[8] Aaron Defazio, et al. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. NIPS 2014.

[9] Fei Gao, et al. http://www.dmtk.io.

[10] Qirong Ho, et al. More effective distributed ml via a stale synchronous parallel parameter server. NIPS 2013.

[11] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. NIPS 2013.

[12] Guolin Ke, et al. A Communication-Efficient Parallel Algorithm for Decision Tree, AAAI 2017.

[13] John Langford, et al. Sparse online learning via truncated gradient. NIPS 2009.

[14] Jason Lee, et a;. Distributed stochastic variance reduced gradient methods. arXiv:1507.07595 (2015)

[15] Mu Li, et al. Parameter server for distributed machine learning. Big Learning Workshop, 2013.

[16] Xiangrui Meng, et al. Mllib: Machine learning in apache spark. JMLR 2016.

[17] Qi Meng, et al. Asynchronous Accelerated Stochastic Gradient Descent, IJCAI 2016.

[18] Qi Meng, et al. Asynchronous Stochastic Proximal Optimization Algorithms with Variance Reduction, AAAI 2017.

[19] Arkadi Nemirovski, et al. Robust stochastic approximation approach to stochastic programming. In SIAM Journal on Optimization, 2009.

[20] Yurii Nesterov. Introductory lectures on convex optimization, Springer Science & Business Media, 2004.

[21] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 2012.

[22] Alexander Rakhlin, et al. Making gradient descent optimal for strongly convex stochastic optimization. ICML 2012.

[23] Peter Richtarik and Martin Takac. Iteration complexity of randomized block coordinate descent methods for minimizing a composite function. Mathematical Programming, 2014.

[24] Shizhao Sun, et al. Ensemble-Compression: A New Method for Parallel Training of Deep Neural Networks, arXiv:1606.00575 (2016)

[25] Shuxin Zheng, et al. Asynchronous Stochastic Gradient Descent with Delay Compensation for Distributed Deep Learning, arXiv preprint (2016)

[26] Eric P. Xing, et al. Petuum: A new platform for distributed machine learning on big data. IEEE Transactions on Big Data, 2015.

[27] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers & Mathematics with Applications, 1976.

[28] DC Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. Mathematical programming, 1989.

[29] R. H. Byrd, S.L. Hansen Jorge Nocedal, Y. Singer, A Stochastic Quasi-Newton Method for Large-Scale Optimization, SIAM Journal on Optimization.

[30] M. Frank, P. Wolfe, An algorithm for quadratic programming, Naval Research Logistics Quarterly, 1952.

[31] Jaggi, Martin, Revisiting Frank–Wolfe: Projection-Free Sparse Convex Optimization, Journal of Machine Learning Research, 2013.

[32] Sutskever, Ilya, et al. On the importance of initialization and momentum in deep learning. ICML 2013.

[33] Ruiliang Zhang, James T. Kwok, Asynchronous Distributed ADMM for Consensus Optimization, ICML 2014.

[34] Reddi, Sashank J., et al. On variance reduction in stochastic gradient descent and its asynchronous variants. NIPS 2015.

[35] Jason Lee, et al. Distributed stochastic variance reduced gradient methods. arXiv:1507.07595 (2015)