

# Assignment 1

Ana-Cristina Rogoz

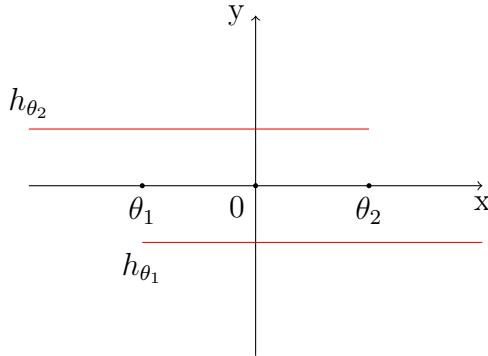
April 24, 2020

**Problem 1** Consider  $\mathcal{H} = \{h_{\theta_1} : \mathbb{R} \rightarrow \{0, 1\}, h_{\theta_1}(x) = \mathbf{1}_{[x \geq \theta_1]}(x) = \mathbf{1}_{[\theta_1, +\infty]}(x), \theta_1 \in \mathbb{R}\} \cup \{h_{\theta_2}(x) = \mathbf{1}_{[x < \theta_2]}(x) = \mathbf{1}_{[-\infty, \theta_2)}(x), \theta_2 \in \mathbb{R}\}$ . Compute  $\text{VCdim}(\mathcal{H})$ .

**Solution** Using the VC-dimension definition, we want to prove that  $\text{VCdim}(\mathcal{H}) = 2$ . Therefore, we want to show that:

1. There  $\exists C \subset \mathbb{R}$ , where  $|C| = 2$ , that is shattered by  $\mathcal{H}$ . ( $\text{VCdim}(\mathcal{H}) \geq 2$ )
2.  $\forall C \subset \mathbb{R}$ , where  $|C| = 3$ ,  $C$  is not shattered by  $\mathcal{H}$  ( $\text{VCdim}(\mathcal{H}) < 3$ )

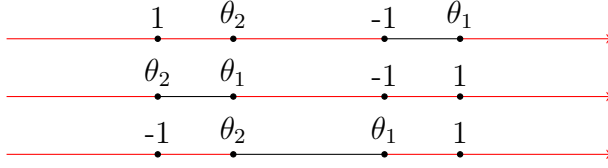
If  $\theta_2 \geq \theta_1$ , then  $\forall x \in \mathbb{R}$  will be labeled 1 (since there is no range not covered by either  $h_{\theta_1}$  or  $h_{\theta_2}$ ). In the image bellow, the red part represents value 1 for both  $\theta$  functions:



So for the following cases we will always consider cases where  $\theta_2 < \theta_1$ . Let's consider  $C = \{-1, 1\}$ , we will show that  $C$  is shattered by  $\mathcal{H}$ , by obtaining all the possible labels:

- label (0,0) – if  $(\theta_2 \leq -1 \text{ and } \theta_1 > 1)$
- label (0,1) – if  $(\theta_2 \leq -1 \text{ and } -1 < \theta_1 \leq 1)$
- label (1,0) – if  $(-1 < \theta_2 \leq 1 \text{ and } \theta_1 > 1)$

- label (1,1) – if  $(\theta_2 > 1 \text{ or } \theta_1 \leq 1 \text{ or } (-1 < \theta_2 \text{ and } \theta_1 \leq 1))$



We've found a set  $C$ ,  $|C| = 2$  that is shattered by  $\mathcal{H}$ , so  $\text{VCdim}(\mathcal{H}) \geq 2$ .

Now we want to show that  $\forall C \subset \mathbb{R}$ , where  $|C| = 3$ ,  $C$  is not shattered by  $\mathcal{H}$ . Consider  $C = \{x_1, x_2, x_3\}$ , with  $x_1 < x_2 < x_3$ . We assume that we can obtain labels  $(0, 1, 0)$ , meaning that label of  $x_2$  is 1. If its label is 1 we are in one of the following two scenarios:

1.  $x_2 < \theta_2$ , so  $h_{\theta_2}(x_2) = 1$  and since  $x_1 < x_2$  it means that  $h_{\theta_2}(x_1) = 1$  as well (**contradiction**,  $x_1$  has label 0)
2.  $x_2 \geq \theta_1$ , so  $h_{\theta_1}(x_2) = 1$  and since  $x_2 < x_3$  it means that  $h_{\theta_1}(x_3) = 1$  as well (**contradiction**,  $x_3$  has label 0)

Since there is no subset  $C \subset \mathbb{R}$  which can output the following labels  $(0,1,0)$ ,  $\text{VCdim}(\mathcal{H}) < 3$ . (We only considered cases with  $\theta_2 < \theta_1$ , because we've showed before that otherwise we will only output label 1 for any input).

To conclude, since  $\text{VCdim}(\mathcal{H}) \geq 2$  and  $\text{VCdim}(\mathcal{H}) < 3 \Rightarrow \text{VCdim}(\mathcal{H}) = 2$ .

