CSCI3230 / ESTR3108 2023-24 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:Fong Long Wai

Student ID :1155177220

Q3C.

By using batch size = 256, learning rate = 0.01, training epochs = 50, and seed is not set, test accuracy of 98.8% is achieved.

It is observed that the training loss can differs with the seed, and it can converge to different value in the training.

The following is the final weight.

W1:

[-1.7558],
[ 1.6814],
[ 0.2296],
[-1.7671]

W2:

[ 0.9549, 0.9903, -0.1146, -0.4299]

```
import torch
from torch import nn
from torch.utils.data import DataLoader
import numpy as np
from torch.utils.data import Dataset
import pandas as pd
from tqdm import tqdm
from matplotlib import pyplot as plt
```

In [ ]:

```
batch_size = 256
learning_rate = 0.01
epoches = 50
```

In [ ]:

```
class Data(Dataset):
    def __init__(self, csv_file, transform=None):

        super(Data, self).__init__()
        file = pd.read_csv(csv_file)
        self.input_data = file[['x']].values
        self.labels = file['y'].values.astype(int)

    def __getitem__(self, index):
        data_item = self.input_data[index]
        data_item = torch.tensor(data_item).float()

        label = self.labels[index]
        return data_item, label

    def __len__(self):
        return len(self.input_data)
```

In [ ]:

```
train_data = Data('train_q3.csv')
test_data = Data('test_q3.csv')
train_dataloader = DataLoader(train_data, batch_size=batch_size)
test_dataloader = DataLoader(test_data, batch_size=batch_size)
```

In [ ]:

```
class NeuralNetwork(nn.Module):
    def __init__(self):
        super(NeuralNetwork, self).__init__()
        self.linear_relu_stack = nn.Sequential(
            nn.Linear(1,4),
            nn.ReLU(),
            nn.Sigmoid(),
        )
        self.initialize_weights()


    def forward(self, x):
        logits = self.linear_relu_stack(x)
        return logits

    def initialize_weights(self):
        for m in self.modules():
            if isinstance(m, nn.Linear):
                nn.init.kaiming_uniform_(m.weight)
                if m.bias is not None:
                    nn.init.zeros_(m.bias)
```

In [ ]:

```
device = "cuda" if torch.cuda.is_available() else "cpu"
model = NeuralNetwork().to(device)
```

In [ ]:

```
loss_fn = nn.CrossEntropyLoss()
optimizer = torch.optim.SGD(model.parameters(), lr=learning_rate)
```

In [ ]:

```
model.train()
epoches_time=0
for epoch in tqdm(range(epoches)):
    epoches_time=epoches_time+1
    train_loss, correct = 0, 0
    for X, y in train_dataloader:
        X, y = X.to(device), y.to(device)

        pred = model(X)
```

```
            loss = loss_fn(pred, y.long())

            # Backpropagation
            optimizer.zero_grad()
            loss.backward()
            optimizer.step()

            # record loss
            train_loss += loss.item()
            correct += (pred.argmax(1) == y).type(torch.float).sum().item()
        size = len(train_dataloader.dataset)
        train_loss /= len(train_dataloader)
        correct /= size
        print(f" Train accuracy: {(100*correct):>0.1f}%, Avg loss: {train_loss:>8f}")
        plt.plot(epoches_time,train_loss,"ro")
plt.xlabel("Epoch")
plt.ylabel("Average Loss")
plt.show()
  2%|▏                 | 1/50 [00:00<00:15,  3.20it/s]
 Train accuracy: 54.7%, Avg loss: 1.340353
  4%|▍                 | 2/50 [00:00<00:12,  3.82it/s]
 Train accuracy: 63.4%, Avg loss: 1.333190
 Train accuracy: 72.1%, Avg loss: 1.326437
  8%|▊                 | 4/50 [00:00<00:10,  4.34it/s]
 Train accuracy: 81.0%, Avg loss: 1.320077
 Train accuracy: 90.1%, Avg loss: 1.314076
 12%|█▏                | 6/50 [00:01<00:10,  4.02it/s]
 Train accuracy: 98.1%, Avg loss: 1.308418
 14%|█▍                | 7/50 [00:01<00:10,  4.14it/s]
 Train accuracy: 99.7%, Avg loss: 1.303068
 16%|█▌                | 8/50 [00:01<00:09,  4.20it/s]
 Train accuracy: 99.6%, Avg loss: 1.297999
 18%|█▊                | 9/50 [00:02<00:10,  3.94it/s]
 Train accuracy: 99.6%, Avg loss: 1.293199
 20%|██                | 10/50 [00:02<00:10,  3.64it/s]
 Train accuracy: 99.6%, Avg loss: 1.288656
 22%|██▏               | 11/50 [00:02<00:11,  3.50it/s]
 Train accuracy: 99.6%, Avg loss: 1.284362
```

