Journées SMAI MODE 2018

Distributed Optimization with Sparse Communications and Structure Identification

Dmitry GRISHCHENKO
Univ. Grenoble Alpes, LJK/LIG

Franck IUTZELER
Univ. Grenoble Alpes, LJK

Jérôme MALICK
CNRS, LJK

Massih-Reza AMINI
Univ. Grenoble Alpes, LIG

Résumé. We propose an efficient distributed algorithm for solving regularized learning problems. In a distributed framework with a master machine coordinating the computations of many slave machines, our proximal-gradient algorithm allows local computations and sparse communications from slaves to master. Furthermore, with the ℓ_1-regularizer, our approach automatically identifies the support of the solution, leading to sparse communications from master to slaves, with near-optimal support. We thus obtain an algorithm with two-way sparse communications.

Mots-clefs : Distributed optimization, sparsity, identification

Learning problem with data on different machines

We consider the following learning optimization problem with composite objective

$$\min_{x\in\mathbb{R}^d} F(x) := \sum_{i=1}^{M} \pi_i f_i(x) + \lambda r(x).$$

with π_i the proportion n_i/n of data locally stored at machine i, and $f_i(x) = \frac{1}{n_i}\sum_{j\in\mathcal{S}_i} \ell_j(x)$ the local empirical risk at machine i. We consider that the machines, also referred to as *slaves*, perform computations separately and communicate with a *master* machine.

Existing methods: synchronous vs asynchronous

Most of the methods for solving the above problem are synchronous, which means that master waits for an update from all slaves and only then make update on it. In this case, the popular stochastic first-order algorithms are methods of choice to solve the problem (stochastic dual coordinate ascent (SDCA)[8], stochastic average gradient (SAG)[7], stochastic variance-reduce gradient (SVRG)[5] and SAGA [3]) But the synchronous aspect is clearly the bottleneck in this situation where asynchronous methods are expected to show better numerical efficiency. In asynchronous methods, the slaves perform computations based on outdated versions of the

main variable, and the master has to gather the slaves inputs into a productive update. Some of the previous stochastic algorithms admit asynchronous counterpart, such that SVRG algorithm [6], ASAGA [4], and the asynchronous parallel optimization algorithm [2]. A significant advantage of last one is a guarantee of sparse communications between slaves and master.

Our approach: random algorithm with support identification

The idea behind our distributed proximal-gradient algorithm is that each agent independently computes a gradient step using its local subset of the data on a randomly drawn subset of coordinates. The master machines keeps track of the weighted average of the most recent values of the agents outputs, computes the proximity operator of the regularizer at this average point and sends this value back to the updating slave.

Thus the master machine does not perform a iteration from the most recent main variable, but rather from a combination of the past main variables associated with the agents inputs. This may appear conservative but it actually performs not worse in theory (as we prove linear convergence in the strongly convex case) and much better in practice (due to the stability of the produced iterations).

We also show that the algorithm identifies the final sparsity pattern in the case of ℓ_1-regularization. This enables us to propose an enhanced algorithm where the current iterates sparsity is used to select the coordinates to be computed, leading to an automatic adaptation to eventual sparsity.

Références

[1] J. Fadili, J. Malick, G. Peyré. *Sensitivity Analysis for Mirror-Stratifiable Convex Functions*. arXiv preprint arXiv:1707.03194.

[2] J. Wangni, J. Wang, J.Liu, T. Zhang. *Gradient Sparsification for Communication-Efficient Distributed Optimization*. arXiv preprint arXiv:1710.09854.

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

[4] R. Leblond, F. Pedregosa, S. Lacoste-Julien. *ASAGA: Asynchronous Parallel SAGA*. 20th International Conference on Artificial Intelligence and Statistics, 46–54, 2017.

[5] R. Johnson, T. Zhang. *Accelerating stochastic gradient descent using predictive variance reduction*. Advances in neural information processing systems, 315–323, 2013.

[6] R. Zhang, S. Zheng, J. T. Kwok. *Asyncronous Distributed Semi-Stochastic Gradient Optimization*. arXiv preprint arXiv:1508.01633.

[7] M. Schmidt, N. Le Roux, F.Bach. *Minimizing finite sums with the stochastic average gradient*. Mathematical Programming: Series A and B, 162(1-2):83–112, 2017.

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