Or Bar Yaacov
ID - 300486396

3. Let $\mathcal{H}$ be a non PAC learnable hypothesis class. Suppose A(S) is an algorithm that always returns the hypothesis $\forall x \in \mathcal{X}, h(x) = 0$.
$\mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)\right]=\mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h\right)\right] = L_{\mathcal{D}}\left(h\right) = \mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_{\mathcal{S}}\left(h\right)\right] = \mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_{\mathcal{S}}\left(A\left(S\right)\right)\right] \leq \mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_{\mathcal{S}}\left(A(S)\right)\right] + \epsilon_m$
Therefore the conditions hold for a non PAC learnable hypothesis class.

4. Let $\mathcal{H}$ be an hypothesis class of binary classifiers.
Suppose that $\mathcal{H}$is agnostic PAC learnable and let A be a learning algorithm that learns $\mathcal{H}$with sample complexity $m_{\mathcal{H}}(.,.)$.
Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}\times\{0,1\}$ and let f be the true function.
Since $\mathcal{H}$is agnostic PAC learnable and A is learning algorithm that learns $\mathcal{H}$with sample complexity we know that for all $\epsilon, \delta \in (0,1)$
$$Pr\left(L_D(h) \leq \underset{h\in\mathcal{H}}{min} L_D\left(h\right) + \epsilon\right) \geq 1 - \delta$$
where
$$L_D\left(h\right) = \mathcal{D}\left(\{(x,y) : h(x) \neq y\}\right)$$
Let us take the realizability assumption, that is,
$$\exists f \in \mathcal{H} s.t. \forall i\left(x_1, ..., x_m\right), y_i = f\left(x_i\right)$$
This means that $\underset{h\in\mathcal{H}}{min} L_D\left(h\right) = 0$
We may further assume w.l.o.g that $Pr\left(y|x\right)$ is determined deterministically by $f(x)$ since the realizability assumption tells us that $y_i = f(x_i)$ and therefore $Pr(y_i = f(x_i)|x_i) = 1$.
This means that $L_D\left(h\right) = \underset{(x,y)\sim\mathcal{D}}{Pr}\left(h(x) \neq y\right) = L_{D,f}\left(h\right) = \underset{x\sim\mathcal{D}}{Pr}\left(h(x) \neq f(x)\right)$
And so, it holds that :
$$Pr\left(L_{D,f}(h) \leq \epsilon\right) \geq 1 - \delta$$
And therefore $\mathcal{H}$ is PAC learnable and A is a successful PAC learner for $\mathcal{H}$.

5. Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{Singleton} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$ where
$$h_z\left(x\right) = \begin{cases} 1 & x = z \\ 0 & x \neq z \end{cases}, h^-\left(x\right) = 0\,\forall x \in \mathcal{X}$$

(a) Let's recall the empirical risk definition:

$$L_S(h) = \frac{1}{m} |\{i : h(x_i) \neq y_i\}|$$

An ERM based algorithm will have an input of a training set $S = (x_1, y_1), ..., (x_m, y_m)$ and output any $h \in \mathcal{H}_{Singleton}$ which minimizes the empirical risk.
I suggest the following:

i. If all labels $y_1, ...y_m = 0$ return h^-
ii. Else find the first label $y_i = 1$ and return h_{x_i}
The realizability assumption assures us the training set will only contain at most one unique sample that is labeled as 1 and therfore a sample labled 1 uniquely defines the true function.

(b) We must show that there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$, for all distributions $\mathcal{D}$ over $\mathcal{X}$and for all labeling function $f : \mathcal{X} \to \{0,1\}$ running the learning algorithm on $m \geq m_{\mathcal{H}}$ i.i.d samples generated by $\mathcal{D}$ the algorithm returns h such that $Pr(L_{\mathcal{D},f}(h) \leq \epsilon) \geq 1 - \delta$.
Let's fix $\mathcal{D}$, and divide into cases:

i. The labeling function $f(x) = h^-(x)$, therefore the training set will not contain a label of 1 and so for any size of a training set my algorithm will return h^- and it's generalization error will be $L_{D,f}(h) = \underset{x\sim\mathcal{D}}{Pr}[h(x) \neq f(x)]$=0
ii. The labeling function $f(x) = h_z(x)$
A. The training set contains a label of 1 and so my algorithm will return h_z and it's generalization error will be $L_{D,f}(h) = \underset{x\sim\mathcal{D}}{Pr}[h(x) \neq f(x)]$=0
B. The training set doesn't contain a label of 1 and so my algorithm will return h^-.
Let $\epsilon, \delta \in (0,1)$, Let's denote $Pr(x = z) = \epsilon'$.
The generalization error in such a case will be

$$L_{D,f}(h) = \underset{x\sim\mathcal{D}}{Pr}[h(x) \neq f(x)] = Pr[x = z] = \epsilon'$$

