HW 3

Everett Brandt
9-18-21

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

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

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

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

w $\qquad$ b

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

$$b = \sum_{i=1}^{d} \log\left(\frac{1-\theta_{i1}}{1-\theta_{01}}\right) + \log\left(\frac{\theta_1}{\theta_0}\right)$$

Q2 $P(y=1 \mid x_1=x_1, x_2=x_2) > P(y=1 \mid x_1=x_1)$

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

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

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

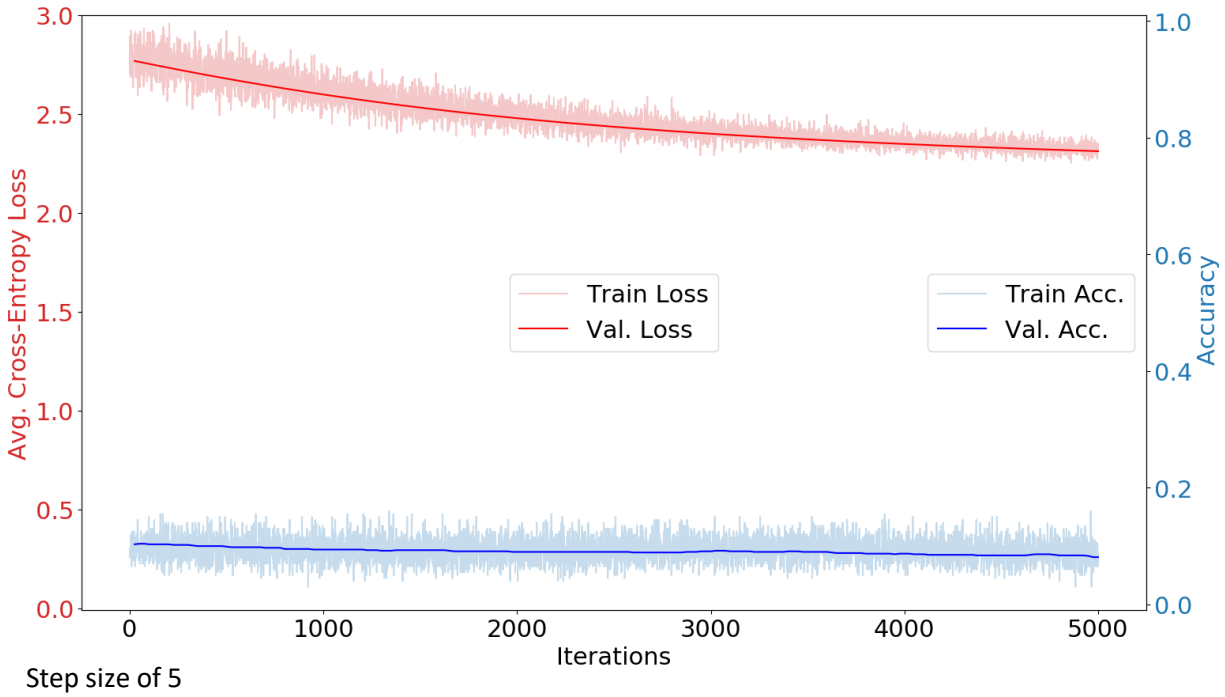
$$P(x_2=x_2 \mid y=1) > P(x_2=x_2)$$

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

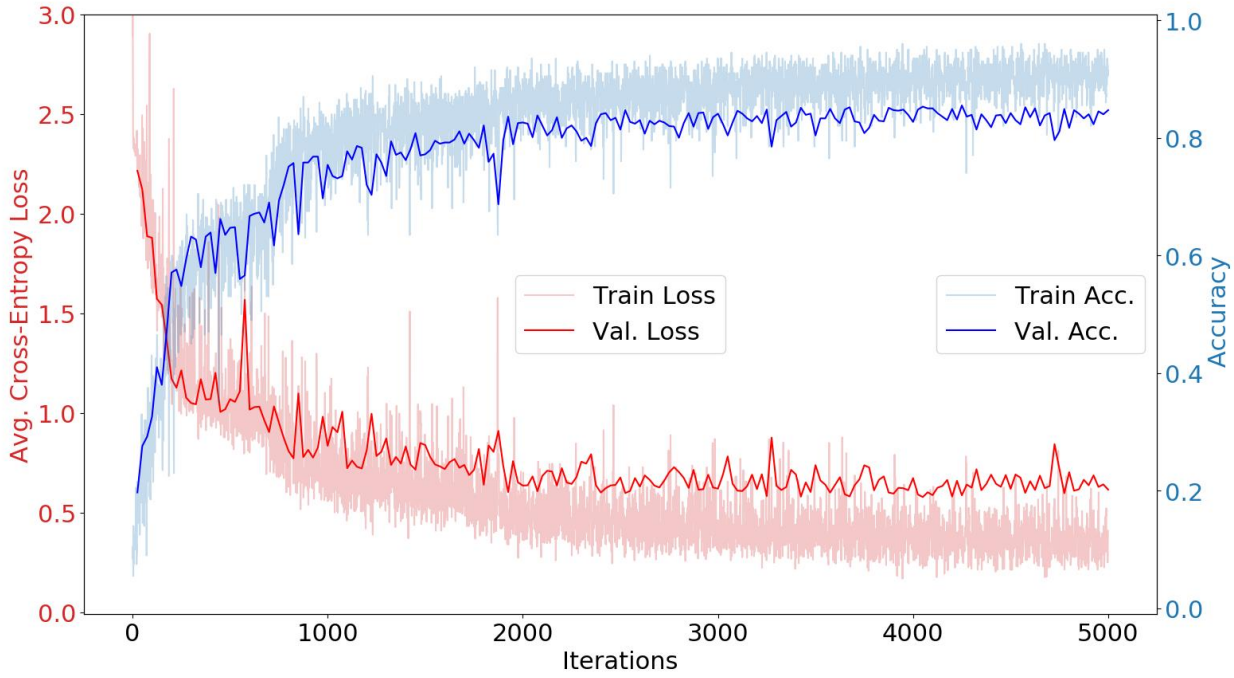
$$\boxed{P(y=1 \mid x_2=x_2) > P(y=1)}$$

