Numerical Optimization

L0. INTRODUCTION

TU Dortmund, Dr. Sangkyun Lee

Course Structure

Everything in English!

Lecture: Mon, 10:15 – 12:00 : optimization theory / methods

Practice: Wed, 10:15 – 12:00 : Julia / demo / homework discussion

Place: OH12, R 1.056

Lecturer: Dr. Sangkyun Lee

Office Hour: By appointment, OH12, R 4.023

Lecture website: check for topics, no lectures, etc.

http://tinyurl.com/nopt-w16

Prerequisite

No prerequisite, but math skills will be helpful

We will cover necessary concepts in class

- We'll review required math concepts next week
- Self-study of unfamiliar concepts is highly encouraged

Homework

HW will be assigned in every 2~3 weeks (total ~5 hw's)

HW will consist of:

- Simple proofs
- Solving optimization problems
- Implementing/using optimization algorithms in Julia

HW's will NOT be graded ☺

Ubung HW sessions, you need to present your answers!

- 2~3 correct solutions will be needed, to pass Ubung and to be qualified for the final exam

Exams:

Exams will be <u>WRITTEN tests</u>, NOT ORAL

Exam questions will be mostly from <u>homework</u> problems

- **Mid-Term (before Christmas: Dec 14th or 21st) : 50%**
- **Final Exam (tentative: Feb 15): 50%**
  - Coverage: midterm ~ the last lecture

Textbook / Lecture Notes

No textbook is required, but the following text is recommended:

Numerical Optimization

J. Nocedal and S. Wright, 2nd Ed, Springer, 2006

Lecture notes will be uploaded after each class

Question?

Optimization

Methods to find solutions of mathematical programs (MPs):

$$\underbrace{\min_{x \in \mathbb{R}^n}}_{\text{Optimization Variable}} \ \underbrace{f(x)}_{\text{Objective Function}} \ \text{ subject to } \ x \in \underbrace{\mathcal{C}}_{\text{Constraint Set}}$$

Why Optimization?

Operations Research

Mathematical Programming

Idea / Problem

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

MP
(Mathematical Program)

$$x^*$$

Solution

Images from shutterstock

Optimizations is a fundamental tool in...

Machine Learning / Statistics

- Regression, Classification
- Maximum likelihood estimation
- Matrix completion (collaborative filtering)
- Robust PCA
- Graphical models (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

Total cost

Separability

- Separable reformulations for parallelization

Relaxations

- Find relaxed formulations that are easier to solve
  - E.g. QP → LP, MIP → SDP

Approximations

- Stochastic approximations to deal with large volume of data

Ex. Data Analysis

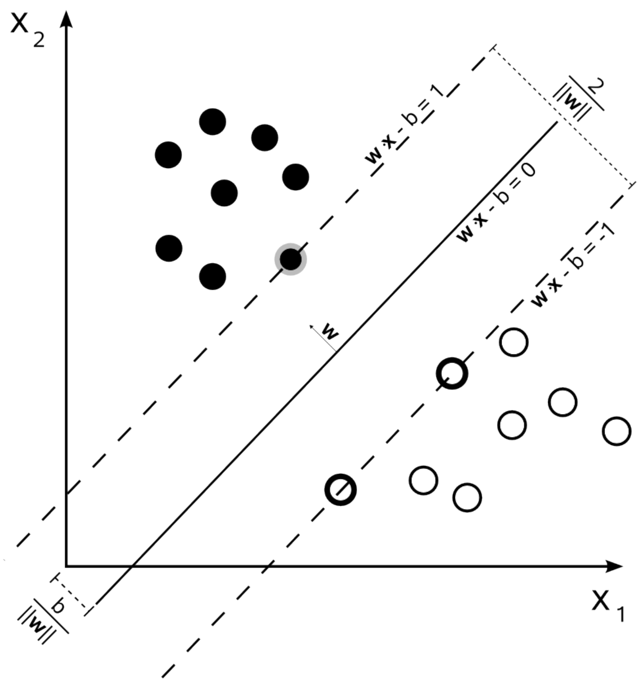
Classification Problem:

We're given m data points (in n dimensions) which belong to two categories. Find a predictor to classify new data point into the two categories, based on the given data.

