Problem 1 – Newtons Method

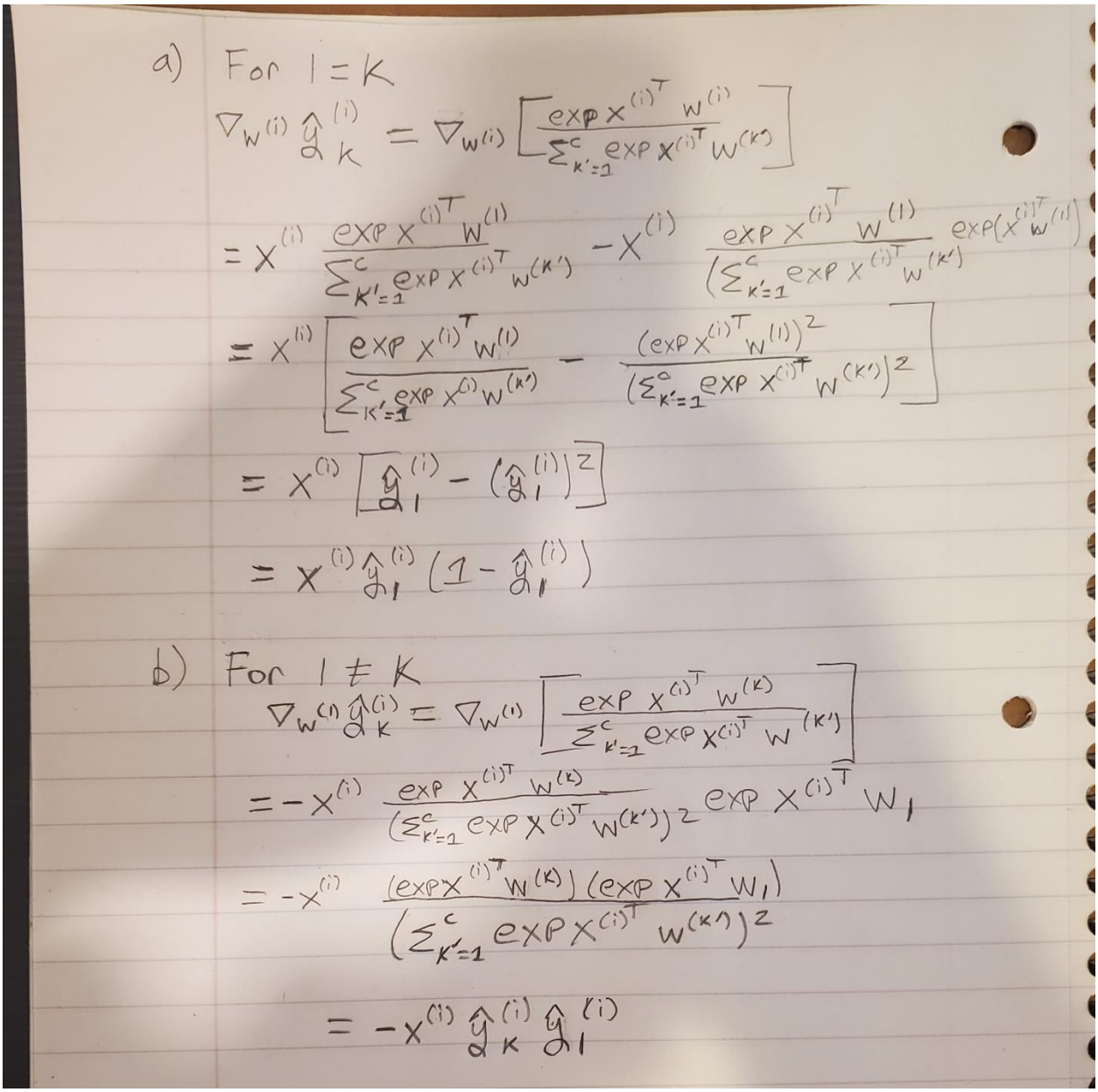
According to the lecture slides the Hessian is a constant: $H = 1/n\ XX\text{^}T$ therefore when updating the weight vectors (with w being the initial weight vector) the optimal solution w* is:

$$w^* = w - H^{-1}\nabla_w Fmse = w - \left(\frac{1}{n}XX^T\right)^{-1}\nabla_w Fmse = w - n(XX^T)^{-1}\nabla_w Fmse$$

$$= w - n(XX^T)^{-}1\,\frac{1}{n}X(X^Tw - y) = w - (XX^T)^{-1}XX^Tw + (XX^T)^{-}1Xy$$

$$= w - Iw + (XX^T)^{-1}Xy = (XX^T)^{-1}Xy = w^*$$

Problem 2 – Derivation of SoftMax Regression Gradient Updates



c) $\nabla_{W^{(l)}} \delta_{CE}(W,b) = -\sum_{i=1}^{m} \sum_{k=1}^{c} y_k^{(i)} \nabla_{W^{(l)}} \log \hat{y}_k^{(i)}$

