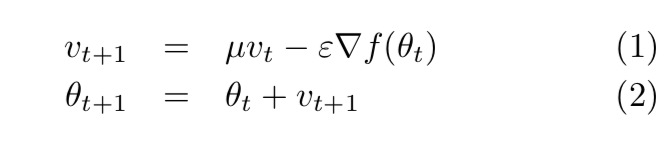
On the importance of initialization and momentum in deep learning

1. Introduction

In particular, we study the effectiveness of SGD when combined with well-chosen initialization schemes and various forms of momentum-based acceleration.

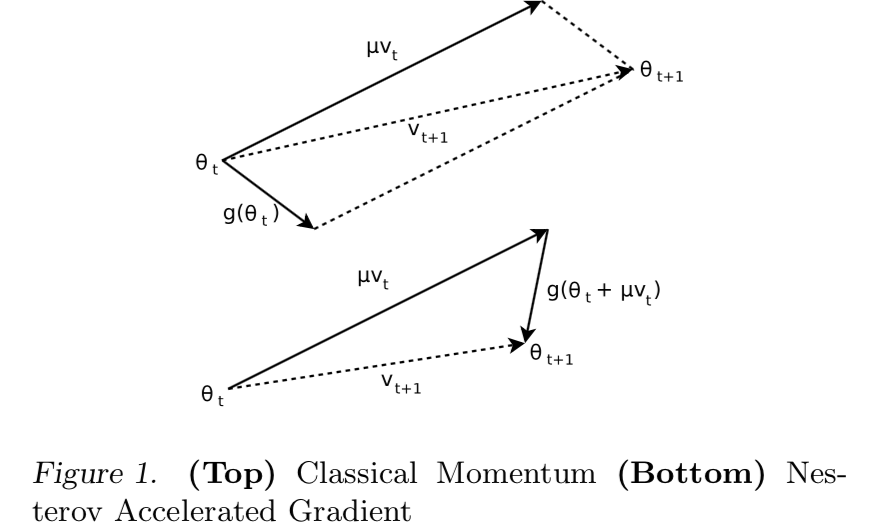
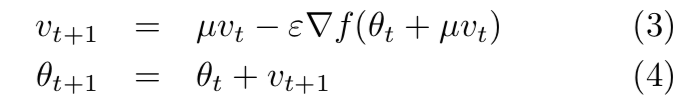
2. Momentum and Nesterov’s Accelerated Gradient

The momentum method, which we refer to as classical momentum (CM), is a technique for accelerating gradient descent that accumulates a velocity vector in directions of persistent reduction in the objective across iterations. Given an objective function f(θ) to be minimized, classical momentum is given by:



where ε > 0 is the learning rate, μ ∈ [0,1] is the momentum coefficient, and ∇f(θt) is the gradient at θt.

Like momentum, NAG is a first-order optimization method with better convergence rate guarantee than gradient descent in certain situations.

The NAG update may be rewritten as: 

2.1. The Relationship between CM and NAG

While CM computes the gradient update from the current position θt, NAG first performs a partial update to θt, computing θt + μvt, which is similar to θt+1, but missing the as yet unknown correction. This benign-looking difference seems to allow NAG to change v in a quicker and more responsive way, letting it behave more stably than CM in many situations, especially for higher values of μ.

3. Deep Autoencoders

The aim of our experiments is three-fold. First, to investigate the attainable performance of stochastic momentum methods on deep autoencoders starting from well-designed random initializations; second, to explore the importance and effect of the schedule for the momentum parameter μ assuming an optimal fixed choice of the learning rate ε; and third, to compare the performance of NAG versus CM.

3.1. Random Initializations

4. Recurrent Neural Networks

Echo-State Networks (ESNs) is a family of RNNs with an unusually simple training method: their hidden-to- output connections are learned from data, but their re- current connections are fixed to a random draw from a specific distribution and are not learned.

4.1. ESN-based Initialization

4.2. Experimental Results

5. Momentum and HF

6. Discussion