TFML HW-2

SIDDHANT CHAUDHARY
BMC201953

Problem 1 (Problem 6.2 of book). Let $\mathcal{X}$ be a finite domain set, i.e $|\mathcal{X}| < \infty$. Let $k \leq |\mathcal{X}|$ be a number. We will figure out the VC dimensions of the given hypothesis classes.

(1) First, consider the class

$$\mathcal{H}_{=k}^{\mathcal{X}} = \left\{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\right\}$$

i.e we are considering the class of all functions that assign the value 1 to exactly k elements of $\mathcal{X}$. We claim that $\text{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = \min(k, |\mathcal{X}| - k)$. To show this, suppose $C = \left\{x_1, ..., x_{|C|}\right\}$ is any subset of $\mathcal{X}$ such that $|C| > \min(k, |\mathcal{X}| - k)$. We consider two cases.

(a) In the first case, suppose $\min(k, |\mathcal{X}| - k) = k$, and hence in this case $|C| > k$. Consider the all 1's function $\mathbf{1} : C \to \{1\}$. Clearly, because $|C| > k$, there is no h in the hypothesis class such that $h|_C = \mathbf{1}$, because h can assign the value 1 to exactly k elements, and not any higher number of elements.

(b) In the second case, suppose $\min(k, |\mathcal{X}| - k) = |\mathcal{X}| - k$, and hence in this case $|C| > |\mathcal{X}| - k$. In this case, consider the all zeroes function $\mathbf{0} : C \to \{0\}$ which assigns 0 to all the elements of C. Note that in this case, $|\mathcal{X}| - |C| < k$; this means that any extension of the function $\mathbf{0}$ to the whole set $\mathcal{X}$ can assign 1 to atmost $|\mathcal{X}| - |C| < k$ elements, and certainly there is no such function in the hypothesis class. So, it follows that there is no function in the hypothesis class which restricts to $\mathbf{0}$ on C.

So, this shows that any set of size $> \min(k, |\mathcal{X}| - k)$ cannot be shattered by the hypothesis class, and this proves our claim.

Next, suppose $|C| \leq \min(k, |\mathcal{X}| - k)$. Let $g : C \to \{0,1\}$ be any function. Let $l = |x \in C : g(x) = 1|$. It is clear that $l \leq \min(k, |\mathcal{X}| - k)$. Also, let $C' = \mathcal{X} - C = \left\{q_1, q_2, ..., q_{|\mathcal{X}|-|C|}\right\}$. Because $|C| \leq |\mathcal{X}| - k$, it is clear that $|\mathcal{X}| - |C| \geq k$. Now, consider the hypothesis h on $\mathcal{X}$ defined as follows: $h|_C = g$; moreover,

$$h(q_1) = h(q_2) = \cdots = h(q_{k-l}) = 1$$

and

$$h(q_{k-l+1}) = \cdots = h(q_{|\mathcal{X}|-|C|}) = 0$$

It is clear that h assigns 1 to exactly k elements of $\mathcal{X}$. Note that even if $l = 0$, we are in good shape because in that case $k - l = k \leq |\mathcal{X}| - |C|$. Since g was

Date: 11 November, 2021.

any arbitrary function, it follows that C can be shattered by the hypothesis class. Hence, it follows that $\text{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = \min(k, |\mathcal{X}| - k)$.

(2) Now, consider the class

$$\mathcal{H}_{at-most-k} = \left\{h \in \{0,1\}^{\mathcal{X}} : |x : h(x) = 1| \leq k \text{ or } |x : h(x) = 0| \leq k\right\}$$

i.e we are considering the class of all functions that either assign 1 to atmost k elements or assign 0 to atmost k elements. We claim that $\text{VCdim}(\mathcal{H}_{at-most-k}) = \min(|\mathcal{X}|, 2k+1)$. To show this, we will consider two cases; the first case will be when $|\mathcal{X}| \leq 2k+1$, and the second case will be when $|\mathcal{X}| > 2k+1$.

