

Summer Course and Workshop
on
Optimization in Machine Learning

UNIVERSITY OF TEXAS AT AUSTIN

UT AUSTIN | PORTUGAL PROGRAM
IN THE AREA OF MATHEMATICS (CoLAB)

May 31 – June 7, 2011

Applied Computational Engineering and Sciences Building

<http://www.ma.utexas.edu/colab/Summer2011>

This Summer Course and Workshop is held under the auspices of the international partnership between the University of Texas at Austin and Portuguese Universities, as part of the UT Austin | Portugal Program in the Area of Mathematics (CoLab) and the Program in Applied and Computational Analysis, Research and Training Group (RTG-NSF), UT Austin, Department of Mathematics and the Institute for Computational Engineering and Sciences (ICES). This event is also part of the programs of the Portuguese Operations Research Society (APDIO) and the Portuguese International Center for Mathematics (CIM).

The Summer Course on Optimization in Machine Learning (May 31 – June 3, 2011) consists of two 10-hour modules given by Katya Scheinberg (Lehigh University) and Nati Srebro (Toyota Technological Institute at Chicago).

The Workshop on Optimization in Machine Learning (June 6-7, 2011) consists of 60-minute plenary talks and a number of contributed talks. The plenary speakers for the workshop are Kristin P. Bennett (Rensselaer Polytechnic Institute), Inderjit S. Dhillon (University of Texas at Austin), Sanjiv Kumar (Google Research), and Lieven Vandenberghé (University of California, Los Angeles).

Organizers: Omar Ghattas (University of Texas at Austin), Katya Scheinberg (Lehigh University), and Luis Nunes Vicente (University of Coimbra).

Summer Course on Optimization in Machine Learning

Course Schedule – all classes in Room ACE 6.304

May 31	9:30am - 12:30pm (coffee break 10:45-11:15) 2:00pm - 5:00pm (coffee break 3:15-3:45)
June 1	9:30am - 12:30pm (coffee break 10:45-11:15) 2:00pm - 5:00pm (coffee break 3:15-3:45)
June 2	9:30am - 12:30pm (coffee break 10:45-11:15) 2:00pm - 5:00pm (coffee break 3:15-3:45)
June 3	9:30am - 12:30pm (coffee break 10:45-11:15) 2:00pm - 5:00pm (coffee break 3:15-3:45)

Course Syllabus

This Summer Course introduces a range of machine learning models and optimization tools that are used to apply these models in practice. For the students with some Machine Learning background the course will introduce what lies behind the optimization tools often used as a black box as well as an understanding of the trade-offs of numerical accuracy and theoretical and empirical complexity. For the students with some Optimization background this course will introduce a variety of applications arising in Machine Learning and Statistics as well as novel optimization methods targeting these applications. The models we will cover include: support vector machines, sparse regression, sparse PCA, collaborative filtering, dimensionality reduction. The optimization methods will include interior point, active set, stochastic gradient, coordinate descent, cutting planes method.

Workshop on Optimization in Machine Learning

June 6

Plenary Talk – Room ACE 6.304

9:30-10:30 **Lieven Vandenberghe**, University of California, Los Angeles
*Convex optimization techniques for topology selection
in graphical models of time series*

10:30-11:00 **Coffee Break**

Plenary Talk – Room ACE 6.304

11:00-12:00 **Sanjiv Kumar**, Google Research
Learning compact hash codes for large-scale matching

12:00-2:00 **Lunch Break** (participants on their own)

Plenary Talk – Room ACE 6.304

2:00-3:00 **Inderjit S. Dhillon**, University of Texas at Austin
Fast and accurate low rank approximation of massive graphs

3:00-3:30 **Coffee Break**

Contributed Talks – Room ACE 6.304

3:30-4:00 **Zhiwei (Tony) Qin**, Columbia University
Structured sparsity via alternating directions methods

4:00-4:30 **Darren Rhea**, SAC Capital Advisors
ADMM on EC2 for GWAS

4:30-5:00 **Afonso Bandeira**, Princeton University
*Computation of sparse low degree interpolating polynomials
and their application to derivative-free optimization*

6:00-7:30 **Reception** (participants are invited)

Workshop on Optimization in Machine Learning

June 7

Contributed Talks – Room ACE 6.304

- 9:30-10:00 **Matthias Chung**, Texas State University, San Marcos
Designing optimal spectral filters for inverse problems
- 10:00-10:30 **Theja Tulabandhula**, MIT
The machine learning and traveling repairman problem
- 10:30-11:00 **Pradeep Ravikumar**, University of Texas at Austin
TBA

