# CS 59200 — Distributed Optimization for Machine Learning

**Instructor:** Brian Bullins
**Email:** bbullins@purdue.edu
**Office:** HAAS 224

## Course Description

As the size of the networks and datasets for modern machine learning models have experienced tremendous growth, so too has there been an increasing need for finding better ways to handle the enormous problems in this setting. Distributed optimization presents one such opportunity, whereby these large datasets/models are split across many machines so that much of the computation occurs in parallel. In this course, we will explore the many recent advances in developing algorithms, many of which are based on stochastic gradient methods, for this important area of research.

We will approach the problem from both theoretical and practical angles. Throughout the course, we will aim to formally describe and analyze various distributed optimization settings in terms of e.g. the number of machines we have, how many rounds of communication are conducted, what oracle models are accessed, etc. At the same time, we will look into the practical implications of these methods' choices for efficiently solving many relevant problems.

## Course Format

The course will be mostly based on reading recent research papers in the distributed optimization literature, as published at conferences such as ICML, NeurIPS, COLT, etc. Students will be expected to present the contributions and results of papers, and will further lead a discussion on the works, with the expectation that the other students will be able to contribute to the discussion. At the beginning of the course, I will present a few introductory lectures on some of the basics of distributed optimization theory.

## Learning Outcomes

By the end of this course, students should be able to:

- Understand the importance and relevance of distributed optimization for modern machine learning, along with their current limitations
- Have a well-informed overview of state-of-the-art results in distributed optimization research

- Critically analyze and interpret the contributions of relevant papers in the research area
- Determine novel directions of research and think effectively about techniques to make progress for such open problems

## Course Materials

While there is no required textbook, the following resources are recommended:

*Convex Optimization, Algorithms and Complexity*, by Sébastien Bubeck
*Convex Optimization*, by Boyd and Vandenberghe
*Numerical Optimization*, by Nocedal and Wright

## Grading Overview

Reading / Paper Discussions: 20%
Paper Presentations: 30%
Final Project: 50%

## Tentative List of Papers

- Dekel, Ofer, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. "Optimal Distributed Online Prediction Using Mini-Batches." Journal of Machine Learning Research 13, no. 1 (2012).

- Zhang, Yuchen, John Duchi, Michael I. Jordan, and Martin J. Wainwright. "Information-theoretic lower bounds for distributed statistical estimation with communication constraints." Advances in Neural Information Processing Systems 26 (2013).

- Shamir, Ohad, Nati Srebro, and Tong Zhang. "Communication-efficient distributed optimization using an approximate newton-type method." In International conference on machine learning, pp. 1000-1008. PMLR, 2014.

- Zhang, Yuchen, and Xiao Lin. "DiSCO: Distributed optimization for self-concordant empirical loss." In International conference on machine learning, pp. 362-370. PMLR, 2015.

- Reddi, Sashank J., Jakub Konečný, Peter Richtárik, Barnabás Póczós, and Alex Smola. "Aide: Fast and communication efficient distributed optimization." arXiv preprint arXiv:1608.06879 (2016).

- Wang, Jialei, Weiran Wang, and Nathan Srebro. "Memory and communication efficient distributed stochastic optimization with minibatch prox." In Conference on Learning Theory, pp. 1882-1919. PMLR, 2017.

- McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. "Communication-efficient learning of deep networks from decentralized data." In Artificial intelligence and statistics, pp. 1273-1282. PMLR, 2017.

- Wang, Shusen, Fred Roosta, Peng Xu, and Michael W. Mahoney. "Giant: Globally improved approximate newton method for distributed optimization." Advances in Neural Information Processing Systems 31 (2018).

- Stich, Sebastian U. "Local SGD converges fast and communicates little." arXiv preprint arXiv:1805.09767 (2018).

- Woodworth, Blake E., Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. "Graph oracle models, lower bounds, and gaps for parallel stochastic optimization." Advances in neural information processing systems 31 (2018).

- Crane, Rixon, and Fred Roosta. "DINGO: Distributed Newton-type method for gradient-norm optimization." Advances in Neural Information Processing Systems 32 (2019).

- Woodworth, Blake, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. "Is local SGD better than minibatch SGD?." In International Conference on Machine Learning, pp. 10334-10343. PMLR, 2020.

- Karimireddy, Sai Praneeth, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. "Scaffold: Stochastic controlled averaging for federated learning." In International Conference on Machine Learning, pp. 5132-5143. PMLR, 2020.

- Woodworth, Blake E., Kumar Kshitij Patel, and Nati Srebro. "Minibatch vs local sgd for heterogeneous distributed learning." Advances in Neural Information Processing Systems 33 (2020).

- Wang, Jianyu, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. "Tackling the objective inconsistency problem in heterogeneous federated optimization." Advances in neural information processing systems 33 (2020).

- Yuan, Honglin, and Tengyu Ma. "Federated accelerated stochastic gradient descent." Advances in Neural Information Processing Systems 33 (2020).

- Gupta, Vipul, Avishek Ghosh, Michal Derezinski, Rajiv Khanna, Kannan Ramchandran, and Michael Mahoney. "LocalNewton: Reducing communication bottleneck for distributed learning." UAI (2021).

- Bullins, Brian, Kshitij Patel, Ohad Shamir, Nathan Srebro, and Blake E. Woodworth. "A Stochastic Newton Algorithm for Distributed Convex Optimization." Advances in Neural Information Processing Systems 34 (2021).

- Derezinski, Michal, Jonathan Lacotte, Mert Pilanci, and Michael W. Mahoney. "Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update." Advances in Neural Information Processing Systems 34 (2021).

- Glasgow, Margalit R., Honglin Yuan, and Tengyu Ma. "Sharp Bounds for Federated Averaging (Local SGD) and Continuous Perspective." In International Conference on Artificial Intelligence and Statistics, pp. 9050-9090. PMLR, 2022.

- Wang, Jianyu, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data." arXiv preprint arXiv:2206.04723 (2022).