CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 144 **May 14, 2024**

SGD vs GD: high noise and rank shrinkage

Mengjia Xu[1,2], Tomer Galanti[1], Akshay Rangamani[1], Lorenzo Rosasco[1,3,4], Andrea Pinto[1], Tomaso Poggio[1,*]

[1]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Data Science, New Jersey Institute of Technology, Newark, NJ, USA
[3]MaLGaCenter - DIBRIS - Universit'a di Genova, Genoa, Italy
[4]MaLGaCenter - DIMA - Universit'a di Genova, Genoa, Italy

Abstract

It was always obvious that SGD with small minibatch size yields for neural networks much higher asymptotic fluctuations in the updates of the weight matrices than GD. It has also been often reported that SGD in deep RELU networks shows empirically a low-rank bias in the weight matrices. A recent theoretical analysis derived a bound on the rank and linked it to the size of the SGD fluctuations [25]. In this paper, we provide an empirical and theoretical analysis of the convergence of SGD vs GD, first for deep RELU networks and then for the case of linear regression, where sharper estimates can be obtained and which is of independent interest. In the linear case, we prove that the component $W^{\perp}$ of the weight matrix W, corresponding to the null space of the data matrix X, converges to zero for both SGD and GD, provided the regularization term is non-zero. In particular, this guarantees recovery of the correct rank, which also means recovery of the support of the function, that is elimination of the weights associated with irrelevant variables. Because of the larger number of updates required to go through all the training data, the convergence rate *per epoch* is much faster for SGD than for GD. In practice, SGD has a much stronger bias than GD towards solutions for weight matrices W with high fluctuations – even when the choice of mini batches is deterministic – and low rank, provided the initialization is from a random matrix. Thus SGD with non-zero regularization, shows the coupled phenomenon of asymptotic noise and a specific low-rank bias– unlike GD.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.