Be robust against memorization (aka overfitting)!

Support Vector Machines

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



$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t. } \xi_i \geq 1 - y_i(\langle w, x_i \rangle + b), \; i = 1, 2, \ldots, m$$

$$\xi_i \geq 0, \; i = 1, 2, \ldots, m.$$

Primal form of the soft-margin SVM

- n+m+1 variables
- 2m constraints

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:

$$\min_{\alpha\in\mathbb{R}^m} \ \frac{1}{2}\alpha^T D_y K D_y \alpha - e^T\alpha \qquad K_{ij} = \langle x_i, x_j\rangle$$

$$\text{s.t. } \ y^T\alpha = 0$$

$$0 \leq \alpha_i \leq C,\ \ \ i = 1, 2, \ldots, m.$$

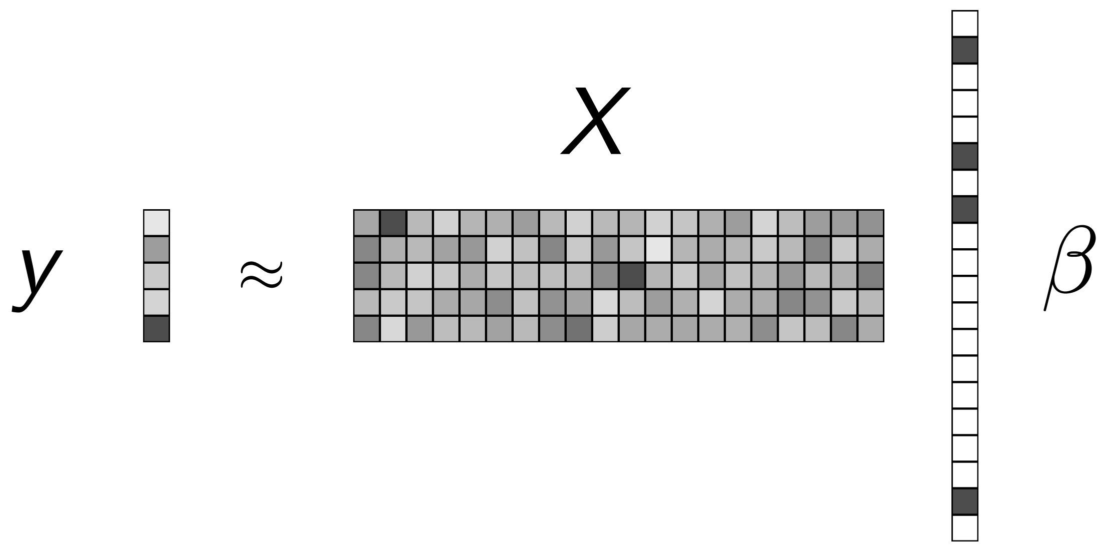
Primal form → dual form

- n+m+1 variables → m variables
- 2m constraints → 2m (simple) + 1 constrains
- Can we solve the dual, instead of the primal ?

Sparse Coding

Data: data (design) matrix X, response y $X \in \mathbb{R}^{m \times n}$ $y \in \mathbb{R}^{m}$

Find a sparse coef vector beta that best predicts responses y



Application: e.g. biomarker discovery from genetic data

Sparse Coding: LASSO

Least Absolute Shrinkage and Selection Operator [Tibshirani, 96]

