# Problem 6.10

May 10, 2024

## 1 Formula

$$\boldsymbol{m}_{t+1} \leftarrow \beta \cdot \boldsymbol{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$
$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \boldsymbol{m}_{t+1}$$

## 2 Problem

Show that the momentum term $\boldsymbol{m}_t$ is an infinite weighted sum of the gradients at the previous iterations and derive an expression for the coefficients (weights) of that sum.

## 3 Answer

$$\boldsymbol{m}_{t+2} \leftarrow \beta \cdot \beta \cdot \left( \boldsymbol{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \right) + (1-\beta) \sum_{i \in \mathcal{B}_{t+1}} \frac{\partial l_i[\boldsymbol{\phi}_{t+1}]}{\partial \boldsymbol{\phi}}$$

$$\boldsymbol{m}_{t+3} \leftarrow \beta \cdot \left( \beta \cdot \beta \cdot \left( \boldsymbol{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \right) + (1-\beta) \sum_{i \in \mathcal{B}_{t+1}} \frac{\partial l_i[\boldsymbol{\phi}_{t+1}]}{\partial \boldsymbol{\phi}} \right) + (1-\beta) \sum_{i \in \mathcal{B}_{t+2}} \frac{\partial l_i[\boldsymbol{\phi}_{t+2}]}{\partial \boldsymbol{\phi}}$$

$$\boldsymbol{m}_{t+n} \leftarrow \beta^n \cdot \boldsymbol{m}_t + \sum_{j=1}^{n} \beta^j \cdot \left( (1-\beta) \sum_{i \in \mathcal{B}_{t+j}} \frac{\partial l_i[\boldsymbol{\phi}_{t+j}]}{\partial \boldsymbol{\phi}} \right)$$

Therefore the weights for the gradients are $w_j$

$$= \beta^j \cdot (1-\beta)$$

# 4 Bard's Explanation

## 4.1 Unfolding the Recursion

We can prove this using mathematical induction and unfolding the recursion for $m_t$

Base Case ($t = 1$): We don't have a previous $m_0$ term in the update rule for $m_1$. So, $m_1$ only considers the gradient at the first iteration:

$$\boldsymbol{m}_1 = (1-\beta) \sum_{i \in \mathcal{B}_1} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Induction Hypothesis

$$\boldsymbol{m}_t = \beta \cdot \boldsymbol{m}_{t-1} + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Induction Step

$$\boldsymbol{m}_{t+1} = \beta \left( \beta \cdot \boldsymbol{m}_{t-2} + (1-\beta) \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial l_i[\boldsymbol{\phi}_{t-1}]}{\partial \boldsymbol{\phi}} \right) + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Rearranging

$$\boldsymbol{m}_{t+1} = \beta^2 \cdot \boldsymbol{m}_{t-2} + (1-\beta)\beta \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial l_i[\boldsymbol{\phi}_{t-1}]}{\partial \boldsymbol{\phi}} + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Weights of the gradient

$$\text{weight}_k = (\beta^{t-k}) \cdot (1-\beta)$$