24%|■■▌ | 12/50 [00:03<00:10, 3.76it/s]
Train accuracy: 99.5%, Avg loss: 1.280300
Train accuracy: 99.5%, Avg loss: 1.276446
28%|■■▉ | 14/50 [00:03<00:09, 3.79it/s]
Train accuracy: 99.5%, Avg loss: 1.272794
30%|■■■ | 15/50 [00:03<00:09, 3.76it/s]
Train accuracy: 99.5%, Avg loss: 1.269322
32%|■■■▏ | 16/50 [00:04<00:09, 3.70it/s]
Train accuracy: 99.5%, Avg loss: 1.266016
34%|■■■▌ | 17/50 [00:04<00:08, 3.71it/s]
Train accuracy: 99.5%, Avg loss: 1.262867
36%|■■■▉ | 18/50 [00:04<00:08, 3.74it/s]
Train accuracy: 99.4%, Avg loss: 1.259869
Train accuracy: 99.4%, Avg loss: 1.257009
40%|■■■■ | 20/50 [00:05<00:07, 4.26it/s]
Train accuracy: 99.4%, Avg loss: 1.254279
Train accuracy: 99.4%, Avg loss: 1.251668
44%|■■■■▌ | 22/50 [00:05<00:06, 4.54it/s]
Train accuracy: 99.4%, Avg loss: 1.249171
46%|■■■■▌ | 23/50 [00:05<00:05, 4.50it/s]
Train accuracy: 99.4%, Avg loss: 1.246782
48%|■■■■▉ | 24/50 [00:06<00:06, 3.96it/s]
Train accuracy: 99.4%, Avg loss: 1.244496
Train accuracy: 99.4%, Avg loss: 1.242306
52%|■■■■■▏ | 26/50 [00:06<00:05, 4.35it/s]
Train accuracy: 99.4%, Avg loss: 1.240206
54%|■■■■■▌ | 27/50 [00:06<00:05, 4.44it/s]
Train accuracy: 99.4%, Avg loss: 1.238191
56%|■■■■■▌ | 28/50 [00:06<00:05, 4.16it/s]
Train accuracy: 99.4%, Avg loss: 1.236254
58%|■■■■■▉ | 29/50 [00:07<00:04, 4.27it/s]
Train accuracy: 99.4%, Avg loss: 1.234388
Train accuracy: 99.4%, Avg loss: 1.232591
62%|■■■■■▏ | 31/50 [00:07<00:04, 4.52it/s]
Train accuracy: 99.4%, Avg loss: 1.230861
Train accuracy: 99.4%, Avg loss: 1.229193
68%|■■■■■■▉ | 34/50 [00:08<00:03, 4.98it/s]
Train accuracy: 99.4%, Avg loss: 1.227584

Train accuracy: 99.4%, Avg loss: 1.226030

70%|■■■■■■■ | 35/50 [00:08<00:03, 4.92it/s]

Train accuracy: 99.4%, Avg loss: 1.224523

Train accuracy: 99.3%, Avg loss: 1.223063

74%|■■■■■■■▌ | 37/50 [00:08<00:02, 5.00it/s]

Train accuracy: 99.3%, Avg loss: 1.221647

Train accuracy: 99.3%, Avg loss: 1.220273

78%|■■■■■■■▉ | 39/50 [00:09<00:02, 5.17it/s]

Train accuracy: 99.3%, Avg loss: 1.218941

Train accuracy: 99.2%, Avg loss: 1.217646

84%|■■■■■■■■▌ | 42/50 [00:09<00:01, 4.84it/s]

