Graduate Topic Course - STOR 893

Selected Methods for Modern Optimization in Data Analysis

(Fall 2018)

Course overview

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 numerical methods for solving modern optimization applications from different areas of data analysis. The content of this course consists of 4 parts ranging from modeling and foundation theory to algorithms and their convergence guarantees. We will focus on different aspects of several optimization methods including the design of algorithms, convergence guarantees, computational complexity, implementation, improvements, and applications. We will also discuss the advantages and disadvantages of each optimization method on different problem classes. We expect to inspire students how to select an appropriate numerical method for a given optimization model 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. Background in convex analysis, numerical linear algebra, algorithms, or statistics is also preferable to better follow the course.

Time and Place

***Lectures*: Tuesdays and Thursdays, 12:30PM - 1:45PM (Hanes 125).**

Instructor

Instructor: Quoc Tran-Dinh (`quoctd@email.unc.edu`)
Personal webpage: `http://quoctd.web.unc.edu`.

Office: 333 Hanes Hall, UNC-Chapel Hill.

Course content

This course consists of four parts:

A. ***Representative optimization models in applications***

B. ***Fundamental concepts and basic theory in optimization***

C. ***Selected first-order methods for convex optimization***

D. ***Selected methods for some classes of nonconvex optimization.***

Depending on time quota, some topics may be skipped, and some may have more emphasis. Regarding these four parts, we plan to cover the following topics:

Part 1: Representative optimization models in applications

Optimization plays a major role in many fundamental areas of statistics including the maximum likelihood principle; of machine learning such as Google page ranks, movie ratings from Netflix, or deep learning; 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 forms of an optimization problem, what we mean by "solving an optimization problem", how 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, and LASSO.
    - Logistic regression and extensions
    - Support vector machine: linear and nonlinear cases
    - Image reconstruction with total variation (TV) norms
    - Matrix completion and robust principal component analysis
    - Sparse inverse covariance selection in graphical models
    - Nonnegative matrix factorization models
    - Optimization models in deep neural networks

We will concentrate on how to formulate these applications into a convex/nonconvex optimization problem. Then, we investigate some explicit properties of these problems to find an appropriate method for efficiently solving them. Depending the available time, some applications are emphasized or just briefly discussed.

Part 2. Fundamental concepts and basic theory in optimization

We will have a very short review on some concepts and tools needed for this course. We do not go deeply in convex analysis or other mathematical tools. Depending on how much students are familiar with convex analysis and numerical 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 to study. More concretely, we will cover the following topics:

1. *Convex sets and convex functions.*
2. *Proximal operators, projections, and monotone operators.*
3. *Fenchel conjugates and Bregman divergences.*
4. *Optimality conditions and Karush-Kuhn-Tucker (KKT) conditions.*
5. *Duality theory.*
6. *Convergence rates and complexity theory.*

Part 3: Selected first-order methods for convex optimization

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 ideas on the design of optimization algorithms. One way to solve these problems is using low-cost optimization methods such as first-order algorithms or stochastic methods. While these methods have low complexity-per-iteration, they often have slow convergent speed. This first half of Part 3 will discuss some recent development in first-order methods for large-scale problems. In the second half, we will provide some advanced methods for large-scale constrained convex problems and min-max saddle-point problems. More specifically, we will look at the following topics:

1. *Gradient and proximal methods, and their accelerated variants:* mathematical view, algorithms, convergence analysis, implementation, and enhancements (e.g., line-search, preconditioning, and restart).
2. *Mirror descent methods and conditional gradient (Frank-Wolfe) methods.*
3. *Coordinate descents for huge-scale convex optimization:* Randomized and cycling coordinate descents, and parallel variants.
4. *Stochastic gradient descent methods:* Empirical risk minimization; basic method; stochastic dual averaging scheme; stochastic variance reduction gradient method (SVRG); and accelerated variants.
5. *Min-max formulation and primal-dual pair.*
6. *Dual ascent and how to recover a primal solution from its dual.*
7. *Penalty and augmented Lagrangian methods.*
8. *Douglas-Rachford's splitting method and alternating direction methods of multipliers (ADMM):* from theory, algorithms to applications.
9. *Primal-dual first-order methods:* Chambolle-Pock's method, and primal-dual hybrid gradient method.

