IE598 Big Data Optimization

Introduction

Instructor: Niao He

Aug 23, 2016

A little about me

- Assistant Professor,
  UIUC, 2016 –
- Ph.D. in Operations Research,
  M.S. in Computational Sci. & Eng.
  Georgia Tech, 2010 – 2015
- B.S. in Mathematics,
  University of Sci. & Tech. of China,
  2006 – 2010

A little about the course

Big Data **Optimization**

- **Explore** modern optimization theories, algorithms, and big data applications
- **Emphasize** a deep understanding of structure of optimization problems and computation complexity of numerical algorithms
- **Expose to** the frontier of research in large-scale optimization and machine learning

Course Details

- **Prerequisites**: no formal ones, but assume knowledge in
  - linear algebra, real analysis, and probability theory
  - mathematical thinking and modeling

- **Textbooks:** no required ones, but recommend to read the listed references on website

- **Evaluation:**
  - *Scribing:* sign up 1~2 lecture per student
  - *Project:* some guidelines on website, details to come

Course Admin

- **Syllabus & Website**

  http://niaohe.ise.illinois.edu/IE598/ (pwd: fall2016)

- **Where to get help**
  - Email: niaohe@illinois.edu with [IE 598] in your subject
  - Office Location: 211 Transport Building
  - Office Hours: Tue. 10:00-11:00 or by appointment via email

- Issue with 3 or 4 credit hours
  - Peggy Regan, TB 104, *plregan@illinois.edu*

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Introduction

Era of Big Data

- Big data heat in academia
  - NIPS’16 conference

  at least 500/2500 submissions are about Big Data

- Big data heat in industry
  - LinkedIn:

  29,203 Data Scientist jobs in United States

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012

Capgemini Source: Davenport, T. H., & Patil, D. J. (2012). Data Scientist. *Harvard business review*
Business Information Management
Big Data & Analytics | October 2013
Copyright © 2013 Capgemini. All rights reserved. 8

Stay tuned for the
UIUC Big Data Symposium 2016
September 23-24

Really, what is *Big Data*?

Why is it so important?

- Big data analytics play a key role in various areas
  - Business and Industry
  - Social Statistics and Natural Recourses
  - Health and Medicine
  - Research and Science

Healthcare

Finance

Environment

Lifestyle

Aerospace

How to do data analysis?

Key Steps

- Pose a problem
- Collect data
- Pre-process and clean data
- Formulate a mathematical model
- Find a solution
- Evaluate and interpret the results

Optimization

What is Optimization?

- Find the optimal solution that minimize/maximize an objective function subject to constraints

$$
\begin{aligned}
\min_{x} \quad & f_0(x) \\
\text{s.t.} \quad & f_i(x) \leq 0, i = 1, \ldots, k \\
& h_j(x) = 0, j = 1, \ldots, \ell \\
& x \in X
\end{aligned}
$$

Why do we care?

Optimization lies at the heart of many fields, especially machine learning.

- Finance
  - Portfolio selection, asset pricing, etc.
- Electrical Engineering
  - Signal and image processing, control and robotics, etc.
- Industrial Engineering
  - Supply chain, revenue management, transportation etc.
- Computer Science
  - Machine learning, computer vision, etc.

Example – Portfolio Selection

- **Markowitz Mean-Variance Model**

$$\min_{w} \; w^T \Sigma w - \lambda \cdot R^T w$$
$$\text{s.t.} \; \sum_i w_i = 1$$

where

- w is a vector of portfolio weights
- R is the expected returns
- Σ is the variance of portfolio returns
- $\lambda > 0$ is the risk tolerance factor

Example – Image Denoising

- **Total Variation Denoising Model**

$$\min_x \sum_{(i,j)\in P} |x_{ij} - O_{ij}|^2 + \lambda \cdot TV(x)$$

where

  - x is image matrix
  - O is the noisy image, P is the observed entries
  - $TV(x)$ is the total variation

$$TV(x) = \sum_{ij} |x_{i+1,j} - x_{ij}| + |x_{i,j+1} - x_{ij}|$$

Example – Revenue Management

- **Newsvendor Model**

$$\max_{q \geq 0} \quad \mathbb{E}_D[p \cdot \min(q, D)] - c \cdot q$$

where

- q is number of newspaper to be stocked
- D is the random demand
- c is the unit purchase price
- p is the sell price