Train accuracy: 99.2%, Avg loss: 1.216385

Train accuracy: 99.2%, Avg loss: 1.215156

86%|■■■■■■■■▌ | 43/50 [00:10<00:01, 4.68it/s]

Train accuracy: 99.1%, Avg loss: 1.213958

Train accuracy: 99.1%, Avg loss: 1.212790

90%|■■■■■■■■■ | 45/50 [00:10<00:01, 4.69it/s]

Train accuracy: 99.1%, Avg loss: 1.211650

Train accuracy: 99.0%, Avg loss: 1.210538

94%|■■■■■■■■■▌ | 47/50 [00:10<00:00, 4.90it/s]

Train accuracy: 99.0%, Avg loss: 1.209453

96%|■■■■■■■■■▌ | 48/50 [00:11<00:00, 4.75it/s]

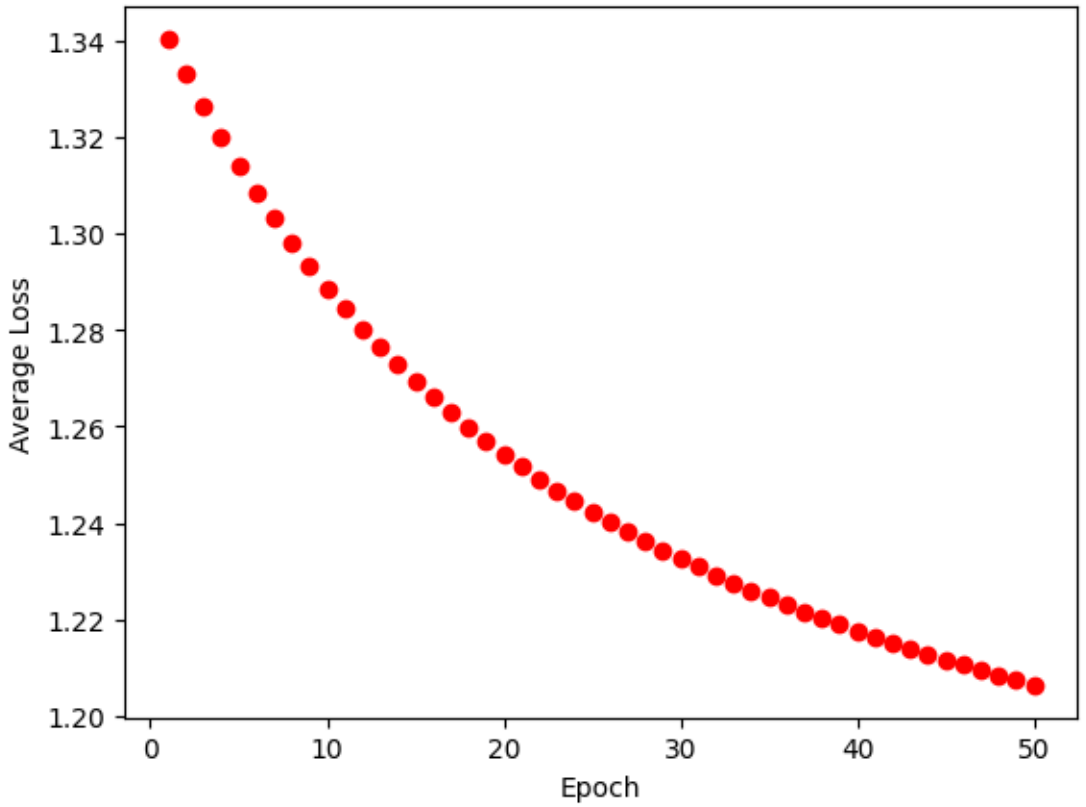
Train accuracy: 99.0%, Avg loss: 1.208394

98%|■■■■■■■■■▉ | 49/50 [00:11<00:00, 4.69it/s]

Train accuracy: 99.0%, Avg loss: 1.207363

100%|■■■■■■■■■■| 50/50 [00:11<00:00, 4.35it/s]

