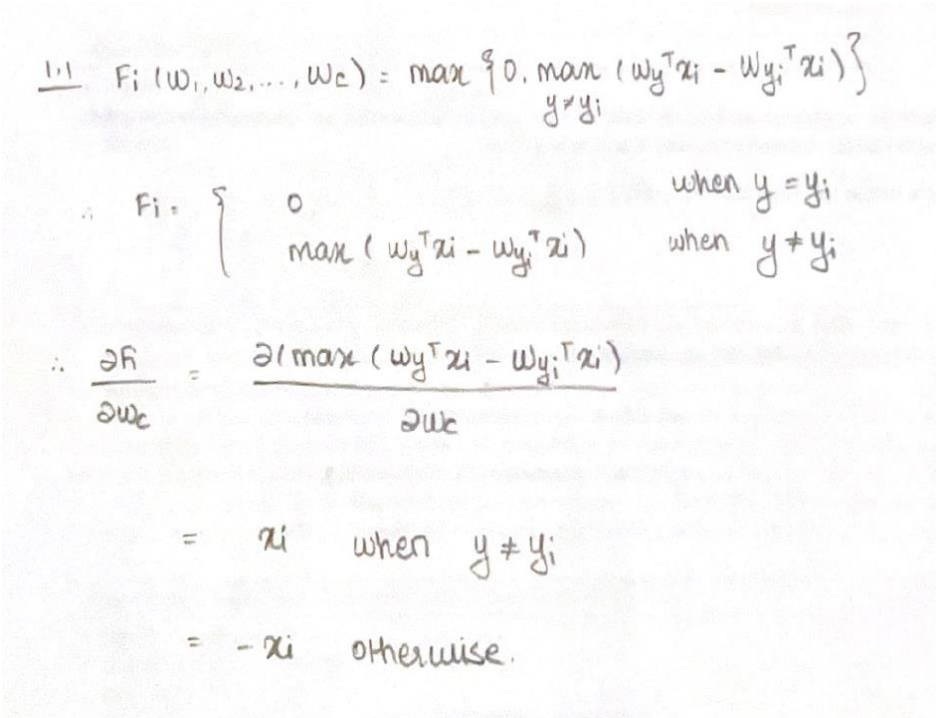
Homework 03

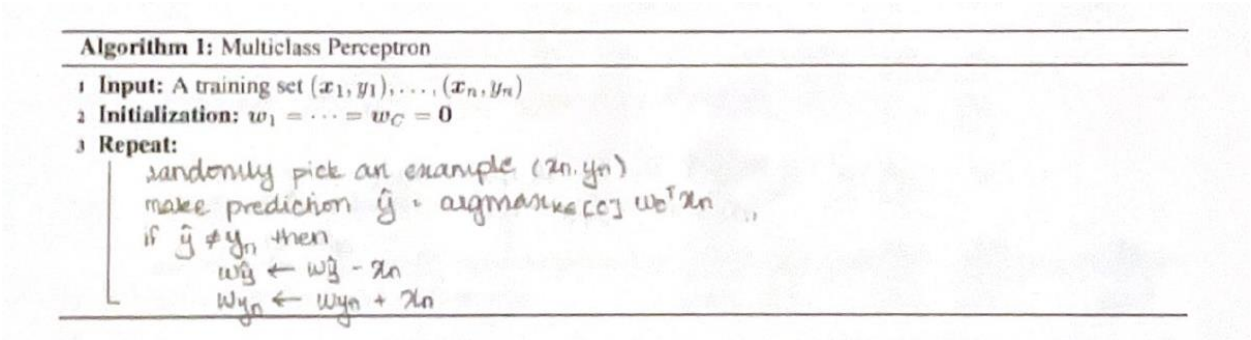
Group Members: Ira Deshmukh (3648692173) and Prem Tibadiya (6351018440)

Question 01:

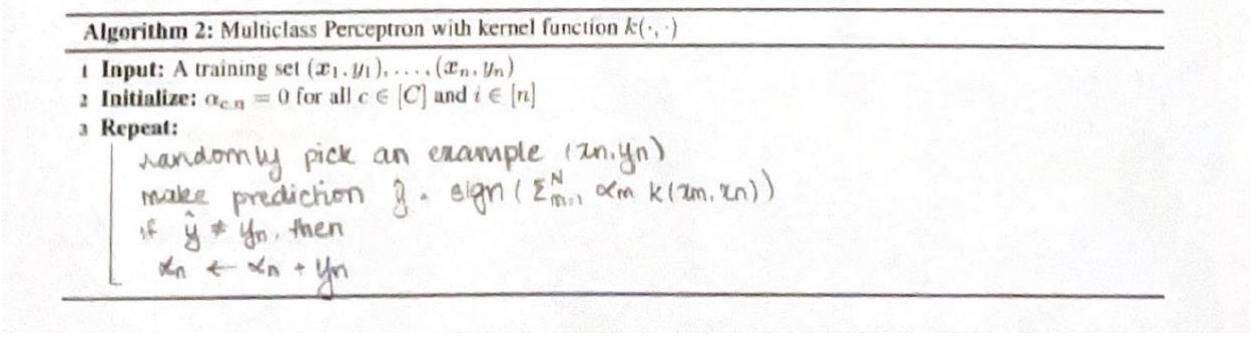
1.1



1.2

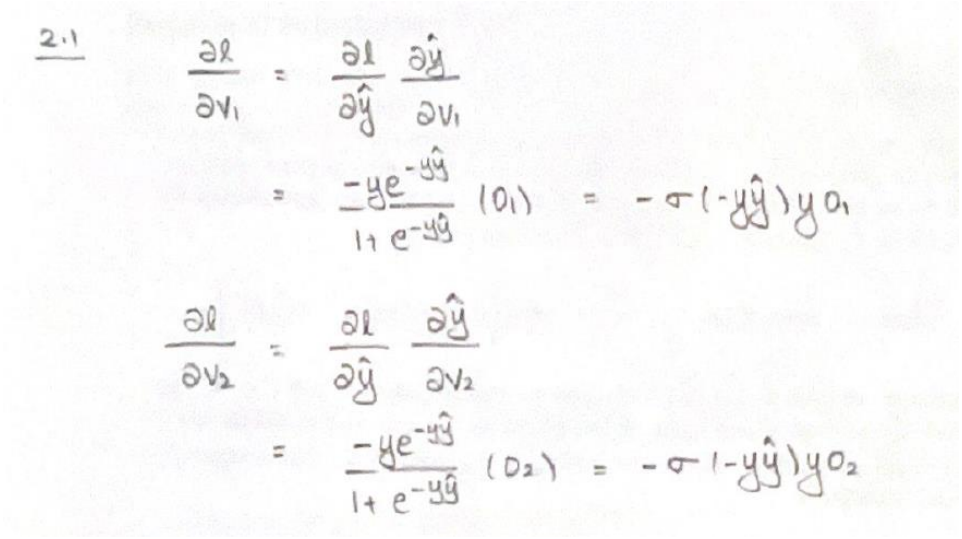


1.3



Question 02:

2.1



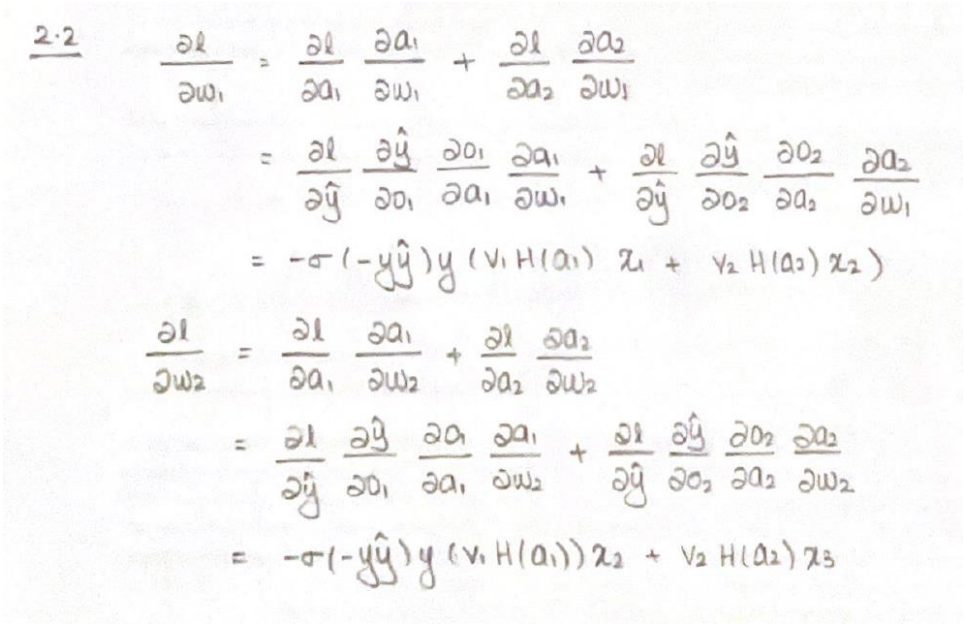
$$\frac{\partial l}{\partial v_1} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v_1}$$

$$= \frac{-ye^{-y\hat{y}}}{1+e^{-y\hat{y}}}(o_1) = -\sigma(-y\hat{y})\, y\, o_1$$

$$\frac{\partial l}{\partial v_2} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v_2}$$

$$= \frac{-ye^{-y\hat{y}}}{1+e^{-y\hat{y}}}(o_2) = -\sigma(-y\hat{y})\, y\, o_2$$

2.2



$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial a_1} \frac{\partial a_1}{\partial w_1} + \frac{\partial l}{\partial a_2} \frac{\partial a_2}{\partial w_1}$$

$$= \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_1} \frac{\partial o_1}{\partial a_1} \frac{\partial a_1}{\partial w_1} + \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_2} \frac{\partial o_2}{\partial a_2} \frac{\partial a_2}{\partial w_1}$$

$$= -\sigma(-y\hat{y})\, y\, (v_1 H(a_1)\, x_1 + v_2 H(a_2)\, x_2)$$

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial a_1} \frac{\partial a_1}{\partial w_2} + \frac{\partial l}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

$$= \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_1} \frac{\partial o_1}{\partial a_1} \frac{\partial a_1}{\partial w_2} + \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_2} \frac{\partial o_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

$$= -\sigma(-y\hat{y})\, y\, (v_1 H(a_1))\, x_2 + v_2 H(a_2)\, x_3$$

2.3

2.3

Forward Propogation:

$$a_1 = x_1 w_1 + x_2 w_2$$

$$a_2 = x_2 w_1 + x_3 w_2$$

$$o_1 = \max\{0, a_1\}$$

$$o_2 = \max\{0, a_2\}$$

$$\hat{y} = o_1 v_1 + o_2 v_2$$

Backward Propogation: update:

$$w_1 \leftarrow w_1 - \eta\left[(\sigma(y\hat{y}) - 1)\, y\, (v_1 H(a_1) x_1 + v_2 H(a_2) x_2)\right]$$

$$w_2 \leftarrow w_2 - \eta\left[\sigma(y\hat{y} - 1)\, y\, (v_1 H(a_1) x_2 + v_2 H(a_2) x_3)\right]$$