Consider the first case, i.e $|\mathcal{X}| \leq 2k+1$. In this case, we see that $\min(|\mathcal{X}|, 2k+1) = |\mathcal{X}|$. We now argue that $\mathcal{X}$ can be shattered. To see this, let $g : \mathcal{X} \to \{0,1\}$ be any function. Now, let $c_1 = |x : g(x) = 1|$ and let $c_2 = |x : g(x) = 0|$. Clearly, we see that

$$c_1 + c_2 = |\mathcal{X}| \leq 2k+1$$

Clearly, one of c_1 or c_2 has to be $\leq k$; if not, then $c_1 \geq k+1$ and $c_2 \geq k+1$, and in that case we will have $c_1 + c_2 \geq 2k+2 > 2k+1$, a contradiction. Without loss of generality, suppose $c_1 \leq k$. But this clearly implies that $g \in \mathcal{H}_{at-most-k}$. The case $c_2 \leq k$ is symmetric to this. So, we see that in this case, $\mathcal{X}$ can be shattered, and hence $\text{VCdim}(\mathcal{H}_{at-most-k}) = |\mathcal{X}| = \min(|\mathcal{X}|, 2k+1)$.

Now, consider the second case, i.e $|\mathcal{X}| > 2k+1$, which means $|\mathcal{X}| \geq 2k+2$. In this case, we see that $\min(|\mathcal{X}|, 2k+1) = 2k+1$. We argue that $2k+1$ is the VC dimension in this case. So, let C be any subset of $\mathcal{X}$ such that $|C| = 2k+2 \leq |\mathcal{X}|$. Consider the hypothesis $g : C \to \{0,1\}$ which assigns 1 to exactly $k+1$ elements of C, and assigns 0 to the rest $k+1$ elements. Clearly, note that g *cannot be* the restriction of any hypothesis $h \in \mathcal{H}_{at-most-k}$, which is clear by the definition of the hypothesis class. So, we've shown that no set of size $2k+2$ can be shattered.

Next, suppose C is any subset of $\mathcal{X}$ with $|C| = 2k+1$. Let $g : C \to \{0,1\}$ be any map. Again, let $c_1 = |x \in C : g(x) = 1|$ and $c_2 = |x \in C : g(x) = 0|$. Clearly, we again have that

$$c_1 + c_2 = |C| = 2k+1$$

As before, one of c_1 or c_2 has to be $\leq k$; without loss of generality, suppose $c_1 \leq k$. Consider the hypothesis $h : \mathcal{X} \to \{0,1\}$ such that: $h|_C = g$ and $h(x) = 0$ for any $x \in \mathcal{X} - C$. Clearly, $|x \in \mathcal{X} : h(x) = 1| = |x \in C : h(x) = 1| = c_1 \leq k$, and by definition, $h \in \mathcal{H}_{at-most-k}$. So, we have shown that g is the restriction of some h in the hypothesis class. The case $c_2 \leq k$ is symmetric to this case. So, it follows that any set of size $2k+1$ can be shattered, and hence $\text{VCdim}(\mathcal{H}_{at-most-k}) = 2k+1 = \min(|\mathcal{X}|, 2k+1)$.

So, in all cases, we have shown that $\text{VCdim}(\mathcal{H}_{at-most-k}) = \min(|\mathcal{X}|, 2k+1)$, and this completes the proof.

Problem 2 (Problem 6.6 of book). In this problem, we will compute the VC dimension of Boolean conjunctions. Let $d \geq 2$ be an integer, and let $\mathcal{H}_{con}^d$ be the class of Boolean conjunctions over the variables $x_1, ..., x_d$. We will do it in the steps given in the problem.

1: We show that

$$|\mathcal{H}^d_{con}| \leq 3^d + 1$$

Note that if Φ is a boolean conjunction over the variables $x_1, ..., x_d$, and if for some variable x_i, both literals x_i and $\neg x_i$ occur in Φ, then Φ can never be satisfied, i.e $\Phi(x_1, ..., x_d) = 0$. So, Φ is just the all negative conjunction. So, we will assume that Φ does not contain both x_i and $\neg x_i$. So now, for each $1 \leq i \leq d$, we have a choice of including either x_i, $\neg x_i$ or none of these in the conjunction Φ. So, there are 3^d such possible conjunctions. Hence, including the all negative conjunction, we see that

$$|\mathcal{H}^d_{con}| = 3^d + 1$$

and this proves the claim.

