Assignment 2
Advanced Machine Learning
Issa Bilal (AI - 407)

Exercise 1 a)

We aim to compute the growth function $\tau_{\mathcal{H}}(m)$ for $m \geq 0$, which provides valuable insights into the capabilities of the hypothesis space H. The growth function measures the maximum number of distinct dichotomies that H can generate over a set C of size m. In other words, it quantifies the different functions that can be obtained by restricting H to C to explore and analyze the specific functions that H can produce within the context of the given subset C, enabling us to understand the growth and expressive power of H with respect to different subsets of the input space. Mathematically, the growth function is defined as

$$\tau_{\mathcal{H}}(m) = max_{C \subseteq X : |C| = m} |HC|$$

In relation to the VC dimension (VCdim), we can make some interesting observations. For H, being an interval class, the VCdim(H) is known to be less than or equal to 2. However, upon examining various scenarios, we find that no matter how we select a set of two points, we cannot achieve certain label combinations such as (1,0) or (0,1).
Suppose we have two points, b1, and b2. We want to determine the VCdim of H based on the possibilities of achieving certain label combinations. we consider the relative ordering of the points on the real number line.

1. If b1 = b2:
   When the two points b1 and b2 are equal, we cannot achieve either the (1,0) or (0,1) classification. Regardless of how we set the intervals in H, it cannot distinguish between the two points.
2. If b1 < b2:
   When b1 is less than b2, it implies that b1 is positioned to the left of b2 on the real number line. In this case, it is not possible to achieve the (0,1) classification using the interval class hypothesis space H.
3. If b1 > b2:
   Similarly, when b1 is greater than b2, it means that b1 is positioned to the right of b2 on the real number line. In this scenario, it is not possible to achieve the (1,0) classification using the interval class hypothesis space H.

Based on these observations, we can conclude that the VCdim(H) is precisely 1. This implies that H can shatter (perfectly classify) a set of one point but fails to shatter a set of two points with certain label combinations.

To ensure the maximum possible number of combinations in a set of m points C = (c1, c2, ..., cm), we need to satisfy the following conditions:

1. The first requirement is that 0 should not be included in the set of points. This is because if 0 is present among the points, it becomes impossible to have a label consisting entirely of 0s. Including 0 would limit the number of distinct label combinations that can be achieved.
2. The second requirement is that the absolute values of any two points ci and cj (where i and j are indices ranging from 1 to m and $i \neq j$) must not be equal. By ensuring distinct absolute values, we enable a larger number of label combinations. When the absolute values of points are unequal, it introduces additional variations in the labels, increasing the overall number of possibilities.
3. Sign irrelevance: The third requirement highlights that the sign of the points does not affect the number of combinations, only the specific labels obtained. The goal is to maximize the number of distinct label combinations, regardless of their specific values. The examples provided illustrate this concept. Regardless of whether the points are all positive, all negative, or scattered on the real axis, the resulting label combinations tend to follow similar patterns. This shows that the sign of the points does not impact the total number of combinations achievable.

From here the rest of the proof can be found in lecture 8, all the previous conditions make our hypothesis class behave exactly like the $H_{thresholds}$. Therefore, we conclude the following:

$$\tau_{\mathcal{H}}(m) = m + 1$$

Exercise 1 b)

Given Sauer's lemma provided in lecture 8 we have the following inequality when the VCdim is ≤ d and < ∞ :

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^d$$

Knowing the shattering coefficient and the growth function results from the previous point a), the Sauer's lemma inequality looks as follows:

$$\sum_{i=0}^{d} C_m^d = C_m^0 + C_m^1 = m + 1$$

In our case, this implies that there is an equivalence between the shattering coefficient and the upper bound provided by the lemma.

$$\tau_{\mathcal{H}}(m) = m + 1 = \sum_{i=0}^{d} C_m^d$$

Exercise 2 a)

In the realizable case, our objective is to obtain an efficient Empirical Risk Minimization (ERM) algorithm, denoted as A, for a given concept class. This algorithm should accurately assign labels to the training set S = $\{(x_1, y_1), ..., (x_m, y_m)\}$, where $x_i \in X$, $|X| = m$. As we are operating in the realizable case, we can assume the existence of hypothesis h_{a^*}(x) that perfectly labels the training set.

