L1. INTRODUCTION
Course Structure

Lecture: Mon, 10:15 – 12:00: optimization methods
Practice: Wed, 10:15 – 12:00: problems where methods are applied

Both @ OH12, R 1.055

Lecturer: Dr. Sangkyun Lee
Office Hour: Thurs 13:00 – 14:00, OH12, R 4.023

Lecture website:
Prerequisite

No prerequisite

But we will use many concepts from numerical optimization, linear algebra, machine learning, and mathematical statistics

- Lecture notes from NOPT SS14 will help for optimization background (more info on lecture website)
- Some will be explained during the lecture briefly
- It is highly recommended to try to learn unfamiliar concepts for yourself, using e.g. Wikipedia
Final Exam

Final exam is an oral presentation of a mini-project

Mini-Project

• Topic: choose Method + Problem of your interest
• Create a point of interest (hypothesis), for example:
  - Will method A perform better on problem P than existing ones?
  - Will a change in method A make it faster for a certain problem?
  - Does the performance of methods A, B, C predicted by theory be well reflected by their real-world performance?
• Design and perform experiments to validate the hypothesis
• Things can be done in one month
Mini-Project Examples

Method Focused

Own implementation of optimization methods

Application Focused

Use available software for optimization

Investigating an optimization method well-suited for your own research problem

• E.g. does the accelerated first-order method performs better on your large-scale regression problem, where the typical method of choice is a second-order method (Newton)?

Benchmarking several opt methods for a certain problem

• Which method will be the best choice, for a certain setting?
• Theoretical performance vs. real performance
Mini-Project: More Ambitious

AR Drone

- 1080p front-facing camera
- Data streams over wifi
- Command streams over wifi

Topics?

- Digital image stabilization
- Object recognition
- Path-planning
- ...
- https://www.youtube.com/watch?v=w2itwFJChgFQ

Risk: I can provide only a single drone!
Mini-Project: More Ambitious

Oculus Rift

- VR (Virtual Reality) device
- Where to use? Find them on youtube!
- My Devkit 2.0 order will arrive in 6~8 weeks
Mini-Project Structure

One Mini-Project per a Small Group

- Small group: 1 or 2 students
- If 2, their contribution to the project should be equal, and total contribution should be twice as much as 1 person case
  - One focusing on a variant of an optimization method + one on reformulation of a research problem

Qualification for the Final Exam

- A satisfactory project proposal, by 02.02.15
- 3~5 single column pages + citation
- Proposal presentation (~10min) in mid Jan 15, to get feedback from class
- No second chance after the deadline
Mini-Project Structure

Methods and problems can be chosen from the lecture

Feel free to discuss with your friends, colleagues, advisors, and the lecturer, to determine the topic for your mini-projects

Once the topic is determined, the work has to be on your own

Final Presentation (tentative: March 19~20, 2015)

• 1 Person project: 20 min + 5 min Q/A
• 2 People project: 30 min + 5 min Q/A
Final Exam / Proposal Grading Criteria

Motivation (3)
  • Why is the problem (setting) is interesting for you or the research community?

Understanding (4)
  • How well do you understand the optimization method you’re using? Also the problem?

Creativity / Contribution (3)
  • What is your reason to choose a particular method, a way of modifying a method, etc., in your project?
  • What is your contribution to the research community?

Evaluation (4)
  • How well experiments are designed to validate the hypothesis

Presentation (3)
  • How well things are organized/presented? How well the talk was practiced/prepared?

Repeatability (3)
  • Open source / a script to run experiments & show results
Homework

Reading Assignments

• Method / Problem papers will be assigned about a week before they’ll be discussed
• Pre-reading of assigned papers is required for class
• Readings on SGD (Stochastic Gradient Descent) are already on the website

Bonus Points

• If you answer my questions or make a good comment/point about the papers in discussion, you get +1 or +2 bonus point(s)
• You can earn up to 10 bonus points, which will add 10% to your final exam grade
Final Exam Grading

Final grade points = total 20 points

<table>
<thead>
<tr>
<th>1.0</th>
<th>1.3</th>
<th>1.7</th>
<th>2.0</th>
<th>2.3</th>
<th>2.7</th>
<th>3.0</th>
<th>3.3</th>
<th>3.7</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-19</td>
<td>18-17</td>
<td>16-15</td>
<td>14-13</td>
<td>12-11</td>
<td>10-9</td>
<td>8-7</td>
<td>6-5</td>
<td>4-3</td>
<td>2-1</td>
</tr>
</tbody>
</table>

+10 bonus points = +2 final grade points
Questions?
What we’ll discuss?