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

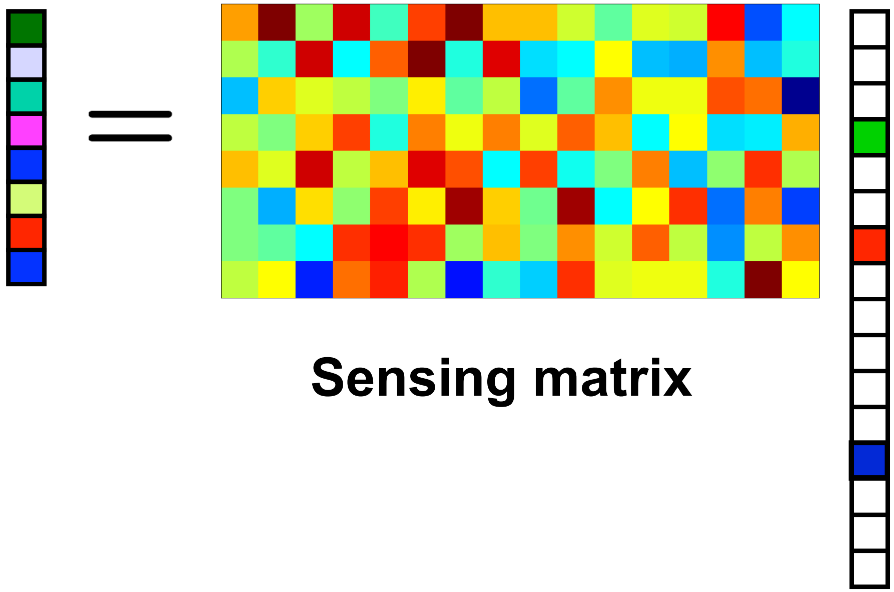
$$\min_{\beta \in \mathbb{R}^n} \|y - X\beta\|^2 + \lambda \|\beta\|_1$$

Properties:

- Convex optimization
- Exact zeros in solution

Compressed Sensing

$y \qquad A \in \mathbb{R}^{k \times n} \qquad x \in \mathbb{R}^n$



An inverse problem of dimensionality reduction:
can we reconstruct the original signal from observations?

(Figure adapted from R.Baraniuk's talk slides)

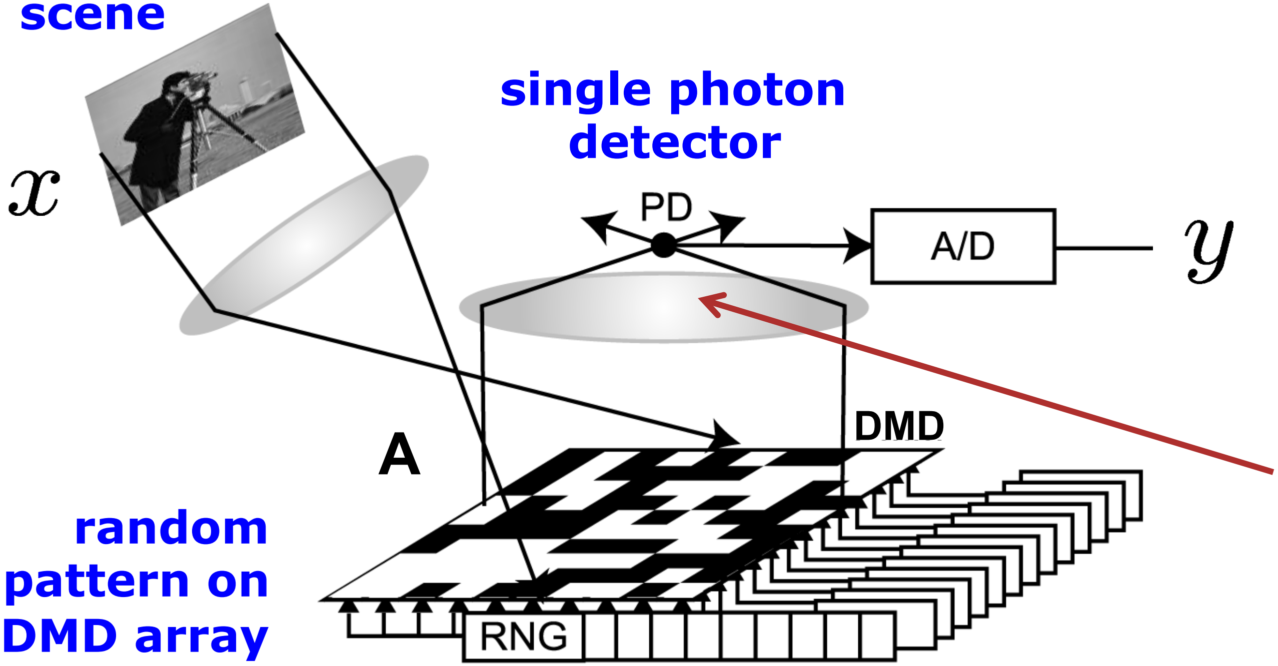
(Slide adapted from R.Baraniuk's talk)

