of each w(h) is given by:

$$w(h) = \frac{1}{|H_n| \cdot 2n^2} \; \forall \; h \in \quad H_n, h \notin H_i \; \forall i[1, n-1]$$

Since each h is only counted once, so the worst-case is when each set is disjunct with the others. So we reformulate w(h) to

$$w(h) = \frac{1}{2} \min_{n:h \in H_n} \frac{1}{|H_n| n^2}$$

Which ends up satisfying all our requirements

7.3 (2)

An infinite, countable set of infinite countable sets is countable if we count like

For a $a \in [1, \infty)$

For $b\ in\ [1, a]$

Count $(H_{a-b}[b])$

Now, we simply have to give a weighting-function that has a sum bounded by 1, even if it is a double infinite sum.

$$w(n, i) = \frac{1}{n^2 i^2} \cdot C$$

Can serve this purpose if we choose C properly.

Since H is a union, we will only count one of the weights of h, and as a result, we may choose arbitrarily among the possible weights that it can have.

In order to formalize this rule, we say:

1) Each H_n has a fixed ordering, so that each h will have a known index $i\ if\ h \in H_n$.
2) $w(h) = w\big(n, i(h, H_n)\big) : n = \min_{n \in H_n} n$

The fact that $\sum_i^\infty \sum_n^\infty \frac{1}{n^2 i^2} \leq C$ can be given by doing the proof from the last step in two steps

1) $\sum_i^\infty \frac{1}{n^2} \leq = 2$
2) $\sum_i^\infty \sum_n^\infty \frac{1}{n^2 i^2} \leq 2 \sum_n^\infty \frac{1}{n^2} \leq 4$
    a. $Since\ the\ two\ variables\ can\ be\ completely\ factorized.$
3)