Monday Workshop Presenters:

Title: BFGS Methods: Block Updates, Adaptive Step Sizes, and Stochastic Variants

Speaker: Don Goldfarb
Affiliation: Columbia University

Abstract: We discuss several variants of the BFGS method that we have recently developed. The primary motivation for developing these methods is the need to solve optimization problems that arise in machine learning, which because of the enormous amounts of data involved in each computation of the function and gradient, usually require a stochastic optimization approach. The issues that we address in this talk are: (i) the use of sketching (i.e., Hessian actions) and block-updates to incorporate (noisy) curvature information; (ii) the determination of adaptive step sizes to avoid line searches for strictly convex self-concordant functions; and (iii) the development of stochastic BFGS variants for stochastic convex optimization problems. Both theoretical and computational results will be presented.

Title: Negative Curvature in Deterministic and Stochastic Nonconvex Optimization

Speaker: Daniel Robinson
Affiliation: Johns Hopkins University

Abstract: Many researchers have designed optimization algorithms for nonconvex problems that guarantee convergence to second-order optimal points. These methods are usually discarded for methods that only ensure convergence to first-order optimal points for multiple reasons: (i) guaranteeing convergence to second-order points necessarily requires the algorithm to become more complicated and computationally expensive; (ii) convergence to a first-order point that is not a second-order point is rare in practice because of the descent nature of the methods; and (iii) no empirical performance gains have been consistently demonstrated, which is especially discouraging since extra computation per iteration is often required (e.g., to approximate the left-most eigenvector/eigenvalue pair of the second-derivate matrix). In this talk I explore options for judiciously including directions of negative curvature in line-search and fixed-step size methods within the context of both deterministic and stochastic optimization.

Title: Adaptive Sampling Strategies for Stochastic Optimization

Speaker: Raghu Bollapragada
Affiliation: Northwestern University

Abstract: In this talk, we present a stochastic optimization method that adaptively controls the sample size used in the computation of gradient approximations. Unlike other variance reduction techniques that either require additional storage or the regular computation of full gradients, the proposed method reduces variance by increasing the sample size as needed. The decision to increase the sample size is governed by an inner product test that ensures that search directions are descent directions with high probability. We show that the inner product test improves upon the well-known norm test, and can be used as a basis for an algorithm that is globally convergent on nonconvex functions and enjoys a global linear rate of convergence on strongly convex functions. Numerical experiments on logistic regression problems illustrate the performance of the algorithm.

Monday Workshop Presenters (continued):

Title: An Inexact Regularized Newton Framework with a Worst-Case Iteration Complexity of *O(eps^(-3/2))* for Nonconvex Optimization

Speaker: Mohammadreza Samadi
Affiliation: Lehigh University

Abstract: An algorithm for solving smooth nonconvex optimization problems is proposed that, in the worst-case, takes *O(eps^(-3/2))* iterations to drive the norm of the gradient of the objective function below a prescribed positive real number *eps* and can take *O(eps^(-3))* iterations to drive the leftmost eigenvalue of the Hessian of the objective above *-eps*. The proposed algorithm is a general framework that covers a wide range of techniques including quadratically and cubically regularized Newton methods, such as the Adaptive Regularisation using Cubics (ARC) method, and the recently proposed Trust-Region Algorithm with Contractions and Expansions (TRACE). The generality of our method is achieved through the introduction of generic conditions that each trial step is required to satisfy, which in particular allow for inexact regularized Newton steps to be used. These conditions center around a new subproblem that can be approximately solved to obtain trial steps that satisfy the conditions. Numerical results demonstrate that an instance of the framework that may be viewed as a hybrid between quadratically and cubically regularized Newton methods, outperforms a cublicly regularized Newton method.

