Solution to Exercises on PAC-Learning and VC-Dimension
CS-526 Learning Theory

Short problems

1. **B and C**.The set Θ parametrizing the hypothesis class must be infinite: if $\mathcal{H}$ has finite cardinality then $\mathrm{VCdim}(\mathcal{H}) \leq \log |\mathcal{H}|$. In the second graded homework, we studied the hypothesis class $\mathcal{H} = \{\lceil \sin(\theta\pi.) \rceil\}_{\theta\in\Theta}$ and proved that it has an infinite VC dimension if $\Theta = \{2n\}_{n\in\mathbb{N}}$ (and by extension $\theta = \mathbb{R}$). Therefore B and C are correct.

2. (a) False. If $\mathcal{H}$ has finite VC dimension then it is PAC learnable due to the Fundamental theorem of Statistical learning.
   (b) True. According to the Fundamental theorem of Statistical learning.
   (c) False. We saw in the homework that there are hypotheses classes with infinite VC dimension that are specified by a single parameter.
   (d) True. If $\mathcal{H}_1, \mathcal{H}_2$ have finite VC dimension then the VC dimension of their union is also finite and therefore $\mathcal{H}$ is also PAC learnable.

3. The VC dimension is 2: A set of size 2 can be shattered by $\mathcal{H}$, but for a set of size 3 with elements $x_1 < x_2 < x_3$ the labeling $(0, 1, 0)$ cannot be obtained by any $h_{a,b} \in \mathcal{H}$. Therefore, the VC dimension is 2.

4. The VC dimension is 4: A set of size 4 can be shattered, but a set of size 5 with elements $x_1 < \ldots < x_5$ with labels $(1, 0, 1, 0, 1)$ cannot be obtained by any $h_{a,b,c,d} \in \mathcal{H}$. Therefore, the VC dimension is 4.

VC dimension of unbiased neurons

Note that tanh does not change the sign of $\alpha_1 x_1 + \alpha_2 x_2$, so we don't need to bother with the tanh in analysis.

$\underline{\mathrm{VCdim}(\mathcal{H}) \geq 2}$: given any two samples $(\mathbf{x}^{(1)}, y^{(1)})$ and $(\mathbf{x}^{(2)}, y^{(2)})$ with $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ linearly independent, we can find valid α_1, α_2 by solving

$$\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b^{(1)} \\ b^{(2)} \end{bmatrix}$$

where $b^{(i)}$ is any real numbers that has the same sign with $(-1)^{1+y^{(i)}}$.

$\underline{\mathrm{VCdim}(\mathcal{H}) \leq 2}$: For any three points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ one can propose $y^{(1)}, y^{(2)}, y^{(3)}$ such that $\mathcal{H}$ does not shatter the 3 points. This amounts to showing that there exists $y^{(1)}, y^{(2)}, y^{(3)}$ such that

$$\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \mathbf{x}^{(3)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b^{(1)} \\ b^{(2)} \\ b^{(3)} \end{bmatrix} \tag{1}$$

has no solutions, with $b^{(i)}$ as defined above. In $\mathbb{R}^2$ any three points are linearly dependent. So (1) is degenerated. We can assume $\mathbf{x}^{(3)} = w_1\mathbf{x}^{(1)} + w_2\mathbf{x}^{(2)}$ for some $w_1, w_2 \in \mathbb{R}$. Suppose $y^{(1)}, y^{(2)}$ allows a solution of α_1, α_2 for the first two equations of (1). However, if one chooses $y^{(3)}$ such that $\sum_{i=1}^{2}\sum_{j=1}^{2} w_i\alpha_j x_j^{(i)}$ has a different sign from $(-1)^{1+y^{(3)}}$, then (1) has no solution.

VC dimension of union

1. Let $\mathcal{H} = \bigcup_{i=1}^{r} \mathcal{H}_i$. By definition of the growth function we have $\tau_{\mathcal{H}}(m) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(m)$ for any set of m points. If $k > d+1$ points are shattered by $\mathcal{H}$ then $2^k = \tau_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(k) \leq rk^d$, where the last inequality follows directly from Sauer's lemma. Taking the logarithm on both sides and using the inequality yields

$$k \leq \frac{4d}{\log(2)} \log\left(\frac{2d}{\log(2)}\right) + 2\frac{\log(r)}{\log(2)} \, .$$

   Note that this inequality is trivially satisfied if $k \leq d+1$.