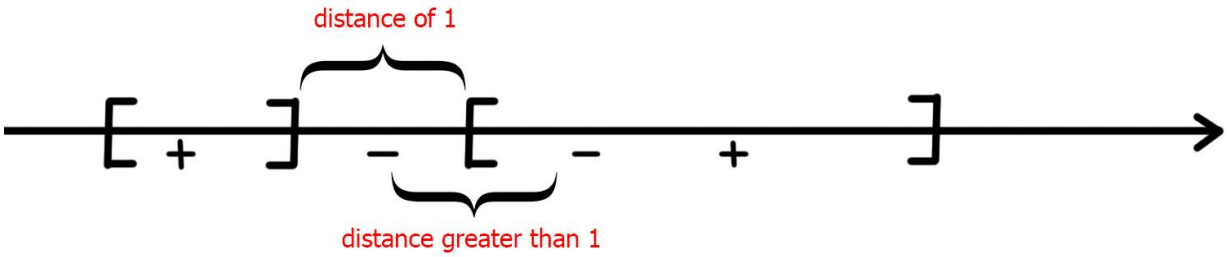
1- Similar to other exercises from the seminar class we start by finding the biggest and smallest positive labeled points, mathematically we denote that as follows:

- $min_p = min_{i=\overline{1,m}} x_i$, where $y_i = 1$
- $max_p = max_{i=\overline{1,m}} x_i$, where $y_i = 1$
- If there are no positive labels in our data set, we set a random point z_i with the condition that it differs from all other points in the dataset.

2- now we look for 3 negative points:

- The biggest negative point smaller than our min positive label
  $max_n = max_{i=\overline{1,m}} x_i$ where $max_n < min_p$ and $y_i = 0$
- The smallest negative point bigger than our max positive label
  $min_n = min_{i=\overline{1,m}} x_i$ where $min_n > max_p$ and $y_i = 0$
- And a negative point between $[min_p+1] \cup [min_p+2]$
  mid_n = any point with label 0 that is smaller than min+2 and bigger than min+1
- If there are no negative labels in our data set then this entire step is pointless and can be skipped.

3- Now to find a suitable position for our interval to settle on the axis, we iterate through all the points and check for negative labels inside of the interval. If we find a negative point inside the first interval, we calculate the distance between it and the mid_n (negative value between the intervals) if the distance is greater than 1 it means we can't fit both of them in the space between the intervals (see the hand-made plot below) and thus we continue our iteration until we find a better fit.



In case we find a good fit where the distance is smaller than 1 between a negative point inside the intervals and the one in between them. We check whether the negative point is inside the first or the second interval and shift the intervals based on that information.

- If the negative point is inside the first interval we shift the intervals to the left with a distance equal to the distance between the negative point and the a+1 bound.
- Else we shift the intervals to the right with a distance equal to the distance between the negative point and the a+2 bound.
- In addition, we have to maintain make sure that the shift doesn't include new negative points, given the fact that we are in the realizable case there should be a good position for the intervals to settle in, all we are doing here is trying to localize it and not create it.

In conclusion, considering that all our steps have a linear complexity of O(n), it is important to note that the combination of multiple O(n) algorithms does not result in an additive complexity. The dominant factor, in this case, remains linear (n), and thus the overall complexity remains **O(n)**.

Exercise 3 b)

Let's consider $\gamma = \min\{\gamma_1, \gamma_2, \gamma_3\}$. Our objective is to demonstrate that the training error of the final classifier h_{final} is at most $\frac{1}{2} - \frac{3}{2}\gamma + 2\gamma^3$ and establish that this bound is strictly smaller than $\frac{1}{2} - \gamma$.

To simplify the proof, we will assume that the final error ϵ_{final} is indeed at most $\frac{1}{2} - \frac{3}{2}\gamma + 2\gamma^3$ and focus on proving the inequality.

$$\frac{1}{2} - \frac{3}{2}\gamma + 2\gamma^3 < \frac{1}{2} - \gamma$$
$$\frac{1}{2} - \frac{1}{2} + \gamma - \frac{3}{2}\gamma + 2\gamma^3 < 0$$
$$\gamma - \frac{3}{2}\gamma + 2\gamma^3 < 0$$
$$-\frac{1}{2}\gamma + 2\gamma^3 < 0$$
$$2\gamma^3 < \frac{1}{2}\gamma$$
$$2\gamma^2 < \frac{1}{2}$$
$$\gamma^2 < \frac{1}{4}$$
$$\gamma < \frac{1}{2}$$

Based on our knowledge that the learners return an error $\epsilon_t \leq \frac{1}{2} - \gamma_t$ for each iteration, and considering that the error is greater than 0, it follows that the inequality $\gamma < \frac{1}{2}$ is actually necessary for the algorithm.

Given that $\epsilon_{final} \leq \frac{1}{2} - \gamma$ and that the inequality $\frac{1}{2} - \frac{3}{2}\gamma + 2\gamma^3 < \frac{1}{2} - \gamma$ holds, we can conclude that our assumption about $\epsilon_{final} \leq \frac{1}{2} - \frac{3}{2}\gamma + 2\gamma^3$ is in turn also true.