Deeplearning HK4. Participants. 1° Hanzhi Zhuang 5578347 2° Rui Song 5576534 3 Tengyunhao Yang 5587382

1. Pen and Paper tasks.

1) $\hat{y}_0 = w_0^T X = [1 \;\; 0] \cdot \left[\begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} \right] = [2 \;\; -1]$

First update step:

1° $W_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \theta_0$

2° sample from: $\left\{ \left(\begin{bmatrix} 2 \\ -1 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} -1 \\ 3 \end{bmatrix}, -1 \right) \right\}$

3° $\hat{g} = \frac{1}{m} \nabla_\theta \sum (\hat{y} - y)^2 = \frac{1}{m} \nabla_\theta \sum (w^T x - y)^2 = \frac{1}{m} \sum 2(w^T x - y) \cdot X = \sum (w^T x - y) X$

$= (2-3) \cdot \begin{bmatrix} 2 \\ -1 \end{bmatrix} + (-1-1) \cdot \begin{bmatrix} -1 \\ 3 \end{bmatrix}$

$= \begin{bmatrix} -2+2 \\ 1-6 \end{bmatrix}$

$= \begin{bmatrix} 0 \\ -5 \end{bmatrix}$

4° $V_1 = \beta \cdot V_0 - \alpha \hat{g} = 0.8 \cdot \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.2 \begin{bmatrix} 0 \\ -5 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

5° $W_1 = W_0 + V_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Second update step:

$\hat{Y}_1 = W_1^T X = [1 \;\; 1] \cdot \left[\begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} \right] = [1 \;\; 2]$.

1° $W_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \theta_1$.

-4 $\quad$ 2 $\quad$ -1 $\quad$ 3.

2° Sample from $\left\{ \left(\begin{bmatrix} 2 \\ -1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} -1 \\ 3 \end{bmatrix}, 2 \right) \right\}$.

3° $\hat{g}_1 = \sum (W^T x - y) X = (1-3) \begin{bmatrix} 2 \\ -1 \end{bmatrix} + (2-1) \cdot \begin{bmatrix} -1 \\ 3 \end{bmatrix} = \begin{bmatrix} -5 \\ 5 \end{bmatrix}$.

4° $V_2 = \beta V_1 - \alpha \hat{g}_1 = 0.8 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 0.2 \begin{bmatrix} -5 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ -0.2 \end{bmatrix}$.

5° $W_2 = W_1 + V_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ -0.2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0.8 \end{bmatrix}$

Third update step:

$\hat{Y}_2 = W_2^T X = [2 \;\; 0.8] \cdot \left[\begin{bmatrix}2\\-1\end{bmatrix} \begin{bmatrix}-1\\3\end{bmatrix}\right] = [3.2 \;\; 0.4].$

1° $W_2 = \begin{bmatrix}2\\0.8\end{bmatrix} = \theta_2.$

2° Sample from $\left\{\left(\begin{bmatrix}2\\-1\end{bmatrix}, 3.2\right), \left(\begin{bmatrix}-1\\3\end{bmatrix}, 0.4\right)\right\}.$

3° $\hat{g}_2 = \sum(w^T x - y) X = (3.2-3)\begin{bmatrix}2\\-1\end{bmatrix} + (0.4-1)\begin{bmatrix}-1\\3\end{bmatrix} = \begin{pmatrix}1\\-2\end{pmatrix}$

4° $V_3 = \beta V_2 - \alpha \hat{g}_2 = 0.8\begin{bmatrix}1\\-0.2\end{bmatrix} - 0.2\begin{pmatrix}1\\-2\end{pmatrix} = \begin{pmatrix}0.6\\0.24\end{pmatrix}$

5° $W_3 = W_2 + V_3 = \begin{bmatrix}2\\0.8\end{bmatrix} + \begin{bmatrix}0.6\\0.24\end{bmatrix} = \begin{bmatrix}2.6\\1.04\end{bmatrix}$

---

2) $\hat{s} = \dfrac{s}{1-\rho_1^t}$, $s_1 = \rho_1 s + (1-\rho_1)\hat{g}.$

proof: for $t=1$: $s_1 = \rho \cdot s_0 + (1-\rho)\hat{g} = (1-\rho)\hat{g}.$

$$\hat{s}_1 = \frac{s_1}{1-\rho^1} = \frac{(1-\rho)\hat{g}}{1-\rho} = \hat{g}.$$

assume it holds true for $t=n$, where $t \geq 1$, such as $\hat{s}_n = \hat{g}$, when $t=n+1$:

$\because \hat{s_n} = \hat{g}$ (under assumption).

also $\hat{s_n} = \dfrac{s_n}{1-\rho^n} = \hat{g}$

$\therefore$ we can get $s_n = (1-\rho^n)\hat{g}$.

for $s_{n+1} = \rho \cdot s_n + (1-\rho) \cdot \hat{g}$

$$= \rho \cdot (1-\rho^n)\hat{g} + (1-\rho)\hat{g}$$

$$= (\rho - \rho^{n+1} + 1 - \rho)\hat{g}$$

$$= (1-\rho^{n+1})\hat{g}.$$

$$\therefore \hat{s}_{n+1} = \frac{s_{n+1}}{(1-\rho^{n+1})} = \frac{(1-\rho^{n+1})\hat{g}}{1-\rho^{n+1}} = \hat{g}$$

$\therefore$ The assumption also holds for $\hat{s}_{n+1}$, when $\hat{s}_n$ holds.

$\therefore$ if $\hat{s}_n = \hat{g}$ holds, then $\hat{s}_{n+1} = \hat{g}$ holds for $n \geq 1$.

$\therefore$ By mathematical induction, $\hat{s}_n = \hat{g}$ for every step.