CSCI3230 / ESTR3108 2021-22 First Term Assignment 3

I declare that the assignment here submitted is original except for source material explicitly acknowledged, and that the same or closely related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the following websites.

University Guideline on Academic Honesty:

`http://www.cuhk.edu.hk/policy/academichonesty/`

Faculty of Engineering Guidelines to Academic Honesty:

`http://www.erg.cuhk.edu.hk/erg-intra/upload/documents/ENGG_Discipline.pdf`

Student Name: ***Lai Man Hin***

Student ID : ***1155136167***

1a)

$$XX^T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

To find XX^T's eigenvalue, we start at $|XX^T - \lambda I| = 0$

$$|XX^T - \lambda I| = 0$$
$$\begin{vmatrix} 2-\lambda & 2 & 2 \\ 2 & 2-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{vmatrix} = 0$$
$$-\lambda^3 + 6\lambda^2 - 12\lambda + 8 + 8 + 8 - 8 + 4\lambda - 8 + 4\lambda - 8 + 4\lambda = 0$$
$$-\lambda^3 + 6\lambda^2 = 0$$
$$\lambda = 0 \text{ (double root) or } 6$$

1b)

To find the first principal axis of X, we should find the eigenvectors associated with largest eigenvalue (i.e. $\lambda = 6$)

Therefore,

$$\begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 6 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

Then, we have

$$\begin{cases} 2v_1 + 2v_2 + 2v_3 = 6v_1 & (1) \\ 2v_1 + 2v_2 + 2v_3 = 6v_2 & (2) \\ 2v_1 + 2v_2 + 2v_3 = 6v_3 & (3) \end{cases}$$

By considering $v_1 = v_2 = v_3$, we have

$$\begin{cases} 6v_1 = 6v_1 \\ 6v_2 = 6v_2 \\ 6v_3 = 6v_3 \end{cases}$$

$$v_1 = k, v_2 = k, v_3 = k \qquad (k \neq 0)$$

$$(\text{e.g. } v_1 = 1, v_2 = 1, v_3 = 1)$$

So, we have $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ as the first principal axis, the first principal component is given by

$$U^{*T}X = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 0 & 3 & 0 \end{bmatrix}$$

1c)
They should be identical by discovering the new XX^T,

$$XX^T = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

We get the same answer as (a), so we will have the same principal axis, ($\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$). No matter how the columns exchange, it will still have the same principal axis. Therefore, it is proven that the order of input record will not affect the resulting principal axis we want to obtain.

2a)
Let $g(\mu) = (X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)$, so

$$\frac{\partial g(\mu)}{\partial \mu} = -I + UU^T$$

Using the chain rule in L2 norm,

$$\begin{aligned} \frac{\partial f(\mu)}{\partial \mu} &= \sum_{i=1}^{m} 2(\frac{\partial g(\mu)}{\partial \mu})^T g(\mu) \\ &= \sum_{i=1}^{m} 2(-I + UU^T)^T((X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)) \\ &= \sum_{i=1}^{m} 2(-I + UU^T)((X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)) \end{aligned}$$

2b)

$$\begin{aligned} \frac{\partial^2 f(\mu)}{\partial \mu^2} &= \sum_{i=1}^{m} 2(-I + UU^T)(-I + UU^T) \\ &= \sum_{i=1}^{m} 2(I - UU^T - UU^T + UU^TUU^T) \\ &= \sum_{i=1}^{m} 2(I - UU^T) \\ &= 2m(I - UU^T) \end{aligned}$$

To prove it is semi positive-definite, we have to prove $x^T(2m(I-UU^T)x \geq 0$, simplify it, $x^T(I-UU^T)x \geq 0$. (Note that $m > 0$, and UU^T must be symmetric because $(UU^T)^T = UU^T$)
Let $x = Vc$, where V is an orthogonal matrix which obtained from adding extra orthonormal basics from U. Therefore,

$$\begin{aligned}
x^T(I-UU^T)x &= (Vc)^T(I-UU^T)(Vc) \\
&= c^TV^T(I-UU^T)Vc \\
&= c^TV^TVc - c^TV^TUU^TVc \\
&= c^Tc - c^T \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} c \\
&= c^Tc - c^T \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} c \\
&= c_1^2 + c_2^2 + ... + c_d^2 + ... + c_D^2 - (c_1^2 + c_2^2 + ... + c_d^2) \\
&= c_{d+1}^2 + c_{d+2}^2 + ... + c_D^2 \\
&\geq 0
\end{aligned}$$

Therefore, $\dfrac{\partial^2 f(\mu)}{\partial \mu^2}$ is semi positive definite.

