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

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph. This latex file is modified from the NIPS 2016 template.

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[3].

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:

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

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[2, 13, 10, 15, 14, 7, 11, 17, 16, 18, 5, 21, 9, 4, 1, 8], and there also exist several parallel or distributed stochastic algorithms[3, 12, 22, 19, 20]. Specifically, S. Shalv-Shwartz and T. Zhang [15] proposed an Stochastic Dual Coordinate Ascent (SDCA) which provided strongly theoretical guarantee regarding the duality gap. T. Yang [19] 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[19], 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

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

gradient descent (SGD)[2, 10], which was designed to avoid the calculation of full gradient and gets faster convergence than gradient descent (GD). To improve the converge rate of SGD, many new algorithms were proposed by exploiting the finite sum structure, including the Stochastic Average Gradient (SAG)[13], Stochastic Dual Coordinate Ascent (SDCA)[15], Stochastic Variance Reduced Gradient (SVRG)[7], and SAGA[6], Prox-SDCA[16], Prox-SVRG[18], Stochastic Primal-dual Coordinate method (SPDC)[21]. Recently, the optimal first-order stochastic optimization method were developed[1, 8]. 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 very difficult to handle it properly.

Now we focus on some related work of distributed optimization.

3 The Proposed Work

Describe your proposed work in this section.

4 Plan

Describe your plan for the project. What data you are going to use to evaluate your methods? What are the baselines that you want to compare? How will you develop your methods? A timeline with important milestones is always perferred.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

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

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

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

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

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

[22] 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.