1)

1)

$$L = -\sum_k p_k \log\left(\frac{e^{a_k}}{\sum_j e^{a_j}}\right)$$

$$= -\sum_k p_k \left(a_k - \log \sum_j e^{a_j}\right).$$

$$\delta^{(L)} = \nabla_a L \cdot \sigma'(z^L) \quad [\text{Last layer error}]$$

$$\delta_i^{(L)} = \frac{\partial}{\partial a} L = -\left(\frac{e^{a_i}}{\sum_j e^{a_j}} - p_i\right)$$

For the other layers, this $\delta^{(L)}$ is propagated backwards

$$\delta^{(l-1)} = (w^{(l)})^T \delta^{(l)} \odot \sigma'(z^l)$$

where $\sigma()$ is the activation function.

Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad \sigma'(x) = \sigma(x)\cdot(1-\sigma(x))$$

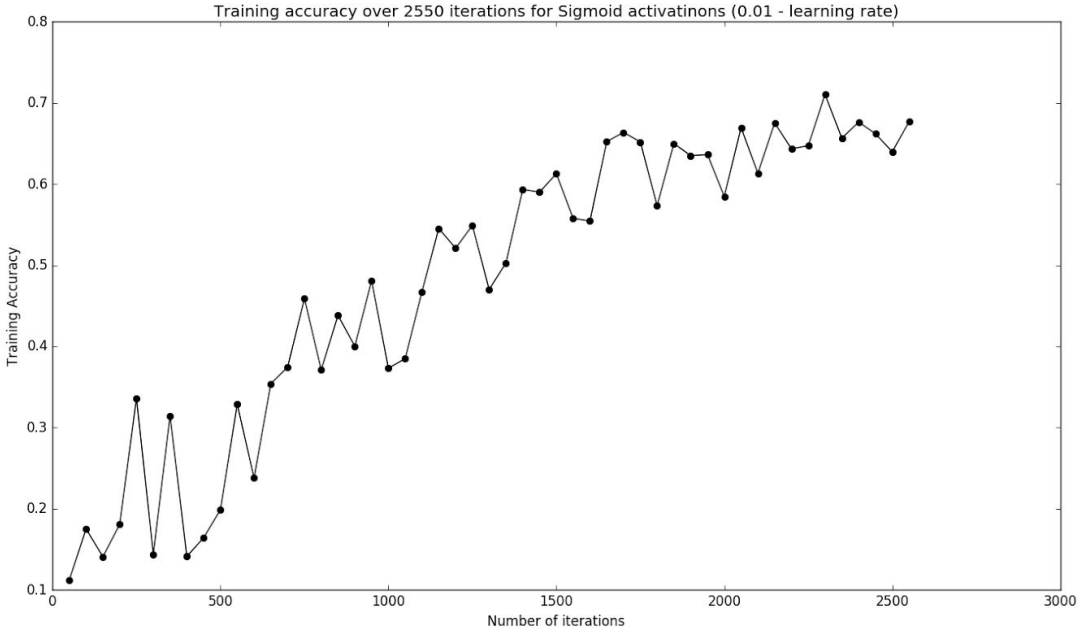
RELU

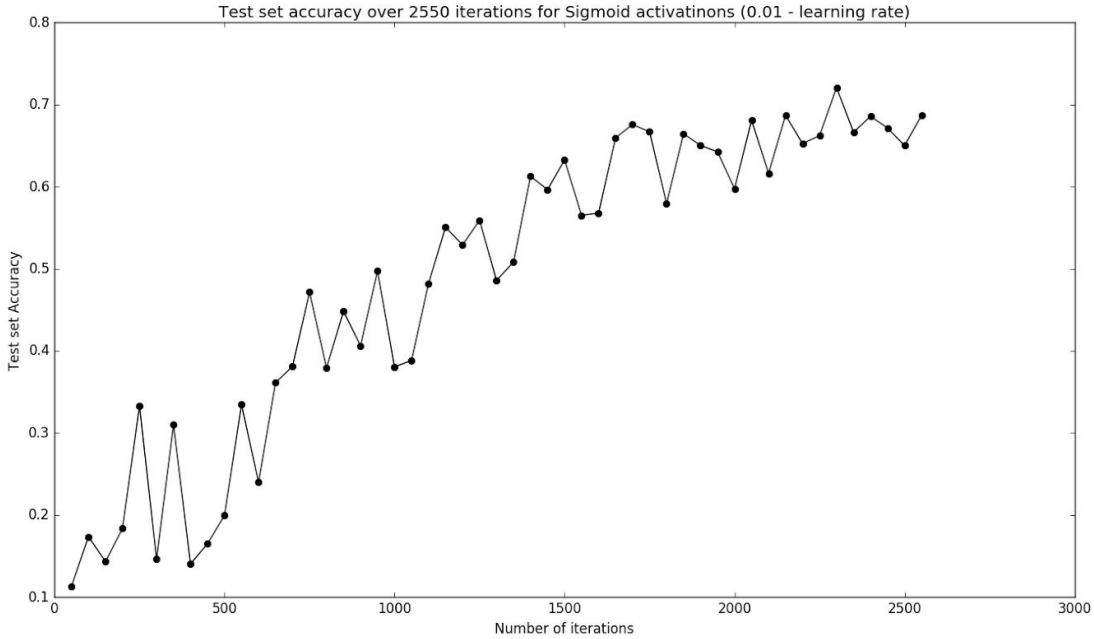
$$\sigma(x) = \begin{cases} x, & x \geq 0 \\ 0, & x \leq 0 \end{cases} \qquad \sigma'(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

$$\frac{\partial L}{\partial w^l_{jk}} = \delta_j^{(l+1)} a_k^{(l)} \qquad \frac{\partial L}{\partial b_j^{(l)}} = \delta_j^{(l+1)}$$

2)