2: Suppose $k = \text{VCdim}(\mathcal{H}^d_{con})$. This means that a size of set k is shattered, i.e we can get all possible 2^k functions by restricting $\mathcal{H}^d_{con}$ to the set. Clearly,

$$2^k \leq 3^d + 1$$

which implies that

$$k \leq \log(3^d + 1)$$

Because k is an integer and $d \geq 2$, we have

$$k \leq \lfloor \log(3^d + 1) \rfloor = \lfloor \log(3^d) \rfloor \leq d \log 3$$

and hence we conclude that

$$\text{VCdim}(\mathcal{H}^d_{con}) \leq d \log 3$$

3: We now show that $\mathcal{H}^d_{con}$ shatters the set of unit vectors $\{\boldsymbol{e}_i : i \leq d\}$. This is actually very easy to see. Let $g : \{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\} \to \{0, 1\}$. Let $\{i_1, ..., i_r\} \subseteq [d]$ be the set of those indices for which $g(\boldsymbol{e}_{i_1}) = \cdots = g(\boldsymbol{e}_{i_r}) = 1$; we have $0 \leq r \leq d$. Consequetly, let $\{j_1, ..., j_{d-r}\} = [d] - \{i_1, ..., i_r\}$ be the set of those indices for which $g(\boldsymbol{e}_{j_1}) = \cdots = g(\boldsymbol{e}_{j_{d-r}}) = 0$. Now, if $d - r = 0$, i.e if $r = d$, then we let h to be the all ones classifier, i.e the empty boolean conjunction. Clearly, $h \in \mathcal{H}^d_{con}$, and

$$h|_{\{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\}} = g$$

So, suppose $r < d$, and in that case, $d - r > 0$. Consider the boolean conjunction

$$h(x_1, ..., x_d) = \overline{x_{j_1}} \wedge \overline{x_{j_2}} \wedge \cdots \wedge \overline{x_{j_{d-r}}}$$

It is now easy to see that

$$h|_{\{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\}} = g$$

Finally, suppose $r = 0$. In that case, simply take h to be the all negative classifier, i.e

$$h(x_1, ..., x_d) = x_1 \wedge \overline{x_1}$$

and again we see that

$$h|_{\{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\}} = g$$

Since g was an arbitrary classifier, we have shown that $\mathcal{H}^d_{con}$ shatters the set $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\}$. Using this, we can conclude that

$$\text{VCdim}(\mathcal{H}^d_{con}) \geq d$$

4: Next, we will show that $\text{VCdim}(\mathcal{H}_{con}^d) \leq d$. For the sake of contradiction, suppose there is a set $C = \{c_1, ..., c_{d+1}\}$ that is shattered by $\mathcal{H}_{con}^d$. Now, let $h_1, ..., h_{d+1}$ be hypothesis in $\mathcal{H}_{con}^d$ that satisfy

$$\forall i, j \in [d+1], h_i(c_j) = \begin{cases} 0 & i = j \\ 1 & \text{otherwise} \end{cases}$$

In simple words, we are considering functions on C which are 0 at exactly one point and 1 at all other points, and such hypothesis $h_1, ..., h_{d+1}$ exist because C is shattered. Now, for each $i \in [d+1]$, this means that the conjunction h_i contains some literal l_i which is false on c_i but is true for all c_j with $j \neq i$. So, we have a set of $d+1$ literals $\{l_1, ..., l_{d+1}\}$. But recall that there are only d variables $x_1, ..., x_d$. So, by the pigeon hole principle, it follows that for some $i < j \leq d+1$, the literals l_i and l_j use the same variable x_k for some $1 \leq k \leq d$. So, we have two cases to consider.

(1) In the first case, suppose $l_i = x_k$. Because c_j satisfies l_i, it must be the case that the value of x_k in c_j is 1. Now, we know that c_j does not satisfy l_j, and hence it must be the case that $l_j = \overline{x_k}$. Now, since $d \geq 2$, we see that $d+1 \geq 3$, and hence there is some index $1 \leq s \leq d+1$ other than i and j. We also know that c_s satisfies l_i and l_j (as $c \neq i, j$); but this is clearly a contradiction as an assignment cannot satisfy both x_k and $\overline{x_k}$.
(2) In the second case, we have $l_i = \overline{x_k}$. This case is symmetric to the above case, as we will have $l_j = x_k$ in this case, and the rest of the reasoning is the same.

