HW3

Q1. $\frac{P(y=1|x_1 \dots x_d)}{P(y=0|x_1 \dots x_d)} = \frac{\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i} (1-\theta_{i1})^{1-x_i}}{\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i} (1-\theta_{i0})^{1-x_i}} > 1$

show:

$$\log \frac{\theta_1}{\theta_0} + \sum_{i=1}^{d} \left(x_i \log \frac{\theta_{i1}}{\theta_{i0}} + (1-x_i) \log \frac{1-\theta_{i1}}{1-\theta_{i0}} \right) > 0$$

~~$\log(\theta_1) - \log(\theta_0) + \sum_{i=1}^{d} \left(x_i (\log(\theta_{i1}) - \log(\theta_{i0})) + (1-x_i)(\log(1-\theta_{i1}) - \log(1-\theta_{i0})) \right) > 0$~~

$$\log \frac{\theta_1}{\theta_0} + \sum_{i=1}^{d} \left[x_i \log \frac{\theta_{i1}}{\theta_{i0}} - x_i \log \frac{1-\theta_{i1}}{1-\theta_{i0}} + \log \frac{1-\theta_{i1}}{1-\theta_{i0}} \right) > 0$$

$\log(\theta_1) - \log(\theta_0) + \sum_{i=1}^{d} \left[x_i \log \theta_{i1} - x_i \log \theta_{i0} - x_i \log(1-\theta_{i1}) + x_i \log(1-\theta_{i0}) + \log(1-\theta_{i1}) - \log(1-\theta_{i0}) \right] > 0$

$$\frac{\log(\theta_1)}{\log(\theta_0)} + \sum_{i=1}^{d} x_i \log(\theta_{i1}) + \sum_{i=1}^{d} (1-x_i) \log(1-\theta_{i1}) - \sum_{i=1}^{d} x_i \log(\theta_{i0}) - \sum_{i=1}^{d} (1-x_i) \log(1-\theta_{01}) > 0$$

$$\frac{\log(\theta_1)}{\log(\theta_0)} + \sum_{i=1}^{d} \left[x_i \left(\frac{\log(\theta_{i1})}{\log(\theta_{i0})} \right] + \sum_{i=1}^{d} \left[x_i \left[\log \left(\frac{\theta_{i1}}{\theta_{i0}} \right) - \log \frac{1-\theta_{i1}}{1-\theta_{i0}} \right) \right] > 0$$

$$w = \log \left(\frac{\theta_{i1}}{\theta_{i0}} \right) - \log \left(\frac{1-\theta_{i1}}{1-\theta_{i0}} \right) = \log \left(\frac{\theta_{i1}(1-\theta_{i0})}{\theta_{i0}(1-\theta_{i1})} \right]$$

$$b = \sum_{i=1}^{d} \log \left(\frac{1-\theta_{11}}{1-\theta_{01}} \right) + \log \left(\frac{\theta_1}{\theta_0} \right)$$

Q: $P(y=1 \mid X_1=x_1) > P(y=0 \mid X_1=x_1)$

$$\frac{P(X_1=x_1, X_2=x_2 \mid y=1) \cdot P(y=1)}{P(X_1=x_1)} > \frac{P(X_1=x_1 \mid y=0)\, P(y=0)}{P(X_1=x_1)}$$

$$P(y=1 \mid x_1, x_2) = \frac{P(x_1, x_2 \mid y=1)\, P(y=1)}{P(x_1, x_2 \mid y=1)\, P(y=1) + P(x_1, x_2 \mid y=0)\, P(y=0)}$$

$$= \frac{P(x_1 \mid y=1)\, P(x_2 \mid y=1)\, P(y=1)}{P(x_1 \mid y=1)\, P(x_2 \mid y=1) P(y=1) + P(x_1 \mid y=0) P(x_2 \mid y=0)\, P(y=0)}$$