Methods to find solutions of optimization problems:

\[
\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } x \in C
\]

In Large-Scale:

- The number of variables \( n \) is large
- Computing the objective function \( f(x) \) involves many data points
- The number of constraints describing the constraint set \( C \) is large
  - \( C \) is described by many data points
  - \( C \) is a discrete set
Why Optimization?

Optimization provides a framework to:

1. Translate the idea/problem into standard forms (OR)

\[
\min_{x \in \mathbb{R}^n} \ f(x) \\
\text{s.t. } x \in C
\]

2. Find solutions of the standard forms (MP)
Optimizations in Data Analysis

Machine Learning / Statistics

• Regression, Classification
• Maximum likelihood estimation
• Matrix completion (collaborative filtering)
• Robust PCA
• Gaussian Markov random field
• Dictionary learning
• …

Signal Processing

• Compressed sensing
• Image denoising, deblurring, inpainting
• Source separation
• …
Considerations for Large-Scale Efficient Algorithms

- Faster convergence rate
- Lower per-iteration cost

Relaxations

- Find alternative formulations that are easier to solve
  - E.g. QP → LP, MIP → SDP
- “Separable” relaxations for better parallelization

Approximations

- Stochastic approximations to deal with large volume of data
Support Vector Machines

Data: \((x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \{+1, -1\}, \ i = 1, 2, \ldots, m\)

Primal form of the soft-margin SVM

- \(n+m+1\) variables
- \(2m\) constraints

\[
\begin{align*}
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{s.t.} & \quad \xi_i \geq 1 - y_i (\langle w, x_i \rangle + b), \ i = 1, 2, \ldots, m \\
& \quad \xi_i \geq 0, \ i = 1, 2, \ldots, m.
\end{align*}
\]
**SVM**

Primal:

\[
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{s.t. } \xi_i \geq 1 - y_i (\langle w, x_i \rangle + b), \ i = 1, 2, \ldots, m \\
\xi_i \geq 0, \ i = 1, 2, \ldots, m.
\]

Dual:

\[
\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \alpha^T D_y K D_y \alpha - e^T \alpha \\
\text{s.t. } y^T \alpha = 0 \\
0 \leq \alpha_i \leq C, \ i = 1, 2, \ldots, m.
\]

**Primal form \(\rightarrow\) dual form**

- \(n+m+1\) variables \(\rightarrow\) \(m\) variables
- \(2m\) constraints \(\rightarrow\) \(m+1\) constrains
SVM: Unconstrained Formulation

Primal: Constrained
\[
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{s.t. } \xi_i \geq 1 - y_i(\langle w, x_i \rangle + b), \ i = 1, 2, \ldots, m \\
\xi_i \geq 0, \ i = 1, 2, \ldots, m.
\]

Primal: Unconstrained
\[
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \max\{1 - y_i(\langle w, x_i \rangle + b), 0\}
\]

Smooth vs. nonsmooth objective functions:
• Which is better for large m?
Sparse Regression

Data: data (design) matrix $X$, response $y$

\[ X \in \mathbb{R}^{m \times n} \quad y \in \mathbb{R}^{m} \]

Find sparse coef vector beta so that

\[ y \approx X \beta \]

Application: e.g. feature selection
Sparse Regression: LASSO

Least Absolute Shrinkage and Selection Operator [Tibshirani, 96]

\[ \min_{\beta \in \mathbb{R}^n} \| y - X \beta \|^2 \quad \text{s.t.} \quad \| \beta \|_1 \leq \gamma \]

\[ \min_{\beta \in \mathbb{R}^n} \| y - X \beta \|^2 + \lambda \| \beta \|_1 \]

Properties:

• Convex optimization
• Exact zeros in solution
• Entire solution path (for the change of gamma or lambda) is piecewise linear
• Solutions are consistent (under certain conditions)
Regularized Convex Formulations

a.k.a. Tikhonov regularization

\[
\min_{x \in \mathbb{R}^n} f(x) + \Psi(x)
\]

Many of our discussions will be about this type of problems

\( f(x) \)
- Convex
- Smooth or nonsmooth
- Evaluates a form of error on data points
- Complicated

\( \psi(x) \)
- Convex
- Nonsmooth
- Regularize predictors
- Simple
How large is Large/Big?

In my grad lectures, teachers said > 10k vars is medium scale, >100k vars large-scale

In this lecture, let’s roughly say $n \times m > 1\text{mil}$
Agenda

Methods

- SGD
- FISTA, SpaRSA
- Nesterov’s accelerated first-order methods
- Projected gradient, conditional gradient (Frank-Wolfe)
- (B)CD
- ADMM
- LBFGS
- Natural gradient
- LD, SDP

Problems/Practice

- LASSO
- SVM
- Compressed sensing
- Matrix completion
- Gaussian MRF
- Optimization Software
- Deep NN
- Robust PCA