So in all cases, we have arrived at a contradiction. Hence, it must be the case that $\text{VCdim}(\mathcal{H}_{con}^d) \leq d$, and combined with **Step 3**, it follows that

$$\text{VCdim}(\mathcal{H}_{con}^d) = d$$

5: Now let $\mathcal{H}_{mcon}^d$ be the class of *monotone* Boolean conjunctions over $\{0,1\}^d$, i.e the conjunctions in $\mathcal{H}_{mcon}^d$ do not contain any negations. Also, we augment $\mathcal{H}_{mcon}^d$ with the all negative hypothesis h^-. We show that

$$\text{VCdim}(\mathcal{H}_{mcon}^d) = d$$

First, note that $|\mathcal{H}_{mcon}^d| = 2^d + 1$; this is true because for every $1 \leq i \leq d$, we have to choose whether to include x_i or not in the conjunction, and we add 1 to include the all negative conjunction. So, if k is the VC dimension of this class, then clearly

$$2^k \leq |\mathcal{H}_{mcon}^d| = 2^d + 1$$

which implies

$$k \leq \log(2^d + 1)$$

Again, since k is an integer, this means

$$k \leq \left\lfloor \log(2^d + 1) \right\rfloor = \left\lfloor \log(2^d) \right\rfloor = d$$

Next, we will show that a set of size d can be shattered by the class. Consider the set

$$C := \{\boldsymbol{o}_j = (1, 1, ..., 1) - \boldsymbol{e}_j : 1 \leq j \leq d\} = \{(0, 1, ..., 1), (1, 0, ..., 1), ..., (1, 1, ..., 0)\}$$

i.e we are considering the set of vectors in which exactly one coordinate is 0. Note that the ith coordinate of $\boldsymbol{o}_i$ is 0, and all the other coordinates are 1. Now, let $g : \{\boldsymbol{o}_i : 1 \leq i \leq d\} \to \{0, 1\}$ be any classifier. Let $\{i_1, ..., i_r\}$ be the set of indices for which $g(\boldsymbol{o}_{i_1}) = \cdots = g(\boldsymbol{o}_{i_r}) = 1$, and let $\{j_1, ..., j_{d-r}\} = [d] - \{i_1, ..., i_r\}$ be the set of all those indices for which $g(\boldsymbol{o}_{j_1}) = \cdots = g(\boldsymbol{o}_{j_{d-r}})$. First, suppose $r = 0$. In that case, we let h be the all negative classifier. Clearly, $h \in \mathcal{H}^d_{mcon}$ and we have

$$h|_{\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_d\}} = g$$

Next, suppose $r = d$, i.e $d - r = 0$. In that case, we let h be the all ones classifier, i.e the conjunction corresponding to h is empty. Again, $h \in \mathcal{H}^d_{mcon}$, and again

$$h|_{\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_d\}} = g$$

So, we assume that $0 < r < d$. In that case, we consider the following conjunction.

$$h(x_1, .., x_d) = x_{j_1} \wedge x_{j_2} \wedge \cdots \wedge x_{j_{d-r}}$$

Clearly, again we have $h \in \mathcal{H}^d_{mcon}$ and again

$$h|_{\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_d\}} = g$$

Since g was arbitrary, we have shown that the class $\mathcal{H}^d_{mcon}$ shatters the set $\{\boldsymbol{o}_1, ..., \boldsymbol{d}_d\}$. Hence, combining all the facts above, we see that

$$\text{VCdim}(\mathcal{H}^d_{mcon}) = d$$

and this proves the claim.

Lemma 0.1. *Suppose* $0.x_1x_2x_3...$ *is the binary representation of* $x \in (0, 1)$*. Then, for any natural number* m*,*

$$\lceil \sin(2^m \pi x) \rceil = (1 - x_m)$$

if there is some $k \geq m$ *s.t* $x_k = 1$*. Here, the convention is* $\lceil -1 \rceil = 0$*.*

Proof. We have the following.

$$\begin{aligned}
\sin(2^m\pi x) &= \sin(2^m\pi(0.x_1x_2x_3...)) \\
&= \sin(2\pi(x_1x_2...x_{m-1}.x_mx_{m+1}...)) \\
&= \sin(2\pi(x_1x_2...x_{m-1}.x_mx_{m+1}...) - 2\pi(x_1x_2...x_{m-1}.0)) \\
&= \sin(2\pi(0.x_mx_{m+1}...))
\end{aligned}$$

where in the second last step we have used the periodicity of sin. Now, we consider two cases.