**Problem 2** Consider  $\mathcal{H}$  to be the class of all centered in origin sphere classifiers in the 3D space. A centered in origin sphere classifier in the 3D space is a classifier  $h_r$  that assigns the value 1 to a point if and only if it is inside the sphere with radius  $r > 0$  and center given by the origin  $\mathbf{O}(0,0,0)$ .

a) show that the class  $\mathcal{H}$  can be  $(\epsilon, \delta)$  – PAC learned by giving an algorithm  $A$  and determining the sample complexity  $m_H(\epsilon, \delta)$  such that the definition of PAC-learnability is satisfied.

b) compute  $\text{VCdim}(\mathcal{H})$ .

**Solution** a) We will define  $\mathcal{H}$ , the class of all centered in origin sphere classifiers in the following manner:

$$\mathcal{H} = \{h_r : \mathbb{R}^3 \rightarrow \{0, 1\}\}, \text{ where}$$

$$h_r(x_1, x_2, x_3) = \begin{cases} 1, & \text{if } x_1^2 + x_2^2 + x_3^2 \leq r^2 \\ 0, & \text{otherwise} \end{cases}$$

Now we want to show that  $\mathcal{H}$  is PAC learnable. From the definition of PAC-learnability we know that  $\mathcal{H}$  is PAC learnable if there exists a function  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$  and a learning algorithm A with the following property: for every  $\epsilon > 0, \delta > 0$ , for every labeling function  $f \in \mathcal{H}$  and for every distribution  $D \sim \mathbb{R}^3$ , when we run the learning algorithm A on the training set S consisting of  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$  examples sampled i.i.d from  $D$  and labelled by  $f$ , the algorithm A returns a hypothesis  $h_S \in \mathcal{H}$  such that with probability at least  $1 - \delta$  (over the choice of examples), the real risk of  $h_S < \epsilon$ .

Since we are under the realizability assumption, there exists  $f \in \mathcal{H}$ ,  $f = h_r^*$ , such that  $L_{D,f}(h_r^*) = 0$ , that labels the training data. We will consider the following training set  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ , where  $y_i = h_r^*(x_i)$  and  $x_i = (x_{i1}, x_{i2}, x_{i3})$  ( $h_r^*$  labels each point drawn from the sphere  $Sp^*$  with label 1 and all other with label 0).

First, we need to find **a learning algorithm A**. Consider the following algorithm A, that takes as input the training set S, prior defined and outputs  $h_S$ .

$$h_S = h_{r_S}, \text{ where } r_S = \max_{i=1,m} \|x_i\|_2$$

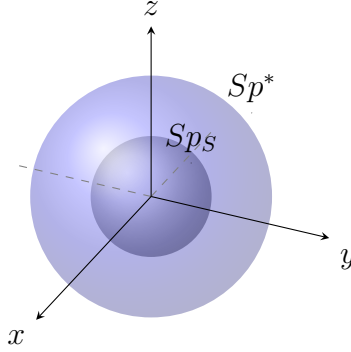
If all the examples from the training set S are labeled 0, then there is no positive example. In this case, we choose the origin (0,0,0) to be labeled 1, so  $r = 0$ , and everything else will be labeled 1.

So for the training set S, the learning algorithm A outputs  $h_S$ , the function of the tightest sphere  $Sp_S$ , enclosing all the positive examples from S.

By construction, A is an ERM, meaning that  $L_S(h_S) = 0$  ( $h_S$  doesn't make any mistakes on the training set S). So after finding A, we want to find the sample complexity  $m_{\mathcal{H}}$  such that:

$$\mathbb{P}_{S \sim D^m} (L_{h_r^*, D}(h_S) \leq \epsilon) \geq (1 - \delta), \text{ when S has } m \geq m_{\mathcal{H}}(\epsilon, \delta) \text{ examples}$$

We notice that  $h_S$  can make errors only in the  $Sp^* \setminus Sp_S$  (the region between the sphere outputted by the algorithm A for training set S, and the sphere determined by  $h_r^*$ ), by labelling points that should have had label 1, with label 0. All other points that are inside  $Sp_S$  and outside  $Sp^*$  will be labeled correctly.



In the image from above, the smaller sphere is the one determined by  $h_S$  and the other one is determined by  $h_r^*$ . By fixing  $\epsilon > 0, \delta > 0$  and consider a distribution  $D \sim \mathbb{R}^3$ , we will consider the following two cases:

1. when  $D(Sp^*) = \mathbb{P}_{x \sim D}(x \in Sp^*) \leq \epsilon$   
 $L_{h_r^*, D}(h_S) = \mathbb{P}_{x \sim D}(h_S(x) \neq h_r^*(x)) = \mathbb{P}_{x \sim D}(x \in Sp^* \setminus Sp_S) \leq \mathbb{P}_{x \sim D}(x \in Sp^*) \leq \epsilon$ , so when sampling m points that leads to  $\Rightarrow \mathbb{P}_{S \sim D^m}(L_{h_r^*, D}(h_S) \leq \epsilon) = 1$
2. when  $D(Sp^*) = \mathbb{P}_{x \sim D}(x \in Sp^*) > \epsilon$ , we construct  $D(S_1) = \mathbb{P}_{x \sim D}(x \in S_1) = \epsilon$ , where  $S_1$  contains only points for which  $r_1 \leq \|x_i\|_2 \leq r^*$ ,  $x_i \in \mathbb{R}^3$ . If the sphere  $Sp_S$  returned by algorithm A for training set S intersects  $S_1$   
 $\Rightarrow L_{h_r^*, D}(h_S) = \mathbb{P}_{x \sim D}(h_S(x) \neq h_r^*(x)) = \mathbb{P}_{x \sim D}(x \in Sp^* \setminus Sp_S) \leq \mathbb{P}_{x \sim D}(x \in S_1) = \epsilon$ , so in this case the result is similar to the first case:  $\mathbb{P}_{S \sim D^m}(L_{h_r^*, D}(h_S) \leq \epsilon) = 1$