Train accuracy: 99.0%, Avg loss: 1.206356




In [ ]:

```
model.eval()
correct = 0
# Turn off gradient descent
with torch.no_grad():
    for X, y in tqdm(test_dataloader):
        X, y = X.to(device), y.to(device)
        pred = model(X)
        correct += (pred.argmax(1) == y).type(torch.float).sum().item()
size = len(test_dataloader.dataset)
correct = correct / size
print(f" Test accuracy: {(100*correct):>0.1f}%")
```

```
100%|■■■■■■■■■■| 8/8 [00:00<00:00, 159.98it/s]
 Test accuracy: 98.8%
```

In [ ]:

```
for param in model.parameters():
    print(param)
```

```
Parameter containing:
tensor([[-1.7558],
        [ 1.6814],
```

[ 0.2296],

[-1.7671]], requires_grad=True)

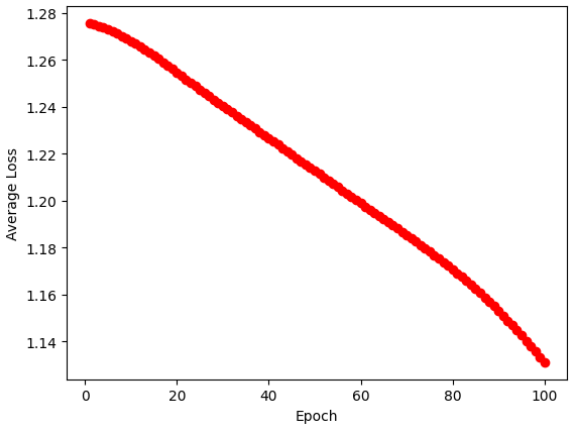
Parameter containing:

tensor([ 0.9549, 0.9903, -0.1146, -0.4299], requires_grad=True)

Q4a.

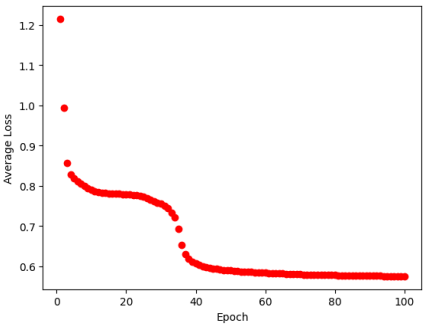
Learning rate=0.001

Accuracy=44.8



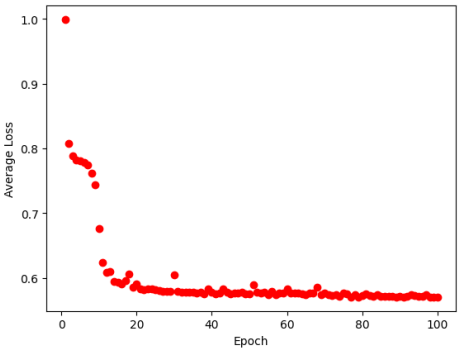

Learning rate= 0.1

Accuracy=82.9




Learning rate=0.4

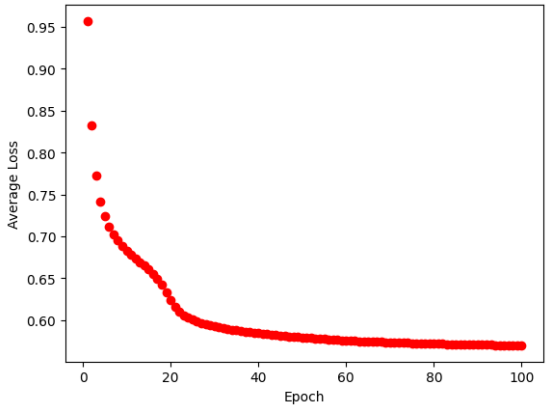
Accuracy=92.2



Q4b.

