Submission for Deep Learning Exercise 4

Team: dl2022-ryd
Students: Yumna Ali, Deepu K Reddy, Rean Fernandes

November 15, 2022

1 Pen and Paper: Stochastic Gradient Descent

1.1 First Update

1.1.1 Forward Pass

$$\hat{\mathbf{y}}^{(0)} = (\mathbf{w}^{(0)})^T \cdot \mathbf{X} = \quad \begin{bmatrix}\begin{bmatrix}1 & 0\\1 & 0\end{bmatrix}\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix} \begin{bmatrix}-1\\3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}2 & -1\end{bmatrix} \tag{1a}$$

$$\mathscr{L}(\hat{\mathbf{y}}^{(0)}, \mathbf{y}) = \quad (\hat{\mathbf{y}}^{(0)} - \mathbf{y})^2 = \quad (\begin{bmatrix}2 & -1\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix})^2 = \quad \begin{bmatrix}1 & 4\end{bmatrix} \tag{1b}$$

1.1.2 Backward Pass

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(0)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(0)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(0)} - \mathbf{y}) = \quad 2 \cdot (\begin{bmatrix}2 & -1\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix}) = \quad \begin{bmatrix}-2 & -4\end{bmatrix} \tag{2a}$$

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(0)}, \mathbf{y})}{\partial \mathbf{w}^{(0)}} = \quad \frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(0)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(0)}} \cdot \frac{\partial \hat{\mathbf{y}}^{(0)}}{\partial \mathbf{w}^{(0)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(0)} - \mathbf{y}) \cdot \mathbf{X} = \quad \begin{bmatrix}\begin{bmatrix}-2\end{bmatrix} & \begin{bmatrix}-4\end{bmatrix}\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix} \begin{bmatrix}-1\\3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}\begin{bmatrix}-4\\2\end{bmatrix}^T & \begin{bmatrix}4\\-12\end{bmatrix}^T\end{bmatrix} \tag{2b}$$

The first column of the matrix in 7b are the gradients for the first input in our batch, and the second column is the gradient for the second input, both with respect to the starting weights. As we perform SGD with batch size equal to our dataset size, there is no need to randomly sample, and we take the average of the gradients wrt each weight as the defined by the algorithm.

1.1.3 Velocity and gradient update

The gradient approximation is defined as

$$\mathbf{g} = \frac{1}{2} \nabla_{\mathbf{w}} \sum_{i=1}^{2} \mathscr{L}(\hat{y}^{(0)}, y) = 0.5 \cdot \begin{bmatrix}-4+4\\2-12\end{bmatrix}^T = \begin{bmatrix}0\\-5\end{bmatrix}^T \tag{3}$$

The new updated velocity is

$$\mathbf{v}^{(1)} = \beta \cdot \mathbf{v}^{(0)} - \alpha \cdot \mathbf{g} = \quad 0.8 \cdot \begin{bmatrix}0\\0\end{bmatrix}^T - 0.2 \cdot \begin{bmatrix}0\\-5\end{bmatrix}^T = \quad \begin{bmatrix}0\\-1\end{bmatrix}^T \tag{4}$$

Adding this velocity to the gradient gives us the first gradient update

$$\mathbf{w}^{(1)} = \mathbf{w}^{(0)} + \mathbf{v}^{(1)} = \begin{bmatrix}1\\0\end{bmatrix}^T + \begin{bmatrix}0\\-1\end{bmatrix}^T = \begin{bmatrix}1\\-1\end{bmatrix}^T \tag{5}$$

1.2 Second update

1.2.1 Forward Pass

$$\hat{\mathbf{y}}^{(1)} = (\mathbf{w}^{(1)})^T \cdot \mathbf{X} = \quad \begin{bmatrix}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\ -1\end{bmatrix} \begin{bmatrix}-1\\ 3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}3 & -4\end{bmatrix} \tag{6a}$$

$$\mathscr{L}(\hat{\mathbf{y}}^{(1)}, \mathbf{y}) = \quad (\hat{\mathbf{y}}^{(1)} - \mathbf{y})^2 = \quad (\begin{bmatrix}3 & -4\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix})^2 = \quad \begin{bmatrix}0 & 25\end{bmatrix} \tag{6b}$$

