Submission for Deep Learning Exercise

Team: shallow_learning_group
Students: Batuhan Karaca

November 18, 2023

Pen and Paper tasks

1)

According to text, the loss function for B number of samples is

$$\mathcal{L}(\hat{y}, y) = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}_i - y_i)^2$$

And the output

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{w}^T X \\
\hat{y}_i &= \sum_{j=1}^{B} w_j X_{ji} \\
\frac{\delta \hat{y}_i}{\delta w_m} &= X_{mi}
\end{aligned}$$

Then, by the chain rule and superposition property of the derivative

$$\begin{aligned}
\frac{\delta \mathcal{L}}{\delta w_m} &= \frac{2}{B}\sum_{i=1}^{B}(\hat{y}_i - y_i)X_{mi} \\
&= \frac{2}{B}\sum_{i=1}^{B}(\hat{y}_i - y_i)X^T_{im}
\end{aligned}$$

The gradient of the loss at time step t becomes

$$
\begin{aligned}
\nabla_{\mathbf{w}^t}\mathcal{L} &= \frac{2}{B}(\hat{\mathbf{y}}^t - \mathbf{y})X^T \\
&= \frac{2}{B}((\mathbf{w}^t)^T X X^T - \mathbf{y}X^T) \\
&= \frac{2}{B}((\mathbf{w}^t)^T X^2 - \mathbf{y}X) \quad \text{since } X \text{ given in the text is a symmetric matrix} \\
X^2 &= (\begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix})^2 \\
&= \begin{bmatrix} 5 & -5 \\ -5 & 10 \end{bmatrix} \\
&= 5(\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}) \\
\mathbf{y}X &= \begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \\
&= \begin{bmatrix} 5 & 0 \end{bmatrix} \\
&= 5(\begin{bmatrix} 1 & 0 \end{bmatrix})
\end{aligned}
$$

Given $B = 2$, we have

$$\nabla_{\mathbf{w}^t}\mathcal{L} = 5((\mathbf{w}^t)^T \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \end{bmatrix})$$

According to the lecture slides, the update rule for the velocity is

$$\mathbf{v}^t = \beta \mathbf{v}^{t-1} - \alpha \nabla_{\mathbf{w}^{t-1}}\mathcal{L}$$

Given $\alpha = 0.2$ and $\beta = 0.8$

$$\mathbf{v}^t = 0.8\mathbf{v}^{t-1} - ((\mathbf{w}^{t-1})^T \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \end{bmatrix})$$

and the update rule for the weights is given as $\mathbf{w}^t = \mathbf{w}^{t-1} + (\mathbf{v}^t)^T$ in the lecture slides. Given $\mathbf{v}_0 = \mathbf{v} = \begin{bmatrix} 0 & 0 \end{bmatrix}$, and $\mathbf{w}_0 = \mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Update step 1

$$
\begin{aligned}
\mathbf{v}_1 &= -(\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \end{bmatrix}) \\
&= \begin{bmatrix} 0 & 1 \end{bmatrix} \\
\mathbf{w}^1 &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\end{aligned}
$$

Update step 2

$$
\begin{aligned}
\mathbf{v}_2 &= \begin{bmatrix}0 & 0.8\end{bmatrix} - \left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & -1\\ -1 & 2\end{bmatrix} - \begin{bmatrix}1 & 0\end{bmatrix}\right)\\
&= \begin{bmatrix}1 & -0.2\end{bmatrix}\\
\mathbf{w}^2 &= \begin{bmatrix}1\\ 1\end{bmatrix} + \begin{bmatrix}1\\ -0.2\end{bmatrix}\\
&= \begin{bmatrix}2\\ 0.8\end{bmatrix}
\end{aligned}
$$

Update step 3

$$
\begin{aligned}
\mathbf{v}_3 &= \begin{bmatrix}0.8 & -0.16\end{bmatrix} - \left(\begin{bmatrix}2 & 0.8\end{bmatrix}\begin{bmatrix}1 & -1\\ -1 & 2\end{bmatrix} - \begin{bmatrix}1 & 0\end{bmatrix}\right)\\
&= \begin{bmatrix}0.6 & 0.24\end{bmatrix}\\
\mathbf{w}^3 &= \begin{bmatrix}2\\ 0.8\end{bmatrix} + \begin{bmatrix}0.6\\ 0.24\end{bmatrix}\\
&= \begin{bmatrix}2.6\\ 1.04\end{bmatrix}
\end{aligned}
$$

2)

The Adam formulae for the first order moments at training time step t are

$$
\begin{aligned}
s_t &= \rho_1 s_{t-1} + (1-\rho_1)\hat{g} \quad (1)\\
\hat{s}_t &= \frac{s_t}{1-{\rho_1}^t} \quad (2)
\end{aligned}
$$

where $\hat{g}$ is the gradient calculated at backpropagation step and assumed constant for all t. It is easy to see the relationship between the estimates of first order moments from equations (1) and (2).

$$
\hat{s}_t = \frac{\rho_1(1-{\rho_1}^{t-1})\hat{s}_{t-1} + (1-\rho_1)\hat{g}}{1-{\rho_1}^t}
$$

At time step $t = 1$

$$
\begin{aligned}
\hat{s}_1 &= \frac{\rho_1(1-{\rho_1}^0)\hat{s}_0 + (1-\rho_1)\hat{g}}{1-\rho_1}\\
&= \frac{(1-\rho_1)\hat{g}}{1-\rho_1}\\
&= \hat{g}
\end{aligned}
$$

Now wee assume that at a time step $t = \tau - 1$, $\hat{s}_{\tau-1} = \hat{g}$. Then

$$
\begin{aligned}
\hat{s}_\tau &= \frac{\rho_1(1-{\rho_1}^{\tau-1})\hat{g} + (1-\rho_1)\hat{g}}{1-{\rho_1}^\tau} \\
&= \frac{(\rho_1 - {\rho_1}^\tau)\hat{g} + (1-\rho_1)\hat{g}}{1-{\rho_1}^\tau} \\
&= \frac{(1-{\rho_1}^\tau)\hat{g}}{1-{\rho_1}^\tau} \\
&= \hat{g}
\end{aligned}
$$

We have shown that for the base step $s_1 = \hat{g}$, and also if $s_{t-1} = \hat{g}$, then $s_t = \hat{g}$ for all t concluding our proof.