CSC 2516

Homework 2

Jane Seo
1001423284

Q1

1.1.1 $\nabla_{w_t} \mathcal{L}_i(x_i, w_t) = \frac{2}{n} x_i^T(x_i w_t - t_i)$

To simplify calculations, assume $\eta = \frac{n}{2}$

(not realistic, but just to simplify calculations)

$w_0 = 0$

$$w_1 \leftarrow w_0 - \eta \frac{2}{n} x_i^T (x_i w_0 - t_i)$$

$$w_1 \leftarrow x_i^T t_i$$

$$w_2 \leftarrow w_1 - \eta \frac{2}{n} x_j^T (x_j w_1 - t_j)$$

$$w_2 \leftarrow x_i^T t_i - x_j^T x_j x_i^T t_i + x_j^T t_j$$

$$w_2 \leftarrow (1 - x_j^T x_j) x_i^T t_i + x_j^T t_j$$

$$w_3 \leftarrow w_2 - \eta \frac{2}{n} x_k^T (x_k w_2 - t_k)$$

$$w_3 \leftarrow (1 - x_j^T x_j) x_i^T t_i + x_j^T t_j - x_k^T x_k [x_i^T t_i - x_j^T x_j x_i^T t_i + x_j^T t_j] + x_k^T t_k$$

$$w_3 \leftarrow (1 - x_j^T x_j - x_k^T x_k + x_k^T x_k x_j^T x_j) x_i^T t_i + (1 - x_k^T x_k) x_j^T t_j + x_k^T t_k$$

= constant

$\therefore$ $\hat{w}$ will be a linear combination of all rows of X

i.e. $\hat{w} = X^T a$ where $a \in \mathbb{R}^n$

$\therefore$ If $X\hat{w} = t$

$XX^T a = t$

$a = (XX^T)^{-1} t$

$\therefore$ $\hat{w} = X^T (XX^T)^{-1} t$

which is the same as w^* from HW1

1.1.2 Assuming $w_0 = 0$ & $\delta_0 = 0$, and $\eta = \frac{n}{2}$ ↰

(not realistic, but just to simplify calculations)

$$\delta_1 \leftarrow -\eta \frac{2}{n} x_i^T (x_i w_0 - t_i)$$

$$\delta_1 \leftarrow x_i^T t_i$$

$$w_1 \leftarrow x_i^T t_i$$

$$\delta_2 \leftarrow -\eta \frac{2}{n} x_j^T (x_j w_1 - t_j) + \alpha x_i^T t_i$$

$$\delta_2 \leftarrow -x_j^T x_j x_i^T t_i + x_j^T t_j + \alpha x_i^T t_i$$

$$\delta_2 \leftarrow (\alpha - x_j^T x_j) x_i^T t_i + x_j^T t_j$$

$$w_2 \leftarrow (1 + \alpha - x_j^T x_j) x_i^T t_i + x_j^T t_j$$

↳ comparing this to w_2 from 1.1.1:

$$w_2 \leftarrow (1 - x_j^T x_j) x_i^T t_i + x_j^T t_j$$

the only difference is the constant in front of $x_i^T t_i$

i.e. SGD with momentum will still converge to a $\hat{w}$ that is a linear combination of all rows of X

∴ as with 1.1.1, SGD with momentum will still converge to the minimum norm solution.

1.2.1 Counter example :

let $X_1 = [1 \quad 2]$, $w_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $t = 3$, $G_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$w_1 = w_0 - \frac{\eta}{\epsilon} \frac{2}{n} X_1^T (X_1 w_0 - t)$$

$$= \frac{2\eta}{\epsilon n} X_1^T t$$

$$= \frac{2\eta}{\epsilon n} \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

$$G_1 = \left(\frac{2\eta}{\epsilon n}\right)^2 \cdot \begin{bmatrix} 9 \\ 36 \end{bmatrix}$$

let $\eta = \epsilon = 0.0001$ to make calculations simpler

$$w_1 = \begin{bmatrix} 6 \\ 12 \end{bmatrix} \qquad G_1 = \begin{bmatrix} 36 \\ 144 \end{bmatrix}$$

$$w_2 = \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \frac{0.0001}{\sqrt{G_1} + 0.0001} \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} [1 \quad 2] \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \begin{bmatrix} 3 \\ 6 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \frac{0.0001}{\sqrt{G_1} + 0.0001} \left(\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \begin{bmatrix} 3 \\ 6 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \frac{0.0001}{\sqrt{G_1} + 0.0001} \begin{bmatrix} 27 \\ 54 \end{bmatrix}$$

$$= \begin{bmatrix} 6 \\ 12 \end{bmatrix} - \begin{bmatrix} \frac{0.0001}{6.0001} \cdot 27 \\ \frac{0.0001}{12.0001} \cdot 54 \end{bmatrix}$$