Therefore, in order to get  $(L_{h_r^*, D}(h_S) > \epsilon)$ , we need that  $Sp_S$ , the sphere returned by algorithm A will not intersect  $S_1$ , so if for a point the probability not to intersect  $S_1 \leq (1 - \epsilon)$ , for m points the probability that none of them will intersect  $S_1 \leq (1 - \epsilon)^m$

$$\Rightarrow \mathbb{P}_{S \sim D^m}(L_{h_r^*, D}(h_S) > \epsilon) \leq (1 - \epsilon)^m$$

We've seen in Lecture 2, slide 15 that  $1 - x \leq e^{-x}$ , so  $1 - \epsilon \leq e^{-\epsilon}$

$$\Rightarrow \mathbb{P}_{S \sim D^m}(L_{h_r^*, D}(h_S) > \epsilon) \leq (1 - \epsilon)^m \leq e^{-m\epsilon}.$$

Now we want the right side to be smaller than  $\delta$ :

$$e^{-m\epsilon} < \delta \iff -m\epsilon < \log(\delta) \iff m > \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$$

So if we choose a training set  $S$  with at least  $\frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$  samples, we obtain the desired result.

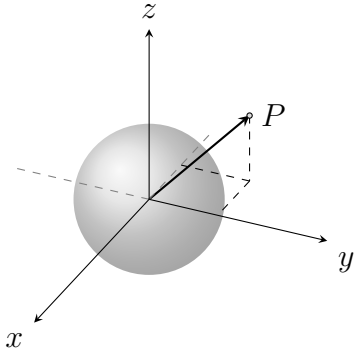
To conclude,  $\mathcal{H}$ , the class of origin centered spheres, is PAC-learnable with the previously presented algorithm and  $\frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$  sample complexity.

b) Using the VC-dimension definition, we want to prove that  $\text{VCdim}(\mathcal{H}) = 1$ . Therefore, we want to show that:

1. There  $\exists C \subset \mathbb{R}^3$ , where  $|C| = 1$ , that is shattered by  $\mathcal{H}$ . ( $\text{VCdim}(H) \geq 1$ )
2.  $\forall C \subset \mathbb{R}^3$ , where  $|C| = 2$ ,  $C$  is not shattered by  $\mathcal{H}$  ( $\text{VCdim}(H) < 2$ )

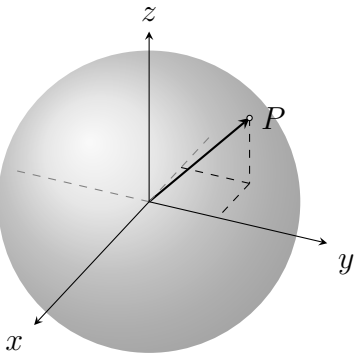
We will consider  $C = \{(-2, 2, 2)\}$ , subset of  $\mathbb{R}^3$ , with  $|C| = 1$ , and we will show that  $C$  is shattered by  $\mathcal{H}$ , by obtaining all the possible labels:

- label 0 (when  $(-2)^2 + 2^2 + 2^2 > r^2 \iff 12 > r^2 \iff 2\sqrt{3} > r$ )



The only point from  $C$ , called  $P = (-2, 2, 2)$  is outside the chosen sphere. Thus, it has label 0.

- label 1 (when  $2\sqrt{3} \leq r$ )



The same point  $P = (-2, 2, 2)$  is now inside the chosen sphere. Thus, it has label 1.

We've found a set  $C$ ,  $|C| = 1$  that is shattered by  $\mathcal{H}$ , so  $\text{VCdim}(\mathcal{H}) \geq 1$ .

Now we want to show that  $\forall C \subset \mathbb{R}$ , where  $|C| = 2$ ,  $C$  is not shattered by  $\mathcal{H}$ . Consider  $C = \{x_1, x_2\}$ , we will have one of the following 3 cases:

**Case 1:**  $\|x_1\|_2 < \|x_2\|_2$ , we assume that we can obtain label  $(0, 1)$ . That means that  $x_1$  has label 0, so the corresponding sphere needs to have its radius  $r < \|x_1\|_2$ . Furthermore,  $x_2$  has label 1, so the corresponding sphere needs to have its radius  $\|x_2\|_2 \leq r$ .

