CSC413 Homework 2

Name: Yulin Wang

Student #: 1003942326

1.1.1 Minimum Norm Solution

$$\nabla_{W_t} L(x_j, W_t) = \nabla_{W_t} \left(\frac{1}{n} \| W_t^T x_j - t_j \|_2^2 \right)$$

$$= x_j \cdot \frac{1}{n} \cdot 2 \left(W_t^T x_j - t_j \right)$$

$$= \frac{2}{n} x_j \left(\underbrace{W_t^T x_j - t_j}_{\in R} \right)$$

Since $x_j \in B$, is randomly sampled without replacement from the data matrix X, so x_j is within the row space of X.

And $\nabla_{W_t} L(x_j, W_t)$ is a constant multiple of x_j, so the update steps of mini-batch SGD are linear combinations of vectors within X, and will never leave the span of X.

Thus, if we assume a solution $\hat{w}$ is obtained, then it must be in the span of X, which could be represented as $\hat{w} = X^T a$ for some $a \in R^n$.

Consider another feasible solution w to the linear regression model such that $Xw = t$, then we have:

$$(\hat{w} - w)^T \hat{w} = (\hat{w} - w)^T X^T a$$

$$= (X\hat{w} - Xw)^T a$$

$$= (t - t) a$$

$$= 0$$

that is, $\hat{w} - w \perp \hat{w} \Rightarrow \|\hat{w} - w\|^2 + \|\hat{w}\|^2 = \|w\|^2 \Rightarrow \|\hat{w}\|^2 \leq \|w\|^2$

Thus, mini-batch SGD solution $\hat{w}$ is the minimum norm solution, i.e., $\hat{w} = w^*$.

1.2.1 Minimum Norm Solution

Consider $X_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $W_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $t = 2$, then $n=1$, $d=2$. Set $\beta = 0.9$, $\varepsilon = 0.1$, $\eta = 0.1$

$$\nabla_{W_0} L(W_0) = \frac{2}{n} X_1 (X_1^T W_0 - t) = 2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} (0-2) = \begin{pmatrix} -8 \\ -4 \end{pmatrix}$$

$$V_0 = (1-0.9) \begin{pmatrix} 64 \\ 16 \end{pmatrix} = \begin{pmatrix} 6.4 \\ 1.6 \end{pmatrix}$$

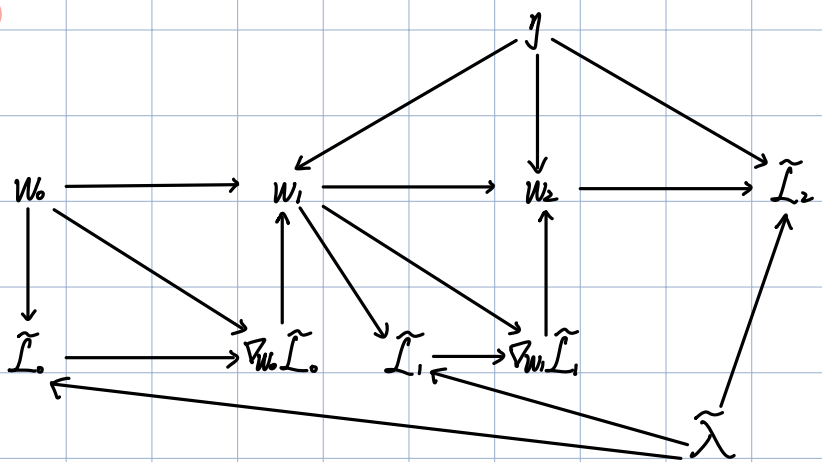
$$\Rightarrow W_1 = W_0 - \begin{pmatrix} \frac{0.1}{\sqrt{64}+0.1} \times (-8) \\ \frac{0.1}{\sqrt{1.6}+0.1} \times (-4) \end{pmatrix} = \begin{pmatrix} \frac{0.8}{\sqrt{6.4}+0.1} \\ \frac{0.4}{\sqrt{1.6}+0.1} \end{pmatrix} \approx \begin{pmatrix} 0.3042 \\ 0.2931 \end{pmatrix}$$

But this solution is not in the span of $\{\begin{pmatrix} 2 \\ 1 \end{pmatrix}\}$, thus it is NOT the minimum norm solution.

Thus, this counterexample violates the assumption.