Title: Worst-Case Complexity and Nonconvex Optimization

Speaker: Frank Curtis
Affiliation: Lehigh University

Abstract: The purpose of this talk is twofold. First, we present a generic framework for achieving optimal complexity for second-order methods for solving smooth nonconvex optimization problems. Second, we propose new ideas for analyzing the worst-case behavior of algorithms for nonconvex optimization, emphasizing that, rather than provide a single global rate, it is more useful to provide guarantees broken down according to different segments of the search space.

Title: A Sub-sampled Semismooth Newton Method for Convex Composite Problems

Speaker: Zaiwen Wen
Affiliation: Peking University

Abstract: The nonsmooth composite minimization problem, whose objective function is the sum of an average of N smooth (possibly nonconvex) functions plus a proper closed convex function, is widely studied in large-scale machine learning, statistics and optimization. Unlike the numerous stochastic first-order methods, the knowledge on stochastic second-order methods for this problem is very limited. In this paper, we propose a stochastic nonsmooth nonconvex Newton-type method which is a hybrid approach of the sub-sampled semismooth Newton steps and the proximal gradient steps. Some growth conditions based on the sub-sampled residual and forward-backward envelope are used to control the acceptance of the sub-sampled semismooth Newton steps. Global convergence to stationary point in expectation and transition to local convergence are established. Under certain assumptions, local superlinear convergence in high probability and in expectation are derived, respectively. Numerical experiments show that the proposed algorithm is effective.

Monday Workshop Presenters (continued):

Title: Nonconvex Optimization in Low-Complexity Data Modeling

Speaker: John Wright
Affiliation: Columbia University

Abstract: Nonconvex optimization plays important role in wide range of areas of science and engineering — from learning feature representations for visual classification, to reconstructing images in biology, medicine and astronomy, to disentangling spikes from multiple neurons. The worst case theory for nonconvex optimization is dismal: in general, even guaranteeing a local minimum is NP hard. However, in these and other applications, very simple iterative methods such as gradient descent often perform surprisingly well.

In this talk, I will discuss examples of nonconvex optimization problems that can be solved to global optimality using simple iterative methods, which succeed independent of initialization. These include variants of the sparse blind deconvolution problem, sparse dictionary learning problem, image recovery from certain types of phaseless measurements. These problems possess a characteristic structure, in which (i) all local minima are global, and (ii) the energy landscape does not have any "flat" saddle points. For each of the aforementioned problems, this geometric structure allows us to obtain new types of performance guarantees. I will motivate these problems from applications in imaging and computer vision, and describe how this viewpoint leads to new approaches to analyzing electron microscopy data.

Joint work with Ju Sun (Stanford), Qing Qu (Columbia), Yuqian Zhang (Columbia), Yenson Lau (Columbia) Sky Cheung, (Columbia), Abhay Pasupathy (Columbia)

Title: Regularized Nonlinear Acceleration.

Speakers: Alexandre d'Aspremont
Affiliation: École Normale Supérieure

Abstract: We describe a convergence acceleration technique for generic optimization problems. Our scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments are detailed on classical classification problems.

Tuesday Workshop Presenters:

Title: Distributionally Robust Stochastic and Online Optimization

Speaker: Yinyu Ye
Affiliation:

Abstract: We present decision/optimization models/problems driven by uncertain and online data, and show how analytical models and computational algorithms can be used to achieve solution efficiency and near optimality.

- First, we describe the so-called Distributionally or Likelihood Robust optimization (DRO) models and their algorithms in dealing stochastic decision problems when the exact uncertainty distribution is unknown but certain statistical moments and samples can be estimated.
- Secondly, when decisions are made in presence of high dimensional stochastic data, handling joint distribution of correlated random variables can present a formidable task, both in terms of sampling and estimation as well as algorithmic complexity. A common heuristic is to estimate only marginal distributions and substitute joint distribution by independent (product) distribution. Here, we study possible loss incurred on ignoring correlations through the DRO approach, and quantify that loss as Price of Correlations (POC).

Thirdly, we describe an online combinatorial auction problem using online linear programming technologies. We discuss near-optimal algorithms for solving this surprisingly general class of online problems under the assumption of random order of arrivals and some conditions on the data and size of the problem.