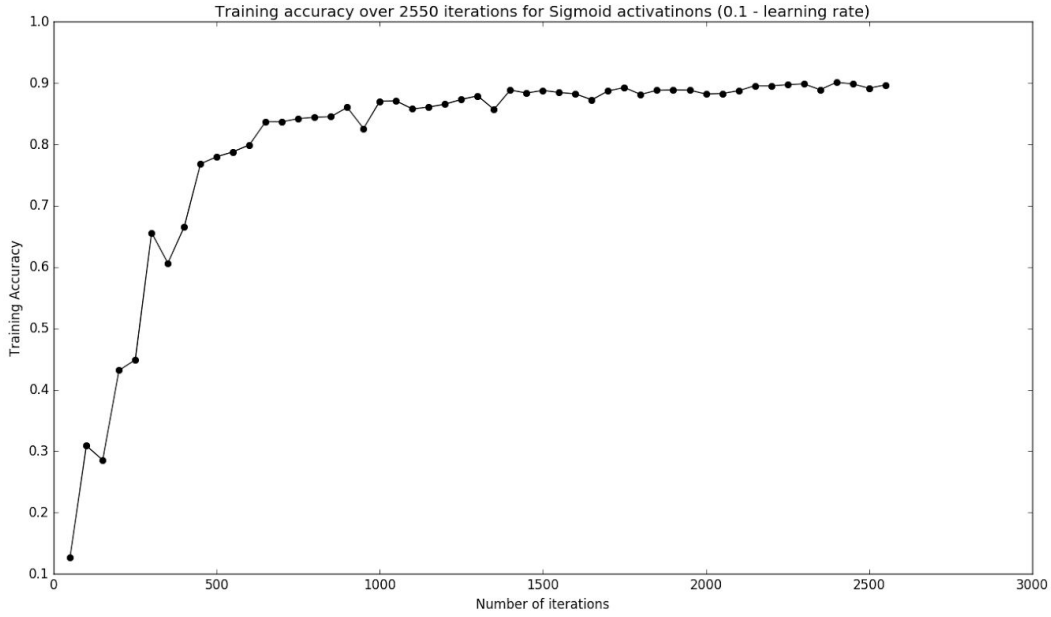
Learning rate of 0.01

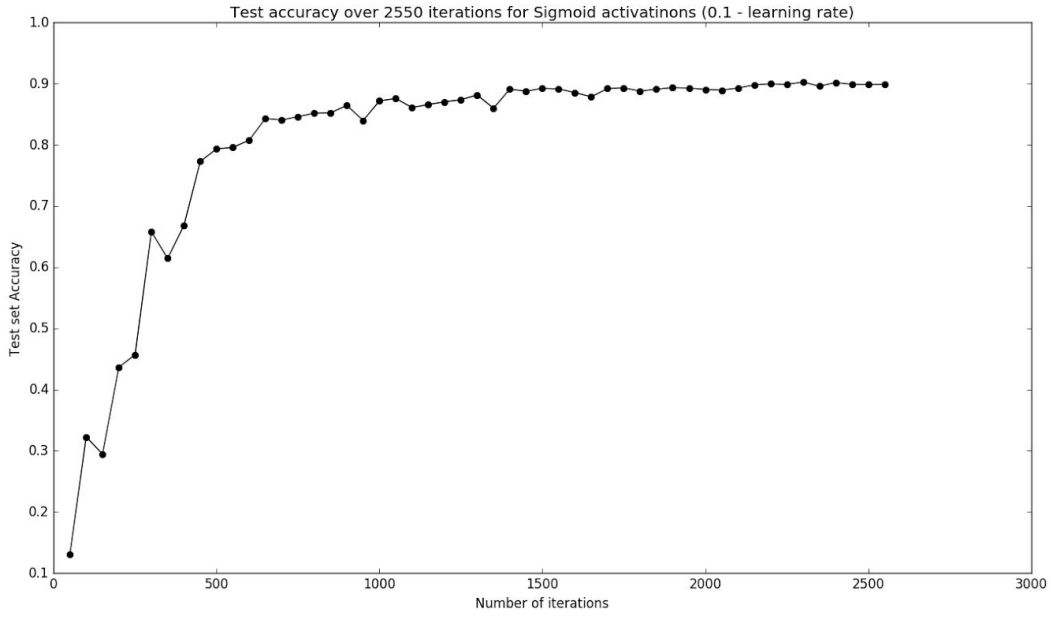




The learning rate is very less and due to computational constraints the number of iterations for which the program was run was only 2550,hence it has to converged properly.