$\Rightarrow \|x_2\|_2 \leq r < \|x_1\|_2$ , but  $\|x_1\|_2 < \|x_2\|_2$  (**contradiction** – there is no sphere centered in origin that can fulfill these constraints)

**Case 2:**  $\|x_1\|_2 = \|x_2\|_2$  we assume that we can obtain label  $(0, 1)$ . That means that  $x_1$  has label 0, so the corresponding sphere needs to have its radius  $r < \|x_1\|_2$ . Furthermore,  $x_2$  has label 1, so the corresponding sphere needs to have its radius  $\|x_2\|_2 \leq r$ .

$\Rightarrow \|x_2\|_2 \leq r < \|x_1\|_2$ , but  $\|x_1\|_2 = \|x_2\|_2$  (**contradiction** – there is no sphere centered in origin that can fulfill these constraints)

**Case 3:**  $\|x_1\|_2 > \|x_2\|_2$  we assume that we can obtain label  $(1, 0)$ . That means that  $x_2$  has label 0, so the corresponding sphere needs to have its radius  $r < \|x_2\|_2$ . Furthermore,  $x_1$  has label 1, so the corresponding sphere needs to have its radius  $\|x_1\|_2 \leq r$ .

$\Rightarrow \|x_1\|_2 \leq r < \|x_2\|_2$ , but  $\|x_1\|_2 > \|x_2\|_2$  (**contradiction** – there is no sphere centered in origin that can fulfill these constraints)

Since there is no subset  $C \subset \mathbb{R}^3$  with  $|C| = 2$  which can output all of the corresponding labels, because it certainly falls into one of the previous 3 scenarios,  $\text{VCdim}(\mathcal{H}) < 2$ .

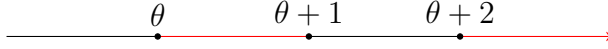
To conclude, since  $\text{VCdim}(\mathcal{H}) \geq 1$  and  $\text{VCdim}(\mathcal{H}) < 2 \Rightarrow \text{VCdim}(\mathcal{H}) = 1$ .

**Problem 3** What is the VC-dimension of the set of subsets  $I_\theta$  of the real line parameterized by a single parameter  $\theta$  where  $I_\theta = [\theta, \theta + 1] \cup [\theta + 2, \infty]$ ?


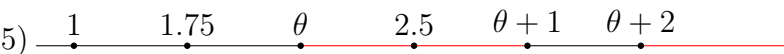
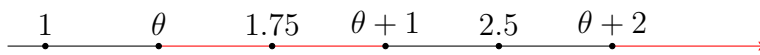
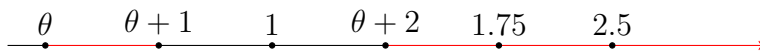
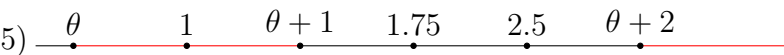
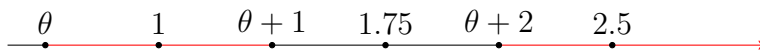
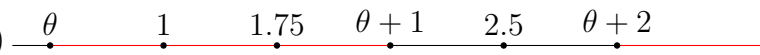
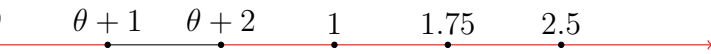
**Solution** We'll consider  $I$  to be the set of all  $I_\theta$ . Using the VC-dimension definition, we want to prove that  $\text{VCdim}(I) = 3$ . Therefore, we want to show that:

1. There  $\exists C \subset \mathbb{R}$ , where  $|C| = 3$ , that is shattered by  $I$ . ( $\text{VCdim}(I) \geq 3$ )
2.  $\forall C \subset \mathbb{R}$ , where  $|C| = 4$ ,  $C$  is not shattered by  $I$  ( $\text{VCdim}(I) < 4$ )

In general, each  $I_\theta$  will look similar to the following graph:



We will consider  $C = \{1, 1.75, 2.5\}$  subset of  $\mathbb{R}$ , with  $|C| = 3$  and we will show that  $C$  is shattered by  $I_\theta$ , obtaining all the possible labels from  $\{0, 1\}^3$ :

- label (0,0,0) – if  $(\theta < 2.5)$  
- label (0,0,1) – if  $(1.75 < \theta \leq 2.5)$  
- label (0,1,0) – if  $(1 < \theta < 1.5)$  
- label (0,1,1) – if  $(-1 < \theta < 0)$  
- label (1,0,0) – if  $(0.5 < \theta < 0.75)$  
- label (1,0,1) – if  $(0 \leq \theta \leq 0.5)$  
- label (1,1,0) – if  $(0.75 \leq \theta \leq 1)$  
- label (1,1,1) – if  $(\theta \leq -1)$  

We've found a set  $C$ ,  $|C| = 3$  that is shattered by  $I$ , so  $\text{VCdim}(I) \geq 3$ .

