Selected Topics in Modern Convex Optimization
Theory, Algorithms and Applications
Spring 2017 – STOR 892

1. Objectives
This is a special topic course taught at the Department of Statistics and Operations Research, UNC-Chapel Hill. The primary goal is to discuss recent development in modern convex optimization and its applications to statistical learning, machine learning, and other areas of sciences. It aims at developing fundamental theory, methodology, and implementation aspects of modern convex optimization for students as a mathematical tool to solve a wide range of real-world problems. We mainly focus on how to design efficient methods, how to implement them and illustrate them on existing convex optimization models. We expect to show students what has been done and give pointers to important open problems in convex optimization. We also expect to inspire students how to select an appropriate optimization method for a given optimization problem in practice.

The course is designed for graduate students who have some background in applied math such as linear algebra, multivariable analysis and computational skills.

2. Time and Place
Lectures: **Mondays and Wednesdays, 1:25 – 2:40PM (Hanes 107).**

3. Staff
Instructor: Quoc Tran-Dinh (quoctd@email.unc.edu)
Office: 333 Hanes Hall, UNC-Chapel Hill.

4. Syllabus
This course consists of three parts:

- ***A. Representative convex optimization models in applications***
- ***B. Fundamental concepts in convex optimization: a brief overview***
- ***C. Selected first-order methods for large-scale convex optimization***
- ***D. Selected primal-dual methods for large-scale constrained convex optimization***

Depending on time quota, some topics may be skipped and some may have more emphasis.

Regarding these three parts, we plan to cover the following topics:

Part 1: Representative convex optimization models in applications (3 lectures):

Optimization plays a major role in many fundamental areas of statistics including the maximum likelihood principle; of machine learning such as Google page rank or movie rating from Netflix; of image processing such as the reconstruction of a clean image from a noisy one; and of control such as the design of a good strategy to stabilize a system. We discuss in this part the mathematical form of an optimization problem; what do we mean by "solving an optimization problem"?; how do we classify an optimization problem into different classes; and provide some well-known and representative examples. Specifically, we discuss the following topics:

1. *Mathematical formulation of an optimization problem*
2. *Classifying optimization problems and the choice of methods.*
3. *Representative applications*
   + Least squares, Basis pursuit, LASSO and beyond
   + Logistic regression and extensions
   + Support vector machine: linear and nonlinear cases
   + Image reconstruction with total variation norms
   + Matrix completion and robust principal component analysis
   + Sparse inverse covariance selection in graphical models
4. *Other related applications*

We concentrate on how to formulate these applications into a convex optimization problem. Then, we investigate some explicit properties of these problems to find an appropriate method for efficiently solving them.

Part 2. Fundamental concepts in convex optimization: A brief overview (5 lectures)

We only give a very short review on some concepts and tools needed for this course. We do not go deeply in convex analysis. If students already had some background in convex analysis and linear algebra, some topics can be skipped. But other topics such as proximal operators and monotone operators are rarely covered in a convex analysis course, and they remain worthy of study. More concretely, we cover the following topics:

1. *Convex sets and convex functions*
2. *Fenchel conjugates*
3. *Proximal operators and projections*
4. *Proximity functions, and Bregman divergences*
5. *Duality theory*
6. *Theory of monotone operators*
7. *Convergence analysis and complexity theory.*

Part 3: Selected first order methods for large-scale convex optimization (10 lectures)

Modern applications require convex optimization on a huge-scale. Traditional approaches such as interior-points and Newton methods are no longer efficient to tackle these models. In addition, not only the size of problems matters, but the structure of problems is also getting more and more complicated. These challenges require new thoughts on the design of optimization algorithms. One opportunity to solve these problems is using low-cost optimization methods such as first-order algorithms. While these methods have low complexity-per-iteration, they often have slow convergent speed. This part discusses some recent development on gradient-type methods for large-scale problems. The topics covered in this part include:

1. *Gradient method and accelerated gradient methods:* mathematical view, algorithms, convergence analysis, implementation, and enhancements (e.g., line-search, preconditioning, and restart).
2. *Proximal gradient and accelerated proximal gradient methods:* algorithms, convergence and complexity analysis, examples and enhancements.
3. *Mirror descent methods* – beyond the Euclidean norm
4. *Conditional gradient (Frank-Wolfe) methods*
5. *Splitting methods:* Forward-backward and Douglas-Rachford splitting
6. *Coordinate descents for huge-scale convex optimization:* Randomized and cycling coordinate descents, and parallel variants.
7. *Stochastic gradient descent methods:* Empirical risk minimization; basic method; stochastic dual averaging scheme; stochastic variance reduction gradient method (SVRG); and accelerated variants.