$= -\sum_{i=1}^{m} x^{(i)} \left[\frac{y_1^{(i)} \hat{y}_1^{(i)} (1-\hat{y}_1^{(i)})}{\hat{y}_1^{(i)}} - \sum_{k \neq 1} \frac{y_k^{(i)} \hat{y}_k^{(i)} \hat{y}_1^{(i)}}{\hat{y}_k^{(i)}} \right]$

$= -\sum_{i=1}^{m} x^{(i)} \left[y_1^{(i)} (1-\hat{y}_1^{(i)}) - \sum_{k \neq 1} y_k^{(i)} \hat{y}_1^{(i)} \right]$

$= -\sum_{i=1}^{m} x^{(i)} \left[y_1^{(i)} (1-\hat{y}_1^{(i)}) + y_1^{(i)} \hat{y}_1^{(i)} - \sum_{k} y_k^{(i)} \hat{y}_1^{(i)} \right]$

$= -\sum_{i=1}^{m} x^{(i)} \left[y_1^{(i)} - y_1^{(i)} \hat{y}_1^{(i)} + y_1^{(i)} \hat{y}_1^{(i)} - \hat{y}_1^{(i)} \sum_{k} y_k^{(i)} \right]$

$= -\sum_{i=1}^{m} x^{(i)} \left[y_1^{(i)} - \hat{y}_1^{(i)} \right]$

Combine into a single vector with the knowledge that b is derived the same way as ∇_W except with no $x^{(i)}$ term

$$\nabla_b \delta_{CE}(W,b) = -\frac{1}{n} \sum_{i=1}^{n} \left[y^{(i)} - \hat{y}^{(i)} \right]$$

Problem 3 – Derivation of Cross-Entropy as Negative Log-Likelihood

Question (3) Given $P(y \mid x, w, b) =$

$$= \prod_{k=1}^{c} \hat{y}_k^{y_k}$$

For entire Dataset

$$-\log P(D \mid w, b) = \prod_{i=1}^{n} \prod_{k=1}^{c} \hat{y}_k^{(i)\, y_k^{(i)}}$$

$$= -\log \prod_{i=1}^{n} \prod_{k=1}^{c} \hat{y}_k^{(i)\, y_k^{(i)}} \quad \left[\because \log(A \cdot B) = \log A + \log B \right]$$

$\therefore$ ~~$\log(A \cdot B) = \log A + \log B$~~.

$$= \prod_{i=1}^{n} \prod_{k=1}^{c} \log \left(\hat{y}_k^{(i)\, y_k^{(i)}} \right)$$

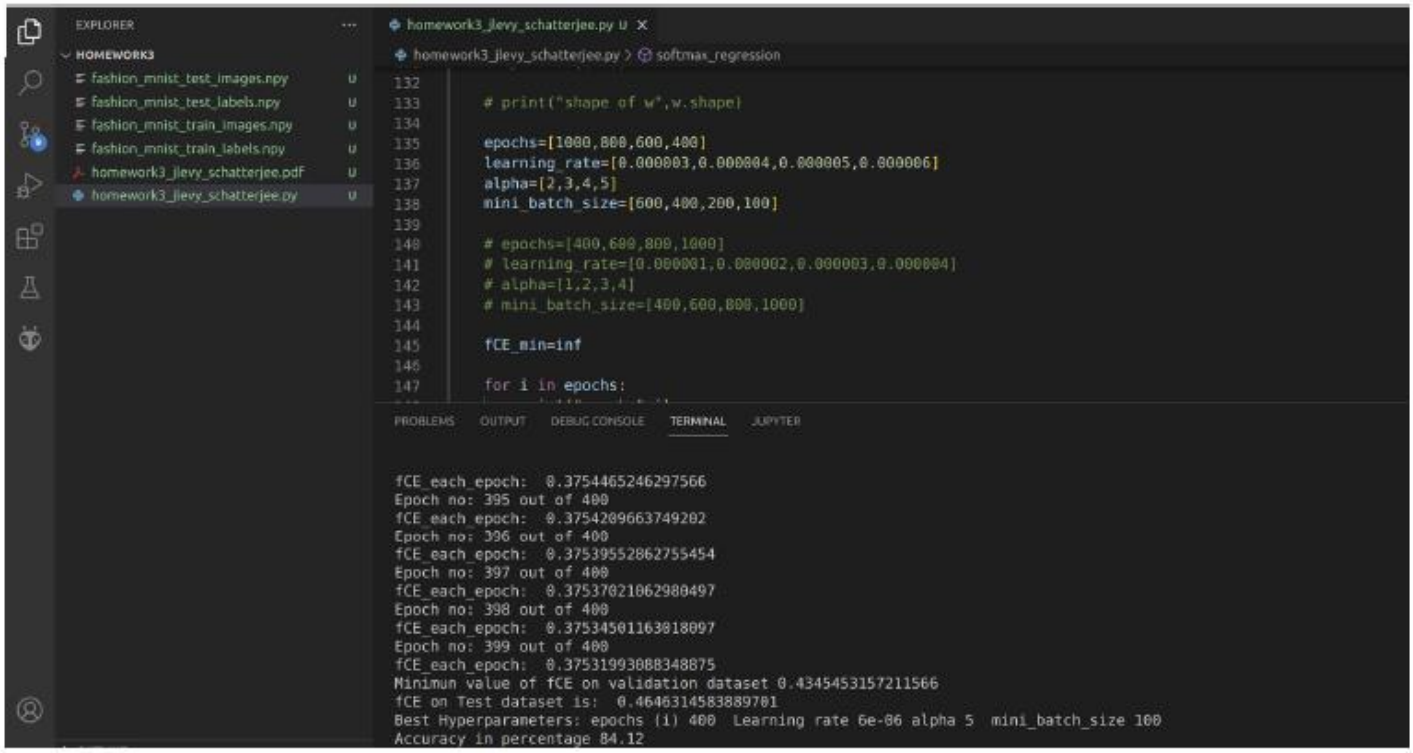
$$= \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}$$