1.2.2 Backward Pass

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(1)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(1)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(1)} - \mathbf{y}) = \quad 2 \cdot (\begin{bmatrix}3 & -4\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix}) = \quad \begin{bmatrix}0 & -10\end{bmatrix} \tag{7a}$$

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(1)}, \mathbf{y})}{\partial \mathbf{w}^{(1)}} = \quad \frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(1)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(1)}} \cdot \frac{\partial \hat{\mathbf{y}}^{(1)}}{\partial \mathbf{w}^{(1)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(1)} - \mathbf{y}) \cdot \mathbf{X} = \quad \begin{bmatrix}[0] & [-10]\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\ -1\end{bmatrix} \begin{bmatrix}-1\\ 3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}\begin{bmatrix}0\\ 0\end{bmatrix}^T & \begin{bmatrix}10\\ -30\end{bmatrix}^T\end{bmatrix} \tag{7b}$$

1.2.3 Velocity and gradient update

The gradient approximation:

$$\mathbf{g} = \frac{1}{2} \nabla_{\mathbf{w}} \sum_{i=1}^{2} \mathscr{L}(\hat{y}^{(0)}, y) = 0.5 \cdot \begin{bmatrix}0 + 10\\ 0 - 30\end{bmatrix}^T = \begin{bmatrix}5\\ -15\end{bmatrix}^T \tag{8}$$

The new updated velocity is

$$\mathbf{v}^{(2)} = \beta \cdot \mathbf{v}^{(1)} - \alpha \cdot \mathbf{g} = \quad 0.8 \cdot \begin{bmatrix}0\\ -1\end{bmatrix}^T - 0.2 \cdot \begin{bmatrix}5\\ -15\end{bmatrix}^T = \quad \begin{bmatrix}1\\ -3.8\end{bmatrix}^T \tag{9}$$

Adding this velocity to the gradient gives us the first gradient update

$$\mathbf{w}^{(2)} = \mathbf{w}^{(1)} + \mathbf{v}^{(2)} = \begin{bmatrix}1\\ -1\end{bmatrix}^T + \begin{bmatrix}1\\ -3.8\end{bmatrix}^T = \begin{bmatrix}2\\ -4.8\end{bmatrix}^T \tag{10}$$

1.3 Third Update

1.3.1 Forward Pass

$$\hat{\mathbf{y}}^{(2)} = (\mathbf{w}^{(2)})^T \cdot \mathbf{X} = \quad \begin{bmatrix}\begin{bmatrix}2 & -4.8\\ 2 & -4.8\end{bmatrix}\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\ -1\end{bmatrix} \begin{bmatrix}-1\\ 3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}8.8 & -16.4\end{bmatrix} \tag{11a}$$

$$\mathscr{L}(\hat{\mathbf{y}}^{(2)}, \mathbf{y}) = \quad (\hat{\mathbf{y}}^{(2)} - \mathbf{y})^2 = \quad (\begin{bmatrix}8.8 & -16.4\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix})^2 = \quad \begin{bmatrix}33.64 & 302.76\end{bmatrix} \tag{11b}$$

1.3.2 Backward Pass

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(2)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(2)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(2)} - \mathbf{y}) = \quad 2 \cdot (\begin{bmatrix}8.8 & -16.4\end{bmatrix} - \begin{bmatrix}3 & 1\end{bmatrix}) = \quad \begin{bmatrix}11.6 & -34.8\end{bmatrix} \tag{12a}$$

$$\frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(2)}, \mathbf{y})}{\partial \mathbf{w}^{(2)}} = \quad \frac{\partial \mathscr{L}(\hat{\mathbf{y}}^{(2)}, \mathbf{y})}{\partial \hat{\mathbf{y}}^{(2)}} \cdot \frac{\partial \hat{\mathbf{y}}^{(2)}}{\partial \mathbf{w}^{(2)}} = \quad 2 \cdot (\hat{\mathbf{y}}^{(1)} - \mathbf{y}) \cdot \mathbf{X} = \quad \begin{bmatrix}[11.6] & [-34.8]\end{bmatrix} \cdot \begin{bmatrix}\begin{bmatrix}2\\ -1\end{bmatrix} \begin{bmatrix}-1\\ 3\end{bmatrix}\end{bmatrix} = \quad \begin{bmatrix}\begin{bmatrix}23.2\\ -11.6\end{bmatrix}^T & \begin{bmatrix}34.8\\ -104.4\end{bmatrix}^T\end{bmatrix} \tag{12b}$$

