

Different Momentum Update Rules

Chih-Jen Lin Cheng-Hung Liu Li-Chung Lin
National Taiwan University

Last updated: November 25, 2022

Introduction

- We found that existing software may use different momentum update rules.
- Here we summarize the differences.
- Both `simpleNN/MATLAB/opt/sgd.m` and `tf.keras.optimizers.SGD` use one update rule, which is the same as the paper (Polyak, 1964; Sutskever et al., 2013).
- PyTorch and `tf.compat.v1.train.MomentumOptimizer` use another.



Two Update Rules

- Suppose \mathbf{g} is the gradient, η is the learning rate, and the parameter $\alpha \in [0, 1)$ are introduced.
- Update rule in simpleNN/MATLAB/opt/sgd.m:

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \mathbf{g}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{v}$$

- PyTorch update rule in torch.optim.SGD:

$$\mathbf{v} \leftarrow \alpha \mathbf{v} + \mathbf{g}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{v}$$



First Rule: $\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \mathbf{g}; \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{v}$

$$\mathbf{v}_1 = \alpha \mathbf{v}_0 - \eta_1 \mathbf{g}_1 = -\eta_1 \mathbf{g}_1$$

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \eta_1 \mathbf{g}_1$$

$$\mathbf{v}_2 = -\alpha \eta_1 \mathbf{g}_1 - \eta_2 \mathbf{g}_2$$

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_0 - (\eta_1 + \alpha \eta_1) \mathbf{g}_1 - \eta_2 \mathbf{g}_2$$

$$\mathbf{v}_3 = -\alpha^2 \eta_1 \mathbf{g}_1 - \alpha \eta_2 \mathbf{g}_2 - \eta_3 \mathbf{g}_3$$

$$\boldsymbol{\theta}_3 = \boldsymbol{\theta}_0 - (\eta_1 + \alpha \eta_1 + \alpha^2 \eta_1) \mathbf{g}_1 - (\eta_2 + \alpha \eta_2) \mathbf{g}_2 - \eta_3 \mathbf{g}_3$$

$$\mathbf{v}_4 = -\alpha^3 \eta_1 \mathbf{g}_1 - \alpha^2 \eta_2 \mathbf{g}_2 - \alpha \eta_3 \mathbf{g}_3 - \eta_4 \mathbf{g}_4$$

$$\begin{aligned} \boldsymbol{\theta}_4 = \boldsymbol{\theta}_0 &- (\eta_1 + \alpha \eta_1 + \alpha^2 \eta_1 + \alpha^3 \eta_1) \mathbf{g}_1 \\ &- (\eta_2 + \alpha \eta_2 + \alpha^2 \eta_2) \mathbf{g}_2 - (\eta_3 + \alpha \eta_3) \mathbf{g}_3 - \eta_4 \mathbf{g}_4 \end{aligned}$$



Second Rule: $\mathbf{v} \leftarrow \alpha \mathbf{v} + \mathbf{g}; \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{v}$

$$\mathbf{v}_1 = \alpha \mathbf{v}_0 + \mathbf{g}_1 = \mathbf{g}_1$$

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \eta_1 \mathbf{g}_1$$

$$\mathbf{v}_2 = \alpha \mathbf{g}_1 + \mathbf{g}_2$$

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_0 - (\eta_1 + \alpha \eta_2) \mathbf{g}_1 - \eta_2 \mathbf{g}_2$$

$$\mathbf{v}_3 = \alpha^2 \mathbf{g}_1 + \alpha \mathbf{g}_2 + \mathbf{g}_3$$

$$\boldsymbol{\theta}_3 = \boldsymbol{\theta}_0 - (\eta_1 + \alpha \eta_2 + \alpha^2 \eta_3) \mathbf{g}_1 - (\eta_2 + \alpha \eta_3) \mathbf{g}_2 - \eta_3 \mathbf{g}_3$$

$$\mathbf{v}_4 = \alpha^3 \mathbf{g}_1 + \alpha^2 \mathbf{g}_2 + \alpha \mathbf{g}_3 + \mathbf{g}_4$$

$$\boldsymbol{\theta}_4 = \boldsymbol{\theta}_0 - (\eta_1 + \alpha \eta_2 + \alpha^2 \eta_3 + \alpha^3 \eta_4) \mathbf{g}_1$$

$$- (\eta_2 + \alpha \eta_3 + \alpha^2 \eta_4) \mathbf{g}_2 - (\eta_3 + \alpha \eta_4) \mathbf{g}_3 - \eta_4 \mathbf{g}_4$$



Compariton of the Fourth Iteration

- Rule 1: weight for \mathbf{g}_k depends only on η_k .