(1) In the first case, suppose $x_m = 0$. In that case, $0.x_mx_{m+1}... < \frac{1}{2}$, and hence $2\pi(0.x_mx_{m+1}...) < \pi$. Also, because there is some $k \geq m$ with $x_k = 1$, we have that $0.x_mx_{m+1}... > 0$. This means that $2\pi(0.x_mx_{m+1}...) \in (0, \pi)$, and hence the sin of this number is positive, implying that

$$\lceil \sin(2^m\pi x) \rceil = 1 = 1 - x_m$$

(2) In the second case, suppose $x_m = 1$. In this case, we see that $2\pi(0.x_mx_{m+1}...) \in [\pi, 2\pi)$, and hence sin of this quantity is non-positive. By our convention, this clearly means that

$$\lceil \sin(2^m\pi x) \rceil = 0 = 1 - x_m$$

So in all cases, the given equality holds, and this completes the proof. ■

Problem 3 (Problem 6.8 of book). Let $\mathcal{X} = \mathbf{R}$, and define

$$\mathcal{H} = \{x \mapsto \lceil \sin(\theta x) \rceil : \theta \in \mathbf{R}\}$$

with the convention that $\lceil -1 \rceil = 0$. We now prove that $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Let $n \in \mathbf{N}$ be any natural number. We will exhibit a set $\{x_1, ..., x_n\} \subset [0, 1]$ shattered by $\mathcal{H}$. To do so, we will use **Lemma** 0.1. Consider all the 2^n possible labellings of n numbers (i.e we consider all vectors in the set $\{0, 1\}^n$); enumerate this set in the usual dictionary order, i.e

$$\{0, 1\}^n = \{v_1, v_2, ..., v_{2^n}\}$$

where $v_1 = (0, 0, ..., 0)$ and $v_{2^n} = (1, 1, 1, ..., 1)$. . The fact that v_{2^n} is the all 1s vector will be important to us.

Define $x_1, ..., x_n \in (0, 1)$ as follows: write down each x_i in a separate line; each x_i will have a binary representation of the form $0.a_{i,1}a_{i,2}a_{i,3} \cdots a_{i,2^n}$; moreover, we choose the binary representations such that for each $1 \leq j \leq 2^n$,

$$(a_{1,j}, a_{2,j}, ..., a_{n,j}) = v_j$$

i.e the jth column of bits is the vector v_j. A pictorial representation of these numbers is given below.

$$\begin{aligned} x_1 &= 0.0 \cdots 1 \\ x_2 &= 0.0 \cdots 1 \\ &\vdots \\ x_n &= 0.0 \cdots 1 \end{aligned}$$

Now, suppose $1 \leq m' \leq 2^n$. Then, by **Lemma** 0.1, we know that

$$\left\lceil \sin(2^{m'} \pi x_i) \right\rceil = (1 - a_{i,m'}) \tag{0.1}$$

where we are using the fact that $x_{i,2^n} = 1$ for each i (i.e the k in the statement of the lemma is $k = 2^n$).

What this means is the following: let v_m for $1 \leq m \leq 2^n$ be any labelling. Consider the labelling $\overline{v_m}$, i.e the labelling obtained by flipping all bits of v_m, or equivalently, applying the function $x \mapsto 1 - x$ to each bit of the vector v_m. Clearly, $\overline{v_m}$ is a labelling too, and hence there is some $1 \leq m' \leq 2^n$ such that $\overline{v_m} = v_{m'}$. So, to obtain the labelling v_m, we just consider the hypothesis

$$h(x) = \left\lceil \sin(2^{m'} \pi x) \right\rceil$$

Then, by equation (0.1) that we showed above, we have

$$h(x_i) = 1 - a_{i,m'} = 1 - (v_{m'})_i = (v_m)_i$$

and hence the hypothesis $h \in \mathcal{H}$ labels the points according to the labelling v_m. So, we have shown that all the labellings can be obtained by restricting functions in $\mathcal{H}$ to these set of points. Since n was arbitrary, it follows that

$$\mathrm{VCdim}(\mathcal{H}) = \infty$$

and this completes the proof.