1.3.3 Velocity and gradient update

The gradient approximation:

$$\mathbf{g} = \frac{1}{2}\nabla_{\mathbf{w}}\sum_{i=1}^{2}\mathscr{L}(\hat{y}^{(2)}, y) = 0.5 \cdot \begin{bmatrix} 23.2 + 34.8 \\ -11.6 - 104.4 \end{bmatrix}^T = \begin{bmatrix} 29 \\ -58 \end{bmatrix}^T \tag{13}$$

The new updated velocity is

$$\mathbf{v}^{(3)} = \beta \cdot \mathbf{v}^{(2)} - \alpha \cdot \mathbf{g} = \quad 0.8 \cdot \begin{bmatrix} 1 \\ -3.8 \end{bmatrix}^T - 0.2 \cdot \begin{bmatrix} 29 \\ -58 \end{bmatrix}^T = \quad \begin{bmatrix} 6.6 \\ -14.64 \end{bmatrix}^T \tag{14}$$

Adding this velocity to the gradient gives us the first gradient update

$$\mathbf{w}^{(3)} = \mathbf{w}^{(2)} + \mathbf{v}^{(3)} = \begin{bmatrix} 2 \\ -4.8 \end{bmatrix}^T + \begin{bmatrix} 6.6 \\ -14.64 \end{bmatrix}^T = \begin{bmatrix} 8.6 \\ -19.44 \end{bmatrix}^T \tag{15}$$

2 Proof

We are given that the gradient doesn't change over a certain number of time steps n. To prove that the bias corrected first order moment equals the gradient when gradient doesnt change, we do the following:
Case 1: $n = 1, s_0 = zeroslike(g)$

$$s_1 = \rho \cdot s_0 + (1-\rho) \cdot g = 0 + (1-\rho) \cdot g \tag{16a}$$

$$\hat{s}_1 = \frac{s_1}{1-\rho^1} = \frac{(1-\rho) \cdot g}{(1-\rho)} = g \tag{16b}$$

Case 2: $n = 2, s_1 = (1-\rho) \cdot g$

$$s_2 = \rho \cdot s_1 + (1-\rho) \cdot g = \rho \cdot (1-\rho) \cdot g + (1-\rho) \cdot g = g \cdot (1-\rho^2) \tag{17a}$$

$$\hat{s}_2 = \frac{s_2}{1-\rho^2} = \frac{(1-\rho^2) \cdot g}{(1-\rho^2)} = g \tag{17b}$$

Case 3: $n = 3, s_2 = (1-\rho^2) \cdot g$

$$s_3 = \rho \cdot s_2 + (1-\rho) \cdot g = \rho \cdot (1-\rho^2) \cdot g + (1-\rho) \cdot g \tag{18a}$$

$$= g \cdot (1-\rho) \cdot (1+\rho+\rho^2) = g(1-\rho^3) \qquad \text{from algebraic identity for cubic polynomials} \tag{18b}$$

$$\hat{s}_3 = \frac{s_3}{1-\rho^3} = \frac{(1-\rho^3) \cdot g}{(1-\rho^3)} = g \tag{18c}$$

Case 4: : $n = 4, s_3 = (1-\rho^3) \cdot g$

$$s_4 = \rho \cdot s_3 + (1-\rho) \cdot g = \rho \cdot g \cdot (1-\rho) \cdot (1+\rho+\rho^2) + (1-\rho) \cdot g \tag{19a}$$

$$= g \cdot (1-\rho) \cdot (1+\rho+\rho^2+\rho^{^3}) = g(1-\rho^4) \qquad \text{on expansion} \tag{19b}$$

$$\hat{s}_4 = \frac{s_4}{1-\rho^4} = \frac{(1-\rho^4) \cdot g}{(1-\rho^4)} = g \tag{19c}$$

Thus by induction we can conclude that for as long as the gradient stays constant over the steps, our bias corrected first moment will be equal to it