STAT 991: OPTIMIZATION METHODS IN MACHINE LEARNING

Classes.

- Tu/Th 3–4:20 pm in F92 JMHH

Instructor. Weijie Su (`suw@wharton.upenn.edu`)

Office: 472 JMHH

Office Hours: by appointment

Course Overview. The course aims to equip students with advanced techniques and methods in optimization that are tailored to large-scale statistics and machine learning problems. Together, we will explore a number of prominent developments in first-order optimization methods in the convex, nonconvex, stochastic, and distributed settings. Upon completing the course, students are expected to be able to better formulate an optimization problem by exploiting desired structural properties (for example, convexity, smoothness, and sparsity), and to select an efficient optimization method under problem constraints (for example, online, distributed, and memory cost).

Prerequisites. Fluency with reasoning and analysis using linear algebra and probability is required. Students are expected to be familiar with computing platforms such as Matlab, Julia, and Python. Students should learn by themselves the basics of the (very user-friendly) convex optimization interpreter `cvx` (`http://cvxr.com/cvx/`) in the Matlab environment. `cvx` is also available in the Julia and Python environments.

Topics. We will cover the following topics in class:

1) Basics of convex optimization
   - convex sets, convexity-preserving operations, examples of convex programs (linear programming (LP), second-order cone programming (SOCP), semidefinite programming (SDP)), convex relaxation, KKT conditions, duality
2) Gradient-based methods
   - gradient descent, subgradient, mirror descent, Frank–Wolfe method, Nesterov's accelerated gradient method, ODE interpretations, primal-dual methods, Nesterov's smoothing, proximal gradient methods, Moreau–Yosida regularization
3) Operator splitting methods
   - augmented Lagrangian methods, alternating direction method of multipliers (ADMM), monotone operators, Douglas–Rachford splitting, dual and primal decomposition
4) Stochastic and nonconvex optimization

- dual averaging, Polyak–Juditsky averaging, stochastic variance reduced gradient (SVRG), Langevin dynamics, escaping saddle points, landscape of nonconvex problems, deep learning

Course Projects. A course project can be either literature review or original research.
Literature review: The instructor will provide a list of prominent machine-learning-flavored optimization papers that are not covered by lectures. One or two students can form a group and choose a paper from the list or a paper of your own interest upon the instructor's approval. This involves two components: (1) an in-class presentation with slides and (2) a written report that is up to 6 pages. The presentation should give a high-level summary of the paper, highlighting the main ideas and describing the approaches/methods in a concise way, whereas the report is expected to contain technical details that are important in understanding the paper, for example, the proof of the main theorem in the paper. In-class presentations are tentatively scheduled to be interspersed with formal lectures starting late February, and the report is due one week after the date of the corresponding lecture.

Original research: A group of up to 3 students can work on a topic on optimization and are encouraged to make a connection with your own research. The instructor can provide guidance on choosing the topics. This involves three components: (1) a one-page proposal due on March 12 (send to the instructor via email), (2) an in-class presentation with slides and (3) a written report that is up to 8 pages (appendix not counted) due on May 14. The proposal should state the problem that you aim to solve and briefly review related references. In-class presentations of original research will be given in the last 2 or 3 lectures of the course. The report should summarize your findings and contributions.

In both cases, turn in a hard copy of your written report (place in the instructor's mailbox on the 4th floor of JMHH) and send an electronic copy to the instructor's email for records.

Scribing. Each lecture will be scribed by one or two students. Scribed notes are due one week after the delivery of the corresponding lecture. A LaTeX template for typesetting notes can be found on Canvas. The sign-up sheet is here (please do not use your own macros in LaTeX).

Textbook. The following books and notes are recommended, though our course will not follow them too closely.

- Stephen Boyd and Lieven Vandenberghe's book: *Convex Optimization*
- Nesterov's old book: *Introductory Lectures on Convex Optimization: A Basic Course*
- Nesterov's new book: *Lectures on Convex Optimization*
- Neal Parikh and Stephen Boyd's monograph: *Proximal Algorithms*
- Sébastien Bubeck's monograph: *Convex Optimization: Algorithms and Complexity*
- Moritz Hardt's Berkeley EE 227C course note
- John Duchi's course note
- Prateek Jain and Purushottam Kar's survey on nonconvex optimization

Course Website. The course website uses both the Canvas and Piazza platforms. Check Canvas for announcements, handouts, and other course materials. Please discuss problems with each other on Piazza.

Evaluation. The performance of students will be evaluated based on in-class presentations, project reports, homework assignments (tentative), scribing, and attendance.

Papers for Literature Review (growing).

1) Maximum likelihood estimation of a multidimensional log-concave density. M Cule, R Samworth, M Stewart, 2010
2) SLOPE—Adaptive variable selection via convex optimization. M Bogdan, E Berg, C Sabatti, WJ Su, EJ Candès, 2015
3) Linear and conic programming estimators in high dimensional errorsinvariables models. A Belloni, M Rosenbaum, AB Tsybakov, 2017
4) A coordinate gradient descent method for nonsmooth separable minimization. P Tseng, S Yun, 2009
5) A variational perspective on accelerated methods in optimization. A Wibisono, AC Wilson, MI Jordan, 2016
6) Understanding the acceleration phenomenon via high-resolution differential equations. B Shi, SS Du, MI Jordan, WJ Su, 2018
7) Composite objective mirror descent. J Duchi, S Shalev-Shwartz, Y Singer, A Tewari, 2010
8) Adaptive piecewise polynomial estimation via trend filtering. RJ Tibshirani, 2014
9) Optimal algorithms for non-smooth distributed optimization in networks. K Scaman, F Bach, S Bubeck, L Massoulié, YT Lee, 2018
10) Robust stochastic approximation approach to stochastic programming. A Nemirovski, A Juditsky, G Lan, A Shapiro, 2009
11) Adaptive online gradient descent. E Hazan, A Rakhlin, PL Bartlett, 2008
12) Near-optimal stochastic approximation for online principal component estimation. CJ Li, M Wang, H Liu, T Zhang, 2018
13) Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. WJ Su, Y Zhu, 2018
14) Asymptotic optimality in stochastic optimization. J Duchi, F Ruan, 2016
15) Natasha 2: Faster non-convex optimization than SGD. Z Allen-Zhu, 2017
16) Stochastic cubic regularization for fast nonconvex optimization. N Tripuraneni, M Stern, C Jin, J Regier, MI Jordan, 2017
17) Phase retrieval via Wirtinger flow: Theory and algorithms. EJ Candès, X Li, M Soltanolkotabi, 2015
18) Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. C Ma, K Wang, Y Chi, Y Chen, 2017

19) Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. P Loh, M Wainwright, 2013
20) The landscape of empirical risk for non-convex losses. S Mei, Y Bai, and A Montanari, 2016
21) Tensor decompositions for learning latent variable models. A Anandkumar, R Ge, D Hsu, SM Kakade, M Telgarsky, 2014
22) Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. M Raginsky, A Rakhlin, M Telgarsky, 2017
23) Sampling can be faster than optimization. YA Ma, Y Chen, C Jin, N Flammarion, MI Jordan, 2018
24) On the optimization of deep networks: Implicit acceleration by overparameterization. S Arora, N Cohen, E Hazan, 2018
25) A mean field view of the landscape of two-layers neural network. S Mei, AMontanari, PM Nguyen, 2018