$$v_1 \leftarrow v_1 - \eta\,(-\sigma)(-y\hat{y})\, y o_1$$

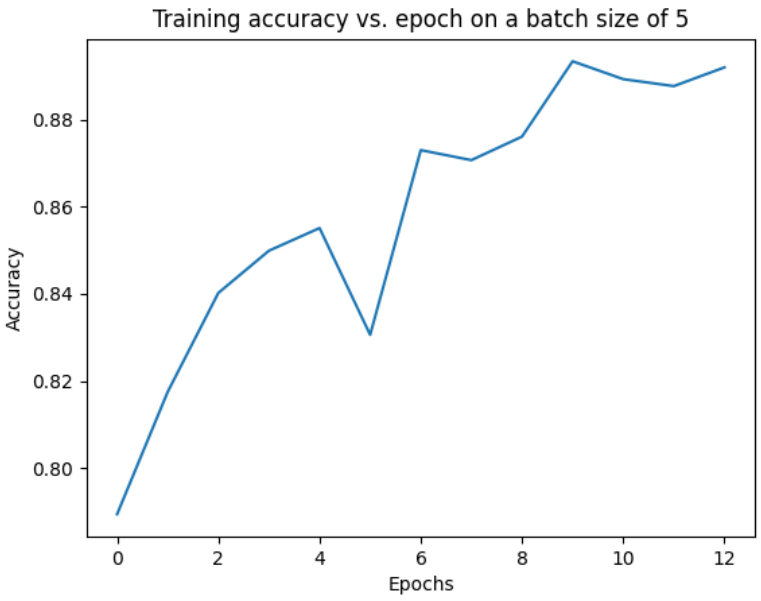
$$v_2 \leftarrow v_2 - \eta\,(\sigma(y\hat{y}) - 1)\, y o_2$$

Question 03:

3.1.6

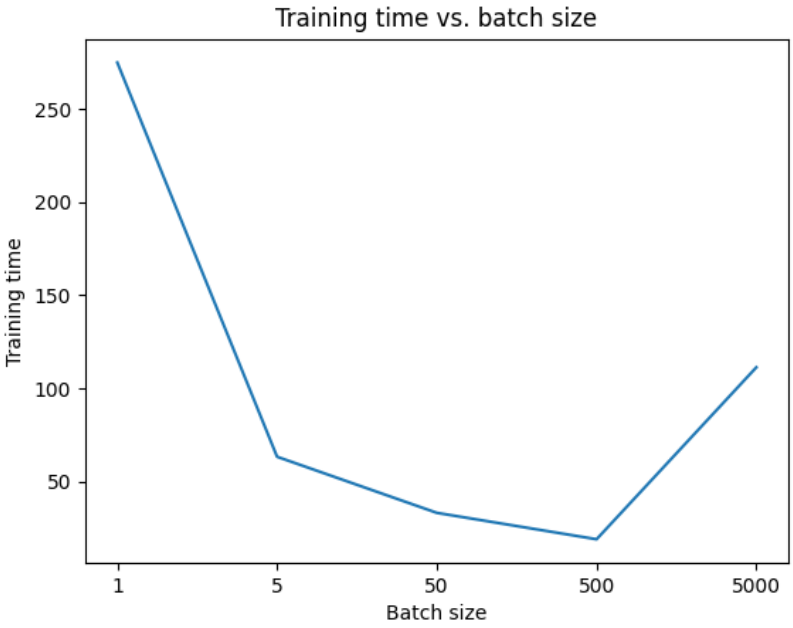
```
PROBLEMS (1)  OUTPUT  DEBUG CONSOLE  TERMINAL  JUPYTER                                                    python3.10

PS C:\Users\idesh\Downloads\University of Southern California\CSCI 567\Homework_03\hw3\startercode> python neural_networks.py --minibatch_size 5 --check_gradient --check_magnitude
Training data size: 10000, Validation data size: 2000, Test data size: 10000
Check the magnitude (L1-norm of layer L1) of gradient with batch size 50: 148.290031 and with batch size 5k: 117.085383
Check the gradient of W in the L1 layer from backpropagation: 0.000000 and from approximation: 0.000000
Check the gradient of b in the L1 layer from backpropagation: 0.129579 and from approximation: 0.129579
Check the gradient of W in the L2 layer from backpropagation: 0.004011 and from approximation: 0.003889
Check the gradient of b in the L2 layer from backpropagation: 0.121975 and from approximation: 0.121980
At epoch 1
100%|| 2000/2000 [00:07<00:00, 268.52it/s]
Training loss at epoch 1 is 6.061195848635396
Training accuracy at epoch 1 is 0.7894
Validation accuracy at epoch 1 is 0.7745
At epoch 2
100%|| 2000/2000 [00:07<00:00, 260.87it/s]
Training loss at epoch 2 is 5.1890470982486585
Training accuracy at epoch 2 is 0.8175
Validation accuracy at epoch 2 is 0.8
At epoch 3
100%|| 2000/2000 [00:07<00:00, 266.95it/s]
Training loss at epoch 3 is 4.581589580677092
Training accuracy at epoch 3 is 0.8402
Validation accuracy at epoch 3 is 0.816
At epoch 4
100%|| 2000/2000 [00:07<00:00, 267.52it/s]
Training loss at epoch 4 is 4.327179321037223
Training accuracy at epoch 4 is 0.8499
Validation accuracy at epoch 4 is 0.821
At epoch 5
 57%|| 1148/2000 [00:04<00:03, 227.75it/s]
```

