Online learning with stochastic gradient descent

Denys Sobchyshak
denys.sobchyshak@tum.de

Technical University Munich

December 3, 2014

EXTENDED ABSTRACT

As of the last century amount of data has shown a steady growth, which exceeds that of processor speed and data transfer bandwidth in both data storage and compute unit interconnects. This tendency has proven "batch" based learning methods, where system learns from the whole data set at once, to be computationally infeasible in the context of large-scale statistical learning, where capabilities of the methods are limited by the computing time and not by the data size, as in [1].

Search for better approach has led to online learning, where our system processes single data instance at a time. This idea is incorporated in stochastic gradient descent(SGD) based methods and in comparison to standard or "batch" gradient descent(GD), while being characterized by a poor optimization efficiency, it shows exceptional performance on large-scale problems. Thus, we will take a look at results where second order and averaged SGDs are shown to be asymptotically efficient already after one pass on training data[1], as compared to GD.

Furthermore, since production scale data sets today reach up to 100TB and beyond, one needs to go above sequential algorithmic limits in order to increase data processing efficiency. However, this task is not trivial since SGD has inherently sequential nature. We will present an overview of this matter and investigate some of the most know remedies. In particular, we will take a look at parallelized asynchronous SGD strategy named Hogwild![2] and a Downpour SGD[3] in addition to making a soft touch on MapReduce usage within online learning setting[4].

Finally, empirical results of former topics will be presented, leading to a conclusion of how well parallelized asynchronous SGD can perform in large-scale setting.

References

[1] Léon Bottou: *Large-Scale Machine Learning with Stochastic Gradient Descent*, Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), 177–187, August 2010.

[2] Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright: *Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent*, Advances in Neural Information Processing Systems, 2011.

[3] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng: *Large Scale Distributed Deep Networks*, Advances in Neural Information Processing Systems, 2012.

[4] Martin Zinkevich, Alexander J. Smola, Markus Weimer, and Lihong Li: *Parallelized stochastic gradient descent*, Advances in Neural Information Processing Systems, 2010.