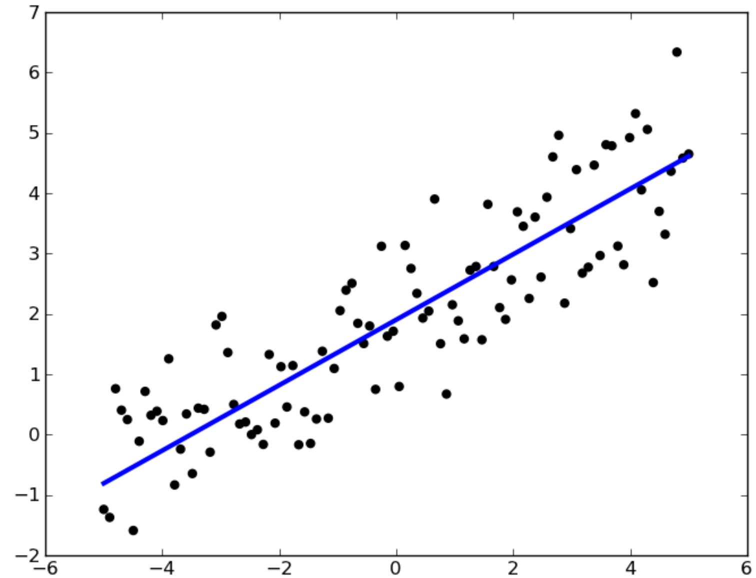
Example – Regression

- **Linear Regression Model**

$$\min_{w \in \mathbf{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

where

– x_i: predictor vector (feature)
– y_i: response vector (label)
– w: parameters to be learned
– n: number of data points



Example – Regularized Regression

- **Ridge Regression Model**

$$\min_{w \in \mathbf{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \frac{\lambda}{2}\|w\|_2^2$$

  - $||w||_2^2 = \sum_{j=1}^{d} w_j^2$ is the L_2-regularization

- **LASSO** (Least Absolute Shrinkage and Selection Operator)

$$\min_{w \in \mathbf{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\|w\|_1$$

  - $||w||_1 = \sum_{j=1}^{d} |w_j|$ is the L_1-regularization

Example – Classification

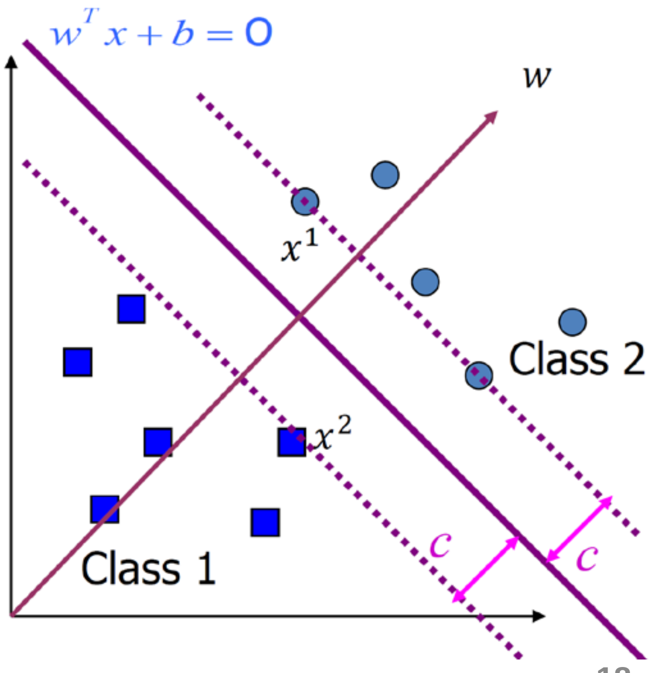
- **Maximum Margin Classifier Model**

$$\max_{w \in \mathbf{R}^d, b \in \mathbf{R}} \quad \frac{2c}{\|w\|_2}$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq c, i = 1, \ldots, n$$

where

- x_i: predictor vector (feature)
- $y_i \in \{1, -1\}$: label/class
- w: parameters to be learned
- n: number of data points



Example – More Classification

- **Soft Margin SVM** (support vector machine)

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \quad \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) + \frac{\lambda}{2} \|w\|_2^2$$

- **Logistic Regression**

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i(w^T x_i + b))) + \frac{\lambda}{2} \|w\|_2^2$$

Example – Maximum Likelihood Estimation

- Assume data points $x_1, \ldots, x_n$ are drawn i.i.d. from some distribution and we want to fit the data with a model $p(x|w)$ with parameter w, the maximum likelihood estimation is to solve

$$\max_{w} \log \prod_{i=1}^{n} p(x_i|w) = \sum_{i=1}^{n} \log(p(x_i|w))$$

  – Least square regression as a special case
  – Logistic regression as a special case

