

# Stochastic Gradient Descent: Going As Fast As Possible But Not Faster

Michele Sebag

Laboratoire de Recherche en Informatique, CNRS, France /

Université Paris Sud, France

sebag@lri.fr

**Abstract:** When applied to training deep neural networks, stochastic gradient descent (SGD) often incurs steady progression phases, interrupted by catastrophic episodes in which loss and gradient norm explode. A possible mitigation of such events is to slow down the learning process.

This paper presents a novel approach to control the SGD learning rate that uses two statistical tests. The first one, aimed at fast learning, compares the momentum of the normalized gradient vectors to that of random unit vectors and accordingly gracefully increases or decreases the learning rate. The second one is a change point detection test, aimed at the detection of catastrophic learning episodes; upon its triggering, the learning rate is instantly halved.

Both abilities of speeding up and slowing down the learning rate allows the proposed approach, called S-AGMA, to learn as fast as possible but not faster. Experiments on real-world benchmarks show that S-AGMA performs well in practice, and compares favorably to the state of the art.

Joint work: Alice Schoenauer-Sebag, Marc Schoenauer.