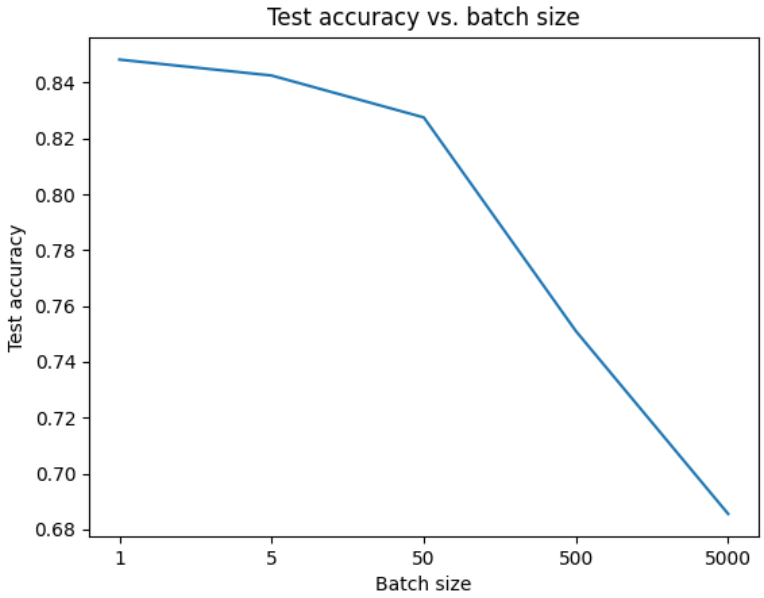
Plot of Training accuracy vs. epochs using plot_train_process.py



3.2.1

Plot of Testing accuracy and training time w.r.t batch size using plot_batch.py





3.2.2

Statictics for each batch size:

Batch Size = 1
Epochs = 13
Best Epoch = 10

Test Accuracy at Epoch = 0.8482

Batch Size = 5
Epochs = 13
Best Epoch = 10
Test Accuracy at Epoch 10 = 0.8425

Batch size = 50
Epochs = 38
Best Epoch = 35
Test Accuracy at Epoch 35 = 0.8275

Batch size = 500
Epochs = 42
Best Epoch = 39
Test Accuracy at Epoch 39 = 0.751

Batch size = 5000
Epochs = 161
Best Epoch = 158
Test Accuracy at Epoch 158 = 0.6856

Thus, for each batch the number of training epochs required to get the best model are:
Batch Size = 1, number of epochs = 13
Batch Size = 5, number of epochs = 13
Batch Size = 50, number of epochs = 38
Batch Size = 500, number of epochs = 42
Batch Size = 5000, number of epochs = 161

Number of gradient updates required to get the best model for each batch size:
Batch Size = 1, number of epochs = 10
Batch Size = 5, number of epochs = 10
Batch Size = 50, number of epochs = 35
Batch Size = 500, number of epochs = 39
Batch Size = 5000, number of epochs = 158