Now we want to show that  $\forall C \subset \mathbb{R}$ , where  $|C| = 4$ ,  $C$  is not shattered by  $I$ . Consider  $C = \{x_1, x_2, x_3, x_4\}$ , with  $x_1 < x_2 < x_3 < x_4$ . We assume that we can obtain labels (1, 0, 1, 0), meaning that label of  $x_4$  is 0 so this should come from one of the following two scenarios:

1.  $x_4 < \theta$ , implying that  $x_1, x_2, x_3$  have also label 0, since they are smaller than  $x_4$  (**contradiction**)
2.  $\theta + 1 < x_4 < \theta + 2$ ,  $\theta \leq x_3 \leq \theta + 1$  so  $x_1$  and  $x_2$  should both have label 0 since they are smaller than  $x_3$  (**contradiction**)

Since there is no subset  $C \subset \mathbb{R}$ , with  $|C| = 4$  which can output the following labels (1, 0, 1, 0),  $\text{VCdim}(I) < 4$ .

To conclude, since  $\text{VCdim}(I) \geq 3$  and  $\text{VCdim}(I) < 4 \Rightarrow \text{VCdim}(I) = 3$ .

**Problem 4** An axis aligned square in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain square. Formally, given the real numbers  $a_1, a_2, r > 0 \in \mathbb{R}$  we define the classifier  $h_{(a_1, a_2, r)}$  by:

$$h_{(a_1, a_2, r)}(x_1, x_2) = \begin{cases} 1, & \text{if } a_1 \leq x_1 \leq a_1 + r, a_2 \leq x_2 \leq a_2 + r \\ 0, & \text{otherwise} \end{cases}$$

The class of all axis aligned squares in the plane is defined as

$$\mathcal{H} = \{h_{(a_1, a_2, r)} : \mathbb{R}^2 \rightarrow \{0, 1\} | a_1, a_2, r \in \mathbb{R}, r > 0\}$$

Consider the realizability assumption.

- a) give a learning algorithm A that returns a hypothesis  $h_s$  from  $\mathcal{H}$ ,  $h_s = A(S)$  consistent with the training set S ( $h_s$  has the empirical risk 0 on S)
- b) find the sample complexity  $m_H(\epsilon, \delta)$  in order to show that  $\mathcal{H}$  is PAC – learnable
- c) compute  $\text{VCdim}(\mathcal{H})$

**Solution** a) Since we are under the realizability assumption, there exists  $f \in \mathcal{H}$ ,  $f = h^*$ , such that  $L_{D, f}(h^*) = 0$ , that labels the training data. We will consider the following training set  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ , where  $y_i = h^*(x_i)$  and  $x_i = (x_{i1}, x_{i2})$  ( $h^*$  labels each point inside the square  $Sq^*$  with label 1 and all other with label 0).

First, we need to find a learning algorithm A. Consider the following algorithm A, that takes as input the training set S, prior defined and outputs  $h_S$ .

$$h_S = h_{(a_{1S}, a_{2S}, r_S)},$$

where  $a_{1S} = \min_{i=1, m}(x_{i1})$  and  $y_i = 1$ ,  $a_{2S} = \min_{i=1, m}(x_{i2})$  and  $y_i = 1$  (representing the lower left positive corner), and for side length we need to compute the upper right positive corner. Let's take  $b_{1S}$  and  $b_{2S}$  to be the coordinates of the upper right corner. So  $b_{1S} = \max_{i=1, m}(x_{i1})$  and  $y_i = 1$ ,  $b_{2S} = \max_{i=1, m}(x_{i2})$  and  $y_i = 1$ . Having these four values, we will now have three possible cases:

1. if  $b_{1S} - a_{1S} = b_{2S} - a_{2S}$  (meaning that we can already encapsulate all our points within a square)  $\Rightarrow r_S = b_{1S} - a_{1S}$



2. if  $b_{1S} - a_{1S} > b_{2S} - a_{2S}$  (meaning that the width is larger than the height)  $\Rightarrow$  we need to extend our rectangle with  $\text{dif} = (b_{1S} - a_{1S}) - (b_{2S} - a_{2S})$  on height
3. otherwise, the height is larger than the width so we will need to extend our rectangle with  $\text{dif} = (b_{2S} - a_{2S}) - (b_{1S} - a_{1S})$  on the width

If we are on the first case, no other errors could be made for the training set S.

If we are on the **second case**, we compute:  $\text{upperLimit} = \min_{i=1,m}(x_{i2} > b_{2S})$ , where  $y_i = 0$  and  $\text{lowerLimit} = \max_{i=1,m}(x_{i2} < a_{2S})$ , where  $y_i = 0$  (i.e. the first point that is right over the rectangle and the first point that is right under the rectangle that have label 0).

1. if there is enough space under the rectangle to extend ( $a_{2S} - \text{lowerLimit} > \text{dif}$ )  $\Rightarrow$   $a_{1S}$  stays the same,  $a_{2S} = a_{2S} - \text{dif}$  and  $r_S = b_{1S} - a_{1S}$
2. if there is enough space above the rectangle to extend ( $\text{upperLimit} - b_{2S} > \text{dif}$ )  $\Rightarrow$   $a_{1S}, a_{2S}$  stay the same and  $r_S = b_{1S} - a_{1S}$
3. if we have to extend both above and below ( $a_{2S} - \text{lowerLimit} \leq \text{dif}$  and  $\text{upperLimit} - b_{2S} \leq \text{dif}$ )  $\Rightarrow$   $a_{1S}$  stays the same,  $a_{2S} = a_{2S} - (\text{dif} - (a_{2S} - \text{lowerLimit})) + \text{eps}$  and  $r_S = b_{1S} - a_{1S}$

If we are on the **third case**, we compute:  $\text{rightLimit} = \min_{i=1,m}(x_{i1} > b_{1S})$ , where  $y_i = 0$  and  $\text{leftLimit} = \max_{i=1,m}(x_{i1} < a_{21})$ , where  $y_i = 0$  (i.e. the first point on the right side of the rectangle and the first point on the left of the rectangle that have label 0).

