Stochastic gradient descent with adaptive step-sizes: from practice to theory

Rachel Ward
University of Texas at Austin, USA
rward@math.utexas.edu

Stochastic gradient descent (SGD) is the foundational optimization algorithm used to train neural networks, and variations of SGD where the step-sizes are updated adaptively based on past stochastic gradient information are a crucial part of this success. Recently, a theoretical understanding of the behavior of stochastic gradient descent with adaptive step-sizes has emerged, shedding light on why adaptivity is so powerful in practice.

We focus on the simplest adaptive gradient algorithm called Adagrad-norm, and recall how Adagrad-norm can match the optimal convergence rates of carefully-tuned SGD. We then discuss recent results which show that even when the optimization landscape is not globally smooth — as arises even in simple neural networks — Adagrad-norm can adapt and converge at an optimal rate without additional hyperparameter tuning.