And so, in this case, the generalization error is not under our control, however, we can avoid this case.
The probability that our training set doesn't contain $(z,1)$ is $Pr[(z,1) \notin S] = (1-\epsilon')^m$. So we can increase our confidence that we will not encounter this case by increasing m.
Any m that satisfies the inequality

$$(1-\epsilon')^m \leq \delta$$

Will give us

$$Pr\left(L_{\mathcal{D},f}(h)=0\leq\epsilon\right)\geq 1-\delta$$

And so, a function $m_{\mathcal{H}}:(0,1)^2\rightarrow\mathbb{N}$ exists and this proves that $\mathcal{H}$ is PAC learnable. To give an upper bound on $m_{\mathcal{H}}$:

$$(1-\epsilon')^m\leq\delta\Rightarrow log\left((1-\epsilon')^m\right)\leq log(\delta)\Rightarrow m{\cdot}log\left(1-\epsilon'\right)\leq log\left(\delta\right)$$

$$m\leq\frac{log\left(\delta\right)}{log\left(1-\epsilon'\right)}$$

7. $L_{\mathcal{D}}\left(h\right)=\underset{(x,y)\sim\mathcal{D}}{Pr}\left[h(x)\neq y\right]=\begin{cases}\underset{(x,y)\sim\mathcal{D}}{Pr}\left[y\neq 0|x\right] & \text{if h(x)=0,}\\ \underset{(x,y)\sim\mathcal{D}}{Pr}\left[y\neq 1|x\right] & \text{if h(x)=1}\end{cases}$

We wish to minimize the function $\phi(x)$ defined below:

$$\phi(x)=\begin{cases}Pr\left[y\neq 0|x\right] & \text{if h(x)=0,}\\ Pr\left[y\neq 1|x\right] & \text{if h(x)=1}\end{cases}=\begin{cases}Pr\left[y=1|x\right] & \text{if h(x)=0,}\\ 1-Pr\left[y=1|x\right] & \text{if h(x)=1}\end{cases}$$

So if $Pr\left[y=1|x\right]<1-Pr\left[y=1|x\right]$ we should choose h(x)=0 otherwise we should choose h(x)=1.
$Pr\left[y=1|x\right]<1-Pr\left[y=1|x\right]\Leftrightarrow Pr\left[y=1|x\right]<\frac{1}{2}$
Therefore, the optimal classifier is given by:

$$f_{\mathcal{D}}(x)=\begin{cases}1 & Pr\left[y=1|x\right]\geq\frac{1}{2}\\ 0 & otherwise\end{cases}$$

8. Let $\mathcal{H}=\{h_1,...,h_N\}$ be a finite hypothesis class over domain $\mathcal{X}$, denote $VC\left(\mathcal{H}\right)=d$. There exists $C\subset\mathcal{X},|C|=d$ such that $\mathcal{H}$ shatters C meaning that $|\mathcal{H}_{\mathcal{C}}|=2^{|C|}\Rightarrow|\mathcal{H}_{\mathcal{C}}|=2^d$ therefore

$$2^d\leq|\mathcal{H}|\Rightarrow d\leq log\left(|H|\right)\Rightarrow\lfloor d\rfloor\leq\lfloor log\left(|H|\right)\rfloor\Rightarrow d\leq\lfloor log\left(|H|\right)\rfloor$$

9. (a) Given a subset $C\subseteq\mathcal{X}$ the domain set, in order to show that $\mathcal{H}$ shatters C I must show that for every possible labeling of the set C there exists an hypothesis $h\in\mathcal{H}$ that explains it.
For $|C|=1,C=\{x\in\mathcal{X}\}$ $\mathcal{H}$ shatters C beacuse $\mathcal{H}_{\mathcal{C}}=\{h_x(x)=1,h^-(x)=0\}\Rightarrow$ $|\mathcal{H}_{\mathcal{C}}|=2$
For every $C\subseteq\mathcal{X}$ such that $|C|=2,C=\{x_1,x_2\in\mathcal{X}\}$ there is no hypothesis in $\mathcal{H}_{Singleton}$ that can explain the labeling $(x_1,1),(x_2,1)$ and therefore $|\mathcal{H}_{\mathcal{C}}|$ =3 $<2^{|C|}$.