Q4

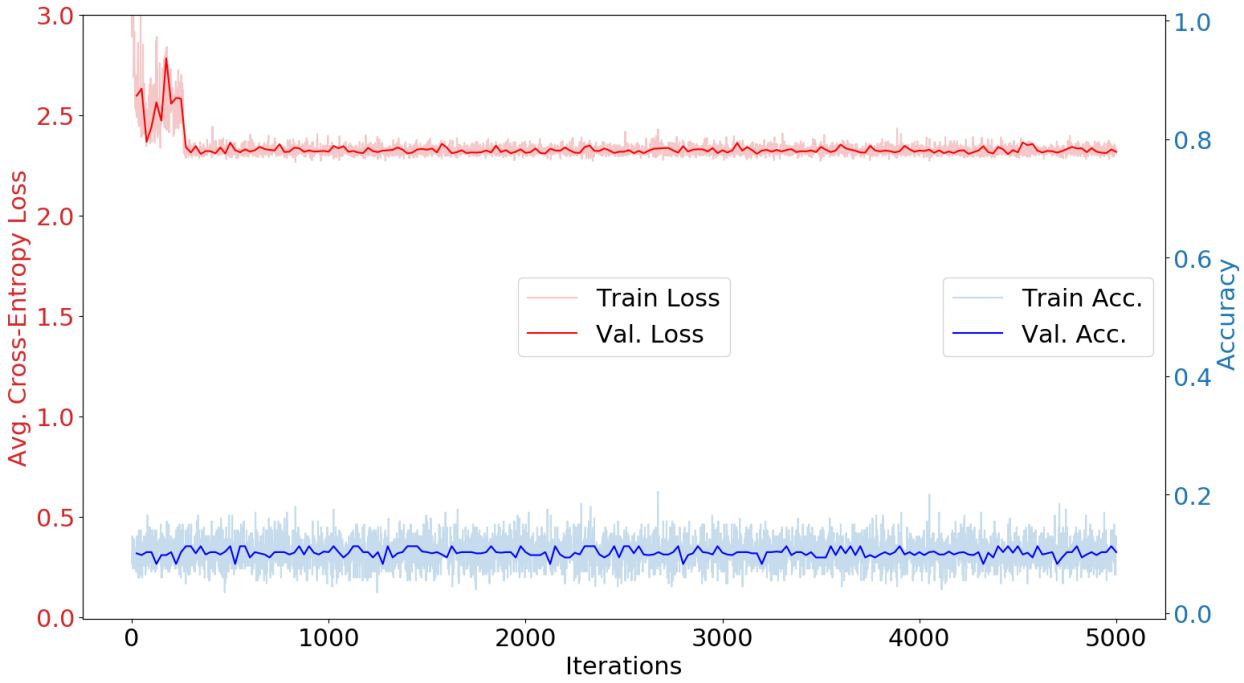
1) Step size of 0.0001



2) Step size of 5



3) Step size of 10

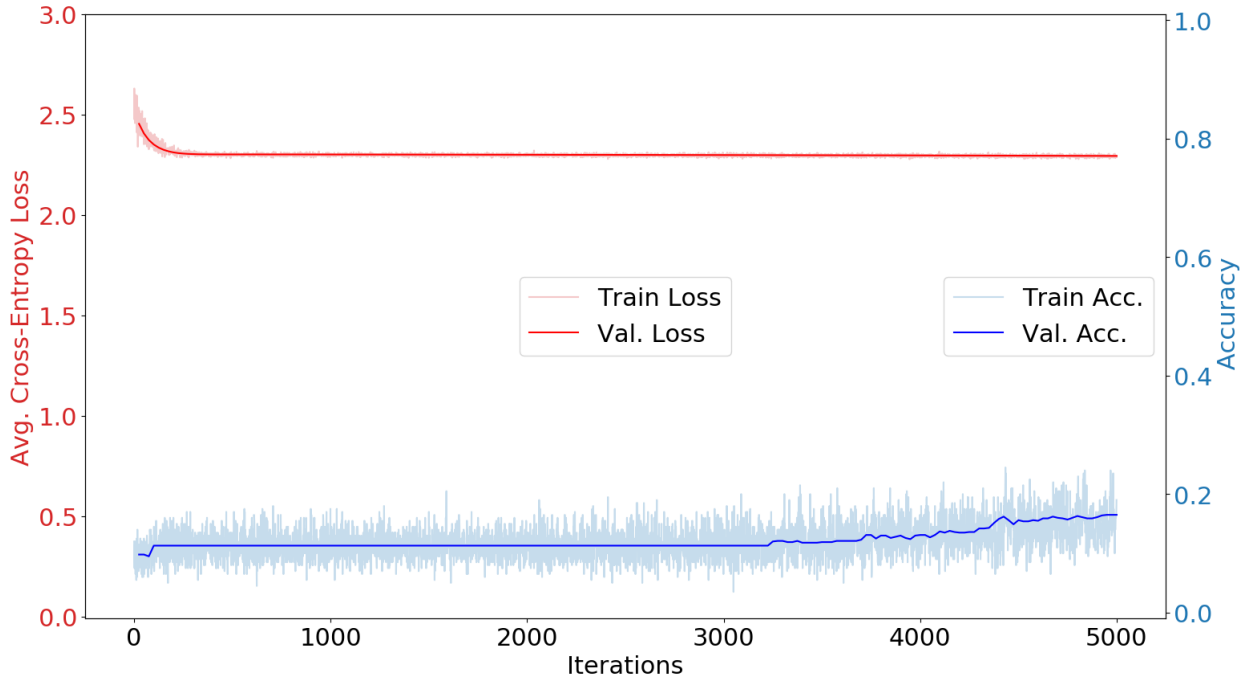


a) As the step size decreases the smoothness increases. The plots for a step size of 10 and .0001 have performed much worse than the plot for 5. It seems the step size is too small at .0001 to reach the optimal weight and with a step size of 10 it's taking too big of steps to get close to the optimal weight. The .0001 and 10 plots don't have much change in shape because they aren't able