$$\mathbf{v}_4 = -\alpha^3\eta_1\mathbf{g}_1 - \alpha^2\eta_2\mathbf{g}_2 - \alpha\eta_3\mathbf{g}_3 - \eta_4\mathbf{g}_4$$

$$\begin{aligned}\boldsymbol{\theta}_4 = \boldsymbol{\theta}_0 &- (\eta_1 + \alpha\eta_1 + \alpha^2\eta_1 + \alpha^3\eta_1)\mathbf{g}_1 \\ &- (\eta_2 + \alpha\eta_2 + \alpha^2\eta_2)\mathbf{g}_2 - (\eta_3 + \alpha\eta_3)\mathbf{g}_3 \\ &- \eta_4\mathbf{g}_4\end{aligned}$$

- Rule 2: weight for \mathbf{g}_k depends on $\eta_k, \eta_{k+1}, \dots$

$$\mathbf{v}_4 = \alpha^3\mathbf{g}_1 + \alpha^2\mathbf{g}_2 + \alpha\mathbf{g}_3 + \mathbf{g}_4$$

$$\begin{aligned}\boldsymbol{\theta}_4 = \boldsymbol{\theta}_0 &- (\eta_1 + \alpha\eta_2 + \alpha^2\eta_3 + \alpha^3\eta_4)\mathbf{g}_1 \\ &- (\eta_2 + \alpha\eta_3 + \alpha^2\eta_4)\mathbf{g}_2 - (\eta_3 + \alpha\eta_4)\mathbf{g}_3 \\ &- \eta_4\mathbf{g}_4\end{aligned}$$



First Update Rule: from θ_{k-1} to θ_k

- Update rule

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \mathbf{g}$$

$$\theta \leftarrow \theta + \mathbf{v}$$

$$\begin{aligned}\theta_k &= \theta_{k-1} + \mathbf{v}_k \\&= \theta_{k-1} - \eta_k \mathbf{g}_k + \alpha \mathbf{v}_{k-1} \\&= \theta_{k-1} - \eta_k \mathbf{g}_k + \alpha(\alpha \mathbf{v}_{k-2} - \eta_{k-1} \mathbf{g}_{k-1}) \\&= \theta_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_{k-1} \mathbf{g}_{k-1} + \alpha^2 \mathbf{v}_{k-2} \\&= \theta_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_{k-1} \mathbf{g}_{k-1} - \alpha^2 \eta_{k-2} \mathbf{g}_{k-2} \\&\quad - \alpha^3 \eta_{k-3} \mathbf{g}_{k-3} - \dots\end{aligned}$$



Second Update Rule: from θ_{k-1} to θ_k

- Update rule

$$\mathbf{v} \leftarrow \alpha \mathbf{v} + \mathbf{g}$$

$$\theta \leftarrow \theta - \eta \mathbf{v}$$

$$\begin{aligned}\theta_k &= \theta_{k-1} - \eta_k \mathbf{v}_k \\&= \theta_{k-1} - \eta_k (\alpha \mathbf{v}_{k-1} + \mathbf{g}_k) \\&= \theta_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k \mathbf{v}_{k-1} \\&= \theta_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k (\alpha \mathbf{v}_{k-2} + \mathbf{g}_{k-1}) \\&= \theta_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k \mathbf{g}_{k-1} - \alpha^2 \eta_k \mathbf{g}_{k-2} \\&\quad - \alpha^3 \eta_k \mathbf{g}_{k-3} - \dots\end{aligned}$$



Differences I

- Rule 1's η_k affects the importance of \mathbf{g}_k , while η_{k-1} affects the importance of $\alpha \mathbf{g}_{k-1}$ and so on

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_{k-1} \mathbf{g}_{k-1} - \alpha^2 \eta_{k-2} \mathbf{g}_{k-2} - \cdots$$

- Rule 2's η_k affects $\mathbf{g}_k, \alpha \mathbf{g}_{k-1}, \alpha^2 \mathbf{g}_{k-2} \cdots$

$$\begin{aligned}\boldsymbol{\theta}_k &= \boldsymbol{\theta}_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k \mathbf{g}_{k-1} - \alpha^2 \eta_k \mathbf{g}_{k-2} - \cdots \\ &= \boldsymbol{\theta}_{k-1} - \eta_k (\mathbf{g}_k + \alpha \mathbf{g}_{k-1} + \alpha^2 \mathbf{g}_{k-2} + \cdots)\end{aligned}$$

- For rule 2, if η_k is decreased from η_{k-1} , weights for past gradients are immediately affected



Differences II

- For rule 1, weights of past gradients changed only because of α ?
- Both formulas are the same if NO learning-rate scheduling



References I

- Polyak, B.T. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964.
- Sutskever, Ilya, Martens, James, Dahl, George, and Hinton, Geoffrey. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning (ICML), pp. 1139–1147, 2013.
- <https://github.com/pytorch/pytorch/issues/1099>



References II

- <https://medium.com/the-artificial-impostor/sgd-implementation-in-pytorch-4115bcb9f02c>
- `tf.keras.optimizers.SGD`
- `tf.compat.v1.train.MomentumOptimizer`
- `simpleNN/MATLAB/opt/sgd.m`
- PyTorch

