Nesterov's Momentum Made Simple

Ivo Danihelka

August 25, 2012

Abstract

A simple update rule is derived. The rule does the same update as Nesterov's momentum.

1 Nesterov's Momentum

A part of Nesterov's accelerated gradient is using of a momentum. Nesterov's momentum uses the following update rule:

$$v_{t+1} = \mu v_t - \epsilon g(\theta_t + \mu v_t)$$
$$\theta_{t+1} = \theta_t + v_{t+1}$$

Used notation:
θ ... a vector of parameters.
v ... velocity vector. The vector has the same shape as θ.
μ ... momentum factor. $\mu \in [0, 1]$
ϵ ... learning rate. $\epsilon \in [0, 1]$
$g(\theta)$ ... gradient at point θ.

The gradient is evaluated at a shifted point: $g(\theta_t + \mu v_t)$

2 Simplification

We can decide to always evaluate the network at the shifted point:

$$\theta'_t = \theta_t + \mu v_t$$

The update rule is then:

$$\begin{aligned} v_{t+1} &= \mu v_t - \epsilon g(\theta'_t) \\ \theta'_{t+1} &= \theta'_t - \mu v_t + v_{t+1} + \mu v_{t+1} \\ &= \theta'_t - v_{t+1} - \epsilon g(\theta'_t) + v_{t+1} + \mu v_{t+1} \\ &= \theta'_t - \epsilon g(\theta'_t) + \mu v_{t+1} \end{aligned}$$

3 Summary

Standard Momentum:

$$v_{t+1} = \mu v_t - \epsilon g(\theta_t)$$
$$\theta_{t+1} = \theta_t + v_{t+1}$$

Nesterov's Momentum:

$$v_{t+1} = \mu v_t - \epsilon g(\theta'_t)$$
$$\theta'_{t+1} = \theta'_t - \epsilon g(\theta'_t) + \mu v_{t+1}$$

4 Resources

1) "On the importance of initialization and momentum in deep learning"