PS – 3

1. (a) $$\frac{\partial l}{\partial w^{(1)}_{(1,2)}} = \frac{\partial l}{\partial o} \cdot \frac{\partial o}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial w^{(1)}_{(1,2)}}$$

$$= \frac{2}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)})\, o^{(i)} (1 - o^{(i)})\, w^{(2)}_{(2,1)}\, h^{(2)} (1 - h^{(2)})\, x_1^{(i)}$$

$$\Rightarrow w^{(1)+}_{(1,2)} := w^{(1)}_{(1,2)} - \alpha \cdot \frac{2}{m} \cdot w^{(2)}_{(2,1)} \sum_{i=1}^{m} (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{(i)})\, h^{(2)} (1 - h^{(2)})\, x_1^{(i)}$$

where $h^{(2)} = \sigma\left(w^{(1)}_{(0,2)} + x_1^{(i)} w^{(1)}_{(1,2)} + x_2^{(i)} w^{(1)}_{(2,2)} \right)$

(b) Yes, understand the three hidden layers as three separate classifiers that are lines which ~~complete~~ generate the bounding triangle. We essentially calibrate slopes / intercepts.

(c) No. The resultant model is no better than a linear classifier.

2. (a) We know that

$$D_{KL}(P \| Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

$$= E\left[\log \frac{P(x)}{Q(x)} \right] = E\left[-\log \frac{P(x)}{Q(x)} \right]$$

$$\geq -\log E\left[\frac{Q(x)}{P(x)} \right] = -\log \sum_{x \in X} P(x) \cdot \frac{Q(x)}{P(x)}$$

$$= -\log \sum_x Q(x) = 0$$

If $P = Q$, $D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = \log 1 \sum_x P(x) = 0$

If $D_{KL}(P \| Q) = 0$, $\frac{Q(x)}{P(x)} = 1 \Rightarrow Q(x) = P(x)$

(b) $$D_{KL}(P(X,Y) \| Q(X,Y)) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x \sum_y P(x) P(y|x) \log \frac{P(x) P(y|x)}{Q(x) Q(y|x)}$$

$$= \sum_x \sum_y P(x)P(y|x)\left(\log \frac{P(x)}{Q(x)} + \log \frac{\cancel{P(x)}\, P(y|x)}{Q(y|x)}\right)$$

$$= \sum_x \sum_y P(x)P(y|x) \log \frac{P(x)}{Q(x)} + \sum_x \sum_y P(x)P(y|x) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} \underbrace{\sum_y P(y|x)}_{1} + \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}$$

$$= D_{KL}(P(X)\,||\,Q(X)) + D_{KL}(P(Y|X)\,||\,Q(Y|X))$$

(c) $$\arg\min_\theta D_{KL}(\hat{P}\,||\,P_\theta) = \arg\min_\theta \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)}$$

$$= \arg\min_\theta \sum_x \hat{P}(x) \log \hat{P}(x) - \cancel{\arg\min} \sum_x \hat{P}(x) \log P_\theta(x)$$

$$= \arg\max_\theta \sum_x \hat{P}(x) \log P_\theta(x)$$

$$= \arg\max_\theta \sum_x \frac{1}{m} \sum_{i=1}^{m} 1\{x^{(i)}=x\} \log P_\theta(x)$$

$$= \arg\max_\theta \sum_{i=1}^{m} \sum_x 1\{x^{(i)}=x\} \log P_\theta(x^{(i)})$$

$$= \arg\max_\theta \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$