Different Momentum Update Rules

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

Last updated: April 10, 2021

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:

$$\begin{aligned} \boldsymbol{v} &\leftarrow \alpha \boldsymbol{v} - \eta \mathbf{g} \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \boldsymbol{v} \end{aligned}$$

- PyTorch update rule in torch.optim.SGD:

$$\begin{aligned} \boldsymbol{v} &\leftarrow \alpha \boldsymbol{v} + \mathbf{g} \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} - \eta \boldsymbol{v} \end{aligned}$$

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

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

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

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

Compariton of the Fourth Iteration

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

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

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

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

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

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

First Update Rule: from $\boldsymbol{\theta}_{k-1}$ to $\boldsymbol{\theta}_k$

- Update rule

$$\begin{aligned} \boldsymbol{v} &\leftarrow \alpha \boldsymbol{v} - \eta \mathbf{g} \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \boldsymbol{v} \end{aligned}$$

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

Second Update Rule: from $\boldsymbol{\theta}_{k-1}$ to $\boldsymbol{\theta}_k$

- Update rule

$$\begin{aligned} \boldsymbol{v} &\leftarrow \alpha \boldsymbol{v} + \mathbf{g} \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} - \eta \boldsymbol{v} \end{aligned}$$

$$\begin{aligned} \boldsymbol{\theta}_k &= \boldsymbol{\theta}_{k-1} - \eta_k \boldsymbol{v}_k \\ &= \boldsymbol{\theta}_{k-1} - \eta_k (\alpha \boldsymbol{v}_{k-1} + g_k) \\ &= \boldsymbol{\theta}_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k \boldsymbol{v}_{k-1} \\ &= \boldsymbol{\theta}_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k (\alpha \boldsymbol{v}_{k-2} + \boldsymbol{g}_{k-1}) \\ &= \boldsymbol{\theta}_{k-1} - \eta_k \mathbf{g}_k - \alpha \eta_k \mathbf{g}_{k-1} - \alpha^2 \eta_k \mathbf{g}_{k-2} \\ &\quad - \alpha^3 \eta_k \mathbf{g}_{k-3} - \cdots \end{aligned}$$

Differences I

- Rule 1's η_k affects the importance of $\boldsymbol{g}_k$, while η_{k-1} affects the importance of $\alpha\boldsymbol{g}_{k-1}$ 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 $\boldsymbol{g}_k$, $\alpha\boldsymbol{g}_{k-1}$, $\alpha^2\boldsymbol{g}_{k-2}\cdots$

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

- For rule 2, 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 $\alpha^?$
- 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