Problem 4 (Problem 9.4 of book). Let $m > 1$ be any integer. Let $R = \sqrt{m} > 1$, and let $\boldsymbol{w}^* = (0, 0, 1)$. We will produce examples $(\boldsymbol{x}_i, y_i)$ for $1 \leq i \leq m$ where each $\boldsymbol{x}_i$ is of the form $(a_i, b_i, 1)$ with $a_i^2 + b_i^2 + 1 = R^2$. Also, observe that for such examples, we have

$$y_i((\boldsymbol{w}^*)^T \boldsymbol{x}_i) = y_i^2 = 1$$

and hence the constant B in the statement of the upper bound is atmost 1. The perceptron algorithm guarantees atmost $(RB)^2 \leq R^2 = m$ mistakes; we will produce these examples so that the perceptron makes exactly $R^2 = m$ mistakes.

Suppose $\boldsymbol{w}_{t-1}$ is the separator vector when we enter time step t. The perceptron initialises $\boldsymbol{w}_0 = \mathbf{0}$. At each round t, we will give an example $(\boldsymbol{x}_t, 1)$ where $\boldsymbol{x}_t = (a_t, b_t, 1)$ such that $a_t^2 + b_t^2 + 1 = R^2$ and $\boldsymbol{w}_{t-1}^T \boldsymbol{x}_t = 0$, i.e the perceptron makes a mistake at time step t on the tth example.

Our first point $\boldsymbol{x}_1$ will be

$$\boldsymbol{x}_1 = (\sqrt{R^2 - 1}, 0, 1)$$

Clearly,

$$w_0^T \boldsymbol{x}_1 = \mathbf{0}^T \boldsymbol{x}_1 = 0$$

So, the perceptron will do the update

$$\boldsymbol{w}_1 \leftarrow \boldsymbol{w}_0 + \boldsymbol{x}_1 = \boldsymbol{x}_1$$

and so observe that $\boldsymbol{w}_1$ is a vector of the form $(\alpha, \beta, 1)$ where α, β are some scalars. Also, note that

$$||\boldsymbol{w}_1||^2 = ||\boldsymbol{x}_1||^2 = R^2 = 1 \cdot R^2$$

Now suppose all the examples till time step $t-1$ have been given, where $R^2 \geq t > 1$ such that $\boldsymbol{w}_{t-1} = (\alpha_{t-1}, \beta_{t-1}, t-1)$ where $\alpha_{t-1}, \beta_{t-1}$ are scalars and

$$||\boldsymbol{w}_{t-1}||^2 = (t-1) \cdot R^2$$

The above equation just means

$$\alpha_{t-1}^2 + \beta_{t-1}^2 + (t-1)^2 = (t-1)R^2$$

which implies

$$\alpha_{t-1}^2 + \beta_{t-1}^2 = (t-1)[R^2 - t + 1]$$

Because $t \leq R^2$ the above quantity is non-negative and makes sense.

We will now give a way to come up with example $\boldsymbol{x}_t$ such that the same equalities continue to hold. Consider the matrix M_{t-1} defined as follows.

$$M_{t-1} = \begin{bmatrix} \frac{\alpha_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & \frac{\beta_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & 0 \\ \frac{-\beta_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & \frac{\alpha_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

M_{t-1} is nothing but the rotation matrix that rotates $\boldsymbol{w}_{t-1}$ about the z-axis to make the y-coordinate of $\boldsymbol{w}_{t-1}$ zero. This will be useful as it will simplify our calculation. It is clear that

$$M_{t-1}^{-1} = \begin{bmatrix} \frac{\alpha_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & \frac{-\beta_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & 0 \\ \frac{\beta_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & \frac{\alpha_{t-1}}{\sqrt{(t-1)[R^2-t+1]}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Now, it is easy to observe that

