Question 1a:

Given: $E(w_1, w_2, v|X) = -\sum_t r^t \log y^t + (1 - r^t)\log(1 - y^t)$

$y^t = \text{sigmoid}(v_2 * z_2 + v_1 * z_1 + v_0)$

$z_1^t = \text{sigmoid}(w_{1,2}x_2^t + w_{1,1}x_1^t + w_{1,0})$

$z_2^t = \text{LReLU}(w_{2,2}x_2^t + w_{2,1}x_1^t + w_{2,0})$

$\text{LReLU}(x) = \begin{cases} 0.01x, & \text{for } x < 0 \\ x, & \text{otherwise} \end{cases}$

$\text{sigmoid} = \frac{1}{1+e^{-x}}$

$$\frac{\partial E}{\partial v} = \frac{\partial E}{\partial y} \cdot \frac{\partial y}{\partial v}$$

$$\frac{\partial E}{\partial y} = -\sum_t \frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}$$

$$= -\sum_t \frac{r^t - r^t y^t - y^t + y^t r^t}{y^t(1-y^t)}$$

$$= -\sum_t \frac{r^t - y^t}{y^t(1-y^t)}$$

$$\frac{\partial y}{\partial v} = \frac{\partial}{\partial v}\left(\frac{1}{1+\exp[v_2 z_2 + v_1 z_1 + v_0]}\right)$$

$$= y^2(z_2 + z_1)$$

$$\frac{\partial E}{\partial v} = \left(-\sum_t \frac{r^t - y^t}{y^t(1-y^t)}\right)\left(y^2(z_2 + z_1)\right)$$

$$= -\sum_t \frac{r^t - y^t}{y^t} y^t(z_2^t + z_1^t)$$

$$\boxed{\Delta v_{ih} = \eta \sum_t \frac{r_i^t - y_i^t}{y_i^t} y_i^t (z_{2i}^t + z_{1i}^t)}$$

$$\frac{\partial E}{\partial w_h} = \frac{\partial}{\partial w}\left(-\sum_t (r^t \log y^t + (1 - r^t)\log(1 - y^t))\right)$$

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w} \qquad \left(\frac{\partial E}{\partial y} \text{ was solved above}\right)$$

$$\frac{\partial y}{\partial z} = y^2(v_2 + v_1)$$

$$\frac{\partial z_h}{\partial w} = \begin{cases} 0.01 w_{2h} x_{2h}^t + w_{1h} x_{1h}^t + w_{0h}, & \text{for } w_{2h} x_{2h}^t + w_{1h} x_{1h}^t + w_{0h} < 0 \\ w_{2h} x_{2h}^t + w_{1h} x_{1h}^t + w_{0h}, & \text{otherwise} \end{cases}$$

$$= \frac{\max(0, w_h^t x^t) + 0.01 \min(w_h^t x^t, 0)}{w_h^t}$$

$$\boxed{\Delta w_{2j} = \eta \left(-\sum_t \frac{r_i^t - y_i^t}{y_i^t} y_i^t (v_2 + v_1) \frac{\max(0, w_{2j}^t x_j^t + w_{2j}^t) + 0.01 \min(w_{2j}^t x_j^t, 0)}{w_{2j}^t} \right)}$$

$$\boxed{\Delta w_{1j} = \eta \left(-\sum_t \frac{r_i^t - y_i^t}{y_i^t} y_i^t (v_2 + v_1) \frac{\max(0, w_{1j}^t x_j^t + w_{1j}^t) + 0.01 \min(w_{1j}^t x_j^t, 0)}{w_{1j}^t} \right)}$$