All these methods often require implementation and application to some specific examples given in Part 1. Note that, we only cover the two last topics if time allows.

Part 4: Selected primal-dual methods for large-scale constrained convex optimization (6 lectures).

So far we have only looked at the methods for unconstrained and simple constrained convex problems. What about problems with complicated constraints, such as problem in networks or graphs where we have flow constraints? These problems require different approaches to solve efficiently.
In this section we discuss some basic and well-know methods for min-max saddle-point and constrained convex optimization problems. More specifically, we consider the following topics:

1. *Minmax formulation and primal-dual pair.*
2. *Dual ascent and how to recover a primal solution from its dual*
3. *Penalty and augmented Lagrangian methods*

4. *Alternating minimization algorithm (AMA), and alternating direction methods of multipliers (ADMM): from theory, algorithms to applications*
5. *Chambolle-Pock primal dual methods and variants.*
6. *Other primal-dual methods*

5. Course material

Lecture notes: Some lecture notes will be provided. It must be used internally in the course. Please do not distribute this material.

Books: Here are some books which contain some parts of the lectures

[B1]. R. T. Rockafellar: Convex Analysis, 1970, Princeton Univ. Press (http://www.convexoptimization.com/TOOLS/ConvexAnalysis.pdf).

[B2]. S. Boyd and L. Vandenberghe: Convex Optimization, 2006, Cambridge Univ. Press (http://stanford.edu/~boyd/cvxbook/)

[B3]. Y. Nesterov: Introductory lectures on Convex Optimization, 2004. (His lectures can be found here: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.855&rep=rep1&type=pdf).

[B4]. H.H. Bauschke and P. Combettes: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, 2011.

[B5]. D. Bertsekas, Convex Optimization Theory/Algorithms, 2009, 2015.

Other material: These are good surveys/lecture notes for the course

[S1]. S. Boyd et al: Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, 3(1):1–122, 2011.

[S2]. N. Parikh and S. Boyd: Proximal algorithms, Foundations and Trends in Optimization, 1(3):123-231, 2014.

[L1]. S. Bubeck, Convex Optimization: Algorithms and Complexity, http://arxiv.org/abs/1405.4980.

Selected papers: These are some remarkable papers:

[P1]. Y. Nesterov, A method of solving a convex programming problem with convergence rate O (1/k2), Soviet Mathematics Doklady, 1983 (translated to English).

This is the original paper on the fast gradient method.

[P2]. A. Beck and M. Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, SIAM J. Imaging Sciences, 2009.

This paper makes fast proximal gradient method become popular, the proof is elementary and easy to read.

[P3]. Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming, 2005.

This paper renews [P1] and makes fast gradient method become a new trend for large-scale convex optimization.

[P4]. P. Tseng, On Accelerated Proximal Gradient Methods for Convex-Concave Optimization, Online paper, 2008.

This paper provides a deep theory for accelerated gradient methods.

[P5]. Y. Nesterov, Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems, SIAM J. Optimization, 2012.

This paper re-popularizes the coordinate descent again for big-data applications

[P6]. M. Jaggi, Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization, ICML 2013.

This paper revisits the classical FW method since 1950, but makes it extremely useful for machine learning (and others) applications.

[P7]. A Nemirovski, A Juditsky, G Lan, A Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM J. Optimization, 2009.

This paper proposes an averaging strategy for stochastic gradient descent, which is the foundation theory for many following works.

[P8]. N. Le Noux, M. Schmidt, F. Bach, A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets, NIPS, 2013.

This paper provides a very efficient method for some machine learning problems.

[P9]. R. Johnson, T. Zhang, Accelerating Stochastic Gradient Descent using Predictive Variance Reduction, NIPS 2014.

A new idea of variance reduction for optimization methods starts from this paper.

[P10]. J Eckstein, DP Bertsekas, On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming, 1992.

This paper is on spitting methods, which become extremely popular nowadays.

<u>References:</u> The references are given at the end of each lecture.

6. Course evaluation:

- **Homework:** A few homework assignments will be given during class (30%).
- **Projects:** Students work on projects (teamwork or individually) (70%) (select one of the following formats).

  + Students are asked to read one or few papers, or book chapters, then write a short report (maximum 8 pages) and present in class.

  + Students are asked to work on an optimization problem, and implement some algorithms to solve it, then test the algorithms on synthetic and/or real datasets, and then write a short report (maximum 8 pages) and present in class.
- **Exam:** There will be no written exam.