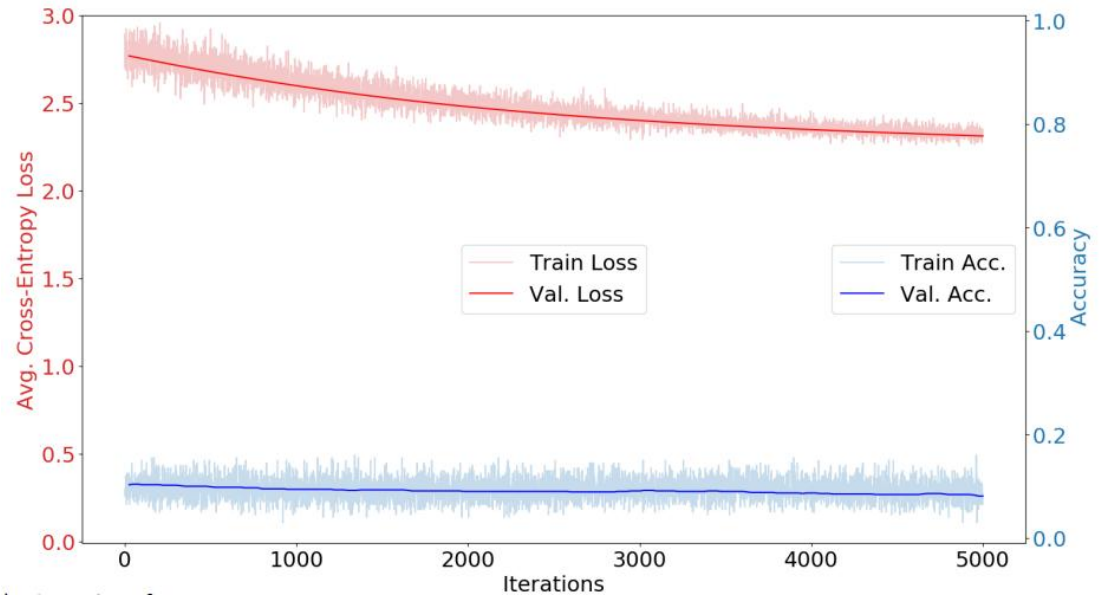
$$> \frac{P(x_2 \mid y=1)\, P(y=1)}{P(x_2 \mid y=1) \cdot P(y=1) + P(x_2 \mid y=0)\, P(y=0)}$$

$$= P(y=1 \mid x_1)$$

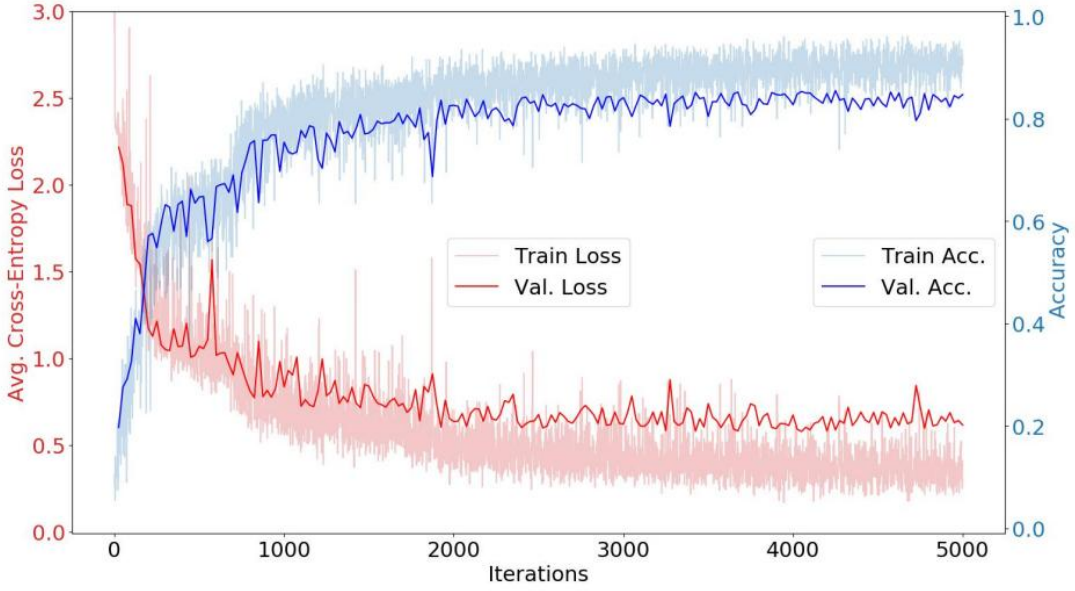
$$\therefore P(y=1 \mid x_1, x_2) > P(y=1 \mid x_1)$$