11:00-11:30 **Coffee Break**

Plenary Talk – Room ACE 6.304

- 11:30-12:30 **Kristin P. Bennett**, Rensselaer Polytechnic Institute
*Nonsmooth nonconvex optimization in machine learning
and data mining*
-

Titles and Abstracts – Plenary Talks

Nonsmooth nonconvex optimization in machine learning and data mining

KRISTIN P. BENNETT (RENSSELAER POLYTECHNIC INSTITUTE)

Many machine learning and data mining problems can be expressed as optimization models. To address problems on massive datasets and in real-time settings, scalable high performance algorithms are required to solve these models. Recently, nonsmooth optimization methods have led to massively scalable algorithms for solution of convex machine learning models such as Support Vector Machines. Nonconvex models can be used to elegantly capture learning tasks such as model selection within support vector machines and multiple instance learning, and visualization tasks such as annotated graph drawing with minimal edge crossings. This talk examines how novel nonsmooth nonconvex optimization methods can be analogously used to create scalable algorithms for solving these nonconvex machine learning and data mining models. Results are illustrated on compelling real-world problems in drug discovery and tuberculosis tracking and control.

Fast and accurate low rank approximation of massive graphs

INDERJIT S. DHILLON (UNIVERSITY OF TEXAS AT AUSTIN)

In this talk, I will present a fast and accurate procedure called clustered low rank matrix approximation for massive graphs. The procedure involves a fast clustering of the graph followed by approximation of each cluster separately using existing methods, e.g. the singular value decomposition, or stochastic algorithms. The clusterwise approximations are then extended to approximate the entire graph. This approach has several benefits: (1) important structure of the graph is preserved due to the clustering; (2) accurate low rank approximations are achieved; (3) the procedure is efficient both in terms of computational speed and memory usage. Further, we generalize stochastic algorithms into the clustered low rank approximation framework and present theoretical bounds for the approximation error. Finally, a set of experiments, using large scale and real world graphs, including massive social networks, show that our methods greatly outperform standard low rank matrix approximation algorithms.

Co-author: B. Savas.

Learning compact hash codes for large-scale matching

SANJIV KUMAR (GOOGLE RESEARCH)

Hashing based Approximate Nearest Neighbor (ANN) search in huge databases has attracted much attention recently due to their fast query time and drastically reduced storage needs. Linear projection based methods are particularly of great interest because of their efficiency and state-of-the-art performance. However, most of these methods either use random projections or extract principal directions from the data to learn hash functions. The resulting embedding suffers from poor discrimination when compact codes are used. In this talk I will describe a family of data-dependent projection learning methods such that each hash function is designed to correct the errors made by the previous one sequentially. These methods easily adapt to both unsupervised and semi-supervised scenarios and show significant performance gains over the state-of-the-art methods. I will also discuss a scalable graph-based hashing method for the data that lives on a low-dimensional manifold.

Convex optimization techniques for topology selection in graphical models of time series

LIEVEN VANDENBERGHE (UNIVERSITY OF CALIFORNIA, LOS ANGELES)

In a Gaussian graphical model, the topology of the graph specifies the sparsity pattern of the inverse covariance matrix. Several sparse topology selection methods based on 1-norm regularization and first-order methods for large-scale convex optimization have been proposed recently. In this talk we will first review some of these results and then discuss extensions to graphical models of Gaussian time series. We discuss the problem of maximum likelihood estimation of autoregressive models with conditional independence constraints and convex techniques for topology selection via nonsmooth regularization.

Titles and Abstracts – Contributed Talks

Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization

AFONSO BANDEIRA

Interpolation-based trust-region methods are an important class of algorithms for Derivative-Free Optimization which rely on locally approximating an objective function by quadratic polynomial interpolation models, frequently built from less points than there are basis components.

Often, in practical applications, the contribution of the problem variables to the objective function is such that many pairwise correlations between variables are negligible, implying, in the smooth case, a sparse structure in the Hessian matrix. To be able to exploit Hessian sparsity, existing optimization approaches require the knowledge of the sparsity structure. The goal of this paper is to develop and analyze a method where the sparse models are constructed automatically.