to converge towards an optimal weight, however the step size of 5 shows a lot of change in performance in the beginning, but as it converges to the optimal weight vector the performance flattens.

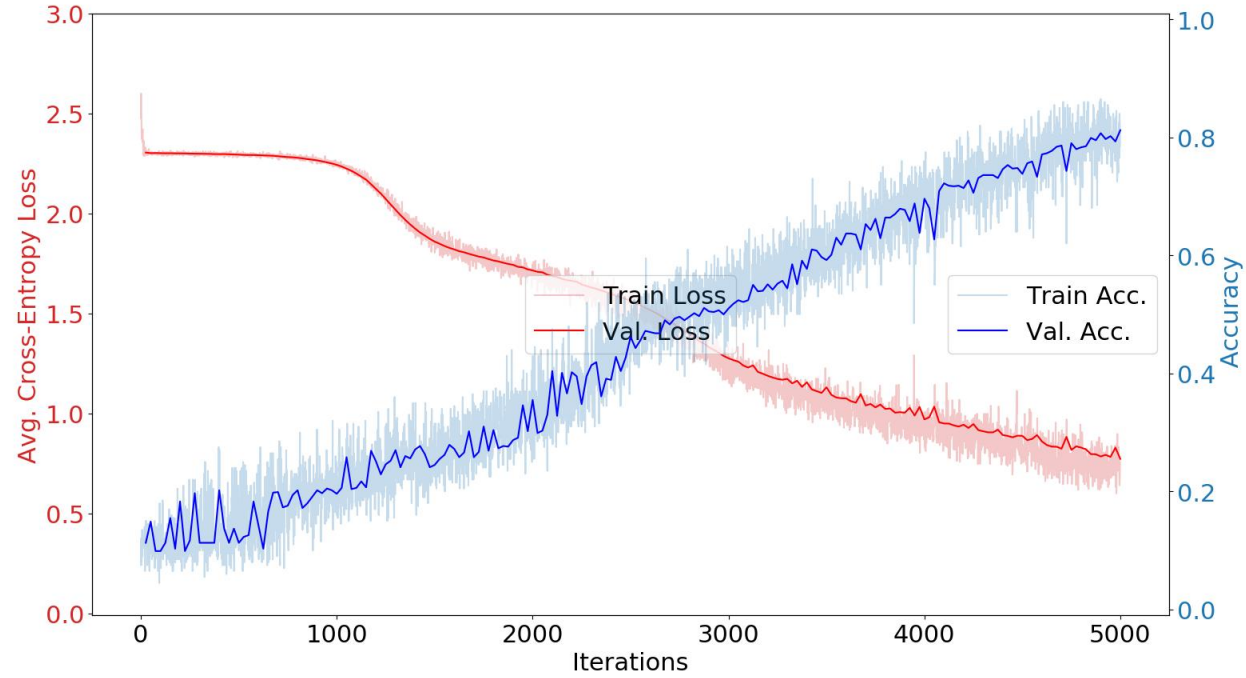
b) If the max epochs increased the smoothness would probably increase. The plot for the .0001 step size would probably gain accuracy since there would be more time to step to the optimal weight vector, however the plots for 5 and 10 wouldn't. The shape would be similar for all of them, just smoother.

