Efficient Distributed Stochastic Dual Coordinate Ascent

Mingrui Liu, Jeff Hajewski
Department of Computer Science
University of Iowa
Iowa City, IA 52242
mingrui-liu@uiowa.edu, jeffery-hajewski@uiowa.edu

Abstract

To be completed after finishing the main contents of the proposal.

1 Introduction

In recent years, we come into the big data era. Many large-scale machine learning problems, most of which are essentially optimization problems with huge magnitude of data size, need to be tackled. Two common countermeasures to deal with this are employing stochastic optimization algorithms, and utilizing computational resources in a parallel or distributed manner[4].

In this paper, we consider a class of convex optimization problems with special structure, whose objective can be expressed as the sum of a finite sum of loss functions and a regularization function:

$$\min_{w\in\mathbb{R}^d} P(w), \text{where } P(w) = \frac{1}{n}\sum_{i=1}^{n}\phi(w^\top x_i, y_i) + \lambda g(w), \tag{1}$$

where $w \in \mathbb{R}^d$ denotes the weight vector, $(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, $i = 1, \ldots, n$ are training data, $\lambda > 0$ is a regularization parameter, $\phi(z, y)$ is a convex function of z, and $g(w)$ is a convex function of w. We refer to the problem in (1) as Regularized Finite Sum Minimization (RFSM) problem. When $g(w) = 0$, the problem reduces to the Finite Sum Minimization (FSM) problem.

Both RFSM and FSM problems have been extensively studied in machine learning and optimization literature. When n is large, numerous sequential stochastic optimization algorithms were proposed[3, 12, 15, 17, 16, 9, 13, 19, 18, 20, 6, 24, 11, 5, 2, 10], and there also exist several parallel or distributed stochastic algorithms[4, 14, 25, 21, 23, 25, 1]. Specifically, S. Shalv-Shwartz and T. Zhang [17] proposed an Stochastic Dual Coordinate Ascent (SDCA) which provided new analysis with strongly theoretical guarantee regarding the duality gap. T. Yang [21, 22] developed a Distributed Stochastic Dual Coordinate Ascent (DisDCA) algorithm and its practical variant, and analyzed the tradeoff between communication and computation. However, to get a more efficient distributed SDCA is a open problem. In this paper, we first provide a GPU implementation of the vanilla distributed SDCA[21], and then give an asynchronous distributed SDCA to make full use of computational resources.

2 Related Work

First we review the related work of sequential stochastic convex optimization for solving FSM and RFSM problems. The first numerical scheme of stochastic optimization stems from stochastic gradient descent (SGD)[3, 12], which was designed to avoid the calculation of full gradient and gets faster convergence than full gradient descent (FGD). To improve the converge rate of SGD,

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

many new algorithms were proposed by exploiting the finite sum structure, including the Stochastic Average Gradient (SAG)[15], Stochastic Dual Coordinate Ascent (SDCA)[17], Stochastic Variance Reduced Gradient (SVRG)[9], Accelerated Proximal Coordinate method (APCG)[11], SAGA[7], Prox-SDCA[18], Prox-SVRG[20], and Stochastic Primal-dual Coordinate method (SPDC)[24]. Recently, the optimal first-order stochastic optimization method were developed[2, 10]. Although there exist rich literature studying sequential stochastic optimization with strong theoretical guarantee, less efforts have been devoted to considering them in a parallel or distributed manner. It constitutes a huge gap between theory and practice, since nowadays the size of data increases at a rapid speed, which makes one-core processor or one computer very difficult to handle it properly.

Then we review several related work of distributed optimization algorithms. In the existing literature, many distributed algorithms were developed on top of stochastic gradient descent (SGD), alternating direction method of multipliers (ADMM), and stochastic dual coordinate ascent (SDCA). The key idea of distributed SGD is calculating the stochastic gradient at the worker nodes according to the local information while performing parameter updates at the master node. On top of this idea, many approaches were proposed[25, 14, 1]. ADMM stems from [8], which was developed to solve equality constrained optimization problem. Recently, two independent work of stochastic ADMM were proposed[13, 19]. A standard reference for distributed ADMM is [4]. The advances of SDCA algorithms[17] and its variant[16, 18, 11] enjoy faster convergence than SGD and ADMM, and the distributed SDCA (DisDCA)[21, 22] was developed along with novel analysis of tradeoff between computation and communication.