Title: Distributionally Robust Optimization via Optimal Mass Transportation: Algorithms, Statistics, and Applications

Speaker: Jose Blanchet
Affiliation: Columbia University

Abstract: We first show that many machine learning algorithms and current adversarial deep learning procedures can be represented as distributionally robust optimization (DRO) formulations. A key aspect in these DRO reformulations is that the distributional uncertainty region is based on optimal transport discrepancies. Given a cost function c(x,y), the optimal transport discrepancy between two distributions is the cheapest way to move all of the mass from one distribution to another assuming, that it costs c(x,y) to move a unit of mass from x to y. Using simply c(x,y) equal to a norm between x and y, recovers many of the machine algorithms involving regularization, as mentioned earlier. But, as we shall explain, more general costs not only makes intuitive sense but one can obtain better empirical performance in machine learning applications. We show that DRO formulations based on a wide range of cost structures c(x,y)'s (beyond just norms) are at least as tractable as their non-robust counterparts (and sometimes even more tractable). Finally, we discuss how to optimally choose the uncertainty size in a data-driven way using sound statistical criteria.

(This presentation is based on joint work with Y. Kang, K. Murthy, and F. Zhang.)

Tuesday Workshop Presenters (continued):

Title: Nonlinear Programming Formulation of Chance-constraints

Speaker: Andreas Waecheter
Affiliation: Northwestern University

Abstract: We present a smooth scenario-based approximation of joint chance constraints. In contrast to current mixed-integer formulations which explicitly count the scenarios that are enforced, the new formulation does not require discrete variables. This is achieved by replacing the probabilistic constraint by a nonparametric value-at-risk estimator. The estimator improves the historical VaR via a kernel, resulting in a better approximation and reducing the effect of spurious local minima introduced by the scenario approximation. In addition, by smoothing the discrete cumulative distribution function, we are able to obtain a constraint that is differentiable and can be handled by nonlinear programming techniques. We propose a tailored trust-region SQP method and present numerical experiments.

This is work in collaboration with Alejandra Pena-Ordieres and James Luedtke.

Title: LP, SOCP, and Optimization-Free Approaches to Polynomial Optimization

Speaker: Amir Ali Ahmadi
Affiliation: Princeton University

Abstract: We propose alternatives to sum of squares optimization that do not rely on semidefinite programming, but instead use linear programming, or second-order cone programming, or are altogether free of optimization. In particular, we present the first Positivstellensatz that certifies infeasibility of a set of polynomial inequalities simply by multiplying certain fixed polynomials together and checking nonnegativity of the coefficients of the resulting product. We also demonstrate the impact of our LP/SOCP-based algorithms on large-scale verification problems in control and robotics.

Joint work in part with Anirduha Majumdar (Princeton) and with Georgina Hall (Princeton).

Title: Stabilized Optimization via an NCL Algorithm

Speaker: Michael Saunders
Affiliation: Stanford University

Optimization problems min f(x) st c(x) >= 0 may have LICQ difficulties. We extend the BCL and LCL approaches of LANCELOT and MINOS to Algorithm NCL, whose subproblems optimize an augmented Lagrangian subject to constraints that satisfy LICQ:

min_{x,r} f(x) + yk'r + 1/2 rhok ||r||^2 st c(x) + r >= 0.

The variables r lead to many superbasic variables with active-set solvers like MINOS and SNOPT, but are easily accommodated by interior methods. We illustrate with Taxation Policy problems modeled in AMPL. Algorithm NCL converges in about 10 major iterations (independent of problem size), and IPOPT is able to warm-start each major iteration.

Joint work with Ding Ma, Kenneth Judd, and Dominique Orban.