All these methods often require implementation and application to some specific examples given in Part 1. Due to time limits, some topics may have a brief discussion.

Part 4: Selected methods for some classes of nonconvex optimization

So far, we have only looked at methods for convex problems. What about nonconvex problems, such as nonnegative matrix factorization, and deep neutral networks? These problems require different approaches to efficiently solve them. Nonconvex optimization is currently an extremely active research field due to the era of data science and deep learning. In this section, we will discuss some well-known and widely used methods for some classes of nonconvex optimization problems. More specifically, we will cover the following topics:

1. *Gradient-type methods for nonconvex problems.*
2. *Newton-type methods and quasi-Newton-type methods.*
3. *DC (different of two convex functions) algorithms (DCAs).*
4. *Alternating optimization methods for nonconvex problems.*

Course materials

Lecture notes

Lecture notes will be provided to students via the Sakai system. They must be used internally in the course. Please do not distribute these materials.

Books

Here are some books which contain some parts of the lectures

[B_1]. R. T. Rockafellar: Convex Analysis, 1970, Princeton Univ. Press. This book is now available online for free and can be downloaded from `http://www.convexoptimization.com/TOOLS/ConvexAnalysisRockafellar.pdf`.

[B_2]. S. Boyd and L. Vandenberghe: Convex Optimization, 2006, Cambridge Univ. Press. This book is available for free at `http://stanford.edu/~boyd/cvxbook/`.

[B_3]. Y. Nesterov: Introductory lectures on Convex Optimization, 2004. The lectures can be found at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.855&rep=rep1&type=pdf`.

[B_4]. H. H. Bauschke and P. Combettes: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer-Verlag, 2017.

[B_5]. D. Bertsekas, Convex Optimization Theory/Algorithms, Athena Scientific, 2009.

[B_6]. J. Nocedal and S. Wright, Numerical Optimization, Springer-Verlag, 2006.

Other materials

These are good surveys/lecture notes for the course

[S_1]. 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.

[S_2]. N. Parikh and S. Boyd: Proximal algorithms, Foundations and Trends in Optimization, 1(3):123-231, 2014.

[L_1]. S. Bubeck, Convex Optimization: Algorithms and Complexity. This lecture note can be downloaded from `http://arxiv.org/abs/1405.4980`.

[S_3]. S. Wright, Optimization Algorithms in Data Analysis. This survey paper is available online at `http://www.optimization-online.org/DB_FILE/2016/12/5748.pdf`

[B_1]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, The MIT Press, 2016.

Selected papers

These are some remarkable papers:

[P_1]. Y. Nesterov, A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, Soviet Mathematics Doklady, 1983 (translated to English). This is the original paper on the fast gradient method.

[P_2]. 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.

[P_3]. Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming, 2005. This paper renews [P_1] and makes fast gradient method become a new trend for large-scale convex optimization.

[P_4]. P. Tseng, On Accelerated Proximal Gradient Methods for Convex-Concave Optimization, Online paper, 2008. This paper provides a deep theory for accelerated gradient methods.

[P_5]. 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.

[P_6]. 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.

[P_7]. 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.

[P_8]. 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.

[P_9]. 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.

[P_{10}]. 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.

[P_{11}]. Y. Nesterov, Cubic regularization of Newton method and its global performance, Math. Program., vol. 108, 2006.

[P_{12}]. L.T. Hoai An, and P.D. Tao, The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems, Annals of Operations Research 133 (1-4), 23-46, 2005.

References

The references are given at the end of each lecture.

Course evaluation

- **Homework assignments:** A few homework assignments will be given during class. They will count for 30% of the final grade.
- **Course projects:** Students work on projects (in teams or individually). They will count for 70% of the final grade. Students can select one of the following two formats:
  - Students are asked to read one or a few papers, or book chapters, then write a short report (between 4 and 8 pages) and present it 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 (between 4 and 8 pages) and present it in class.
- **Exams:** There will be no written exam.