Q4

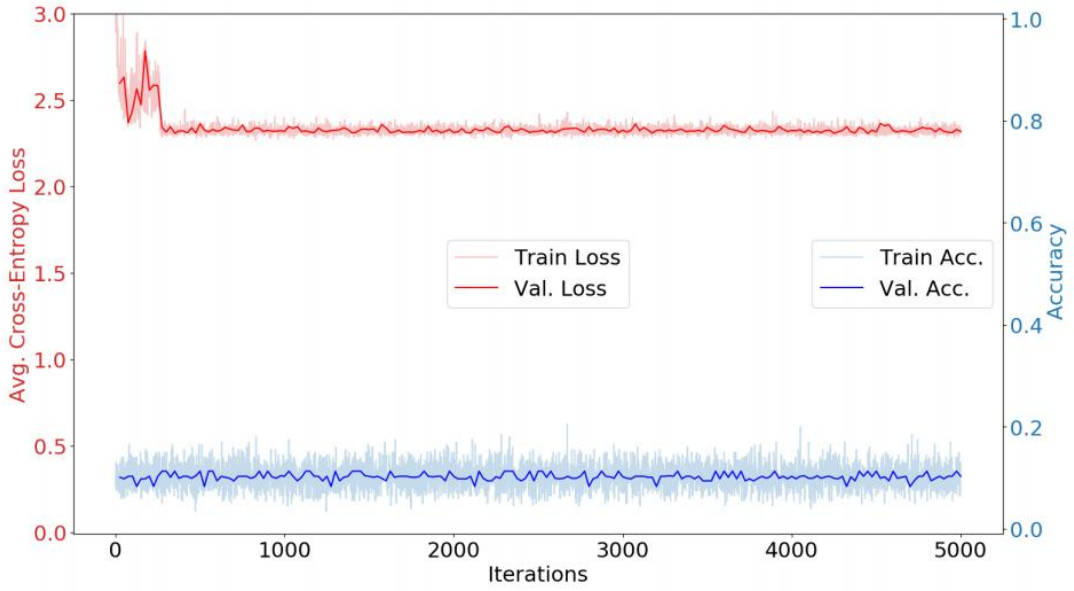
1) Step size of 0.0001



2) Step size of 5



3) Step size of 10

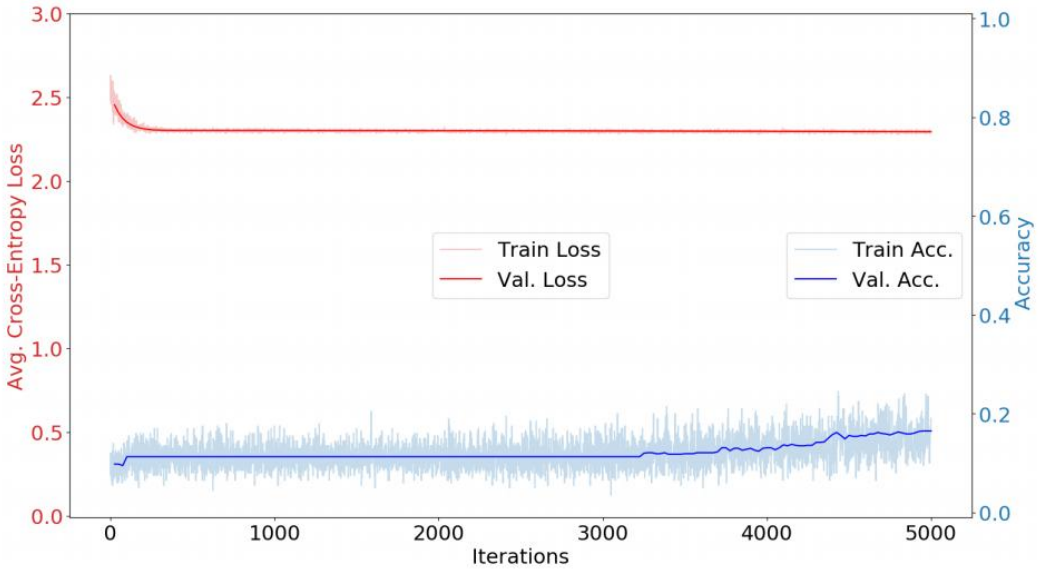


The smoothness gets smoother as the step size gets smaller. The plot for a step size of 5 fared significantly better than the plots for 10 and.0001. At.0001, it appears that the step size is too small to reach the ideal weight, while at 10, it requires too many steps to approach the ideal weight. However, the step size of 5 shows a lot of change in performance at first, but as it converges to the optimal weight vector, the performance flattens. The.0001 and 10 plots don't show much change in shape because they can't converge towards an optimal weight.