Single-Pixel Camera

scene

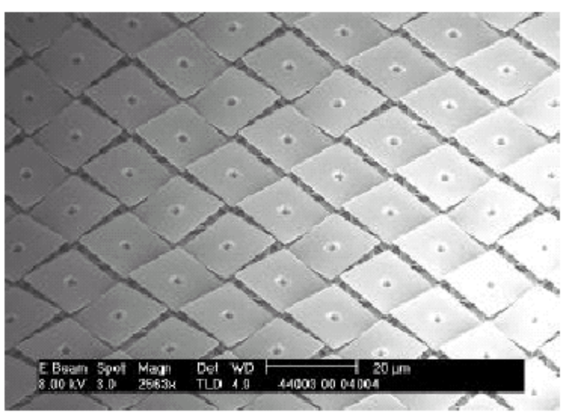
single photon detector

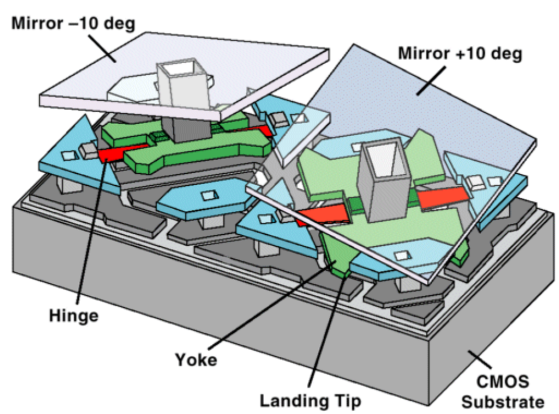
image reconstruction or processing



"inner product"