1. if there is enough space on the left side of the rectangle to extend ( $a_{1S} - \text{leftLimit} > \text{dif}$ )  $\Rightarrow$   $a_{2S}$  stays the same,  $a_{1S} = a_{1S} - \text{dif}$  and  $r_S = b_{2S} - a_{2S}$
2. if there is enough space on the right side of the rectangle to extend ( $\text{rightLimit} - b_{1S} > \text{dif}$ )  $\Rightarrow$   $a_{1S}, a_{2S}$  stay the same and  $r_S = b_{2S} - a_{2S}$
3. if we have to extend both on the left and on the right ( $a_{1S} - \text{leftLimit} \leq \text{dif}$  and  $\text{rightLimit} - b_{1S} < \text{dif}$ )  $\Rightarrow$   $a_{2S}$  stays the same,  $a_{1S} = a_{1S} - (\text{dif} - (a_{2S} - \text{leftLimit})) + \text{eps}$  and  $r_S = b_{2S} - a_{2S}$

So for the training set  $S$ , the learning algorithm  $A$  outputs  $h_S$ , the function of the tightest square  $Sq_S$ , enclosing all the positive examples from  $S$ , without incorporating any negative examples with its extension from an initial rectangle to a square.

By construction,  $A$  is an ERM, meaning that  $L_S(h_S) = 0$  ( $h_S$  doesn't make any mistakes on the training set  $S$ ).

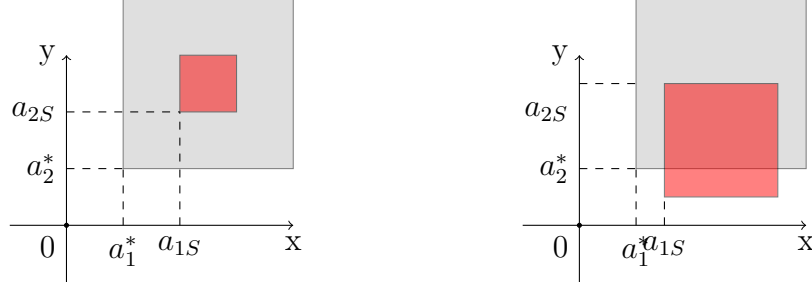
b) In order to show that  $\mathcal{H}$  is PAC learnable, we will start from the definition of PAC-learnability. We know that  $\mathcal{H}$  is PAC learnable if there exists a function  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$  and a learning algorithm  $A$  with the following property: for every  $\epsilon > 0, \delta > 0$ , for every labeling function  $f \in \mathcal{H}$  and for every distribution  $D \in \mathbb{R}^3$ , when we run the learning algorithm  $A$  on the training set  $S$  consisting of  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$  examples sampled i.i.d from  $D$  and labelled by  $f$ , the algorithm  $A$  returns a hypothesis  $h_S \in \mathcal{H}$  such that with probability at least  $1 - \delta$  (over the choice of examples), the real risk of  $h_S < \epsilon$ .

Since we are under the realizability assumption, there exists  $f \in \mathcal{H}$ ,  $f = h_{(a_1, a_2, r)}^*$ , such that  $L_{D, f}(h_{(a_1, a_2, r)}^*) = 0$ , that labels the training data. We will consider the following training set  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ , where  $y_i = h_{(a_1, a_2, r)}^*(x_i)$  and  $x_i = (x_{i1}, x_{i2})$  ( $h_{(a_1, a_2, r)}^*$  labels each point drawn from the square  $Sq^*$  with label 1 and all other with label 0).

We will use the algorithm earlier presented in the first part of the exercise, which describes exactly how for a given training set  $S$ , we get the corresponding hypothesis  $h_S \in \mathcal{H}$ . This  $h_S$  does no mistake on the training set.

$$\mathbb{P}_{S \sim D^m} (L_{h_{a_1, a_2, r}^*, D}(h_S) \leq \epsilon) \geq (1 - \delta), \text{ when } S \text{ has } m \geq m_{\mathcal{H}}(\epsilon, \delta) \text{ examples}$$

We notice that  $h_S$  can make errors in two ways: 1) in the  $Sq^* \setminus Sq_S$  (the region between the square outputted by the algorithm  $A$  for training set  $S$ , and the square determined by  $h_{(a_1, a_2, r)}^*$ ), by labelling points that should have had label 1, with label 0. All other points that are inside  $Sq_S$  and outside  $Sq^*$  will be labeled correctly. 2) in the  $Sq_S \setminus Sq^*$  (the region of  $Sq_S$  that goes outside of  $Sq^*$ , because of the way we've previously extended the rectangle into a square).



