Jacob Snyder

CS156a Set 6

1. B
   As stated in lecture (Lecture 11 - Overfitting), using a more sophisticated hypothesis set will cause the deterministic noise to decrease in general. This means that using $\mathcal{H}'$ instead of $\mathcal{H}$ will cause a general increase in deterministic noise because $\mathcal{H}'$ is contained within $\mathcal{H}$ (so therefore $\mathcal{H}$ is a more sophisticated hypothesis set). Therefore, because deterministic noice will in general increase, the answer is B.

2. A
   Justification in Jupyter Notebook

3. D
   Justification in Jupyter Notebook

4. E
   Justification in Jupyter Notebook

5. D
   Justification in Jupyter Notebook

6. B
   Justification in Jupyter Notebook

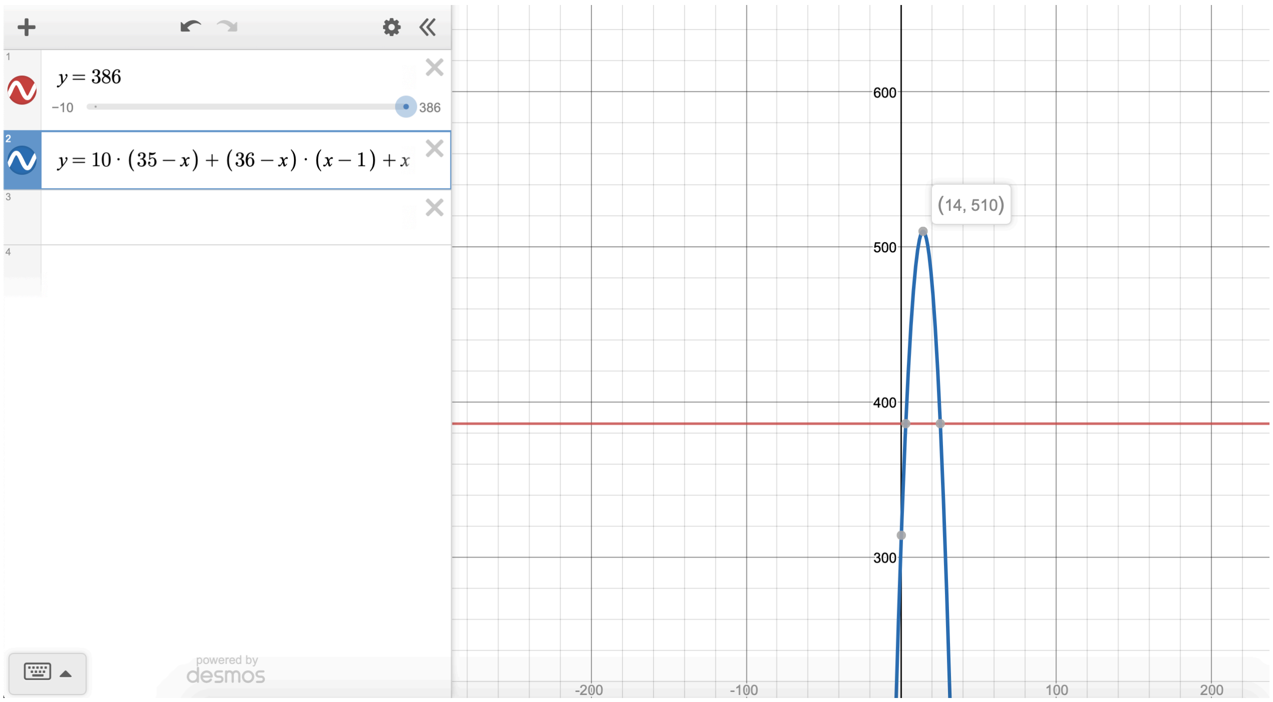
7. C
   First, notice that $\mathcal{H}_2 = \mathcal{H}(10,0,3)$, because all weights that are greater or equal to 3 will be equal to zero; this means only terms with $q \leq 2$ contribute to the hypothesis set.
   Furthermore, notice that $\mathcal{H}(10,0,3) \cap \mathcal{H}(10,0,4) = \mathcal{H}(10,0,3)$. This is because $\mathcal{H}(10,0,3) \subset \mathcal{H}(10,0,4)$; the only difference between the sets of hypotheses is that $\mathcal{H}(10,0,4)$ can include hypotheses with nonzero weights for $q = 3$. But because $\mathcal{H}(10,0,3)$ cannot do this, we conclude that $\mathcal{H}(10,0,3) \cap \mathcal{H}(10,0,4) = \mathcal{H}(10,0,3)$. Therefore, $\mathcal{H}(10,0,3) \cap \mathcal{H}(10,0,4) = \mathcal{H}_2$ so the answer is C.

8. D
   According to the formulas from lecture we have:
   $$x_j^{(l)} = \theta\left(\sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}\right)$$
   Which gives us $6 * 3 + 4 * 1 = 22$ evaluations of $w_{ij}^{(l)} x_i^{(l-1)}$.
   $$\delta_i^{(l-1)} = \left(1 - \left(x_i^{(l-1)}\right)^2\right) \sum_{j=1}^{d^{(l)}} w_{ij}^{(l)} \delta_j^{(l)}$$
   Which gives us $3 * 1 = 3$ evaluations of $w_{ij}^{(l)} \delta_j^{(l)}$.
   And because $d^{(0)} = 5$ and $d^{(1)} = 3$ and $d^{(2)} = 1$ we have a total of $6 * 3 + 4 * 1 = 22$ weights to calculate.
   Therefore, in total we have $22 + 3 + 22 = 47$ operations in a single iteration of backpropagation for this network. Therefore the answer is D because 47 is closest to 45.

9. A

We minimize the number of weights by putting 2 hidden units in each layer. This is the minimum because every layer (other than layer 0 and the final layer) must have at least 2 hidden units in it, so that one of the units can take inputs in order to be fully connected. This yields 1 weight per unit which is a minimum for a fully connected network. Then we have 10 weights from layer 0 to layer 1, then 2 weights for each pair of hidden units: $2 * 18 = 36$. So in total we have $10 + 36 = 46$ weights as a minimum.

10. E



Plotted is the number of weights, either by putting all of the hidden units all in the same layer (red line) or by splitting them (not necessarily evenly) between 2 layers (blue line). Splitting them between more than 2 layers will not yield more weights because as we create more and more layers we lose weights (intuitively: $a * a > \left(\frac{a}{2}\right) * \left(\frac{a}{2}\right) + \left(\frac{a}{2}\right) * \left(\frac{a}{2}\right)$). The maximum is at 22 hidden units in layer 1 and 14 units in layer 2, with a total of 510 weights. Therefore the answer is E.