2.1 Computation Graph

2.1.1



2.1.2

forward-propagation : $O(1)$

standard back-propagation: $O(t)$

2.2 Optimal Learning Rates

2.2.1

$$w_1 = w_0 - \eta \nabla_{w_0} \tilde{L}_0$$

$$= w_0 - \eta \cdot \frac{2}{n} \cdot X^T(Xw_0 - t)$$

$$= w_0 - \frac{2\eta}{n} \cdot X^T a \quad ; \text{ set } a = Xw_0 - t$$

$$L_1 = \frac{1}{n} \| Xw_1 - t \|_2^2$$

$$= \frac{1}{n} \| X(w_0 - \frac{2\eta}{n} X^T a) - t \|_2^2$$

$$= \frac{1}{n} \| Xw_0 - \frac{2\eta}{n} XX^T a - t \|_2^2$$

$$= \frac{1}{n} \| a + t - \frac{2\eta}{n} XX^T a - t \|_2^2 \quad ; \text{ since } Xw_0 = a + t$$

$$= \frac{1}{n} \| a - \frac{2\eta}{n} XX^T a \|_2^2$$

$$= \frac{1}{n} \| (I - \frac{2\eta}{n} XX^T) a \|_2^2$$

$$= \frac{1}{n} \cdot a^T (I - \frac{2\eta}{n} XX^T)^2 a$$

2.2.3

$$\frac{dL_1}{d\eta} = \frac{d}{d\eta} \left(\frac{1}{n} \cdot a^T (I - \frac{2\eta}{n} XX^T)^2 a \right)$$

$$= \frac{2}{n} \cdot a^T (I - \frac{2\eta}{n} XX^T)(-\frac{2}{n} XX^T) a$$

$$= \frac{2}{n} \cdot a^T (-\frac{2}{n} XX^T + \frac{4\eta}{n^2} XX^T XX^T) a$$

$$\underline{\text{set}} \ 0$$

$$\Rightarrow a^T (-\frac{2}{n} XX^T + \frac{4\eta^*}{n^2} XX^T XX^T) a = 0$$

$$a^T (-nXX^T + 2\eta^* XX^T XX^T) a = 0$$

$$\eta^* (2a^T XX^T XX^T a) = n a^T XX^T a$$

$$\eta^* = \frac{n}{2} \cdot \frac{a^T XX^T a}{a^T XX^T XX^T a}$$

$$\eta^* = \frac{n}{2} \cdot \frac{(X^T a)^2}{(XX^T a)^2}$$

2.3 Weight decay and L_2 regularization

2.3.1

1) first expression using $\tilde{L}$

$$\nabla_{w_0}\tilde{L} = \frac{2}{n}X^T(Xw_0 - t) + 2\tilde{\lambda}w_0$$

$$w_1 = w_0 - \eta \cdot \nabla_{w_0}\tilde{L}$$

$$= w_0 - \eta \cdot \left[\frac{2}{n}X^T(Xw_0 - t) + 2\tilde{\lambda}w_0\right]$$

$$= w_0 - \frac{2\eta}{n}X^T(Xw_0 - t) - 2\eta\tilde{\lambda}w_0$$

2) seond expression using L and weight decay

$$\nabla_{w_0}L = \frac{2}{n}X^T(Xw_0 - t)$$

$$w_1 = (1-\lambda)w_0 - \eta\nabla_{w_0}L$$

$$= (1-\lambda)w_0 - \frac{2\eta}{n}X^T(Xw_0 - t)$$

2.3.2

Since each corresponding update step should be the same, then:

$$w_0 - \frac{2\eta}{n}X^T(Xw_0 - t) - 2\eta\tilde{\lambda}w_0 = (1-\lambda)w_0 - \frac{2\eta}{n}X^T(Xw_0 - t)$$

$$w_0 - 2\eta\tilde{\lambda}w_0 = (1-\lambda)w_0$$

$$2\eta\tilde{\lambda}w_0 = \lambda w_0$$

$$\tilde{\lambda} = \frac{\lambda}{2\eta}$$

3.1 Convolutional Filters

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

This convolutional filter detects edges in horizontal direction.

3.2 Size of Conv Nets

1> CNN:

① number of neurons	② number of trainable parameters
Conv1: $32 \times 32 \times 1$	Conv1: $3 \times 3 + 1$
Pool1: $16 \times 16 \times 1$	Pool1: 0
Conv2: $16 \times 16 \times 1$	Conv2: $3 \times 3 + 1$
Pool2: $8 \times 8 \times 1$	Pool2: 0
Conv3: $8 \times 8 \times 1$	Conv3: $3 \times 3 + 1$
⇒ Total: 1664	⇒ Total: 30

2) FCNN:

	① number of neurons	② number of trainable parameters
FC1	$32 \times 32 \times 1$	$(32 \times 32) \times (32 \times 32) + (32 \times 32)$
Pool1	$16 \times 16 \times 1$	0
FC2	$16 \times 16 \times 1$	$(16 \times 16) \times (16 \times 16) + (16 \times 16)$
Pool2	$8 \times 8 \times 1$	0
FC3	$8 \times 8 \times 1$	$(8 \times 8) \times (8 \times 8) + (8 \times 8)$
⇒ Total:	1664	1119552

3) More trainable parameters would increase computation complexity.

3.3 Receptive Fields

1) the receptive field of a neuron:

after Conv 1: 5×5

after Pool1: 6×6

after Conv 2: 14×14

2) 2 other things can affect the size of the receptive field of a neuron:

① the stride value of max pooling layer

② the depth of the layer