Csc413 Homework 2

Name: Kwan Kiu Choy
Student no: 1005005879
UtorId: choykwan

1.1.1

To find the $\hat{w}$, we know that $\sum_{x_j \in B} \nabla_{w_t} L(x_j, w_t) = 0$

$$\nabla_{w_t} L(x_j, w_t) = \frac{2}{n} X_j^T (X_j w_t - t)$$
$$= \frac{2}{n} X_j^T X_j w_t - \frac{2}{n} X_j^T t$$

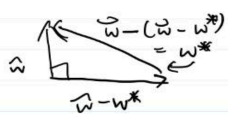
This means that $\nabla_{w_t} L(x_j, w_t)$ is always a member of the span of X. Since we know that $w_0 = 0$, thus the weight update process never exit the span of X. As a result, we can express $\hat{w} = X^T A$.

$$\hat{w} = X^T A \quad \text{since we know that} \quad X\hat{w} = t$$
$$XX^T A = t$$
$$A = (XX^T)^{-1} t$$
$$\hat{w} = X^T (XX^T)^{-1} t.$$

Next Part: we need to prove that $\hat{w}$ is identical to the minimum norm solution w^*.

$$\hat{w}^T (\hat{w} - w^*) = (X^T A)^T (\hat{w} - w^*)$$
$$= AX(\hat{w} - w^*)$$
$$= A(X\hat{w} - Xw^*)$$
$$= A(t - t)$$
$$= 0$$

So, we have shown that $\hat{w}^T$ and $\hat{w} - w^*$ is othogonal



By the properties of norm, we know that $\|\hat{w}\| < \|w^*\|$

Given the assumption that w^* is the minimum solution, the only possible norm value of $\hat{w}$ is $\|w^*\|$.
Therefore, I have shown that the 2 solutions are identical.

1.2.1

Let $\eta = 0.05$, $\beta = 0.9$, $\epsilon = 0.001$.

$$W_1 = 0 - \frac{0.05}{\sqrt{V_{1,0}} + 0.001} \, 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} (0 - [2])$$

$$V_{1,-1} = 0$$
$$V_{1,0} = 0.1 \begin{pmatrix} -8 \\ -4 \end{pmatrix}^2 = \begin{pmatrix} 6.4 \\ 1.6 \end{pmatrix}$$

$$= 0 - 0.05 \begin{bmatrix} \frac{-8}{\sqrt{6.4} + 0.001} \\ \frac{-4}{\sqrt{1.6} + 0.001} \end{bmatrix}$$

$$= \begin{bmatrix} 0.158 \\ 0.158 \end{bmatrix} \longrightarrow \text{this is not in the span of } X.$$

Since the weight update process exit the span of X, it would not always give the minimum norm solution

1.2.1

Let $\eta = 0.05$, $\beta = 0.9$, $\epsilon = 0.001$.

$$W_1 = 0 - \frac{0.05}{V_{1,0} + 0.001}\, 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} (0 - [2])$$

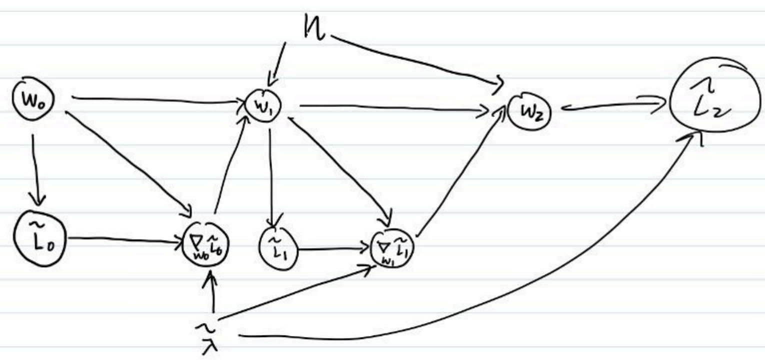
$$V_{1,-1} = 0$$
$$V_{1,0} = 0.1 \begin{pmatrix} -8 \\ -4 \end{pmatrix}^2 = \begin{pmatrix} 6.4 \\ 1.6 \end{pmatrix}$$

$$= 0 - 0.05 \begin{bmatrix} \frac{-8}{\sqrt{6.4} + 0.001} \\ \frac{-4}{\sqrt{1.6} + 0.001} \end{bmatrix}$$

$$= \begin{bmatrix} 0.158 \\ 0.158 \end{bmatrix} \longrightarrow \text{this is not in the span of } X.$$

Since the weight update process exit the span of X, it would not always give the minimum norm solution

2.1.1



2.1.2

Forward Propagation: $O(1)$

The number of values the network need to store during the forward propagation is independent of the number of iterations of gradient descent in back propagation

Backward Propagation: $O(t)$

In each iteration of the gradient, the network has to store the $\nabla_{w_t}\tilde{L}_t$ value and the original weight w_t value to compute the next weight value. Thus, it is $O(t)$.