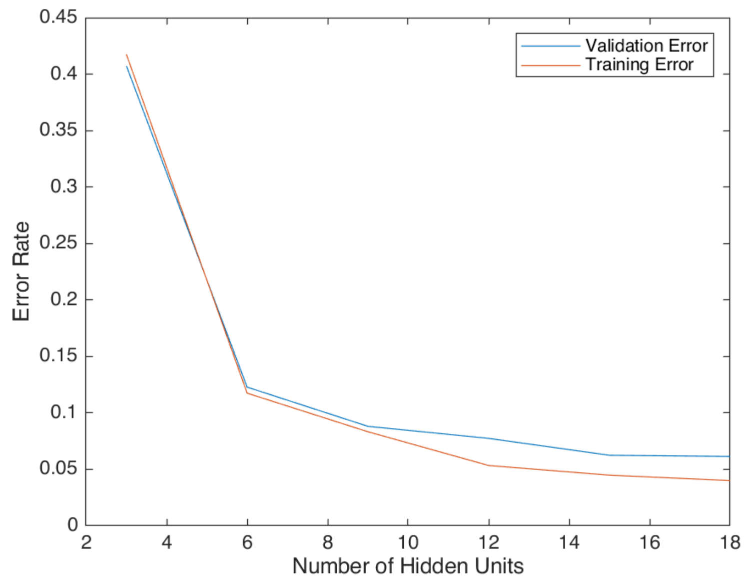
Question 1b:

For $w = w_1 = w_2$:

$$\boxed{\Delta v_{ih} = \eta \sum_t \frac{r_i^t - y_i^t}{y_i^t} y_i^t (z_{2i}^t + z_{1i}^t)}$$

$$\boxed{\Delta w_{hj} = \eta \left(-\sum_t \frac{r_i^t - y_i^t}{y_i^t} y_i^t (v_2 + v_1) \frac{\max(0, w_{hj}^t x_j^t) + 0.01 \min(w_{hj}^t x_j^t, 0)}{w_{hj}^t} \right)}$$

Question 2a:

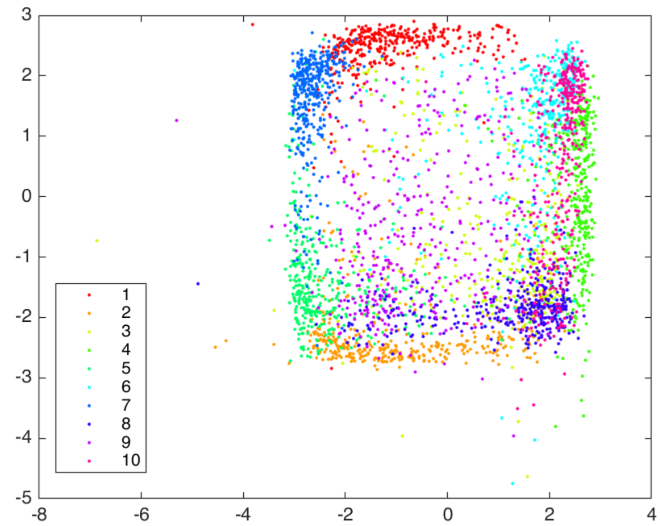


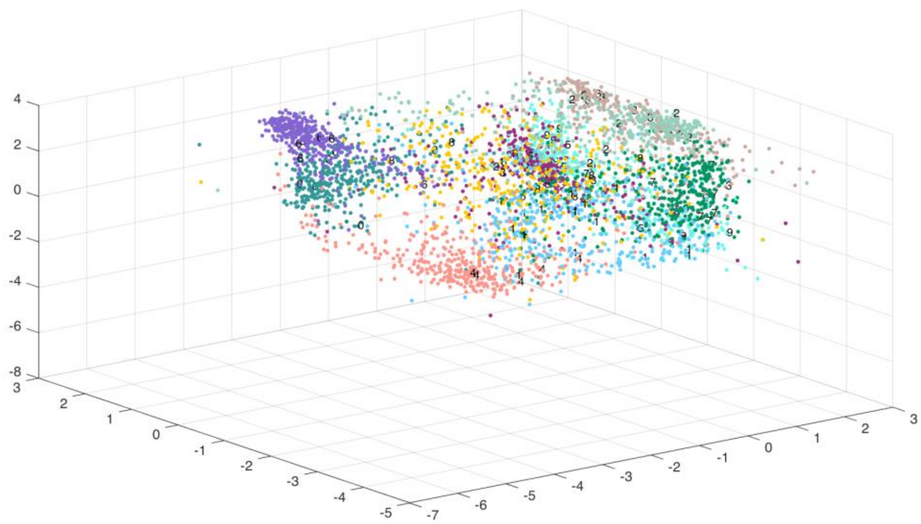
# Hidden Units:	3	6	9	12	15	18
Training Error Rate (%)	0.41751	0.11746	0.083289	0.05339	0.044848	0.040043
Validation Error Rate (%)	0.40737	0.1228	0.088094	0.077416	0.062467	0.061399
Testing Error Rate (%)	0.42956	0.14248	0.091249	0.067236	0.065101	0.060299

