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

We propose the design and analysis of an efficient, distributed, SDCA algorithm that uses GPUs to improve compute efficiency an asynchronous communication for model updates. This work builds off the previous work of T. Yang [23], M. Li [11], and A. Agarwal [1] creating a more scalable and efficient SDCA algorithm. Specifically, we propose to use CUDA to improve the runtime efficiency of the gradient and parameter update calculations along with the use of MPI for network communication. The asynchronous nature of the communication will be handled using standard MPI facilities combined with threading in C++.

1 Introduction

In recent years, the amount and size of available data has grown at an incredible rate. As the size of the data grows, the challenge of applying standard machine learning algorithms to the data has become increasingly complex. Two common countermeasures used 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 have been proposed [3, 14, 17, 19, 18, 9, 15, 21, 20, 22, 6, 26, 13, 5, 2, 10], and there also exist several parallel or distributed stochastic algorithms [4, 16, 27, 23, 25, 27, 1]. Specifically, S. Shalv-Shwartz and T. Zhang [19] proposed the Stochastic Dual Coordinate Ascent (SDCA) which provided new analysis with strong theoretical guarantees regarding the duality gap. T. Yang [23, 24] developed two Distributed Stochastic Dual Coordinate Ascent (DisDCA) algorithms and analyzed the tradeoff between network communication between nodes and the difficulty of the performed computation (task). However, the problem of developing a more efficient distributed SDCA algorithm is still open. In this paper, we first provide a GPU implementation of the vanilla distributed SDCA [23], and then give an asynchronous distributed approach to SDCA to make full use of computational resources that scale well.

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

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, 14], 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, many new algorithms were proposed by exploiting the finite sum structure, including the Stochastic average gradient (SAG) [17], stochastic dual coordinate ascent (SDCA) [19], stochastic variance reduced gradient (SVRG) [9], accelerated proximal coordinate gradient method (APCG) [13], SAGA [7], Prox-SDCA [20], Prox-SVRG [22], and stochastic primal-dual coordinate method (SPDC) [26]. 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 have been developed on top of stochastic gradient descent (SGD), alternating direction method of multipliers (ADMM), and stochastic dual coordinate ascent (SDCA). The two main approaches used in developing parallel algorithsm for SGD are based on shared memory and distributed memory architectures. Some work [12] looks at both settings, in addition to removing the syncrhonization requirement in the distributed memory setting. A number of approaches [27, 16, 1] consider the asynchronous or lock-free setting, making use of parameter servers [11], sparsity [16], as well as unique data-flow architectures for parameter updates [1]. ADMM stems from [8], which was developed to solve the equality constrained optimization problem. Recently, two independent works of stochastic ADMM were proposed [15, 21]. A standard reference for distributed ADMM is [4]. The advances of SDCA algorithms [19] and its variant [18, 20, 13] enjoy faster convergence than SGD and ADMM, and the distributed SDCA (DisDCA) [23, 24] was developed along with novel analysis of tradeoff between computation and communication. [23] serves as the starting point for our work.

We will build off of work from T. Yang's work on distributed SDCA [23], first add GPU capabilities and then incorporating M. Li's work on parameter servers [11] to handle communication updates in the distributed setting. In [23], 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. [11] 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). Additionally, [1] explores the theoretical ramifications of delayed parameter updates in general distributed stochastic optimization setting. By combining the work of T. Yang, M. Li, and A. Agarwal, we hope to develop a robust, distributed SDCA solution.

3 GPU Acceleration for Sequential SDCA

Algorithm 1 Sequential SDCA

Require: $\alpha^{(0)}$
Ensure: $\bar{w}$
1: Let $w^{(0)} = w(\alpha^{(0)})$, where $w(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i x_i$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Randomly pick i
4: Find $\triangle\alpha_i$ to maximize $-\phi_i^*(-(\alpha^{(t-1)} + \triangle\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\triangle\alpha_i x_i\|^2$
5: $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \triangle\alpha_i e_i$
6: $w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1}\triangle\alpha_i x_i$
7: **end for**
8: Output (Random option):
Let $\bar{\alpha} = \alpha^{(t)}$ and $\bar{w} = w^{(t)}$ for some random $t \in T_0 + 1, \ldots, T$
9: **return** $\bar{w}$

The traditional SDCA algorithm [19] is described in Algorithm 1, where $g(w) = \frac{1}{2}\|w\|_2^2$. In that procedure, the most expensive work is in line 4–6. We employ the GPU acceleration technique to make those lines run in a faster manner.

3.1 Naive GPU Implementation

3.2 Refined GPU Implementation

4 GPU Acceleration for Distributed SDCA

The distributed SDCA [23] is described in Algorithm 2. We restrict our implementation over the case when $g(w) = \frac{1}{2}\|w\|_2^2$. The procedure **SDCA-mR** is the procedure on k-th machine (process).

Algorithm 2 Distributed SDCA

1: Start K processes by calling the following procedure **SDCA-mR** with input m and T.

Procedure **SDCA-mR**

Require: Number of Iterations T, number of samples m at each iteration
Ensure: w^T
2: Let $\alpha_k^{(0)} = 0$, $v^{(0)} = 0$, $w^0 = 0$
3: **Read Data:** $(x_{k,i}, y_{k,i})$, $i = 1, \ldots, n_k$
4: **for** $t = 1, \ldots, T$ **do**
5: **for** $j = 1, \ldots, m$ **do**
6: Randomly pick $i \in \{1, \ldots, n_k\}$ and let $i_j = i$
7: Find $\Delta\alpha_{k,i}$ by

$$\Delta\alpha_{k,i} = \max_{\Delta\alpha} -\phi_{k,i}^*(-(\alpha_{k,i}^{t-1} + \Delta\alpha)) - \Delta\alpha x_{k,i}^\top w^{t-1} - \frac{mK}{2\lambda n}(\Delta\alpha)^2\|x_{k,i}\|_2^2$$

8: Set $\alpha_{k,i} = \alpha_{k,i}^{t-1} + \Delta\alpha_{k,i}$
9: **end for**
10: **Reduce:** $v^t : \frac{1}{\lambda n}\sum_{j=1}^m \Delta\alpha_{k,i_j} x_{k,i_j} \to v^{t-1}$
(Remark: the reduce step is implemented in two steps
- The master node adds the increment from each machine (process) together
- Each machine (process) receives v^t from the broadcasting process from master node

)
11: **Update:** $w^t = v^t$
12: **end for**
13: **return** w^T

5 Experimental Results

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] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su. Scaling distributed machine learning with the parameter server. In *OSDI*, volume 14, pages 583–598, 2014.

[12] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

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

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

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

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

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

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

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

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

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

[27] 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.