2.2.1

$$W_1 = w_0 - \eta \nabla w_0$$
$$= w_0 - \eta\left(\tfrac{2}{n} X^T (X w_0 - t)\right)$$

Let $a = Xw_0 - t$

$$= w_0 - \eta \tfrac{2}{n} X^T a$$

$$L_1 = \tfrac{1}{n} \| X\hat{w} - t \|_2^2$$
$$= \tfrac{1}{n} \| X(w_0 - \eta \tfrac{2}{n} X^T a) - t \|_2^2$$
$$= \tfrac{1}{n} \| Xw_0 - \tfrac{2}{n}\eta X X^T a - t \|_2^2$$
$$= \tfrac{1}{n} \| (I - \tfrac{2}{n}\eta X X^T) a \|_2^2$$

2.2.3

$$L_1 = \tfrac{1}{n} \| Ia - \tfrac{2}{n}\eta X X^T a \|_2^2 = \tfrac{1}{n}\left[\|a\|^2 - 2a^T(\tfrac{2}{n}\eta X X^T a) + \tfrac{4}{n^2}\eta^2 \| X X^T a \|^2 \right]$$

$$\frac{\partial L_1}{\partial \eta} = \frac{8}{n^2}\eta \|XX^Ta\|^2 - \frac{4}{n} a^T X X^T a$$

$$\cancel{\tfrac{8}{n^2}}\eta^* \|XX^Ta\|^2 = \overset{\frac{n}{2}}{\cancel{\tfrac{4}{n}}} a^T X X^T a$$

$$\eta^* = \frac{n a^T X X^T a}{2 \|XX^Ta\|^2}$$

2.3.1. In terms of $\hat{L}_1$

$$w_1 = w_0 - \eta \nabla \hat{L}_1(X)$$
$$= w_0 - \eta\left(\tfrac{2}{n} X^T(Xw_0 - t) - 2\hat{\lambda} w_0\right)$$
$$= (1 - 2\hat{\lambda}\eta) w_0 - \eta \tfrac{2}{n} X^T(Xw_0 - t)$$

In terms of L_1:

$$w_1 = (1-\lambda)w_0 - \eta \nabla L_1(X)$$
$$= w_0 - \lambda w_0 - \eta\left[\tfrac{2}{n} X^T(Xw_0 - t)\right]$$

2.3.2

$$w_1 = (1 - 2\eta\hat{\lambda}) w_0 - \tfrac{2\eta}{n}(X^T X w_0 - X^T t)$$
$$\lambda = 2\eta\hat{\lambda}$$
$$\hat{\lambda} = \frac{\lambda}{2\eta}$$

3.1

$$I = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$J = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$J'' = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

It detect Edges.

$I * J$

$$= \begin{bmatrix} 0 & -1 & -2 & -3 & -2 \\ -2 & -3 & -3 & -2 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ 2 & 2 & 2 & 1 & 1 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix}$$

3.2

CNN:

\# neurons:

① Conv 3×3 = 32×32

① Pooling 2×2 = 16×16

② Conv 3×3 = 16×16

③ pooling 2×2 = 8×8

④ Conv 3×3 = 8×8

\# neurons: $32\times32 + 16\times16\times2 + 8\times8\times2$
$= 1664$ //

\# Parameters (with bias)

$((3\times3\times1)+1)\times1\times3$

$= 10\times3$

$= 30$ parameters ← Assuming there's no training parameters in pooling layer.

Fully-Connected:

\# neurons

① Conv 3×3 = 32×32

① Pooling 2×2 = 16×16

② Conv 3×3 = 16×16

③ pooling 2×2 = 8×8

④ Conv 3×3 = 8×8

\# neurons: $32\times32 + 16^2\times2 + 8^2\times2$
$= 1664$ //

\# Parameters: (including bias)

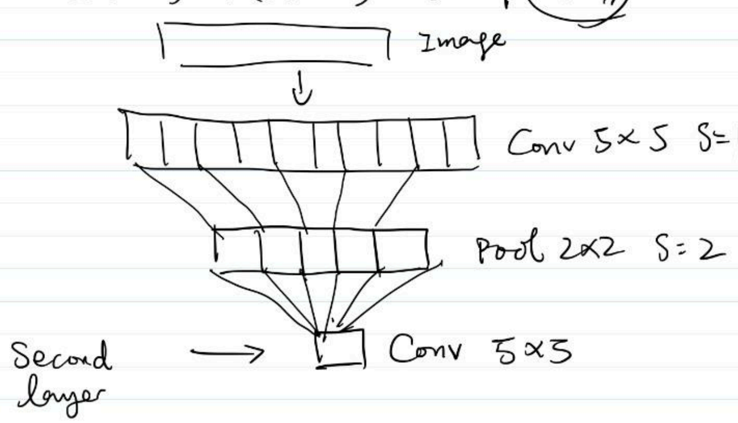
$(32\times32)^2 + 32^2 + ((16\times16)^2 + 16^2 + (8\times8)^2 + 8^2$

$= 1119552$ //

Disadvantage: the memory load will be too large, the time needed to compute both the forward propagation and the backward propagation will be significantly greater.

3.3.3 Receptive Fields

receptive fields $= (1\times5) + (10-1) = 5 + 9 = 14$ //



Two other things

① Stride

② the number of layers (depth of the neural network)