(b) Let's complete the proof by showing that $d \leq VC(\mathcal{H})$, in other words, we wish to give a set $C \subseteq \mathcal{X}, |C| = d$ such that $\mathcal{H}$ shatters C.

Let's consider the set $C = \{e_1, ..., e_d\}$ where $e_i = \begin{bmatrix} e_{i1} = 0 \\ \vdots \\ e_{ii} = 1 \\ \vdots \\ e_{id} = 0 \end{bmatrix}$ is the unit vector. Let $l_1, ..., l_d \in \{0, 1\}$ be some labeling of $e_1, ..., e_d$ respectively.

The hypothesis $r_1 \hat{} r_2 \hat{} ... \hat{} r_d \in \mathcal{H}$ where $r_i = \begin{cases} e_i & l_i = 1 \\ \bar{e}_i & l_i = 0 \end{cases}$ explains this labeling for each member $e_j \in C$ and therefore $\mathcal{H}$ shatters C.
In recitation we have seen that $d \geq VC(\mathcal{H})$ and we can conclude that $VC(\mathcal{H}) = d$

10. (a) The hypothesis class defind in question 11 has a VCdim that is equal to the upper bound as I have proved in question 11.

(b) Consider the domain $\mathcal{X} = \{1, ..., k\}$ and consider the hypothesis class of threshold functions

$$\mathcal{H}_{th} = \{h_\theta(x) = sign(x - \theta) : \theta \in \mathbb{R}\}$$

$|H| = k$ but VC($\mathcal{H}$) = 1 and since k can be arbitrarily large, the gap between $log_2(|\mathcal{H}|)$ and VC($\mathcal{H}$) can be arbitrarily large.

11. First let's note that $|H_{parity}| = 2^n$ and therefore, by q8 we know that

$$VC(H_{parity}) \leq \lfloor log(|H_{parity}|) \rfloor = n$$

Let's show that $VC(H_{parity}) \geq n$ and conclude that $VC(H_{parity}) = n$.

Consider the set $C \subseteq \{0, 1\}^n$,$|C| = n, C = \{e_1, ..., e_n\}$ where $e_i = \begin{bmatrix} e_{i1} = 0 \\ \vdots \\ e_{ii} = 1 \\ \vdots \\ e_{id} = 0 \end{bmatrix}$ is the unit vector. Let $l_1, ..., l_d \in \{0, 1\}$ be some labeling of $e_1, ..., e_d$ respectively.
Consider the hypothesis h_I where $I = \{j : j \in [n], l_j = 1\}$ In words, the hypothesis will sum over all j's where the labeling of e_j is 1.
Now $h_I(e_i) = \left(\sum_{j \in I} e_{ij}\right) mod2 = \begin{cases} 1mod2 = 1 & l_i = 1 \\ 0mod2 = 0 & l_i = 0 \end{cases}$. So the hypothesis h_I labels correctly all elements in C and so $\mathcal{H}_{parity}$ shatters C.

