Yin Junyu G2101985L

Question 1

Layer (type)	Output Shape	Number of Parameters
Conv2d-1	[-1, 6, 28, 28]	(1x5x5+1)x6 = 156
ReLU-2	[-1, 6, 28, 28]	0
MaxPool2d-3	[-1, 6, 14, 14]	0
Conv2d-4	[-1, 16, 10, 10]	(6x5x5+1)x16 = 2416
ReLU-5	[-1, 16, 10, 10]	0
MaxPool2d-6	[-1, 16, 5, 5]	0
Conv2d-7	[-1, 120, 1, 1]	(16x5x5+1)x120 = 48120
ReLU-8	[-1, 120, 1, 1]	0
Linear-9	[-1, 84]	120x84+84 = 10164
ReLU-10	[-1, 84]	0
Linear-11	[-1, 10]	84x10+10 = 850
LogSoftmax-12	[-1, 10]	0
Total Number of Parameters	61706	

Question 2

Rerange the kernel as a sparse Toeplitz circulant matrix:

$$W = \begin{pmatrix} -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 \end{pmatrix}$$

Flatten the input row by row, from top to bottom:

$$\boldsymbol{x} = (10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0)^{\mathrm{T}}$$

Then we have:

$$W\boldsymbol{x} = (-40 \quad -40 \quad -40 \quad -40)^{\mathrm{T}}$$

Reshape it to a 2 by 2 matrix:

$$\boldsymbol{w} \star \boldsymbol{x} = \begin{pmatrix} -40 & -40 \\ -40 & -40 \end{pmatrix}$$

Question 3

i)

When using the MSE loss, we implicitly admit that the conditional probability distribution of Y given X is a Gaussian distribution. It may be true in some regression tasks, but in this binary classification problem, the random variable Y only has two possible values (0 or 1). So it is unreasonable to use the MSE loss here which means a Gaussian prior.

ii)

$$l(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

iii)

$$J = -\frac{1}{3}(\log 0.2 + \log 0.5 + \log 0.9) = 1.158$$

iv)

The norm of model A's weights will be less than which of model B's.

Question 4

When using L2 loss, the squaring operation amplifies the difference of the data points of which the predicted value differs from the true value by more than 1. And if the difference between the predicted value and the true value is less than 1, the squaring operation will shorten the difference. This means that the L2 loss is more sensitive to outliers. If there are outliers in the given data set, L2 loss will give higher weights to them, which will sacrifice the prediction accuracy of normal data points, and ultimately reduce the overall model performance.

When using L1 loss, the absolute value operation calculates the error by only taking absolute value of the difference between the predicted value and the true value. This means that the penalty is fixed for any size difference.

Therefore, when we are dealing with datasets containing outliers, we prefer to use the L1 loss as the cost function rather than the L2 loss.

Question 5

A very small mini-batch size can result in a large variation in gradients computed per batch. If there are some outliers, they will dominate the resulting gradient. This means that the loss function can decrease very slowly or even goes up in some extreme situations.