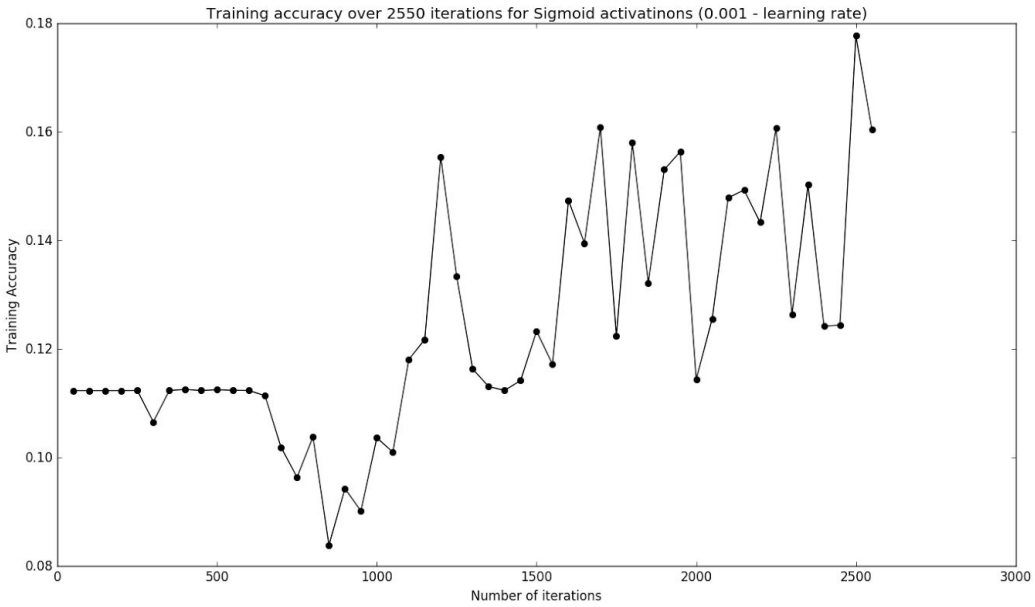
Learning rate of 0.1

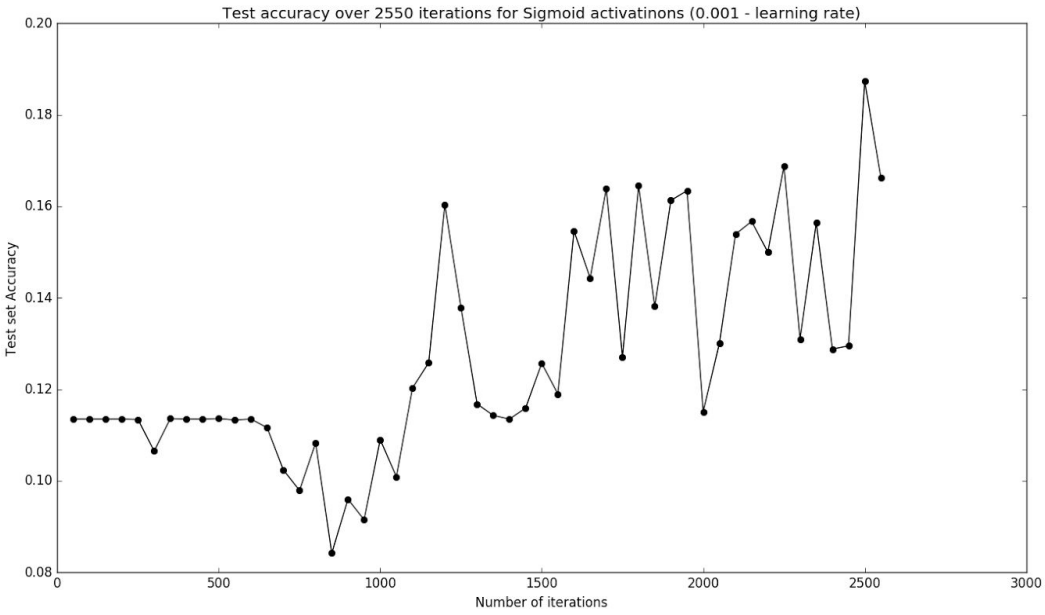




The learning rate of 0.1 seems to be the ideal among the three learning rate,it has converged within 2550 iterations.