$$\boldsymbol{w}'_{t-1} = M_{t-1}\boldsymbol{w}_{t-1} = M_{t-1}(\alpha_{t-1}, \beta_{t-1}, 1) = (\sqrt{(t-1)[R^2 - t + 1]}, 0, (t-1))$$

where the above equation is matrix multiplication. Let P_{t-1} be the quantity

$$P_{t-1} = \sqrt{(t-1)[R^2 - t + 1]}$$

So, we see that

$$\boldsymbol{w}'_{t-1} = (P_{t-1}, 0, (t-1))$$

Now consider the *rotated vector* $\boldsymbol{w}'_{t-1}$. Based on this vector, we will choose our new point $\boldsymbol{x}_t$. Suppose the point $\boldsymbol{x}_t$ is $(a'_t, b'_t, 1) = \boldsymbol{x}'_t$ in the *rotated coordinate system.* We choose

$$a'_t = \frac{-(t-1)}{P_{t-1}}$$

Then, observe that

$$(\boldsymbol{w}'_{t-1})^T(a'_t, b'_t, 1) = \frac{-(t-1)}{P_{t-1}} \cdot P_{t-1} + 0 + (t-1) = 0$$

i.e perceptron will make a mistake at the point $(a'_t, b'_t, 1)$. Now, observe that

$$a'^2_t + 1 = \frac{(t-1)^2}{P^2_{t-1}} + 1 = \frac{(t-1)}{R^2 - t + 1} + 1 = \frac{R^2}{R^2 - t + 1} \leq R^2$$

where we have used the fact that $t \leq R^2$. So, the quantity

$$\sqrt{R^2 - a'^2_t - 1}$$

makes sense, and if we put

$$b'_t = \sqrt{R^2 - a'^2_t - 1}$$

then we will have

$$a'^2_t + b'^2_t + 1 = R^2$$

So, the coordinates of the point $\boldsymbol{x}_t$ in the rotated coordinate system are

$$\boldsymbol{x}'_t = \left(\frac{-(t-1)}{P_{t-1}}, \sqrt{R^2 - \frac{(t-1)^2}{P^2_{t-1}} - 1}, 1\right)$$

So in the original coordinate system, the coordinates of $\boldsymbol{x}_t$ are

$$\boldsymbol{x}_t = M^{-1}_{t-1}\boldsymbol{x}'_t$$

Since rotations preserve norm, we see that

$$||\boldsymbol{x}_t||^2 = ||\boldsymbol{x}'_t||^2 = a'^2_t + b'^2_t + 1 = R^2$$

So, as promised initially, $\boldsymbol{x}_t$ is a point of the form $(a_t, b_t, 1)$ with $a^2_t + b^2_t + 1 = R^2$. Moreover, since rotations preserve inner products, we see that

$$0 = (\boldsymbol{w}'_{t-1})^T\boldsymbol{x}'_t = \boldsymbol{w}^T_{t-1}\boldsymbol{x}_t$$

i.e the perceptron will make a mistake at time step t. Also, the above equation means that $\boldsymbol{w}_{t-1}$ and $\boldsymbol{x}_t$ are orthogonal to each other. The update will be

$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} + \boldsymbol{x}_t$$

and hence $\boldsymbol{w}_t$ will be a vector of the form $\boldsymbol{w}_t = (\alpha_t, \beta_t, t)$ as we wanted.

Finally by **Pythagoras Theorem**, we have

$$||\boldsymbol{w}_t||^2 = ||\boldsymbol{w}_{t-1}||^2 + ||\boldsymbol{x}_t||^2 = (t-1) \cdot R^2 + R^2 = t \cdot R^2$$

and hence we have successfully shown how to construct the tth point $\boldsymbol{x}_t$. This way, for all $1 \leq t \leq R^2 = m$, we have produced examples $\boldsymbol{x}_t$ such that the perceptron makes a mistake at every step, i.e the perceptron makes exactly m mistakes. This completes the construction.