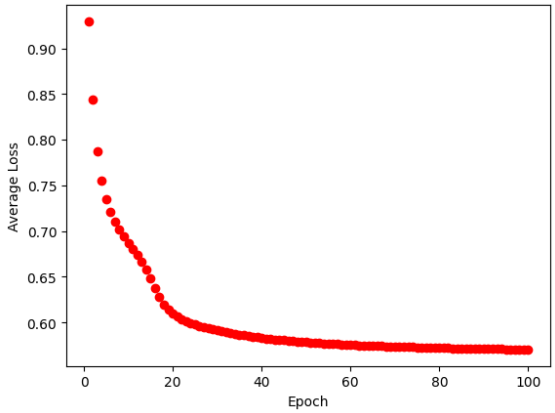
kaiming_normal

Testing accuracy=95.0%



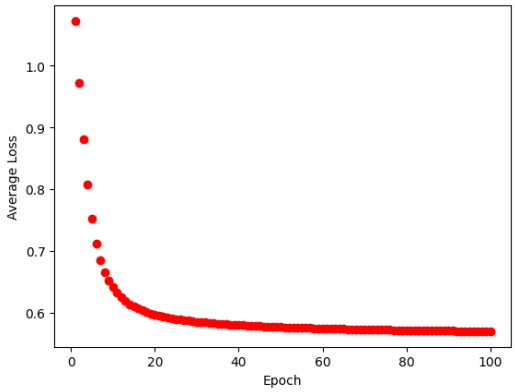

xavier_normal

Testing accuracy=94.5%




xavier_normal, but set the input tensor as m.weight*1.1

Testing accuracy=97%



```
Q4 Code
import torch
from torch import nn
from torch.utils.data import DataLoader
import numpy as np
from torch.utils.data import Dataset
import pandas as pd
from tqdm import tqdm
import torch.backends.cudnn as cudnn
import random
import numpy as np
from matplotlib import pyplot as plt
seed = 1443
cudnn.benchmark = False
cudnn.deterministic = True
random.seed(seed)
np.random.seed(seed)
torch.manual_seed(seed)
torch.cuda.manual_seed(seed)


# %%
batch_size = 256
learning_rate = 0.1
epoches = 100



# %%
class Data(Dataset):
    def __init__(self, csv_file, transform=None):

        super(Data, self).__init__()
        file = pd.read_csv(csv_file)
        self.input_data = file[['x1', 'x2']].values
        self.labels = file['y'].values.astype(int)
```

```
    def __getitem__(self, index):
        data_item = self.input_data[index]
        data_item = torch.tensor(data_item).float()

        label = self.labels[index]
        return data_item, label

    def __len__(self):
        return len(self.input_data)

# %%
train_data = Data('train_q4.csv')
test_data = Data('test_q4.csv')
train_dataloader = DataLoader(train_data, batch_size=batch_size)
test_dataloader = DataLoader(test_data, batch_size=batch_size)

# %%
class NeuralNetwork(nn.Module):
    def __init__(self):
        super(NeuralNetwork, self).__init__()
        self.linear_relu_stack = nn.Sequential(
           nn.Linear(2,4),
           nn.ReLU(),
           nn.Linear(4,4),
           nn.ReLU(),
           nn.Linear(4,3),
           nn.Softmax(),
        )
        self.initialize_weights()


    def forward(self, x):
        logits = self.linear_relu_stack(x)
        return logits

    def initialize_weights(self):
        for m in self.modules():
            if isinstance(m, nn.Linear):
```

```
                nn.init.kaiming_normal_(m.weight)
                if m.bias is not None:
                    nn.init.zeros_(m.bias)

# %%
device = "cuda" if torch.cuda.is_available() else "cpu"
model = NeuralNetwork().to(device)

# %%
loss_fn = nn.CrossEntropyLoss()
optimizer = torch.optim.SGD(model.parameters(), lr=learning_rate)

# %%
model.train()
epoches_time=0
for epoch in tqdm(range(epoches)):
    epoches_time=epoches_time+1
    train_loss, correct = 0, 0
    for X, y in train_dataloader:
        X, y = X.to(device), y.to(device)

        pred = model(X)

        loss = loss_fn(pred, y.long())

        # Backpropagation
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

        # record loss
        train_loss += loss.item()
        correct += (pred.argmax(1) == y).type(torch.float).sum().item()
    size = len(train_dataloader.dataset)
    train_loss /= len(train_dataloader)
    correct /= size
    print(f" Train accuracy: {(100*correct):>0.1f}%, Avg loss:
{train_loss:>8f}")
```

```
    plt.plot(epoches_time,train_loss,"ro")
plt.xlabel("Epoch")
plt.ylabel("Average Loss")
plt.show()


# %%
model.eval()
correct = 0
# Turn off gradient descent
with torch.no_grad():
    for X, y in tqdm(test_dataloader):
        X, y = X.to(device), y.to(device)
        pred = model(X)
        correct += (pred.argmax(1) == y).type(torch.float).sum().item()
size = len(test_dataloader.dataset)
correct = correct / size
print(f" Test accuracy: {(100*correct):>0.1f}%")
```

No.