Tuesday Workshop Presenters (continued):

Title: Reduced-Hessian Methods for Constrained Optimization

Speaker: Phillip Gill
University of California, San Diego

Abstract: Reduced-Hessian (RH) methods for unconstrained optimization are based on approximating the Hessian on an expanding sequence of subspaces spanned by the gradient vectors. In a limited-memory reduced-Hessian (L-RH) method the reduced Hessian is restricted in dimension by using a gradient space spanned by a subset of the preceding search directions. A projected-search method is proposed that extends the limited-memory variant of the reduced-Hessian method to problems with upper and lower bounds on the variables. The method uses a modified Wolfe line search that extends the conventional Wolfe line search to piecewise continuous functions. Numerical results are presented for the software package L-RH-B, which implements a limited-memory reduced-Hessian method based on the BFGS approximate Hessian. Finally, the use of L-RH-B in the sequential quadratic programming (SQP) package SNOPT is discussed. This is joint work with Michael Ferry and Elizabeth Wong.

Title: Primal-dual algorithms for the sum of two and three functions

Speaker: Ming Yan
Affiliation: Michigan State University

Abstract: There are several primal-dual algorithms for minimizing f(x)+g(x)+h(Ax), where f, g, and h are convex functions, f is differentiable with a Lipschitz continuous gradient, and A is a bounded linear operator. Two examples for minimizing the sum of two functions are Chambolle-Pock (f=0) and Proximal Alternating Predictor-Corrector (PAPC) (g=0). In this talk, I will introduce a new primal-dual algorithm for minimizing the sum of three functions. This new algorithm has the Chambolle-Pock and PAPC as special cases. It also enjoys most advantages of existing algorithms for solving the same problem. In addition, I will show that the parameters for PAPC can be relaxed. Then I will give some applications in decentralized consensus optimization.

Title: Multi-Agent Constrained Optimization of a Strongly Convex Function OverTime-Varying Directed Networks

Speaker: Serhat Aybat
Affiliation: Pennsylvania State University

Abstract: We consider cooperative multi-agent consensus optimization problems over possibly directed, time-varying communication networks, where only local communications are allowed. The objective is to minimize the sum of agent-specific possibly non-smooth composite convex functions over agent-specific private conic constraint sets; hence, the optimal consensus decision should lie in the intersection of these private sets. Assuming the sum function is strongly convex, we provide convergence rates in sub-optimality, infeasibility and consensus violation; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithm.

Tuesday Workshop Presenters (continued):

Title: Toward Conveniently Handling Bi-Level Optimization Problems

Speaker: David M. Gay
Affiliation: AMPL Optimization, Inc.

Abstract: Sometimes in reaching a decision that affects another party, one must consider how that party will react. Modeling this situation may give rise to a bi-level optimization problem in which an inner optimization problem models how the affected party will react and the overall problem optimizes the decision in light of this reaction. Extending the *problem* facility of the AMPL modeling language to allow expressing bi-level and multilevel optimization problems should make stating such problems straightforward. This talk describes such a possible extension and considers how to extend the AMPL/solver interface library to automatically formulate the necessary conditions for inner problems as constraints whose first and possibly second derivatives are computed by automatic differentiation. These computations would use machinery analogous to that used now for gradients and Hessians of ordinary algebraic objectives and constraints.

Wednesday Workshop Free Day- no presentations

Thursday Workshop Presenters:

Title: Max-k-sums

Speaker: Michael Todd,
Affiliation: Cornell University

Abstract: The max-k-sum of a set of real scalars is the maximum sum of a subset of size k, or alternatively the sum of the k largest elements. We study two extensions:
First, we show how to obtain smooth approximations to functions that are pointwise max-k-sums of smooth functions. Second, we discuss how the max-k-sum can be defined on vectors in a finite-dimensional real vector space ordered by a closed convex cone.

