Assignment 1
Zheng Weixiang
G2103278G

Question 1:
Number of parameters:

Conv2d-1:	(5*5*1+1) * 6 = 156
ReLu-2:	0
MaxPool2d-3:	0
Conv2d-4:	(5*5*6+1) *16 = 2416
ReLu-5:	0
MaxPool2d-6:	0
Conv2d-7:	(5*5*16+1) * 120 = 48120
ReLu-8:	0
Linear-9:	120*84+84 = 10164
ReLu-10:	0
Linear-11:	84*10+10 = 850
LogSoftmax-12:	0

Total: 156 + 2416 + 48120 + 10164 + 850 = 61706

Question 2:

```
        -1 0 1     10 10 0 0
w * x = -2 0 2  *  10 10 0 0
        -1 0 1     10 10 0 0
                   10 10 0 0
```

```
        -1 0 1 0      -2 0  2 0      -1  0 1 0      0  0 0 0
W =     0 -1 0 1      0 -2  0 2       0 -1 0 1      0  0 0 0
        0  0 0 0     -1 0 1  0      -2  0 2 0     -1 0 1 0
        0 0 0  0      0 -1 0  1       0 -2 0 2      0 -1 0 1
```

V(x) = (10 10 0 0 10 10 0 0 10 10 0 0 10 10 0 0)(transposed)

W*V(x) = (-1)*10+0+0+0 + (-2)*10+0+0+0 + (-1)*10+0+0+0 + 0+0+0+0 = -40
0+(-1)*10+0+0 + 0+(-2)*10+0+0 + 0+(-1)*10+0+0 + 0+0+0+0 = -40
0+0+0+0 + (-1)*10+0+0+0 + (-2)*10 + 0+0+0 + (-1)*10+0+0 = -40
0+0+0+0 + 0 + (-1)*10+0+0+ 0 + (-2)*10+0+0 + 0+(-1)*10+0+0 = -40

W*V(x) = (-40 -40 -40 -40)(transposed)
Can be reshaped to -40 -40
-40 -40

Question 3:

i) MSE might not be a good choice for the loss function because MSE is a non-convex function for binary classification. Therefore, the using MSE as loss function might not direct the trained label towards the underlying true label.

ii) $Loss = -(y * \log(y_{hat}) + (1 - y) * \log(1 - y_{hat}))$

iii) $Loss(Y, Y_{hat}) = [0.69897, 0.30102, 0.04575]$ (log base 10)
J = 1.046

$Loss(Y, Y_{hat}) = [1.60943, 0.69314, 0.10536]$ (log base e)
J = 2.408

$Loss(Y, Y_{hat}) = [2.32192, 1, 0.15200]$ (log base 2)
J = 3.474

iv) The model that uses L2 Regularization adds squared magnitude of coefficient as penalty term to the loss function. Therefore, while minimizing the loss, we are also minimizing the weights of the model to prevent model from becoming too complex. Then the weights of model A will be smaller in general than model B.

Question 4:

1. Vulnerable to outliers. Some outliers will have extreme squared differences which will overwhelm other valid square sums.
2. The step size might get very small when getting closer enough to the target value, therefore the model with L2 loss require more iterations to reach the target.

Question 5:

1. Suffer from outliers. The normalized value might be affected more than models with larger batch size, and this bias will be counted towards the final result. Therefore, the accuracy will be worse than the model with larger batch size while others things being equal.
2. Suffer from sample bias. You are overfitting the mini-batch and not the actual distribution of the dataset. This cause reduced accuracy.