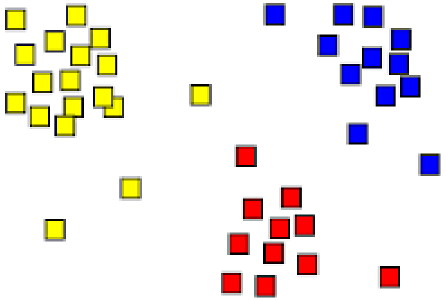
Example – Clustering

- **K-Means Model**

$$\min_{\substack{\mu_1,\ldots,\mu_k \\ C_1,\ldots,C_k}} \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - c_i\|^2$$

where

- $x_1, \ldots, x_n$: data
- $\mu_1, \ldots, \mu_k$: cluster centers to be learned
- $C_1, \ldots, C_k$: clusters to be assigned to



Many More Examples in ML

- Supervised learning (predictive models)
  - Regression
  - Classification
  - Neural networks
  - Boosting
- Unsupervised learning (data exploration)
  - Clustering (K-means)
  - Dimension reduction (PCA)
  - Density estimation
- Reinforcement learning
- Collaborative filtering
- Graphical models
- Active learning
- … …

Theme of This Course

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \ldots, k \\ & h_j(x) = 0, j = 1, \ldots, \ell \\ & x \in X \end{aligned}$$

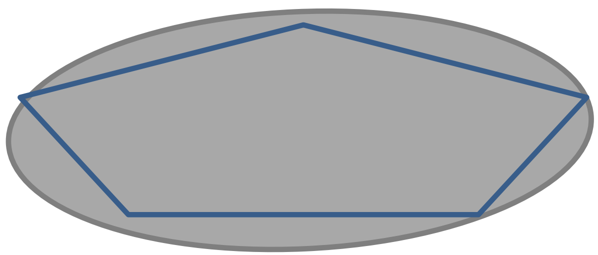
How to solve optimization problems efficiently in the new Big Data environment?

Structure of Optimization

- Linear vs. Nonlinear
- Deterministic vs. Stochastic
- Continuous vs. Combinatorial
- Smooth vs. Nonsmooth
- Convex vs. Nonconvex
- Separable vs. Non-separable
- Low-dimensional vs. High-dimensional
- Static vs. Online
- Single vs. Sequential Decision Making

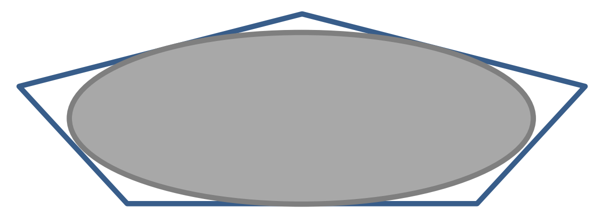
Easy or Hard?

What makes an optimization problem easy or hard?



Find minimum volume ellipsoid

NP-hard



Find maximum volume ellipsoid

Polynomial solvable

Example from L. Xiao, CS286 seminar

Easy or Hard?

What makes an optimization problem easy or hard?