By fixing  $\epsilon > 0, \delta > 0$  and consider a distribution  $D \sim \mathbb{R}^2$ , we will consider the following two cases:

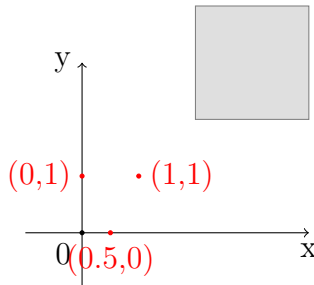
1. when  $D(Sq^*) = \mathbb{P}_{x \sim D}(x \in Sq^*) \leq \epsilon$  and  $Sq_S$  is inside  $Sq^*$   
 $L_{h_{a_1, a_2, r}, D}^*(h_S) = \mathbb{P}_{x \sim D}(h_S(x) \neq h_{a_1, a_2, r}^*(x)) = \mathbb{P}_{x \sim D}(x \in Sq^* \setminus Sq_S) \leq \mathbb{P}_{x \sim D}(x \in Sq^*) \leq \epsilon$ , so  
when sampling  $m$  points that leads to  $\Rightarrow \mathbb{P}_{S \sim D^m}(L_{h_{a_1, a_2, r}, D}^*(h_S) \leq \epsilon) = 1$
2. when  $D(Sp^*) = \mathbb{P}_{x \sim D}(x \in Sp^*) > \epsilon$  and  $Sq_S$  is inside  $Sq^*$
3. when  $D(Sp^*) = \mathbb{P}_{x \sim D}(x \in Sp^*) \leq \epsilon$ , but a part of  $Sq_S$  is outside  $Sq^*$
4. when  $D(Sp^*) = \mathbb{P}_{x \sim D}(x \in Sp^*) < \epsilon$ , but a part of  $Sq_S$  is outside  $Sq^*$

c) Using the VC-dimension definition, we want to prove that  $\text{VCdim}(\mathcal{H}) = 3$ . Therefore, we want to show that:

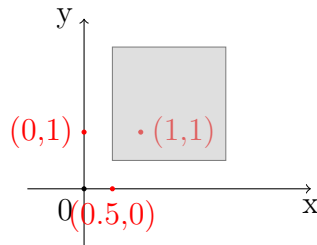
1. There  $\exists C \subset \mathbb{R}^2$ , where  $|C| = 3$ , that is shattered by  $\mathcal{H}$ . ( $\text{VCdim}(H) \geq 3$ )
2.  $\forall C \subset \mathbb{R}^2$ , where  $|C| = 4$ ,  $C$  is not shattered by  $\mathcal{H}$  ( $\text{VCdim}(H) < 4$ )

For the first assumption, we will choose the following subset  $C = \{(0.5, 0), (0, 1), (1, 1)\}$ , subset of  $\mathbb{R}^2$  with  $|C| = 3$  and will show that  $C$  is shattered by  $\mathcal{H}$ , by obtaining all the possible labels:

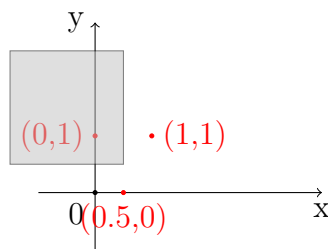
- label  $(0, 0, 0)$ , when every point is outside the square



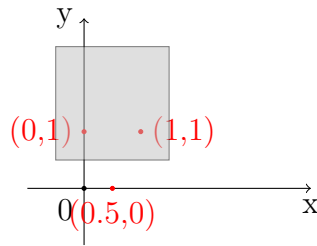
- label  $(0, 0, 1)$  - when only the third point is inside the square



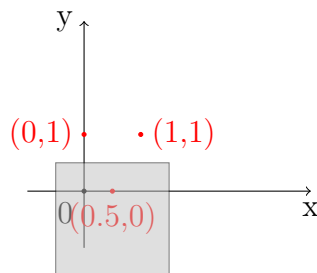
- label  $(0, 1, 0)$  - when only the second point is inside the square



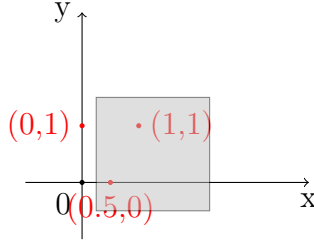
- label  $(0, 1, 1)$  - when only the second and the third point are inside the square



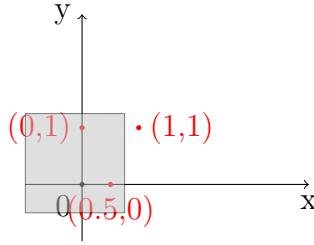
- label  $(1, 0, 0)$  - when only the first point is inside the square



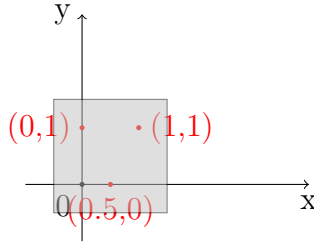
- label  $(1, 0, 1)$  - when only the first and the third point are inside the square



- label  $(1, 1, 0)$  - when only the first and the second point are inside the square



- label  $(1, 1, 1)$  - when all the points are inside the square



We've found a set  $C$ ,  $|C| = 3$  that is shattered by  $\mathcal{H}$ , so  $\text{VCdim}(\mathcal{H}) \geq 3$ .