3.2.3

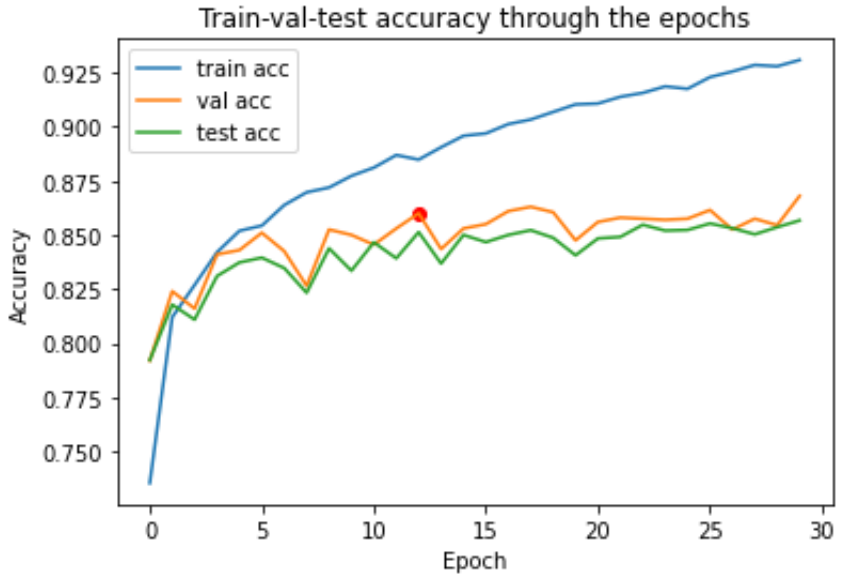
(i) No smaller batch size does not mean faster training. This is because smaller batch size means taking smaller steps and thereby it won't train the model faster but might infact be a bit slower as compared to bigger batch sizes.

(ii) Increasing the batch size does not guarantee better test accuracy. This is because the batch size has a direct relation to the variation of the gradient estimator. Batch size controls the accuracy of the estimate of the error gradient when training neural networks.

(iii) large batch size means the model makes very large gradient updates and takes more number of epochs to reach a convergence to minimizer. This does not necessarily mean that large batch sizes require less gradient updates

3.3

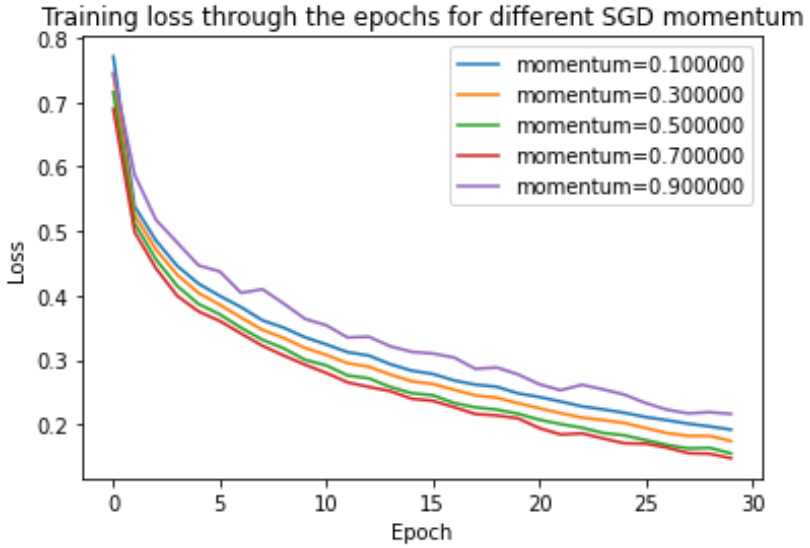
(i)



(ii) The trend after early stopping point: training accuracy increases greatly for each epoch whereas the testing and validation accuracy don't show any improvement but have a fluctuation within a certain range
(iii) Setting it to zero or really small means that as soon as the performance measure gets worse from one epoch to the next, the training is terminated. This might not be ideal since, the model's performance is noisy and might go up or down from one epoch to the next. What we really care about is that the general trend should be improving.

3.4

(i)



(ii)

Based on this the suitable value of alpha is 0.7.