2. Assume that $k \geq 2d+2$. It is enough to prove that $\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) < 2^k$.

$$\begin{aligned}
&\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) \leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \leq \sum_{i=0}^{d}\binom{k}{i} + \sum_{i=0}^{d}\binom{k}{i} = \\
&= \sum_{i=0}^{d}\binom{k}{i} + \sum_{i=0}^{d}\binom{k}{k-i} = \sum_{i=0}^{d}\binom{k}{i} + \sum_{i=k-d}^{k}\binom{k}{i} \leq \\
&\leq \sum_{i=0}^{d}\binom{k}{i} + \sum_{i=d+2}^{k}\binom{k}{i} < \sum_{i=0}^{d}\binom{k}{i} + \sum_{i=d+1}^{k}\binom{k}{i} = \\
&= \sum_{i=0}^{k}\binom{k}{i} = 2^k
\end{aligned}$$

Lemma *(Sauer-Shelah-Perles)* Let $\mathcal{H}$ be a hypothesis class with $VCdim(H) \leq d < \infty$ and growth function $\tau_{\mathcal{H}}$. Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d}\binom{m}{i}$. In particular, if $m > d+1$ and $d > 2$ then $\tau_{\mathcal{H}}(m) < m^d$.

Stability implies Generalization

1. Note that since $\tilde{S}$ is composed of n i.i.d. samples $L_{\mathcal{D}}(h_S) = E_{(\tilde{x_i},\tilde{y_i})\sim\mathcal{D}}[l(h_S(\tilde{x_i}), \tilde{y_i})]$ for all i. Thus, by linearity of expectation $L_{\mathcal{D}}(h_S) = E_{\tilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})]$.

2.

$E_{S,\tilde{S}}[l(h_S(\tilde{x_i}), \tilde{y_i})] = E_{S,(\tilde{x_i},\tilde{y_i})}[l(h_S(\tilde{x_i}), \tilde{y_i})] =$
(since $(x_1, y_1), \ldots, (x_n, y_n), (\tilde{x_i}, \tilde{y_i})$ are i.i.d. we can interchange (x_i, y_i) with $(\tilde{x_i}, \tilde{y_i})$)
$= E_{S^{(i)},(x_i,y_i)}[l(h_{S^{(i)}}(x_i), y_i)]$

3.

$$
\begin{aligned}
&|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| \overset{(1)}{=} |E_S\left[L_S(h_S) - E_{\tilde{S}}\left[\tfrac{1}{n}\textstyle\sum_{i=1}^n l(h_S(\tilde{x_i}), \tilde{y_i})\right]\right]| = \\
&= |E_S\left[L_S(h_S)\right] - E_{S,\tilde{S}}\left[\tfrac{1}{n}\textstyle\sum_{i=1}^n l(h_S(\tilde{x_i}), \tilde{y_i})\right]| = \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^n E_{S,\tilde{S}}\left[l(h_S(\tilde{x_i}), \tilde{y_i})\right]| \overset{(2)}{=} \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^n E_{S^{(i)},(x_i,y_i)}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |E_S\left[\tfrac{1}{n}\textstyle\sum_{i=1}^n l(h_S(x_i), y_i))\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^n E_{S,S^{(i)}}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |\tfrac{1}{n}\textstyle\sum_{i=1}^n E_{S,S^{(i)}}\left[l(h_S(x_i), y_i)) - l(h_{S^{(i)}}(x_i), y_i)\right]| \overset{(\text{ }\epsilon\text{-uniform stability})}{\leq} \\
&\leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon = \epsilon
\end{aligned}
$$

VC dimension of decision trees with binary features

1. For each feature i, there exist two trivial decision trees (that both return zero or both return one) and two non-trivial ones (the one that returns 0 if $x_i = 1$ and 1 otherwise and the one that returns 1 if $x_i = 1$ and 0 otherwise). Therefore, with d features we can have at most $2d + 2$ distinct labelings. In order to shatter m samples, we need to obtain all 2^m possible labelings, hence we have the bound

$$2d + 2 \geq 2^m.$$

Resolving for m we get the stated upper bound.