Title: Finding infimum point with respect to the second order cone, applications and extensions

Speaker: Farid Alizadeh,
Affiliation: Rutgers University

Abstract: We define the notion of infimum and supremum of a set of points with respect to the second order cone. These problems can be formulated as second order cone optimization and thus solvable by interior point methods in polynomial time. We present an extension of the simplex method to solve these problems. We outline both primal and dual versions of the simplex method. We also show some applications of infimum and supremum problems. In particular, application to the minimum ball containing a set of balls, and the maximum balls contained in the intersection of a set of balls, will be examined.
Joint work with Marta Cavaleiro, Rutgers University.

Title: Douglas-Rachford Splitting and ADMM for Pathological Conic Programs

Speaker: Wotao Yin
Affiliation: University of California; Los Angeles

Abstract: First-order methods such as ADMM and Douglas-Rachford splitting are known for their easy implementations and low per-iteration costs. What is less known is their usefulness for “solving” infeasible problems and unbounded (though feasible) problems. In this talk, we present a method for classifying infeasible, unbounded, and pathological conic programs based on Douglas-Rachford splitting, or equivalently ADMM. When an optimization program is infeasible, unbounded, or pathological, the z^k iterates of Douglas-Rachford splitting or ADMM diverge. Surprisingly, such divergent iterates still provide useful information, which our method uses for classification. They help us identify some of the cases where existing solvers cannot do reliably. When the problem is infeasible or weakly feasible, it is useful to know how to minimally modify the problem data to achieve strong feasibility, where “strong” makes it easier to find a solution. We also get this information via the divergent iterates.

This is joint work with Yanli Liu and Ernest Ryu.

Thursday Workshop Presenters (continued):

Title: Optimal performance of the steepest descent algorithm for quadratic functions

Speaker: Clovis C. Gonzaga
Affiliation: Federal University of Santa Catarina, Brasil

Abstract: In recent years there has been much interest on the most classical of all methods for minimizing differentiable functions in Rn, the steepest descent method. Each iteration takes a step along the negative gradient direction, and the only difference between specific algorithms is in the step lengths. Quadratic problems are used to study their performance. The existing choices for the step lengths require in the worst case O(C log(1/∈)) iterations to achieve a precision ∈, where *C* is the Hessian condition number, but modern strategies are much better in practice. We discuss these new methods and then use properties of Chebyshev polynomials to construct a sequence of step lengths with which the algorithm stops in O($\sqrt{C}$ log(1/∈))iterations, with a bound very similar to that of the Krylov space methods.

Title: Derivative-Free Robust Optimization by Outer Approximations

Speaker: Stefan Wild
Affiliation: Argonne National Laboratory

Abstract: We develop an algorithm for minimax problems that arise in robust optimization in the absence of objective function derivatives. The algorithm utilizes an extension of methods for inexact outer approximation in sampling a potentially infinite-cardinality uncertainty set. Clarke stationarity of the algorithm output is established alongside desirable features of the model-based trust-region subproblems encountered. We demonstrate the practical benefits of the algorithm on a new class of test problems. Joint work with Matt Menickelly

Title: Optimization of Noisy Functions via Quasi-Newton Methods

Speakers: Jorge Nocedal, Albert Berahas, Richard Byrd
Affiliation: Northwestern University

Abstract: We present a finite difference quasi-Newton method for the minimization of noisy functions. The method takes advantage of the scalability and power of BFGS updating, and employs an adaptive procedure for choosing the differencing interval h at every iteration, based on the noise estimation techniques of Hamming, as extended by More' and Wild. This noise estimation procedure and the selection of h are inexpensive but not always accurate, and to prevent failures the algorithm incorporates a recovery procedure that takes appropriate action in the case when the line search procedure is unable to produce an acceptable point. A novel convergence analysis is presented that considers the effect of a (noisy) line search procedure. We report results of numerical experiments comparing the method to a model based trust region method.

Thursday Workshop Presenters (continued):

Title: Directly (and quickly) optimizing prediction error and AUC of linear classifiers.