Date.

1a. $Y = (XX^T)$

$$= \begin{pmatrix} 1+(-1)(-1) & 0 & 1\times(-1)+(-1)\times 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1\times 1+1\times(-1) & 0 & -1\times(-1)+1\times 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$\text{Det}(Y - \lambda I) = 0$

$$\begin{vmatrix} 2-\lambda & 0 & -2 & 0 \\ 0 & -\lambda & 0 & 0 \\ -2 & 0 & 2-\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{vmatrix} = 0$$

$$(2-\lambda)(-\lambda)(2-\lambda)(-\lambda) - 2(-(-\lambda)\times(-2)(-\lambda)) = 0$$

$$(4\lambda^2 - 4\lambda^3 + \lambda^4) - 2(2\lambda^2) = 0$$

$$\lambda^4 - 4\lambda^3 = 0$$

$$\lambda = 0 \quad \text{or} \quad \lambda = 4$$

$\therefore \lambda_1 = 0, \lambda_2 = 4$

1b. ~~largest~~ eigenvector with largest eigenvalue of X.

For $\lambda = 4$,

$$\begin{pmatrix} 2-4 & 0 & -2 & 0 \\ 0 & -4 & 0 & 0 \\ -2 & 0 & 2-4 & 0 \\ 0 & 0 & 0 & -4 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

~~First principle axis~~ = eigenvector with largest eigenvalue of X $= (1, 0, -1, 0)^T = e$

First principle axis $= \frac{1}{\|e\|} e = \left(\frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2}, 0\right)^T$

First component $= e^T X = \left(\frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2}, 0\right) \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

$= (\sqrt{2} \quad 0 \quad -\sqrt{2})$

The result is a 1x3 matrix. Each of the 3 samples are compressed to 1 dimension from 4 dimension

[c. $X'X'^T = \begin{pmatrix} 2 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

Let $A$ be a $m \times n$ data matrix.

$A'$ is generated by swapping column of $A$.

$A' = AP$, $P$ = a permutation combination of $n \times 1$ vector $v$ $a_i$,

~~which has all 3~~

$ap_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$ $ap_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ ...

$P$ = a $n \times n$ matrix.

$\therefore A'A'^T v = APP^T A^T v$

$P^T = P^{-1}$,

$\therefore APP^T A^T v = AA^T v$.

$\therefore$ eigenvector is unchanged after swapping column.

Proof that $P^T = P^{-1}$ :

$P = (p_1 \dots p_n)$

$P^T = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$

$P^T P = \begin{pmatrix} p_1 \cdot p_1 & p_1 \cdot p_2 & \dots & p_1 \cdot p_n \\ \vdots & \ddots & & \\ p_n \cdot p_1 & & & p_n \cdot p_n \end{pmatrix}$

$p_i \cdot p_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

$\therefore P^T P = I_n$

$P^T P P^{-1} = I_n P^{-1}$

$P^T = P^{-1}$

No.

Date.