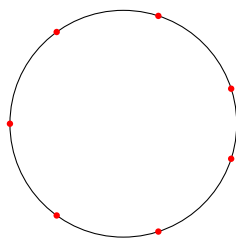
Now we want to show that there is no subset  $C$ , with  $|C| = 4$  such that all the labels can be obtained. Considering  $C = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$ , we can compute the convex-hull for these points. If the convex hull does not contain all the points, it means that there is at least one point inside the convex-hull. Therefore, we can not obtain the label where the interior points are 0 and the points that form the convex-hull are labeled 1.

Anyway, if all the 4 points form the convex-hull, we will not be able to obtain labels such as  $(1,0,1,0)$  or  $(0,1,0,1)$  (each diagonal corresponds to one of the two labels). Thus, for any subset  $C$  in  $\mathbb{R}^2$ , with  $|C| = 4$ , there is at least one label that can not be obtained  $\Rightarrow C$  is not shattered by  $\mathcal{H}$  ( $\text{VCdim}(\mathcal{H}) < 4$ ). To conclude, since  $\text{VCdim}(\mathcal{H}) \geq 3$  and  $\text{VCdim}(\mathcal{H}) < 4 \Rightarrow \text{VCdim}(\mathcal{H}) = 3$ .

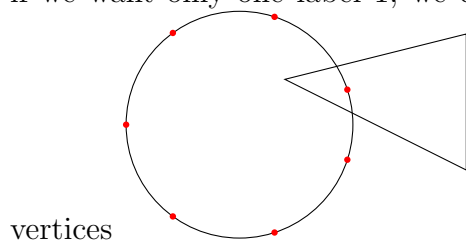
**Problem 5** Compute the VC-dimension of the class of convex  $d$ -gons (convex polygons with exactly  $d$  sides) in the plane. Provide a detailed proof of your result.

**Solution** We want to prove that the VC-dimension of the class of convex  $d$ -gons is  $2d + 1$ .

We'll start by considering the case of a triangle and we want to show that the VC-dimension for 3-gons is 7, so we'll place those 7 points on a circle. If those 7 points are not placed in such a manner, the convex hull of that set will not contain all of them so we will never be able to make label 1 for the interior points and 0 for the points that form the convex-hull. Also, for any other subset we need to have no interior point, and this property can happen only if they are placed on a circle. This rule applies in general, so from now on we will use only datasets of points placed in such a manner.

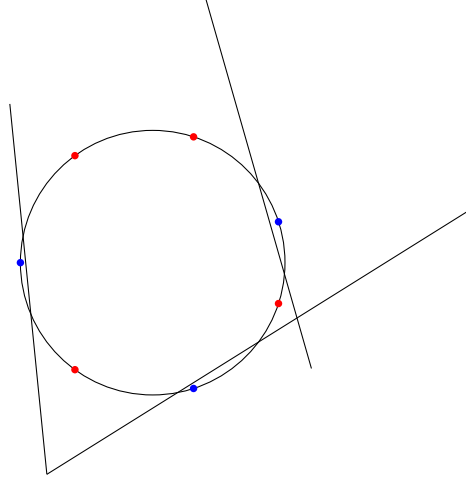


- if we want only one label 1, we can place that specific point in one of the triangle's



- if we want two labels 1, we can place an edge of the triangle exactly on those two points
- if we want three labels 1, we can place each triangle vertex in one of those points (triangle being in the interior of the circle)
- now, when the number of positive examples becomes greater than the number of negative examples (meaning at most 3 negative examples) – we can draw the tangent lines for those negative examples on the circle, creating the polygon which will contain exactly all the other points (the positive examples)

In general, since we want to prove that VC-dimension of a  $d$ -gon is equal to  $2d+1$ , we will always have a majority of labels (there are either more 1 labels, or 0 labels). If there are more 0 labels, it means that the number of positive labels is  $\leq d$ , so we can place each positive labeled point to be one of the vertices (the  $d$ -gon is formed inside the circle). If there are more 1 labels, it means that the number of negative labels is  $\leq d$ , so we will draw the tangents to the circle from each point and, without actually including those negative examples, the polygon formed by each intersection will keep within exactly the points with label 1. For the triangle example, if we have 4 positive examples and 3 negative ones, we will have following scenario:



So the VC-dimension for a  $d$ -gon is  $\geq (2d + 1)$ . In order to show that we can not obtain all the labels for  $2(d + 1)$  points, we will consider the case where labels alternate on the circle (positive, negative, positive,...). If this is the case, we have  $d + 1$  positive examples, so between each two of them we need one edge  $\Rightarrow$  a total of  $d + 1$  edges, which can not be obtained with a  $d$ -gon.

To conclude, since the VC-dimension of a  $d$ -gon is  $\geq (2d + 1)$  and  $< (2d + 2) \Rightarrow$  VC-dimension of a  $d$ -gon is equal to  $2d + 1$ .