Problem 4 – Implementation of SoftMax Regression

Question 4)

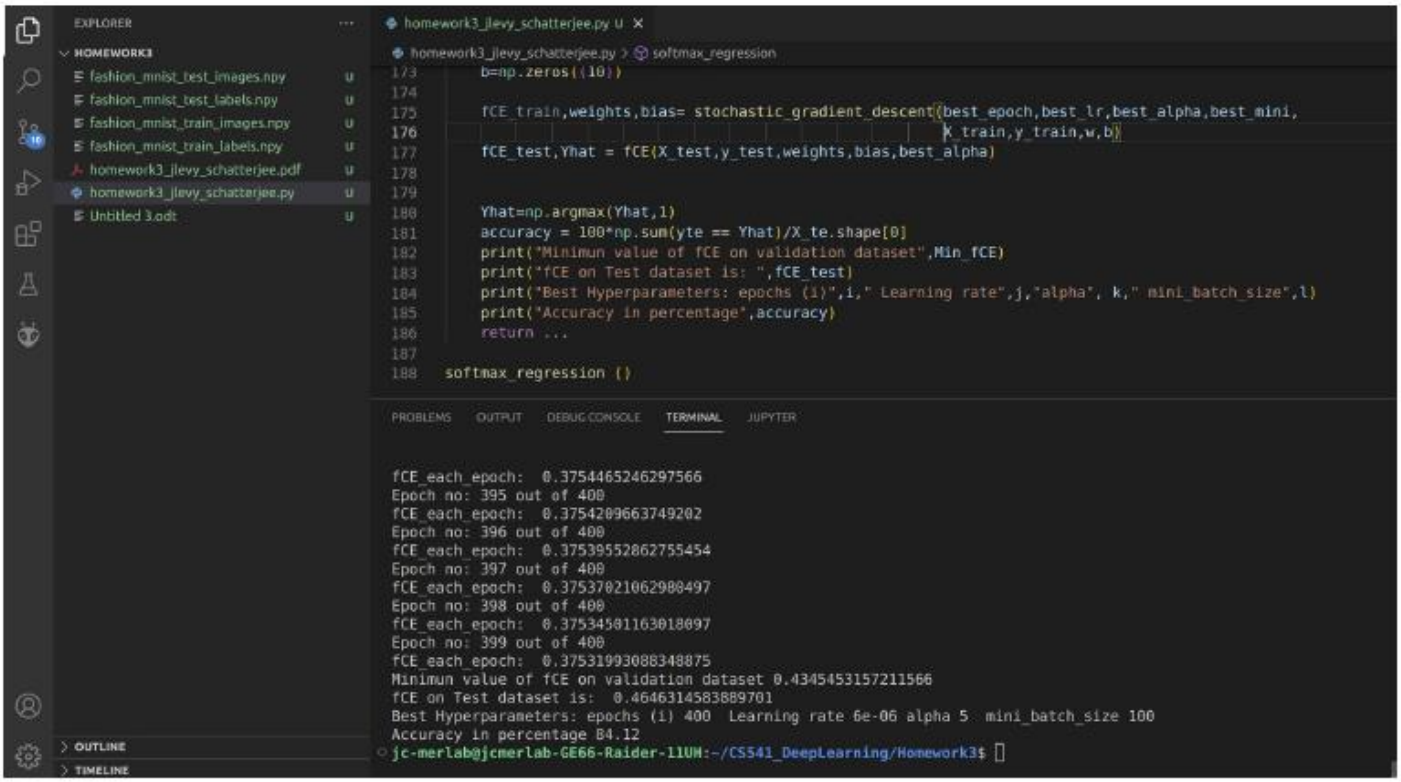
Hyperparameters:

```
epochs=[1000,800,600,400]
learning_rate=[0.000003,0.000004,0.000005,0.000006]
alpha=[2,3,4,5]
mini_batch_size=[600,400,200,100]
```



Results:

```
Minimun value of fCE on validation dataset 0.4345453157211566
fCE on Test dataset is:  0.4646314583889701
Best Hyperparameters: epochs (i) 400  Learning rate 6e-06 alpha 5  mini_batch_size 100
Accuracy in percentage 84.12
```



fCE on Test data: **0.4345**
Optimized Hyperparameter values -
Epochs: **400**
Learning rate: **0.000006**
alpha: **5**
Minibatch size: **100**
Final Accuracy : **84.12**