We will build off of work from [CITE YANG 1 & 2], incorporating work from [CITE LI PS]. In [CITE YANG], SDCA is implemented in a distributed, synchronized fashion. While the achieved results were quite promising, they did not take advantage of hardware acceleration or asynchronous communication. [CITE LI PS] builds an asynchronous communication framework using the concept of parameter servers, which are central data stores for model parameters, and distributed workers working in an asynchronous fashion (i.e., communication and parameter updates are non-blocking operations).

3 The Proposed Work

We will approach this problem from both theoretical and implementation perspectives. On the theoretical side we hope to make guarantees on convergence of our proposed approach, while on the implementation side we hope to build a scalable system that can take advantage of its hardware.

3.1 Theory

The theory part will try to establish the convergence result of the proposed asynchronous distributed SDCA and analyze the tradeoff between computation and asynchronous communication.

3.2 Implementation

There are two key aspects to the implementation: taking advantage of the GPU and working in a distributed setting.

3.2.1 GPU

When available, GPU acceleration will be used via CUDA. We will start by using CUDA libraries such as `cuBLAS` for efficient math operations. We will also add CUDA kernels for portions of our algorithms that are easily parallelized in a SIMD fashion.

3.2.2 Asynchronous Communication

In addition to using GPUs for math operations, we will distribute the workload over a cluster of computers. Each comptuer will work on a small subproblem used in the parameter update and send this result to a parameter server, which will handle parameter updates and synchronoization across workers. The architecture used will be similar to that of [CITE PS], using a central parameter server that communicates asynchronously with distributed workers. This asyncrhonous communication framework will be built using MPI and C++ threading facilities.

4 Plan

We will work in parallel, making progress on both the theory and implementation aspects of the project.

4.1 Theory

We will first study the theories and analysis techniques in the existing literature and then try to prove the correctness of naive asynchronous approach. If it works, we will try to establish the convergence rate and analyze the computation and communication cost respectively.

4.2 Implementation

The first part of the implementation approach will be adding CUDA support for the core math operations. The focus is on correctness, rather than optimizing runtime. This is also the simpler portion of the project and will require less time that the asynchronous communciation frmaework.

The second part of the implementation is building the asynchronous communication framework. This will be the most complex part of the implementation and require the most time and effort.

4.3 Comparisions with Other Methods

We plan to conduct the same experiments (with same baselines, same datasets) in [21] to compare our approach with existing approaches.

References

[1] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[2] Z. Allen-Zhu. Katyusha: Accelerated variance reduction for faster sgd. *ArXiv e-prints, abs/1603.05953*, 2016.

[3] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] A. Defazio. A simple practical accelerated method for finite sums. In *Advances In Neural Information Processing Systems*, pages 676–684, 2016.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[7] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *Nips*, pages 1–12, 2014.

[8] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[9] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[10] G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

[11] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pages 3059–3067, 2014.

[12] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[13] H. Ouyang, N. He, L. Tran, and A. G. Gray. Stochastic alternating direction method of multipliers. *ICML (1)*, 28:80–88, 2013.

[14] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[15] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[16] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 378–385, 2013.

[17] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[18] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *ICML*, pages 64–72, 2014.

[19] T. Suzuki et al. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *ICML (1)*, pages 392–400, 2013.

[20] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[21] T. Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 629–637, 2013.

[22] T. Yang, S. Zhu, R. Jin, and Y. Lin. Analysis of distributed stochastic dual coordinate ascent. *arXiv preprint arXiv:1312.1031*, 2013.

[23] R. Zhang and J. T. Kwok. Asynchronous distributed admm for consensus optimization. In *ICML*, pages 1701–1709, 2014.

[24] Y. Zhang and X. Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML*, pages 353–361, 2015.

[25] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.