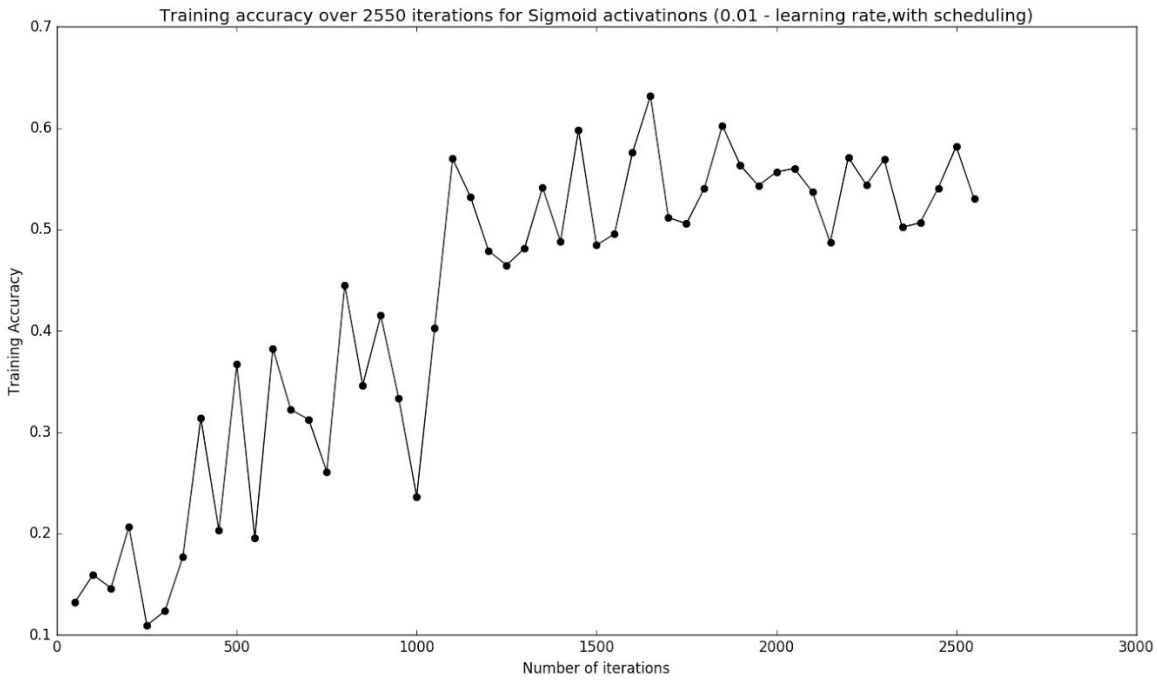
Learning rate of 0.001

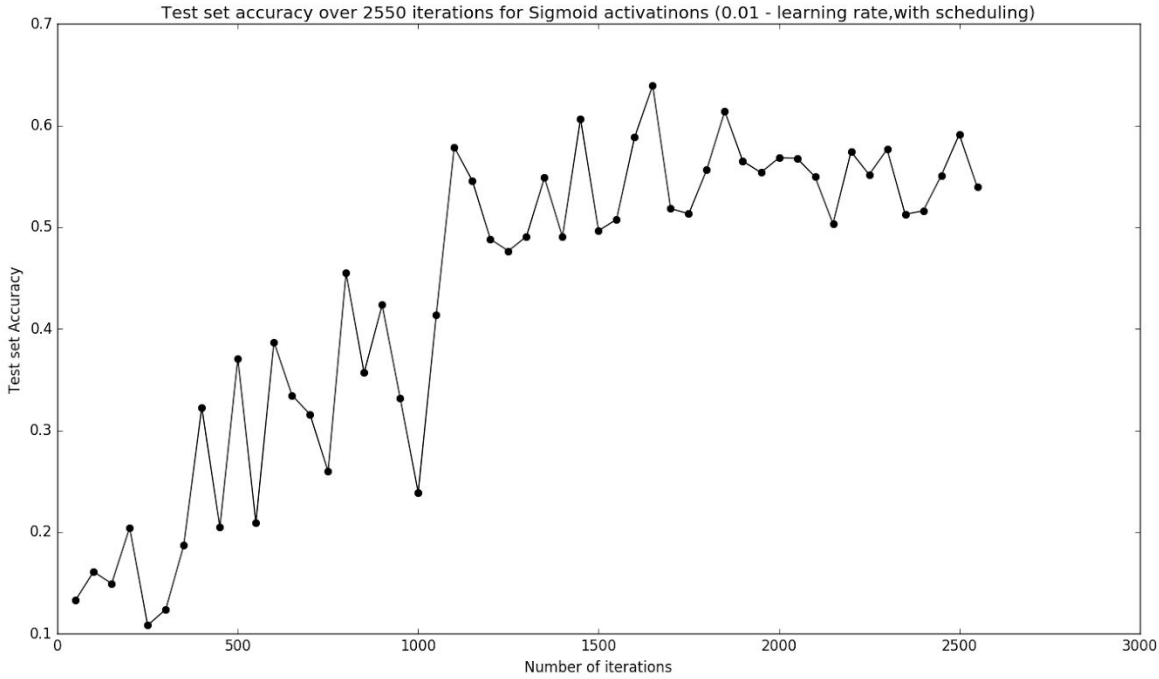




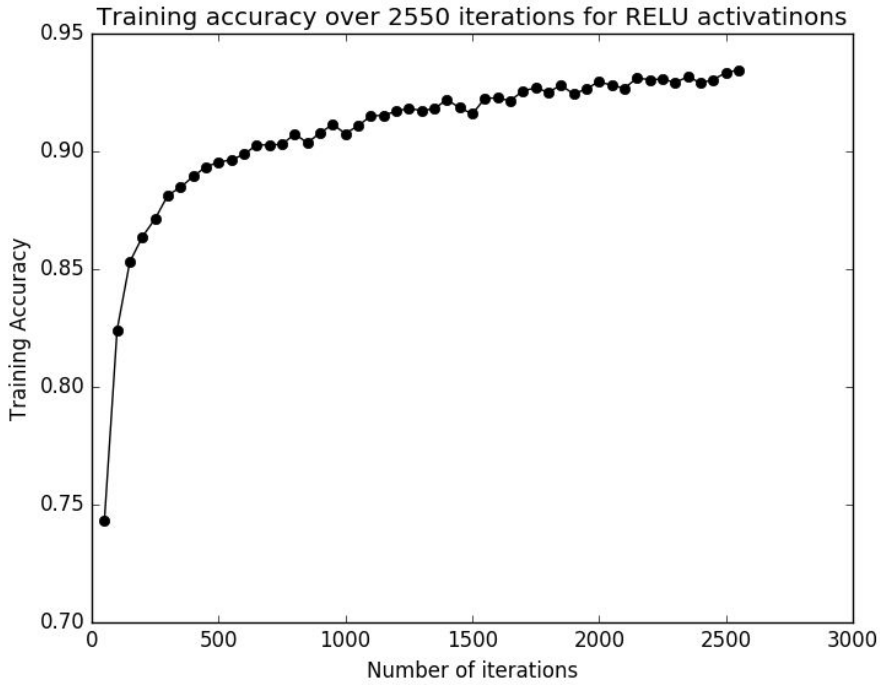
The learning rate is very less and due to computational constraints the number of iterations for which the program was run was only 2550,hence it has to converged properly.

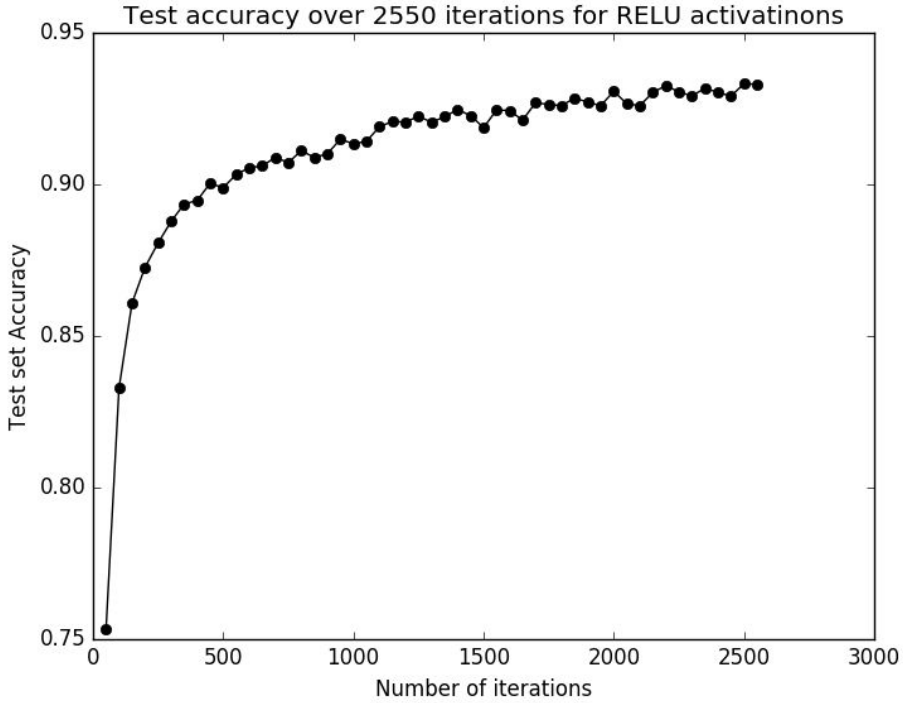
3)


Training accuracy over 2550 iterations for Sigmoid activatinons (0.01 - learning rate,with scheduling)
Training Accuracy
Number of iterations
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
500
1000
1500
2000
2500
3000


Test set accuracy over 2550 iterations for Sigmoid activatinons (0.01 - learning rate,with scheduling)
Test set Accuracy
Number of iterations
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
500
1000
1500
2000
2500
3000

4) RELU performs better than sigmoid activation function for the given learning rate.It could be accounted by the lesser gradients for sigmoid function at the extremes.





5)

RELU

7 - 7,3,2
2 - 2,6,3
1 - 1,2,6
0 - 0,2,6
4 - 4,7,2

SIGMOID

4 - 4,9,7
3 - 3,2,8
7 - 7,9,4
9 - 7,9,4
5 - 3,5,8