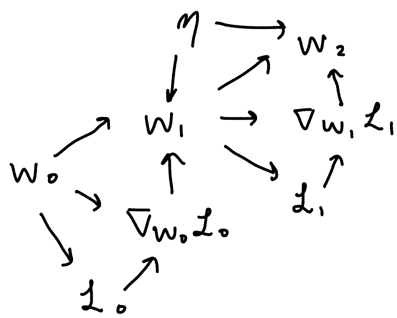
$$= \begin{bmatrix} 5.99955 \\ 11.99955 \end{bmatrix}$$

$\leftarrow$ cannot be written as linear combination of X_1

$\therefore$ will not converge to minimum norm solution

Q 2

2.1.1



2.1.2 Forward propagation : $O(1)$

Back-propagation : $O(t)$ because we need to store all w_t & $\nabla_{w_t}\mathcal{L}_t$

2.2.1 $\nabla_{w_0}\mathcal{L}_0 = \frac{2}{n} X^T (X w_0 - t)$ Let $a = Xw_0 - t$

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

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

$$\mathcal{L}_1 = \frac{1}{n} \| X w_1 - t \|_2^2$$

$$\therefore \mathcal{L}_1 = \frac{1}{n} \| X (w_0 - \frac{2\eta}{n} X^T a) - t \|_2^2$$

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

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

2.2.2 $$\frac{\partial \mathcal{L}_1}{\partial \eta} = -\left(\frac{2}{n}\right)^2 a^T \left(\mathbb{I} - \frac{2\eta}{n} X X^T \right) X X^T a$$

$$\frac{\partial^2 \mathcal{L}_1}{\partial \eta^2} = \left(\frac{2}{n}\right)^3 a^T X X^T X X^T a > 0 \quad \therefore \text{ convex}$$

2.2.3
$$\frac{\partial \mathcal{L}_1}{\partial \eta} = -\left(\frac{2}{n}\right)^2 a^T \left(\mathbb{I} - \frac{2\eta}{n} X X^T\right) X X^T a$$
$$= -\left(\frac{2}{n}\right)^2 a^T X X^T a + \left(\frac{2}{n}\right)^3 \eta\, a^T X X^T X X^T a$$

$$-\left(\frac{2}{n}\right)^2 a^T X X^T a + \left(\frac{2}{n}\right)^3 \eta^* a^T X X^T X X^T a = 0$$
$$\left(\frac{2}{n}\right)^3 \eta^* a^T X X^T X X^T a = \left(\frac{2}{n}\right)^2 a^T X X^T a$$
$$\therefore \quad \eta^* = \frac{n}{2} \frac{a^T X X^T a}{a^T X X^T X X^T a}$$

2.3.1
$$\nabla_{w_0} \mathcal{L}_0 = \frac{2}{n} X^T (X w_0 - t) \qquad \text{Let } a = X w_0 - t$$
$$w_1 = w_0 - \eta \nabla_{w_0} \mathcal{L}_0$$
$$= w_0 - \frac{2\eta}{n} X^T a$$
$$\mathcal{L}_1 = \frac{1}{n} \left\| a - \frac{2\eta}{n} X X^T a \right\|_2^2$$
$$\nabla_{w_1} \mathcal{L}_1 = \frac{2}{n} X^T (X w_1 - t)$$
$$= \frac{2}{n} X^T \left[X \left(w_0 - \frac{2\eta}{n} X^T a\right) - t\right]$$
$$= \frac{2}{n} X^T \left[a - \frac{2\eta}{n} X X^T a\right]$$

$$w_2 = w_1 - \eta \nabla_{w_1} \mathcal{L}_1$$
$$= w_0 - \frac{2\eta}{n} X^T a - \frac{2\eta}{n} X^T \left(a - \frac{2\eta}{n} X X^T a\right)$$
$$= w_0 - \frac{2\eta}{n} X^T \left[\frac{2\eta}{n} X X^T\right] a$$
$$\mathcal{L}_2 = \frac{1}{n} \left\| X \left(w_0 - \frac{2\eta}{n} X^T \left[\frac{2\eta}{n} X X^T\right] a\right) - t \right\|_2^2$$
$$= \frac{1}{n} \left\| a - \left(\frac{2\eta}{n} X X^T\right)^2 a \right\|_2^2$$