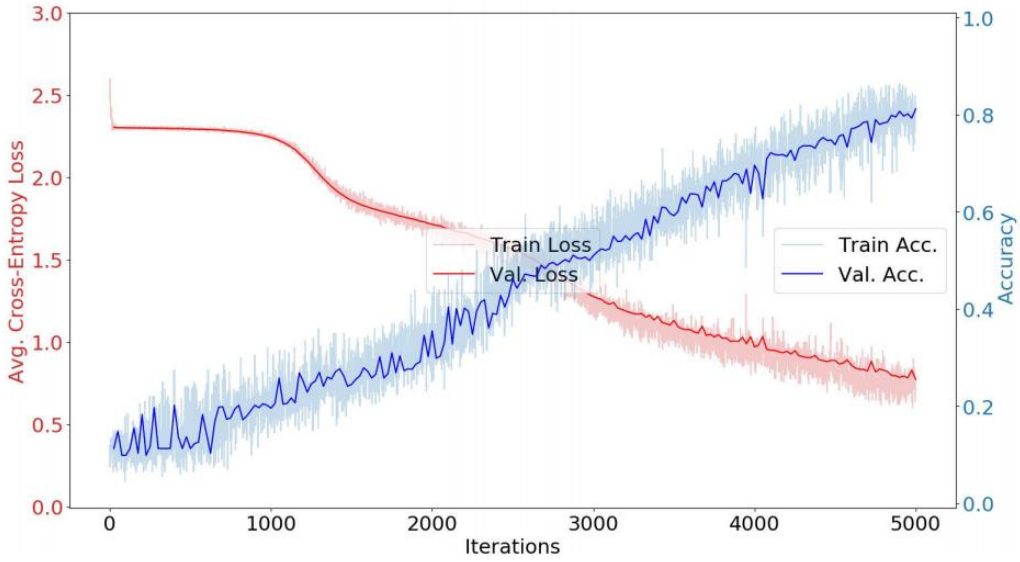
If the max epochs increased the smoothness would probably increase.

Q5

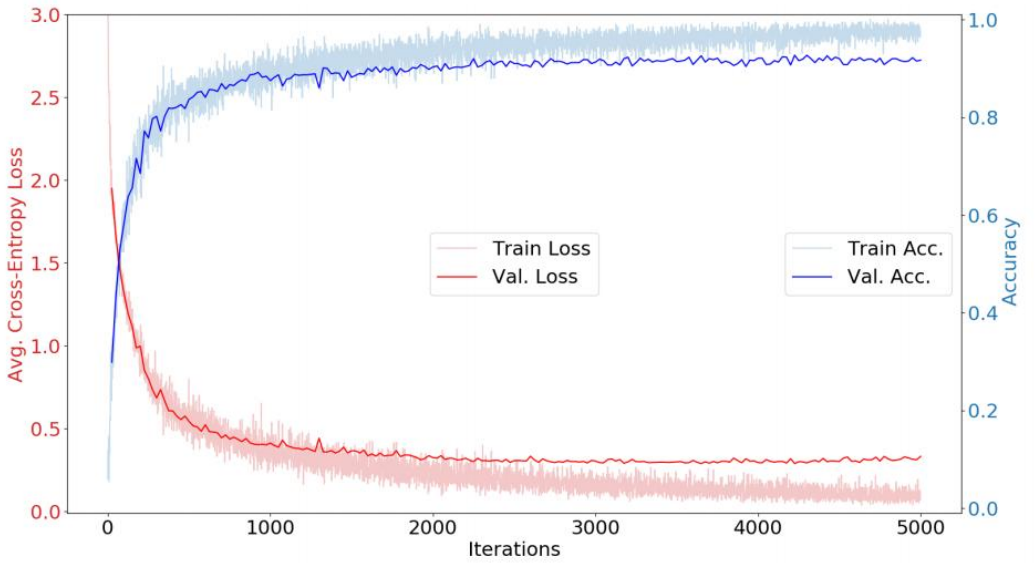
1. 5-layer with Sigmoid Activation



2. 5-layer with Sigmoid Activation with 0.1 step size



3. 5-layer with ReLu Activation



a) Compare and contrast the learning curves you observe and the curve for the default parameters in termsof smoothness, shape, and what performance they reach. Do you notice any differences in the relationship between the train and validation curves in each plot?
For(1), it is least smooth and has a bad performance on iterations.
For(2), more accurate than (1).
For(3), with high accuracy and learns quickly.

b) If you observed increasing the learning rate in (2) improves over (1), why might that be?
The first network will have a slow converge speed and with very low learning rate.

c) If (3) outperformed (1), why might that be? Consider the derivative of the sigmoid and ReLU functions. At x = 0, the ReLU's derivative is a step function, while the Sigmoid function's derivative is a gaussian. The gaussian will have smaller steps than the ReLu because it is more evenly spread around x = 0. On the other hand, the ReLu is equally distributed at x > 0, allowing for the taking of larger steps with smaller step sizes.

Q6 5 times yields the following validation accuracies: 88.7, 87.2, 87.6,88.5, 88.3