2c)

From (b), we can claim that $f(\mu)$ yields global minimum when $\dfrac{\partial f(\mu)}{\partial \mu} = 0$. Therefore, by using result of (a),

$$\sum_{i=1}^{m} 2(-I + UU^T)((X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)) = 0$$

$$2m(-I + UU^T)\sum_{i=1}^{m}((X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)) = 0$$

$$\sum_{i=1}^{m}((X^{(i)} - \mu) - UU^T(X^{(i)} - \mu)) = 0$$

$$\sum_{i=1}^{m}(I - UU^T)(X^{(i)} - \mu) = 0$$

$$m(I - UU^T)\sum_{i=1}^{m}(X^{(i)} - \mu) = 0$$

$$\sum_{i=1}^{m}(X^{(i)} - \mu) = 0$$

$$m\mu = \sum_{i=1}^{m} X^{(i)}$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} X^{(i)}$$

Therefore, $\mu^* = \frac{1}{m}\sum_{i=1}^{m} X^{(i)}$ yields global minimum of $f(\mu)$.

3a)

$$
\begin{aligned}
a_1 &= x \\
a_2 &= ReLU(w_{1,2}a_1) \\
a_3 &= ReLU(w_{1,3}a_1) \\
a_4 &= ReLU(w_{1,4}a_1) \\
\hat{y} = a_5 &= \frac{1}{1+e^{-w_{2,5}a_2 - w_{3,5}a_3 - w_{4,5}a_4}} \\
&= \frac{1}{1+e^{-w_{2,5}ReLU(w_{1,2}a_1) - w_{3,5}ReLU(w_{1,3}a_1) - w_{4,5}ReLU(w_{1,4}a_1)}} \\
&= \frac{1}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}
\end{aligned}
$$

3b)

The specific loss function based on the formula in (a) is (I keep the $(1-y)$ bracket because it will be useful for binary case ($y = 0$ *or* $y = 1$) this time.

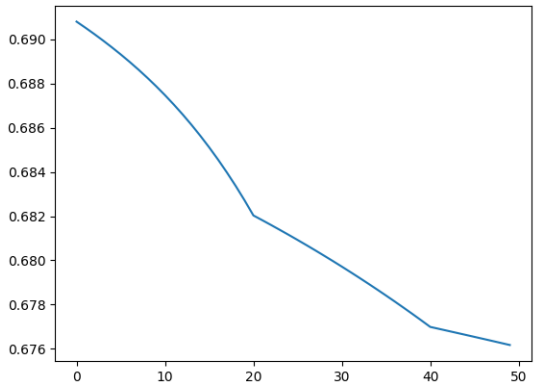
$$
\begin{aligned}
\ell_{CE} &= -y_i ln(\hat{y_i}) - (1-y_i) ln(1-\hat{y_i}) \\
&= -y_i ln(\frac{1}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}) \\
&- (1-y_i) ln(1 - \frac{1}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}) \\
&= y_i ln(1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}) \\
&- (1-y_i)(-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x) \\
&- ln(1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}))
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell_{CE}}{\partial w_{1,2}} &= y\frac{-w_{2,5}xe^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}} \\
&- (1-y)(-w_{2,5}x - \frac{-w_{2,5}xe^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}) \\
\frac{\partial \ell_{CE}}{\partial w_{2,5}} &= y\frac{-ReLU(w_{1,2}x)e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}} \\
&- (1-y)(-ReLU(w_{1,2}x) - \frac{-ReLU(w_{1,2}x)e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}}{1+e^{-w_{2,5}ReLU(w_{1,2}x) - w_{3,5}ReLU(w_{1,3}x) - w_{4,5}ReLU(w_{1,4}x)}})
\end{aligned}
$$

3c)

Final optimized weights for $W_1 = \begin{bmatrix} 0.2656 \\ 0.4745 \\ -0.1379 \end{bmatrix}, W_2 = \begin{bmatrix} 0.2613 \\ 0.4189 \\ -0.0420 \end{bmatrix}$

Testing Accuracy: 100%
The figure is shown at right hand side

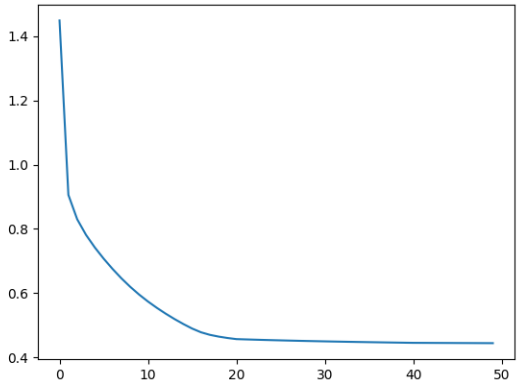


4a)

$$W_1 = \begin{bmatrix} -0.1698 & 0.2748 & 0.2536 \\ 0.0057 & 0.0080 & 0.0358 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 1.0186 & -0.2970 & 0.9568 \\ -0.0795 & 1.0447 & 1.0025 \\ -0.4609 & 0.4832 & 0.3152 \end{bmatrix}$$

Test Accuracy: 93.6 %



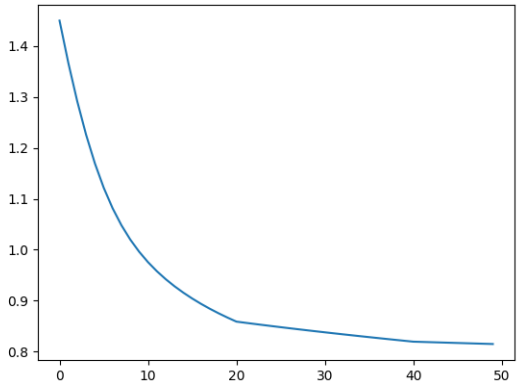
4b)