2. To prove the lower bound, we need to construct the set of $m = \lfloor \log_2(d+1) \rfloor + 1$ samples that is shattered. To do this, take the set of all possible labelings except all-zero and all-one and for each labeling $(y_1, \dots, y_m)$ remove its complement from the set. This leaves $2^{m-1} - 1$ distinct labelings $y^{(i)}$. Now we create the samples $x^{(1)}, \dots, x^{(m)}$ s.t. $x_i^{(j)} = y_j^{(i)}$ for $1 \leq j \leq m, 1 \leq i \leq 2^{m-1} - 1 = d$. It remains to notice that a tree with node $x_i = 0$? gives either the labeling $y^{(i)}$ or its complement (if we reverse the labels on branches) and in addition all-one and all-zero labelings if both branches return the same label, which completes the proof.

3. We need to construct the set of $m = \lfloor \log_2(d - N + 2) \rfloor + N$ samples on which we get all 2^m possible labels. We start from the case of one bottom node, with $d = 2^{m-1} - 1$ features for m samples. Now assume we get an extra feature x_{d+1} and an extra sample s.t. $x_{d+1}^{(m+1)} = 1$ and $x_i^{(m+1)} = 0$ for $i \neq d+1$ ($x_{d+1}^{(i)} = 0$ for $i < m+1$). We create a parent node that contains the existing node and our new sample as children and the splitting rule is the new feature. The new splitting rule allows to label $x^{(m+1)}$ independently of other $x^{(i)}$, so we get all possible labelings on $m + 1$ samples. This procedure can be performed $N - 1$ times since we have N decision nodes in the tree. Therefore, for m samples we have $d = 2^{m-1-(N-1)} - 1 + (N - 1) = 2^{m-N} + N - 2$ features that generate all 2^m possible labelings.

Expectation Learnability

1. Set $\gamma = \epsilon\delta$. By the E learnability, the algorithm running on $m \geq m_{\mathcal{H}}^{(\mathrm{E})}(\epsilon\delta)$ samples returns a hypothesis h so that $\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] \leq \epsilon\delta$. Using the Markov inequality, we have:
$$\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \geq \epsilon\big] \leq \frac{\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big]}{\epsilon} \leq \frac{\epsilon\delta}{\epsilon} = \delta.$$
Moreover, the number of samples needed to generate h is bounded by a function in $\epsilon\delta$, which is a function in ϵ, δ. Therefore, the requirements of the PAC learnability are satisfied.

2. Set $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$, then by PAC learnability, we have an algorithm that running on $m \geq m_{\mathcal{H}}^{(\mathrm{PAC})}\big(\frac{\gamma}{2}, \frac{\gamma}{2}\big)$ samples returns a hypothesis h so that $\mathbb{P}\big[L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big] \leq \frac{\gamma}{2}$. We have
$$\begin{aligned}\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] &= \mathbb{E}\big[L_{(\mathcal{D},f)}(h)|L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big]\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big] \\ &\qquad + \mathbb{E}\big[L_{(\mathcal{D},f)}(h)|L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big]\mathbb{P}\big[L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big] \\ &\leq \frac{\gamma}{2}\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big] + \mathbb{E}\big[L_{(\mathcal{D},f)}(h)|L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big]\frac{\gamma}{2} \\ &\leq \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma\end{aligned}$$
where the last inequality is due to the boundedness of $L_{(\mathcal{D},f)}(h)$, since probability is bounded by 1.

   Moreover, the number of samples needed to generate h is bounded by a function in $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$ which is a function in γ. Therefore, the requirements of the E learnability are satisfied.

3. From the course, we know that every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}^{(\mathrm{PAC})}(\epsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon} \right\rceil$. Setting $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$, we get the result.