12. $\mathcal{X} = \mathbb{R}$.Let's prove that $VC\left(\mathcal{H}_{k-intervals}\right) = 2k$
First let's show that $C \subseteq \mathcal{X}, |C| = 2k, C = \{1, ..., 2k\}$ is shattered by $\mathcal{H}_{k-intervals}$. Let $l_1, ..., l_{2^k} \in \{0,1\}$ be some labeling of $1, ..., 2k$ respectively.
Let $\epsilon << 1$, Define the hypothesis $h_A\left(x\right)$,where $A = \cup_{i=1}^{k}\left[a_i, b_i\right]$ and $a_i = \begin{cases} 2i-1-\epsilon & l_{2i-1} = 1 \\ 2i-1+\epsilon & l_{2i-1} = 0 \end{cases}$ and $b_i = \begin{cases} 2i-1+\epsilon & l_{2i-1} = 1, l_{2i} = 0 \\ 2i+\epsilon & l_{2i-1} = 1, l_{2i} = 1 \\ 2i+\epsilon & l_{2i-1} = 0, l_{2i} = 1 \\ 2i-\epsilon & l_{2i-1} = 0, l_{2i} = 0 \end{cases}$ clearly $\forall j \in [2k]\,, h_A\left(j\right) = \begin{cases} 1 & l_j = 1 \\ 0 & l_j = 0 \end{cases}$
In words, we treat each pair of adjacent points in C seperately and explain each pair with it's own interval.
Now let's prove that for all $C \subseteq \mathcal{X}, |C| = 2k+1, C = \{c_1, ..., c_{2k+1}\}$, C is not shattered by $\mathcal{H}_{k-intervals}$.
Let's consider the following labeling for C: $l_1 = 1, l_2 = 0, l_3 = 1, ..., l_{2k} = 0, l_{2k+1} = 1 \in \{0,1\}$, i.e we take the alternating labeling of the elements of C starting with a positive labeling. Since 2k+1 is an odd number we know that we have k+1 elements labeled as 1 and they are all seperated by elements labeled as 0 therefore there is no hypothesis in $\mathcal{H}_{k-intervals}$ that can explain this labeling, and therefore C is not shattered by $\mathcal{H}_{k-intervals}$.

If k in unlimited then for any $C \subseteq \mathcal{X}, |C| = p$ and for any labeling of C we can take k=p. The hypothesis where there's an interval for each $c \in C$ that is labeled as 1 and that interval contains only c and no other c' in C clearly explains this labeling of C. Therefore in this case $VC(\mathcal{H}_{intervals}) = \infty$

13. Consider the class of homogenous halfspaces in $\mathbb{R}^d$:$\mathcal{H} = \left\{h_w : h_w(x) = sgn\left(\langle w, x\rangle\right), w \in \mathbb{R}^d\right\}$, as we have seen in recitation 5, VCdim($\mathcal{H}$) = d
$\mathcal{H}S_d = \left\{h_{w,b} : h_{w,b}(x) = sgn\left(\langle w, x\rangle + b\right), w \in \mathbb{R}^d, b \in \mathbb{R}\right\}$, let's prove that $VCdim\left(\mathcal{H}S_d\right) = d+1$
Let's first show that there exists $C \subseteq \mathcal{X}, |C| = d+1$,which $\mathcal{H}S_d$ shatters.
Let's consider $C = \{e_1, ...e_d, 0\}$ where e_i is the i'th unit vector. Given some labeling of C $l_1, ..., l_{d+1} \in \{-1, 1\}$ take the hypothesis $w = \begin{bmatrix} l_1 \\ \vdots \\ l_d \end{bmatrix}, b = \frac{1}{2}l_{d+1}$

$$\forall i \in [d]\,, h_{w,b}(e_i) = sgn\left(\langle w, x\rangle + b\right) = sgn\left(l_i + \frac{1}{2}l_{d+1}\right) = sgn\left(l_i\right)$$

$$h_{w,b}(0) = sgn\left(\langle w, 0\rangle + b\right) = sgn\left(b\right) = sgn\left(\frac{1}{2}l_{d+1}\right) = sgn\left(l_{d+1}\right)$$

Now it's left to prove that for all $C \subseteq \mathcal{X}, |C| = d+2$ $\mathcal{H}S_d$ doesn't shatter C, let's assume by contradiction that there exists $C \subseteq \mathcal{X}, |C| = d+$

2, $C = \{c_1, ..., c_{d+2}\}$. that $\mathcal{H}S_d$ shatters, i.e for every possible label of C $l_1, ..., l_{d+2} \in \{-1, 1\}$ there exists a $w = (w_1, ..., w_d) \in \mathbb{R}^d, b \in \mathbb{R}$such that $h_{w,b}$ predicts correctly the labeling.

$$h_{w,b}(c_i) = \langle c_i, w \rangle + b = l_i$$

The set $C' = \{(1, c_1), ..., (1, c_{d+2})\}$ is shattered by the class of homogenous halfspaces in $\mathbb{R}^{d+1}$ because given a labeling $l_1, ..., l_{d+2}$ the hypothesis $h_w(x) = \langle (b, w), x \rangle$ predicts correctly on all elements of C'

$$h_w((1, c_i)) = \langle (b, w), (1, c_i) \rangle = b + \langle w, c_i \rangle = l_i$$

In contradiction to the VCdim of the class of homogenous halfspaces in $\mathbb{R}^{d+1}$ being d+1.