For learning rate = 0.01,

$$W_1 = \begin{bmatrix} 0.8627 & -0.0390 & 0.8312 \\ -2.0662 & -1.3738 & -0.3055 \end{bmatrix}$$

$$, W_2 = \begin{bmatrix} 1.4603 & -0.8739 & -0.7407 \\ -0.1039 & 0.9317 & 0.6105 \\ -0.3664 & 0.5522 & 0.7302 \end{bmatrix}$$

Testing Accuracy: 54.9 %

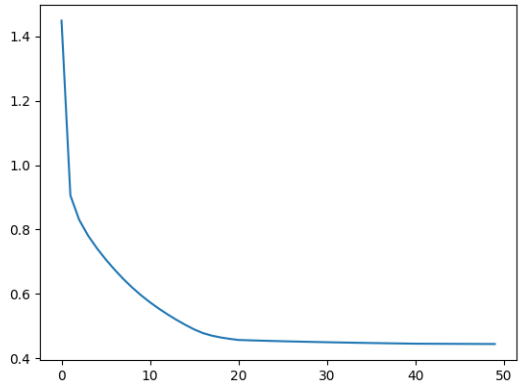


For learning rate = 0.1,

$$W_1 = \begin{bmatrix} -0.1698 & 0.2748 & 0.2536 \\ 0.0057 & 0.0080 & 0.0358 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 1.0186 & -0.2970 & 0.9568 \\ -0.0795 & 1.0447 & 1.0025 \\ -0.4609 & 0.4832 & 0.3152 \end{bmatrix}$$

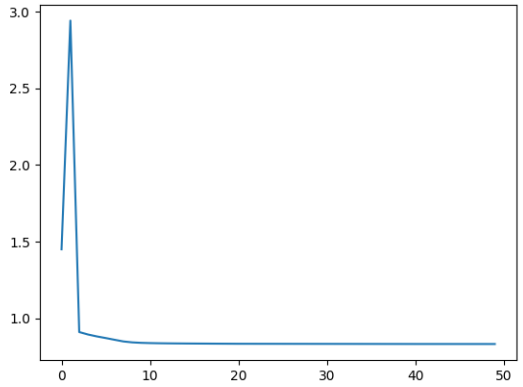
Testing Accuracy: 93.6 %



For learning rate = 1,

$$W_1 = \begin{bmatrix} 0.2537 & -0.4833 & -0.3027 \\ -3.1115 & -1.5151 & -0.7082 \end{bmatrix}, W_2 = \begin{bmatrix} 1.5082 & -0.6465 & -0.1343 \\ 0.4220 & 0.9842 & 1.1584 \\ -0.9402 & 0.2722 & -0.4241 \end{bmatrix}$$

Testing Accuracy: 50.25 %



From these 3 results, we can find out that 0.1 learning rate gives the best testing accuracy among 3 different learning rate values. Therefore, either too high or too low learning rate will not increase the testing accuracy, but an appropriate value will. (Too high will lead to loss explode and too low will make the weights update too slow!)

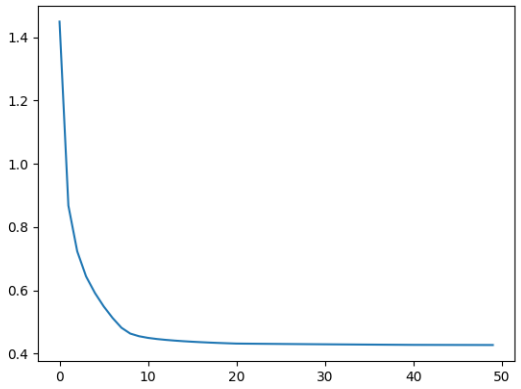
4c)

In the following test, I will run the following 100 times to get a optimal learning rate for the question:

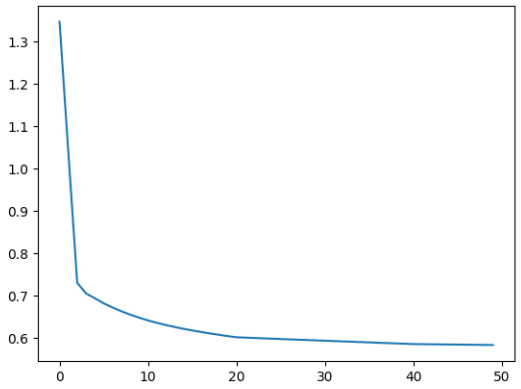
Each time I will set W_1, W_2 into some random value, and then I will find the optimal learning rate range from (0.05 to 0.5) that gives the highest test accuracy for that random case.

Finally, I get a mean value for these 100 cases: 0.21845
Then I will put it into the test set,

Use the weight given in (a), testing accuracy: 93.55% (See figure below)



A random weight, testing accuracy: 72.95% (See figure below)



(10 random weights, average testing accuracy: 66.86%)