Problem 6 (Problem 10.1 of book). In this problem, we will use **Corollary 4.6** of the book, which states the following (we have also proven this in class): *let $\mathcal{H}$ be a finite hypothesis class, Z a domain, and $l : \mathcal{H} \to Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ is agnostically PAC learnable using ERM with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

We now solve the problem. Let A be an algorithm such that the following is true: there is some $\delta_0 \in (0,1)$ and a function $m_{\mathcal{H}} : (0,1) \to \mathbf{N}$ such that for every $\epsilon \in (0,1)$, if $m \geq m_{\mathcal{H}}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds that with probability atleast $1-\delta_0$,

$$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

We will come up with a procedure that uses A and learns $\mathcal{H}$ in the usual agnostic PAC learning model, i.e we will boost the confidence parameter δ. We will also show that to do this the sample complexity has the following upper bound.

$$m_{\mathcal{H}}(\epsilon, \delta) \leq k m_{\mathcal{H}}(\epsilon) + \left\lceil \frac{2\log\left(4k/\delta\right)}{\epsilon^2} \right\rceil$$

Above,

$$k = \left\lceil \frac{\log(\delta)}{\log(\delta_0)} - \frac{1}{\log(\delta_0)} \right\rceil$$

We do the following: we divide our data into $k+1$ chunks. The first k chunks will consist of $m_{\mathcal{H}}(\epsilon/2)$ examples. We will describe the last chunk later.

Now, we run the algorithm A on the first k chunks to obtain outputs $h_1, ..., h_k$. Note that by the guarantees of algortihm A, we know that for each i,

$$\mathbf{P}\left[L_{\mathcal{D}}(h_i) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}\right] \geq 1 - \delta_0$$

These means that

$$\begin{aligned}
&\mathbf{P}\left[L_{\mathcal{D}}(h_i) > \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}, \forall 1 \leq i \leq k\right] \leq \delta_0^k && (0.2)\\
&\leq \delta_0^{\frac{\log(\delta)}{\log(\delta_0)} - \frac{1}{\log(\delta_0)}} && (0.3)\\
&= 2^{\log(\delta) - 1} && (0.4)\\
&= \frac{\delta}{2} && (0.5)
\end{aligned}$$

The above inequality implies that

$$\mathbf{P}\left[\min_{1 \leq i \leq k} L_{\mathcal{D}}(h_i) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}\right] \geq 1 - \frac{\delta}{2}$$

Now let us describe what we do with the $k+1$th chunk. We let the size of this chunk be q

$$\left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$$

Then, we will run ERM with this chunk over the hypothesis class $\{h_1, ..., h_k\}$. Suppose the output of this is $\hat{h}$. Clearly, this is a finite hypothesis class of size k. Now, note that **Corollary 4.6** (mentioned in the very beginning) guarantees that

$$m_{\{h_1,...,h_k\}}(\epsilon/2, \delta/2) \leq \left\lceil \frac{2\log(4|\{h_1, ..., h_k\}|/\delta)}{\epsilon^2} \right\rceil = \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$$

This means that with probability atmost $\frac{\delta}{2}$, running ERM over the class $\{h_1, ..., h_k\}$ on the $k+1$th chunk results in $\hat{h}$ such that

$$L_{\mathcal{D}}(\hat{h}) > \min_{i\in[k]} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2} \tag{0.6}$$

Using (0.5) and (0.6) and a simple union bound, we see that

$$\mathbf{P}\left[L_{\mathcal{D}}(h_i) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}, \forall 1 \leq i \leq k \text{ or } L_{\mathcal{D}}(\hat{h}) > \min_{i\in[k]} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2}\right] \leq \delta$$

This means that

$$\mathbf{P}\left[L_{\mathcal{D}}(h_i) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}, \forall 1 \leq i \leq k \text{ and } L_{\mathcal{D}}(\hat{h}) \leq \min_{i\in[k]} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2}\right] \geq 1-\delta$$

which is equivalent to saying that

$$\mathbf{P}\left[L_{\mathcal{D}}(\hat{h}) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right] \geq 1-\delta$$

and this is nothing but the requirement in the definition of agnostic PAC learning. So, we've shown a successful PAC learner.

Now, the sample complexity is simply $m_{\mathcal{H}}(\epsilon/2)$ times k, plus the size of the $k+1$th chunk, i.e the sample complexity is

$$m_{\mathcal{H}}(\epsilon, \delta) \leq k m_{\mathcal{H}}(\epsilon/2) + \left\lceil \frac{2\log(2k/\delta)}{\epsilon^2} \right\rceil$$