Q5

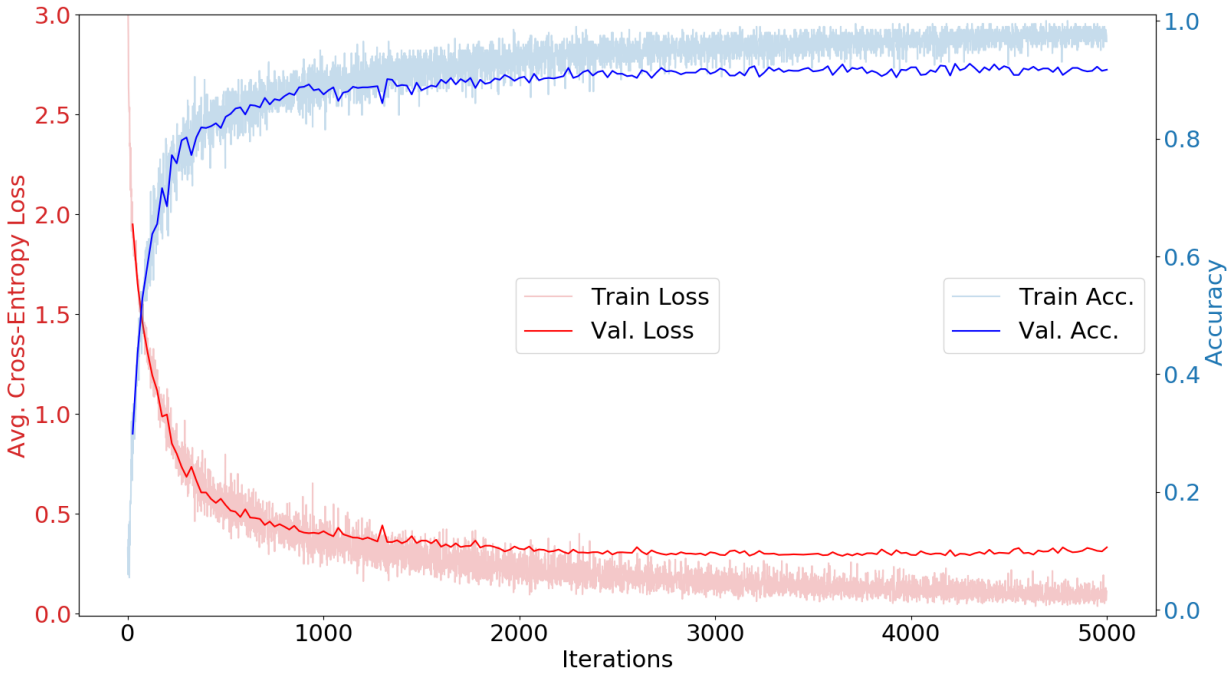
1. 5-layer with Sigmoid Activation



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



3. 5-layer with ReLu Activation



a) The default step size is the smoothest, with the ReLu taking second and the .1 step size sigmoid taking third. The default sigmoid has rather poor performance and does not increase much with more iterations. The Sigmoid with .1 step size has a linear plot, however it is the least smooth, and the performance is good but not great. The ReLu has the best performance and is pretty smooth. The plot flattens around 1000 iterations with around 90% validation accuracy.

b) The increase learning rate allows the weights to be more accurate, but less precise. With the smaller step size the weights never change much from their initial values. With the larger step size it gets close to the optimal weights.

c) The derivative of the ReLU is a step function at x = 0 and the derivate of the Sigmoid function is a gaussian. Since the gaussian is distributed more around x = 0 it will have smaller steps than the ReLu, even when the step sizes are the same. The ReLu on the other hand is equally distributed at x > 0 which means larger steps can be taken with smaller step sizes.

Q5

1. 11.3%
2. 11.3%
3. 11.3%
4. 20.0%
5. 15.8%

The 5 tests had quite a lot of variance so it would be harder to trust a single trained model's performance. For the previous questions I am still fairly confident since most of the training models had a large disparity in their performance that my answers are pretty accurate, however for picking

the ideal model I would need to run them with multiple random numbers in training to choose between them.

1. 8 hours
2. Easy
3. Alone
4. 70%