The sparse recovery theory developed recently in the field of compressed sensing characterizes conditions under which a sparse vector can be accurately recovered from few random measurements. Such a recovery is achieved by minimizing the ℓ_1 -norm of a vector subject to the measurements constraints. We suggest an approach for building sparse quadratic polynomial interpolation models by minimizing the ℓ_1 -norm of the entries of the model Hessian subject to the interpolation conditions. We show that this procedure recovers accurate models when the function Hessian is sparse, using relatively few randomly selected sample points.

Motivated by this result, we developed a practical interpolation-based trust-region method using deterministic sample sets and minimum ℓ_1 -norm quadratic models. Our computational results show that the new approach exhibits a promising numerical performance both in the general case and in the sparse one.

Co-authors: K. Scheinberg and L. N. Vicente

Designing optimal spectral filters for inverse problems

MATTHIAS CHUNG

Spectral filtering suppresses the amplification of errors when computing solutions to ill-posed inverse problems; however, selecting good regularization parameters is often expensive. In many applications, data is available from calibration experiments. In this talk, we describe how to use this data to pre-compute optimal spectral filters. We formulate the problem in an empirical Bayesian risk minimization framework

and use efficient methods from stochastic and numerical optimization to compute optimal filters.

Co-authors: J. Chung and D.P. O’Leary

Structured sparsity via alternating directions methods

ZHIWEI (TONY) QIN

We consider a class of sparse learning problems in high dimensional feature space regularized by a structured sparsity-inducing norm which incorporates prior knowledge of the group structure of the features. Such problems often pose a considerable challenge to optimization algorithms due to the non-smoothness and non-separability of the regularization term. In this paper, we focus on two commonly adopted sparsity-inducing regularization terms, the overlapping Group Lasso penalty ℓ_1/ℓ_2 -norm and the ℓ_1/ℓ_∞ -norm. We propose a unified framework based on the augmented Lagrangian method, under which problems with both types of regularization and their variants can be efficiently solved. As the core building-block of this framework, we develop new algorithms using an alternating partial-linearization/splitting technique, and we prove that the accelerated versions of these algorithms have an iteration complexity of $O(1/\sqrt{\epsilon})$. To demonstrate the efficiency and relevance of our algorithms, we test them on a collection of data sets and apply them to two real-world examples to compare the relative merits of the two norms.

Co-author: D. Goldfarb

On the use of variational inference for learning discrete graphical models

PRADEEP RAVIKUMAR

A natural estimator for learning the structure of a discrete graphical model is through the minimization of its ℓ_1 regularized negative log-likelihood. This objective while convex is however not available in closed-form. We thus study a class of “approximate” estimators that optimize ℓ_1 -regularized surrogate log-likelihoods instead, that are based on tractable variational approximations of the partition function of the graphical model. Indeed many state of the art methods can be shown to fall into this category. The resulting optimization problem can be solved naturally using composite gradient descent, since the gradients of the surrogate log-likelihood take the form of approximate marginals which can be obtained through an appropriate approximate inference procedure. However this involves performing graphical model inference for each of the gradient steps, which could be expensive.

To address this, we provide a message-passing algorithm that *directly* computes the solution of the ℓ_1 regularized approximate log-likelihood. Further, in the

case of certain reweighted entropy approximations to the partition function such as the tree-reweighted approximation, we show that surprisingly the ℓ_1 regularized approximate MLE estimator has a closed-form, so that we would no longer need to run through many iterations of approximate inference and message-passing. We show that in many cases such simple closed-form estimators have comparable statistical performance to the state of the art even though their computational complexity scales as $O(p^2)$, where p is the number of nodes in the graphical model, in contrast to more complex state of the art methods which scale as $O(p^5)$.

Co-author: E. Yang

The alternating directions method of multipliers on Amazon Elastic Compute Cloud with applications to Genome-Wide Association Studies

DARREN RHEA

As described in great detail in (S. Boyd et al 2011), the Alternating Direction Method of Multipliers (ADMM) is a simple to implement convex optimization algorithm which gives good results on a wide range of problems that occur in machine learning, and which can readily take advantage of Amazon’s Elastic Compute Cloud (EC2). Genome-wide association studies (GWAS) are a kind of genetic study that simultaneously measures hundreds of thousands of genetic loci in many individuals in order to find a small subset that are important for some trait of interest. We use ADMM and EC2 to implement a lasso penalized regression method similar to the one first proposed for GWAS in (Wu et al, Bioinformatics 2009). For a variety of continuous traits in the plant model species *Arabidopsis thaliana* our method identifies numerous important genes.