2a. When $u_i^T u_i = 1$

$$\left(\sum_{j=1}^{D} (c_j v_j^T)\right)\left(\sum_{k=1}^{D} (c_j v_j)\right) = 1$$

Let $v_{ij}$ be $j^{th}$ entries of $i^{th}$ vector.

$$\sum_{i=1}^{D}\left[\left(\sum_{j=1}^{D} c_j v_{ji}\right)^2\right] = 1$$

$$\sum_{i=1}^{D} c_i^2 \|v_i\|^2 + \sum_{k=1}^{D}\sum_{k\neq j} c_k c_j v_k^T v_j = 1$$

Since $v_i$ are orthonormal vectors,

$$v_k^T v_j = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$$

$$\sum_{i=1}^{D} c_i^2 = 1 \;//$$

2b. $E(XX^T) = \sum_{k=1}^{D} \lambda_k v_k v_k^T$, $u \in \mathbb{R}^D$, $u^T u = 1$

$u_1^T E(XX^T) u_1$

$= u_1^T \left( \sum_{k=1}^{D} \lambda_k v_k v_k^T \right) u_1$

~~$\leq \lambda_{max} u_1^T \|V\| u_1$~~

$\leq \lambda_1 u_1^T \left( \sum_{k=1}^{D} v_k v_k^T \right) u_1$

$= \lambda_1 u_1^T u_1 = \lambda_1$.

$c_1 = 1, c_{i \neq 1} = 0$

When $u_1 = v_1$, $u_1^T E(XX^T) u_1$

$= u_1^T \left( \sum_{k=1}^{D} \lambda_k v_k v_k^T \right) u_1$

~~$= u_1^T \sum_{k=1}^{D} \lambda_k u_k$~~

$= u_1^T \left[ (\lambda_1 \cdot 0) v_k v_k^T + \sum_{k=2}^{D} \lambda_k v_k v_k^T \right] u_1$

$= u_1^T \left[ \lambda_1 + \sum_{k=2}^{D} \lambda_k \|v_k\|^2 \right] u_1$

$\geq \lambda_1$, since $\lambda_i$ are non-negative.

when $u_1 = v_1$, $u_1^T E(XX^T) u_1 \geq \lambda_1$

$\leq \lambda_1$

$\therefore u_1^T E(XX^T) u_1 = \lambda_1$ //

No.

Date.

2b $\lambda_K \geq 0$ 's proof:

$$X^T X v = \lambda v$$

$$v^T X^T X = \lambda v^T$$

$$v^T X^T X v = \lambda v^T v$$

$$(Xv)^T Xv = \lambda v^T v$$

$$\|Xv\|^2 = \lambda \|v\|^2$$

$$\lambda = \frac{\|Xv\|^2}{\|v\|} \geq 0 .$$

3a. $\hat{y} = \text{Sigmoid}\left(\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right)$,

$a_i = \text{ReLU}(W_{1,i}, a_1)$ for $i = 2,3$

b. Loss function

$l_{CE} = -y \times \log\left(\text{Sigmoid}\left(\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right)\right)$

Gradient for $W_{1,3}$

$= \frac{\partial l_{CE}}{\partial W_{3,6}} \times \frac{\partial W_{3,6}}{\partial W_{1,3}} = \frac{\partial l_{CE}}{\partial W_{1,3}}$

$= \frac{\partial l_{CE}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_{3,6}} \frac{\partial W_{3,6}}{\partial W_{1,3}}$

$\frac{\partial l_{CE}}{\partial \hat{y}} = -y \times \log\left(\text{sigmoid}\left(\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right)\right)$

$\times \left[\frac{1}{\text{Sigmoid}\left(\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right)}\right]$

$$\frac{\partial \hat{y}}{\partial W_{3,6}} = \frac{\partial \text{Sigmoid}(a_3 W_{3,6} + C)}{\partial W_{3,6}}$$

$$= a_3 \times \text{Sigmoid}(a_3 W_{3,6} + C) \hat{y} \times (1 - \hat{y})$$

$$\frac{\partial W_{3,6}}{\partial W_{1,3}} = \frac{\partial (a_3 W_{3,6} + C)}{\partial W_{1,3}}$$

$$= \frac{\partial (W_{3,6} \times \text{ReLU}(W_{1,3} a_1))}{\partial W_{1,3}}$$

$$= \begin{cases} W_{3,6} \times a_1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$\therefore \frac{\partial L_{CE}}{\partial W_{1,3}} = -y \times \log\left(\text{Sigmoid}\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right) \times \left(1 - \log\left(\text{Sigmoid}\begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} \cdot W_2\right)\right) \times W_{3,6} \times a_1 \quad \text{for } x \geq 0$$

$$= 0 \quad \text{for } x < 0$$