Proposal for CS798, 2016 Fall
Optimization for Machine Learning

Hemant Saxena
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3E6
h2saxena@uwaterloo.ca

Royal Sequiera
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3E6
rdsequie@uwaterloo.ca

Abstract

The main idea of the project is to discuss the convergence and accuracy of the accelerated proximal gradient algorithm [1] over distributed computing framework. The project will focus on two types of distributed computing frameworks: bulk synchronous parallel (BSP) systems, and stale synchronous parallel (SSP) system. The convergence rate and the accuracy over the BSP and SSP systems will be compared against the single node implementation of the algorithm. The implementation of the BSP version will be tested over Spark [2] and the SSP version will be tested over Petuum [3].

1 Introduction

Due to the increasing volume of data most of the computation is being parallelized across multiple computing compute nodes. This project will discuss the convergence and the accuracy of parallelizing one of the gradient descent algorithms, the accelerated proximal gradient(APG) algorithm. APG algorithm is the optimal gradient descent algorithm with respect to convergence rate, which is $O(1/t^2)$ where t is iteration count. The parallel implementation of the algorithm will be tested over bulk synchronous parallel (BSP) systems, and stale synchronous parallel (SSP) system. BSP systems enforces synchronization across worker nodes while progressing through intermediate stages, which makes the system slow due to stragglers. On the other hand, SSP systems compromise consistency between workers and allow them to operate asynchronously, which makes the system fast but with bounded errors.

The project is motivated by the recent interest in parallelizing gradient descent algorithms. In one of the recent works [4], authors discussed the parallel implementation of proximal gradient algorithm. Following that work, the next natural step would be to test the parallel implementation of the APG algorithm.

2 Related Work

The work on Stale Synchronous can be broadly divided into two major sets: (i) SSP systems, where individual machines skip updates while solving an optimization problem [5–9]. (ii) SSP systems, where machines do not allowed skip updates [10–14].

Early research in this field, however, started in the late 1980s [6–9]. Recently, Zhou et al. proposed **msPG**, an extension to the proximal gradient algorithm to the model parallel and stale synchronous setting[4]. The authors showed that **msPG** converges to a critical point under mild assumptions and such a critical point is optimal under convexity assumptions.

3 Proposed Work

We divide the work into three subtasks: first, implement parallel APG for Spark, second, implement parallel APG for Petuum, third, and compare the performance of both the implementations with respect to a single node implementation. We will test the APG algorithm over a non-convex Lasso problem. The data will be generated from $\mathcal{N}(0, 1)$ withe normalized columns.

4 Team

Hemant: APG implementation for Spark. Royal: APG implementation for Petuum. Experiments will be designed and conducted jointly.

Acknowledgement

We thank Professor Yaoliang Yu for his invaluable guidance, which led to the conception of this project.

References

[1] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11): 2419–2434, 2009.

[2] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association. URL http://dl.acm.org/citation.cfm?id=1863103.1863113.

[3] Eric P. Xing, Qirong Ho, Wei Dai, Jin-Kyu Kim, Jinliang Wei, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanu Kumar, and Yaoliang Yu. Petuum: A new platform for distributed machine learning on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1335–1344, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783323. URL http://doi.acm.org/10.1145/2783258.2783323.

[4] Yi Zhou, Yaoliang Yu, Wei Dai, Yingbin Liang, and Eric P Xing. On convergence of model parallel proximal gradient algorithm for stale synchronous parallel system. *algorithms*, 6(11): 18, 2016.

[5] Hamid Reza Feyzmahdavian and Mikael Johansson. On the convergence rates of asynchronous iterations. In *53rd IEEE Conference on Decision and Control*, pages 153–159. IEEE, 2014.

[6] Dimitri P Bertsekas and John N Tsitsiklis. Convergence rate and termination of asynchronous iterative algorithms. In *Proceedings of the 3rd international conference on Supercomputing*, pages 461–470. ACM, 1989.

[7] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[8] Paul Tseng. On the rate of convergence of a partially asynchronous gradient projection algorithm. *SIAM Journal on Optimization*, 1(4):603–619, 1991.

[9] John N Tsitsiklis, Dimitri P Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. In *1984 American Control Conference*, pages 484–489, 1984.

[10] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 583–598, 2014.

[11] Mu Li, David G Andersen, and Alexander Smola. Distributed delayed proximal gradient methods. In *NIPS Workshop on Optimization for Machine Learning*, 2013.

[12] Mu Li, Li Zhou, Zichao Yang, Aaron Li, Fei Xia, David G Andersen, and Alexander Smola. Parameter server for distributed machine learning. In *Big Learning NIPS Workshop*, volume 6, page 2, 2013.

[13] Hamid Reza Feyzmahdavian, Arda Aytekin, and Mikael Johansson. A delayed proximal gradient method with linear convergence rate. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

[14] Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in neural information processing systems*, pages 1223–1231, 2013.