Based on the testing results, **18 hidden units** is optimal out of the {3,6,9,12,15,18} set. With 18 hidden units, a testing error rate of 0.060299% was achieved.

Brian Cooper (5310799)
coope824@umn.edu

Question 2b:





The 3-D plot allows for a better perception of the distribution of the data. In the 2-D plot, there is a lot of overlap among the data points, so it's harder to see the separation.

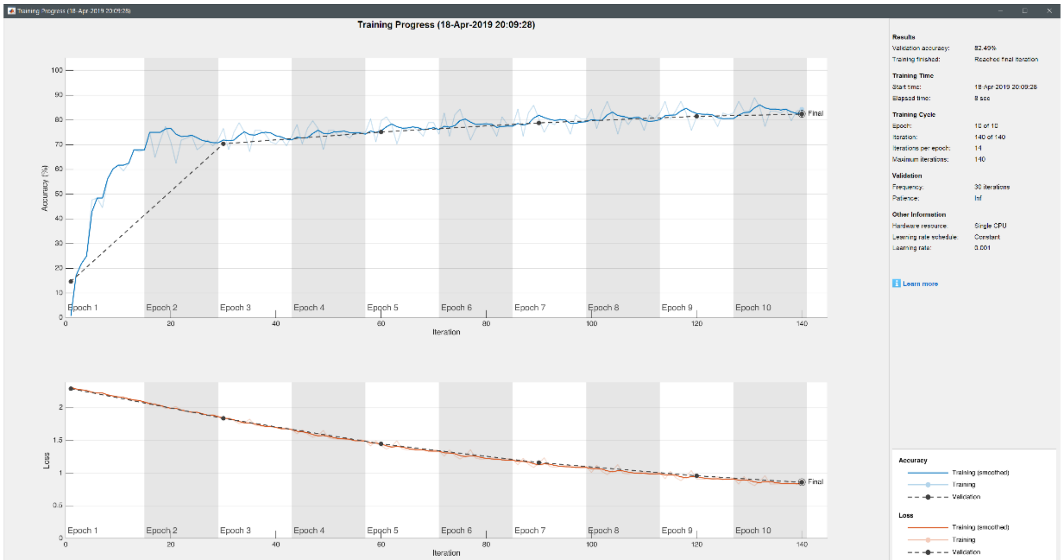
The imlementation of the ReLU function in 2a dictates that the value will either be zero or its current value, depending on whether or not it is below zero. In the given data, most of the values are positive, so they are displayed as they are in input on the graphs. The negative values are shown because of my usage of the logarithm in the MATLAB code: negative values are converted to positive, and then the log of this value is taken and finally multipled by -1.

Brian Cooper (5310799)
coope824@umn.edu

Question 3c:

CNN structure 1: Input layer → 2D convolution layer (1 filter w/ size 4) → Batch normalization layer → LReLU layer → Fully connected layer → Softmax layer → Classification layer

Accuracy: 0.8212



CNN structure 2: Input layer → 2D convolution layer (20 filter w/ size 3) → Batch normalization layer → LReLU layer → Pooling layer (max pooling, poolsize 3 and stride size 2) → 2D convolution layer (32 filter with size 3) → Batch normalization layer → LReLU layer → Fully connected layer → Softmax layer → Classification layer

Accuracy: 0.9146