High-probability Convergence Bounds for Non-convex Stochastic Gradient Descent, with applications to learning

Stephen Becker
Department of Applied Mathematics, University of Colorado Boulder
stephen.becker@colorado.edu

Stochastic gradient descent is one of the most common iterative algorithms used in machine learning. While being computationally cheap to implement, recent literature suggests it may have implicit regularization properties that prevent over-fitting. This paper analyzes the properties of stochastic gradient descent from a theoretical standpoint to help bridge the gap between theoretical and empirical results. We specifically tackle the case of heavy-tailed noise, since recent results have shown empirically that noise due to mini-batch sampling can be non-Gaussian.

Most theoretical results either assume convexity or only provide convergence results in mean, while this paper proves convergence bounds in high probability without assuming convexity. By high-probability, we mean that our bounds are of the form "with probability at least $1 - \delta$, $\text{error}_k \leq g(k, \delta)$", for some function g (decreasing in the number of iterations k) that depends at most polynomially on $\log(\delta^{-1})$, rather than on δ^{-1}.

Assuming strong smoothness, we prove high probability convergence bounds in two settings:

1. assuming the Polyak-Łojasiewicz inequality and norm sub-Gaussian gradient noise, and
2. assuming norm sub-Weibull gradient noise.

In the first setting, in the setting of statistical learning, we combine our convergence bounds with existing generalization bounds based on algorithmic stability in order to bound the true risk and show that for a certain number of epochs, convergence and generalization balance in such a way that the true risk goes to the empirical minimum as the number of samples goes to infinity.

In the second setting, as an intermediate step to proving convergence, we prove a probability result of independent interest. The probability result extends Freedman-type concentration beyond the sub-exponential threshold to heavier-tailed martingale difference sequences.

Joint work with: Liam Madden, Emiliano Dall'Anese.

References

[1] L. Madden, E. Dall'Anese, S. Becker. High-probability Convergence Bounds for Non-convex Stochastic Gradient Descent. *arXiv* https://arxiv.org/abs/2006.05610, 2021.