Speaker: Hiva Ghanbari
Affiliation: Lehigh University

Abstract: The predictive quality of most machine learning models is measured by expected prediction error or so-called Area Under the Curve (AUC). However, these functions are not used in the empirical loss minimization, because their empirical approximations are nonconvex and nonsmooth, and more importantly have zero derivative almost everywhere. Instead, other loss functions are used, such as logistic loss. In this work, we show that in the case of linear predictors, and under the assumption that the data has normal distribution, the expected error and the expected AUC are not only smooth, but have well defined derivatives, which depend on the first and second moments of the distribution. We show that these derivatives can be approximated and used in empirical risk minimization, thus proposing a gradient-based optimization methods for direct optimization of prediction error and AUC. Moreover the proposed algorithm has no dependence on the size of the dataset, unlike logistic regression and all other well-known empirical risk minimization techniques.

Friday Workshop Presenters (half day):

Title: Investigation of Crouzeix's Conjecture via Nonsmooth Optimization

Speaker: Michael Overton
Affiliation: New York University

Abstract: Crouzeix's conjecture is among the most intriguing developments in matrix theory in recent years. Made in 2004 by Michel Crouzeix, it postulates that, for any polynomial p and any matrix A, ||p(A)|| <= 2 max(|p(z)|: z in W(A)), where the norm is the 2-norm and W(A) is the field of values (numerical range) of A, that is the set of points attained by v*Av for some vector v of unit length. Crouzeix proved in 2007 that the inequality above holds if 2 is replaced by 11.08, and very recently this was greatly improved by Palencia, replacing 2 by 1+sqrt(2). Furthermore, it is known that the conjecture holds in a number of special cases, including n=2. We use nonsmooth optimization to investigate the conjecture numerically by attempting to minimize the "Crouzeix ratio", defined as the quotient with numerator the right-hand side and denominator the left-hand side of the conjectured inequality. We present numerical results that lead to some theorems and further conjectures, including variational analysis of the Crouzeix ratio at conjectured global minimizers. All the computations strongly support the truth of Crouzeix's conjecture. This is joint work with Anne Greenbaum and Adrian Lewis.

Title: The Ascendance of the Dual Simplex Method: A Geometric View

Speaker: Robert Fourer
Affiliation: Northwestern University

Abstract: First described in the 1950s, the dual simplex evolved in the 1990s to become the method most often used in solving linear programs. Factors in the ascendance of the dual simplex method include Don Goldfarb's proposal for a steepest-edge variant, and an improved understanding of the bounded-variable extension. The ways that these come together to produce a highly effective algorithm are still not widely appreciated, however. This talk employs a geometric approach to the dual simplex method to provide a unified and straightforward description of the factors that work in its favor.

Title:Optimal Decision Trees

Speaker: Oktay Gunluk
Affiliation: IBM Research

Abstract: Decision trees have been a very popular class of predictive models for decades due to their interpretability and good performance on categorical features. However, they are not always robust and tend to overfit the data. Additionally, if allowed to grow large, they lose interpretability. In this talk, we present a novel mixed integer programming formulation to construct optimal decision trees of a prescribed size. We take the special structure of categorical features into account and allow combinatorial decisions (based on subsets of values of features) at each node. We show that very good accuracy can be achieved with small trees using moderately sized training sets. The optimization problems we solve are tractable with modern solvers.

Friday Workshop Presenters (continued):

Title: When Does Stochastic Gradient Algorithm Work Well?

Speaker: Katya Scheinberg
Affiliation: Lehigh University

Abstract: We consider a general stochastic optimization problem which is often at the core of supervised learning, such as deep learning and linear classification. We analyze a standard stochastic gradient descent method with a fixed step size and propose a set of assumptions on the objective function, under which this method has the same convergence rate (to a neighborhood of the optimal solutions) as gradient descent and variance reduction methods. We then empirically demonstrate that these assumptions hold for logistic regression and standard deep neural networks on classical data sets. Thus our analysis helps explain when efficient behavior can be expected from the SGD method in training classification models and deep neural networks.