Co-author: M. Blom

The machine learning and traveling repairman problem

THEJA TULABANDHULA

The goal of the Machine Learning and Traveling Repairman Problem (ML&TRP) is to determine a route for a “repair crew,” which repairs nodes on a graph. The repair crew aims to minimize the cost of failures at the nodes, but as in many real situations, the failure probabilities are not known and must be estimated. We introduce two formulations for the ML&TRP, where the first formulation is sequential: failure probabilities are estimated at each node, and then a weighted version of the traveling repairman problem is used to construct the route from the failure cost. We develop two models for the failure cost, based on whether repeat failures are considered, or only the first failure on a node. Our second formulation is

a multi-objective learning problem for ranking on graphs. Here, we are estimating failure probabilities simultaneously with determining the graph traversal route; the choice of route influences the estimated failure probabilities. This is in accordance with a prior belief that probabilities that cannot be well-estimated will generally be low. It also agrees with a managerial goal of finding a scenario where the data can plausibly support choosing a route that has a low operational cost.

Co-authors: C. Rudin

List of Participants

Evan Archer
ICES, University of Texas at Austin
archer@ices.utexas.edu

Anna Ayzenshtat
Mathematics, University of Texas at Austin
anna.ayzenshtat@gmail.com

Afonso Bandeira
PACM, Princeton University and University of Coimbra
afonsobandeira@gmail.com

Caleb Bastian
PACM, Princeton University
cbastian@math.princeton.edu

Kristin P. Bennett
Mathematical and Computer Sciences, Rensselaer Polytechnic Institute
bennek@rpi.edu

James Blondin
Computer Science, Rensselaer Polytechnic Institute
blondj@rpi.edu

Marta Cavaleiro
Mathematics, University of Coimbra
martacav@gmail.com

Matthias Chung
Mathematics, Texas State University, San Marcos
mc85@txstate.edu

Inderjit S. Dhillon
Computer Sciences, University of Texas at Austin
inderjit@cs.utexas.edu

Russell Foltz-Smith
russell.foltzsmith@gmail.com

Patrick W. Gallagher
Computer Science, University of California, San Diego
patrick.w.gallagher@gmail.com

Rohollah (Nima) Garmanjani
Mathematics, University of Coimbra
nima@mat.uc.pt

Susana Gomes
Mathematics, University of Coimbra
susanatngomes@gmail.com

Prashanth Harshangi
Electrical Engineering, University of Southern California
harshang@usc.edu

Jason Jo
Computer Sciences, University of Texas at Austin
jjo@math.utexas.edu

Tipaluck Krityakierne
Applied Mathematics, Cornell University
tk338@cornell.edu

Sanjiv Kumar
Google Research, New York
sanjivk@google.com

Min Li
Mathematics, University of Coimbra
limin@mat.uc.pt

Weichang Li
Advanced Sensing and Analytics, Corporate Strategic Research, ExxonMobil
weichang.li@exxonmobil.com

Fahim Mannan
Computer Science, McGill University
fmannan@cim.mcgill.ca

Zhiwei (Tony) Qin
Industrial Engineering and OR, Columbia University
zq2107@columbia.edu

Pradeep Ravikumar
Computer Sciences, University of Texas at Austin
pradeepr@cs.utexas.edu

Darren Rhea
SAC Capital Advisors
darren.rhea@gmail.com

Kiran Sajjanshetty
Electrical Engineering, Texas A&M University
kiran_26@neo.tamu.edu

Katya Scheinberg
Industrial and Systems Engineering, Lehigh University
katyas@lehigh.edu

Vaibhav Shah
Production & Systems, Minho University
vaibhav.shah@dps.uminho.pt

Nati Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Georg Stadler
ICES, University of Texas at Austin
georgst@ices.utexas.edu

Joaquim Tinoco
Civil Engineering, Minho University
jabinoco@civil.uminho.pt

Seda Tolun
Quantitative Methods, School of Business, Istanbul University
tolunseda@gmail.com

Theja Tulabandhula
EECS, MIT
theja@mit.edu

Lieven Vandenbergh
Electrical Engineering and Mathematics, University of California, Los Angeles
vandenbe@ee.ucla.edu

Xu Wang
Computer Sciences, University of Texas at Austin
wangxu@cs.utexas.edu