Efficient Distributed Stochastic Dual Coordinate Ascent

Jeff Hajewski
Mingrui Liu

May 3, 2017

University of Iowa

Overview

Problem Description and Related Work

Outline

Problem of Interest

Many machine learning problems can be formulated as the Regularized Finite Sum Minimization (RFSM) problem.

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

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.

Approaches to solve RFSM problem

- **The difficulty:**
  When the data size n is very large, it is difficult to use full gradient method or even fit all data on one single machine.

Approaches to solve RFSM problem

- **The difficulty:**
  When the data size n is very large, it is difficult to use full gradient method or even fit all data on one single machine.
- **The countermeasures:**
  - Stochastic Optimization
  - Distributed Optimization

Outline

Stochastic Optimization

- Stochastic Gradient Descent (SGD)
  [Bottou, 2010, Nemirovski et al., 2009]
- Stochastic Variance Reduced Gradient (SVRG)
  [Johnson and Zhang, 2013, Xiao and Zhang, 2014]
- Stochastic Dual Coordinate Ascent (SDCA)
  [Shalev-Shwartz and Zhang, 2013,
  Shalev-Shwartz and Zhang, 2014]
- ...

Distributed Optimization

- Distributed SGD [Lian et al., 2015]
- Distributed Stochastic ADMM [Boyd et al., 2011]
- Distributed SDCA [Yang, 2013, Yang et al., 2013]

Our Contribution

- The current SDCA and distributed SDCA are implemented by CPU
- Our contribution is to implement a practically more efficient GPU-based implementation, in both sequential setting and distributed setting.

Practical GPU-version of SDCA

Outline

GPU-version vanilla SDCA

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

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

Johnson, R. and Zhang, T. (2013).
Accelerating stochastic gradient descent using predictive variance reduction.

In *Advances in Neural Information Processing Systems*, pages 315–323.

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

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

Shalev-Shwartz, S. and Zhang, T. (2013).

Stochastic dual coordinate ascent methods for regularized loss minimization.
Journal of Machine Learning Research, 14(Feb):567–599.

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

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

Yang, T. (2013).
Trading computation for communication: Distributed stochastic dual coordinate ascent.

In *Advances in Neural Information Processing Systems*, pages 629–637.

Yang, T., Zhu, S., Jin, R., and Lin, Y. (2013).
Analysis of distributed stochastic dual coordinate ascent.
arXiv preprint arXiv:1312.1031.