$$\min_{x} \ c^T x$$
$$\text{s.t.} \ \ Ax \leq b$$

$$\min_{x} \ P(x)$$
$$\text{where } P(x) \text{ is polynomial}$$

Linear Optimization

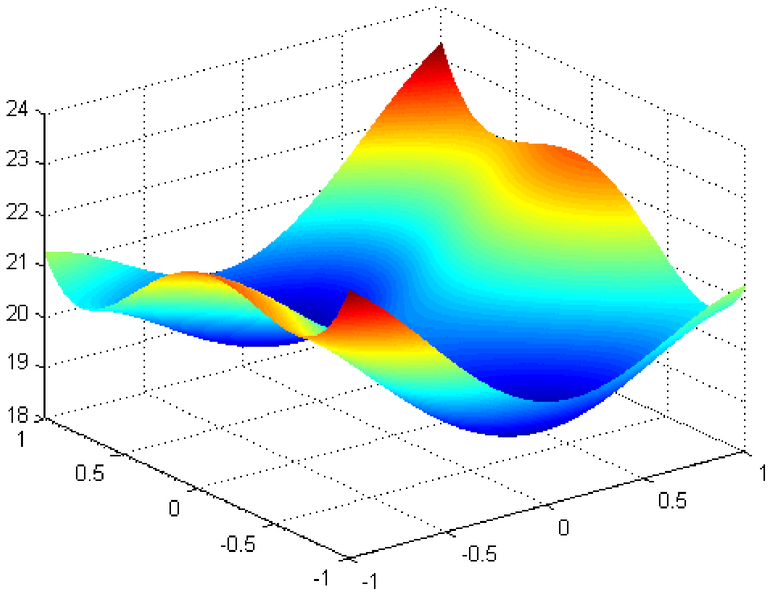
Polynomial Optimization

Polynomial solvable

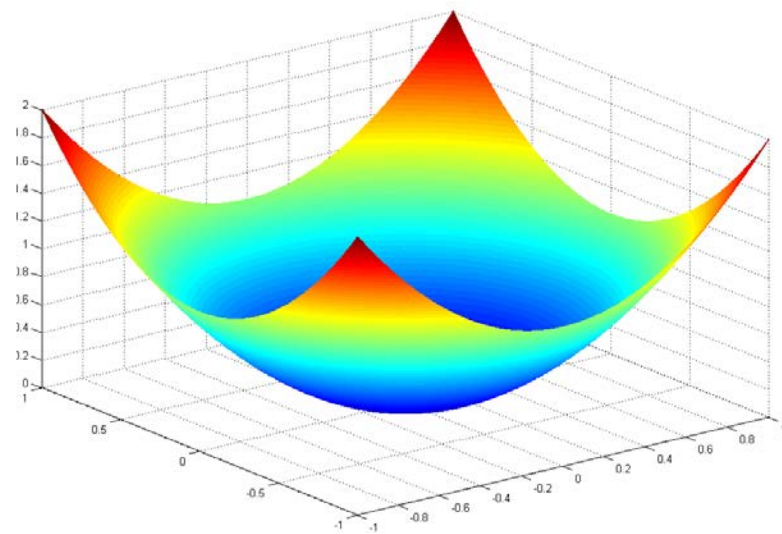
P ~ NP-hard

Complexity

> “ The great watershed in optimization isn’t between linearity and nonlinearity, but convexity and nonconvexity.”
>
> — R. Rockafellar, SIAM Review 1993



Non-Convex Optimization



Convex Optimization

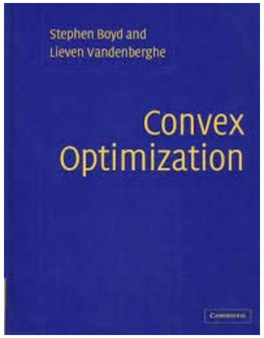
Types of Algorithms

- **Polynomial-time algorithms** (dates back to 1970s or so)
  - E.g., interior point method (IPM)
- **First-order algorithms** (dates back to 1900s, resurrection since 1980)
  - E.g., gradient descent method (GD)
- **Second-order algorithms**
  - E.g., Newton method, L-BFGS
- **Stochastic algorithms** (dates back to 1950s, resurrection since 2004)
  - E.g., stochastic approximation (SA)

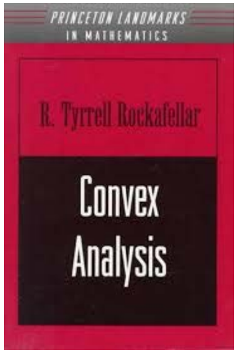
Central Topics

Convex Optimization

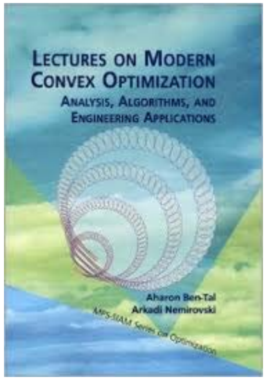
First-Order Methods



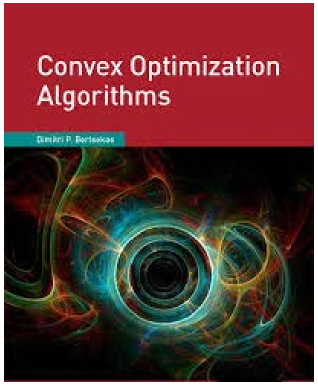
Boyd & Vandenberghe



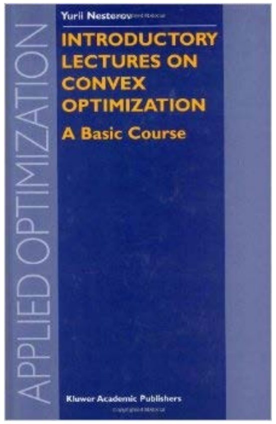
Rockafellar



Ben-Tal & Nemirovski



Bertsekas



Nesterov

Index Card

1. Name
2. Major
3. Class year (M.S. or Ph.D.?)
4. Have you taken an optimization course before? If so, what course?
5. What is your background in optimization? (none/limited/moderate/strong, etc.)
6. What are your expectations for this course? What do you hope to learn?
7. Tell me two interesting facts about yourself / hobbies.