2.3.1 (cont'd)

$$\nabla_{w_2}\mathcal{L}_2 = \frac{2}{n}X^T(Xw_2 - t)$$
$$= \frac{2}{n}X^T\left(X\left[w_0 - \frac{2\eta}{n}X^T\left(\frac{2\eta}{n}XX^T\right)a\right] - t\right)$$
$$= \frac{2}{n}X^T\left(a - \left[\frac{2\eta}{n}XX^T\right]^2 a\right)$$

$$w_3 = w_2 - \eta\nabla_{w_2}\mathcal{L}_2$$
$$= w_0 - \frac{2\eta}{n}X^T\left[\frac{2\eta}{n}XX^T\right]a - \frac{2\eta}{n}X^T\left(a - \left[\frac{2\eta}{n}XX^T\right]^2 a\right)$$
$$= w_0 - \frac{2\eta}{n}X^T\left[\frac{2\eta}{n}XX^T\right]^2 a$$
$$\mathcal{L}_3 = \frac{1}{n}\left\|X\left(w_0 - \frac{2\eta}{n}X^T\left[\frac{2\eta}{n}XX^T\right]^2 a\right) - t\right\|_2^2$$
$$= \frac{1}{n}\left\|a - \left[\frac{2\eta}{n}XX^T\right]^3 a\right\|_2^2$$

$$\therefore\ \mathcal{L}_t = \frac{1}{n}\left\|a - \left[\frac{2\eta}{n}XX^T\right]^t a\right\|_2^2$$

2.3.2 $$\mathcal{L}_t = \frac{1}{n}\left\|a - \left[\frac{2\eta}{n}XX^T\right]^t a\right\|_2^2$$
$$= \frac{1}{n}a^T\left(\mathbb{I} - \frac{2\eta}{n}XX^T\right)^{2t} a$$
$$\frac{\partial\mathcal{L}_t}{\partial\eta} = \frac{-4t}{n^2}a^T\left(\mathbb{I} - \frac{2\eta}{n}XX^T\right)^{2t-1}XX^T a$$
$$\frac{\partial^2\mathcal{L}_t}{\partial\eta^2} = \frac{8t(2t-1)}{n^3}a^T\left(\mathbb{I} - \frac{2\eta}{n}XX^T\right)^{2t-2}XX^TXX^T a$$

$> 0 \quad \therefore$ convex

Q3

3.1

$$I * J = \begin{array}{|c|c|c|c|c|} \hline -1 & 2 & 2 & -2 & 0 \\ \hline -2 & 1 & 0 & 2 & -1 \\ \hline 3 & 0 & 0 & 1 & -1 \\ \hline -2 & 2 & 0 & 2 & -1 \\ \hline 0 & -2 & 3 & -2 & 0 \\ \hline \end{array}$$

This is an edge detector

3.2

layer	# neurons	# parameters	# input connections
conv3-64	$112\times112\times64$ = 802816	$3^2\cdot3\cdot64+64$ = 1792	$112\cdot112\cdot3^2\cdot64\cdot3$ = 21676032
Max Pool	$56\times56\times64$ = 200704	0	$56\times56\times2^2\times64$ = 802816
conv3-128	$56\times56\times128$ = 401408	$3^2\times64\times128+128$ = 73856	$56\times56\times3^2\times64\times128$ = 231211008
Max Pool	$28\times28\times128$ = 100352	0	$28\times28\times2^2\times128$ = 401408
conv 3-256	$28\times28\times256$ = 200704	$3^2\times128\times256+256$ = 295168	$28\times28\times3^2\times128\times256$ = 231211008
conv3-256	$28\times28\times256$ = 200704	$3^2\times256\times256+256$ = 590080	$28\times28\times3^2\times256\times256$ = 462422016
max pool	$14\times14\times256$ = 50176	0	$14\times14\times2^2\times256$ = 200704
FC-1024	1024	$14\times14\times256\times1024+1024$ = 51381248	$14\times14\times256\times1024$ = 51380224
FC-100	100	$1024\times100+100$ = 102500	1024×100 = 102400
Total	1957988	52444644	999407616

3.3. 1) stride : greater stride can lead to larger receptive field ie more information can be 'seen' with the same # of strides

2) kernel : larger kernel can lead to larger receptive field ie more information 'seen' with each stride

3) number of layers : more layers can lead to larger receptive field e.g. if a pooling layer had been added to 3.2, a 224×224 3-channel image could have been 'seen'