ECE 284 HW2 (Jing-Hua Chang A69040678)

Q2:

MSE Loss: $L(w,b) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 = \frac{1}{2}\left[(Y3_1 - T_1)^2 + (Y3_2 - T_2)^2\right]$

$\phi(x) = \begin{cases} x, & x>0 \\ 0, & \text{otherwise} \end{cases}$

$S2 = [1,2]\begin{bmatrix}2&1\\0&1\end{bmatrix} = [2,3]$

$Y2 = \phi([2,3]) = [2,3]$

$S3 = [2,3]\begin{bmatrix}1&2\\1&2\end{bmatrix} = [5,10]$

$Y3 = \phi([5,10]) = [5,10]$

$L = \frac{1}{2}\left[(5-0)^2 + (10-0)^2\right] = \frac{1}{2}(25+100) = 62.5$

```
print(W1.grad)
print(W2.grad)
print(Y3)
```

```
tensor([[25., 25.],
        [50., 50.]])
tensor([[10., 20.],
        [15., 30.]])
tensor([ 5., 10.], grad_fn=<ReluBackward0>)
```

$\frac{\partial L}{\partial S3} = \frac{\partial L}{\partial Y3} \odot ReLU'(S3) = (Y_3 - T) \odot [1,1] = [5,10]$

$S3 = Y2\,W2 \Rightarrow \frac{1}{W2} = Y2\,\frac{1}{S3}\,\frac{\partial L}{\partial L}\frac{\partial}{\partial} \Rightarrow \frac{\partial L}{\partial W_2} = Y2^T \cdot \frac{\partial L}{\partial S3}$

$= \begin{bmatrix}2\\3\end{bmatrix} \cdot [5\ 10] = \begin{bmatrix}10&20\\15&30\end{bmatrix}$

* gradient of W2

$S3 = Y2\,W2 \Rightarrow \frac{\partial L}{\partial S3} = \frac{\partial L}{\partial Y2} W_2 \Rightarrow \frac{\partial L}{\partial Y_2} = \frac{\partial L}{\partial S_3} W_2^T$

$= [5,10]\begin{bmatrix}1&1\\2&2\end{bmatrix} = [25\ 25]$

$\frac{\partial L}{S2} = \frac{\partial L}{\partial Y2} \odot ReLU'(S_2) = [25,25] \odot [1,1] = [25,25]$

$S2 = X\,W1 \Rightarrow \frac{\partial L}{\partial W1} = X^T \frac{\partial L}{\partial S_2} = \begin{bmatrix}1\\2\end{bmatrix} \cdot [25\ 25] = \begin{bmatrix}25&25\\50&50\end{bmatrix}$

* gradient of W1

Q3:

$$loss = \frac{1}{2}\left[(Y3_1 - T_1)^2 + (Y3_2 - T_2)^2\right]$$

$$S2 = X W1 = [1\ 2]\begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix} = [2\ -1]$$

$$Y2 = [2\ 0]$$

$$S3 = [2\ 0]\begin{bmatrix} -1 & 2 \\ 1 & 2 \end{bmatrix} = [-2\ 4]$$

$$Y3 = [0\ 4]$$

$$S3 = Y2\,W2 \Rightarrow \frac{\partial L}{\partial W2} = Y2^T \frac{\partial L}{\partial S3} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \cdot [0\ 4] = \begin{bmatrix} 0 & 8 \\ 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial S3} = \frac{\partial L}{Y3} \odot ReLU'(S3) = (Y3 - T) \cdot [0\ 1] = [0\ 4]$$

$$S2 = X W1 \Rightarrow \frac{\partial L}{\partial W1} = X^T \frac{\partial L}{\partial S_2} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}[8\ 0] = \begin{bmatrix} 8 & 0 \\ 16 & 0 \end{bmatrix}$$

$$S3 = Y2\,W_2 \Rightarrow \frac{\partial L}{\partial Y_2} = \frac{\partial L}{\partial S3} W_2^T = [0\ 4]\begin{bmatrix} -1 & 1 \\ 2 & 2 \end{bmatrix} = [8\ 8]$$

$$\frac{\partial L}{\partial S2} = \frac{\partial L}{\partial Y2} \odot ReLU'(S2) = [8\ 8] \odot [1\ 0] = [8\ 0]$$

Q4:

Parameter

	weights	bias
fc1	$(28\times28)\times512 = 401408$	512
fc2	$512\times512 = 262144$	512
fc3	$512\times10 = 5120$	10

Total parameter: $(401408+262144+5120)+(512+512+10) = 669706$ #

Operation

	MAC (includes input, weight, bias) (input × output)	ReLU
fc1	$(28\times28)\times512 = 401408$	512
fc2	$512\times512 = 262144$	512
fc3	$512\times10 = 5120$	10
total operations	668672×2	$1034 = 1338378$

MAC has 2 operations (add & multiplication)

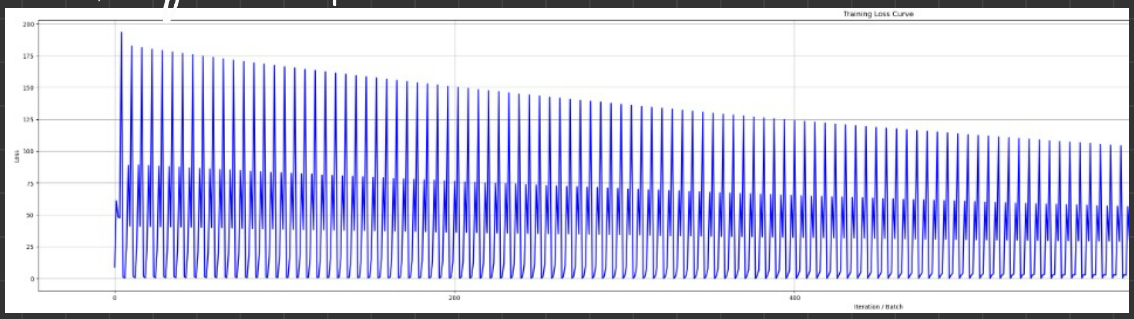
The default bias is set to True

ReLU is counted as 1 operation

There are 1338378 operations/image #

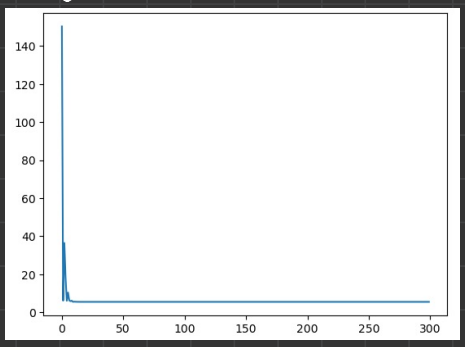
Q6:

SGD, single data point



Weights are updated using only one data at each iteration. So the gradient goes up and down to update different points, making the oscillation sharp and produce noisy gradient. But the overall gradient is still smooth

Batch-based



Loss drops and stablize fast, for it update all data point a once. Allowing it to stablize faster. The gradient looks smooth