random pattern on DMD array



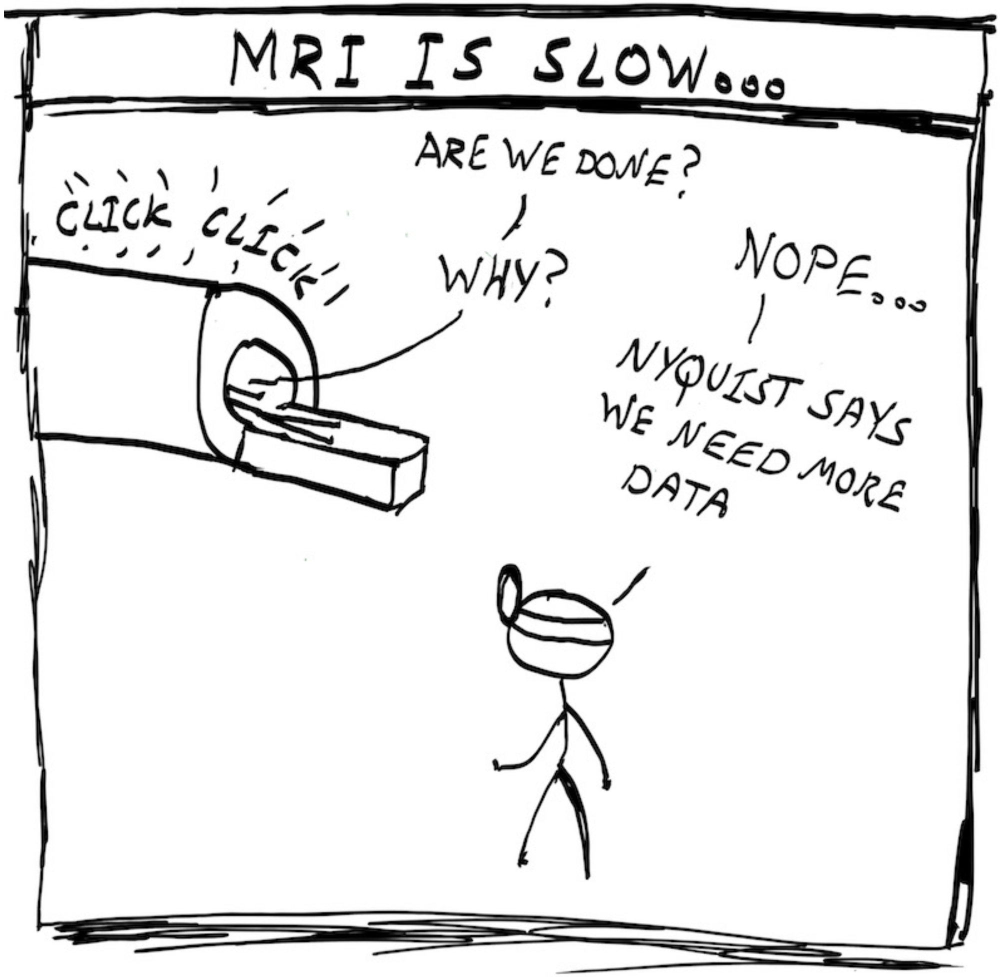


w/ Kevin Kelly

DLP
TEXAS INSTRUMENTS

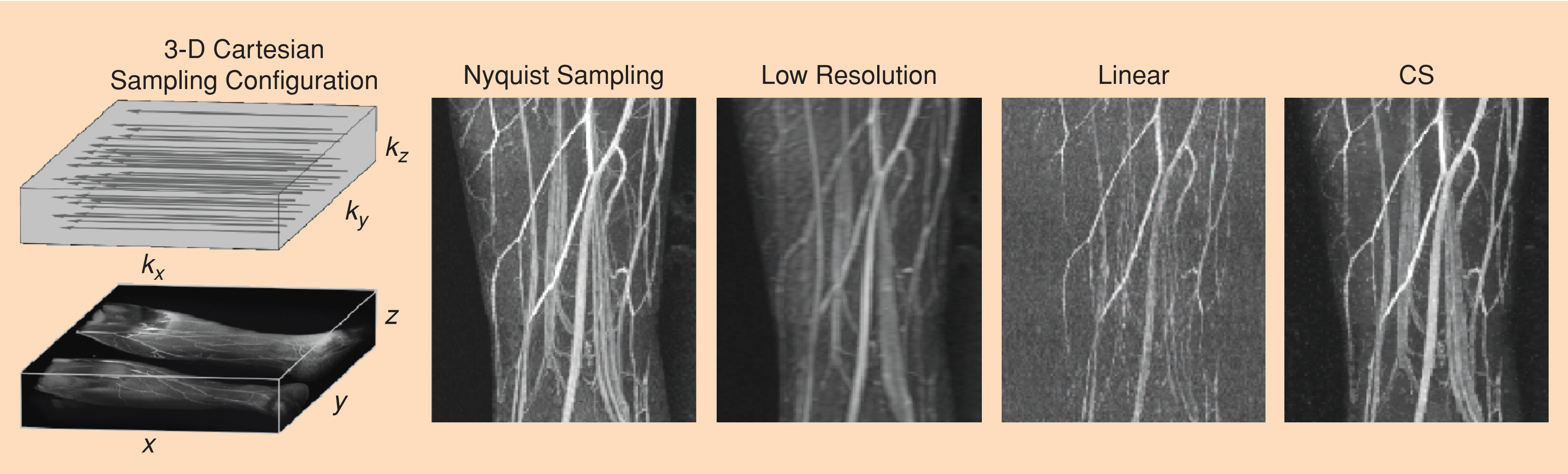
INVIEW

Magnetic Resonance Imaging



http://www.eecs.berkeley.edu/~mlustig/CS.html

Speeding up MRI by CS



[FIG8] 3-D Contrast enhanced angiography. Right: Even with 10-fold undersampling CS can recover most blood vessel information revealed by Nyquist sampling; there is significant artifact reduction compared to linear reconstruction; and a significant resolution improvement compared to a low-resolution centric k-space acquisition. Left: The 3-D Cartesian random undersampling configuration.

Compressed Sensing MRI, Lustig, Donoho, Santos, and Pauly, IEEE Signal Processing Magazine, 72, 2008

A Bigger Picture

Programming Language

Parallel computing (e.g. GPGPU)

Distributed data

Idea / Problem

$$\min_{x \in \mathbb{R}^n} \; f(x)$$
$$\text{s.t.} \;\; x \in \mathcal{C}$$

$$x^*$$

Data structure

Computation cost

Energy usage

Machine Learning
Statistical Data Analysis

Agenda

Theory

- Optimality Conditions, KKT
- Rate of Convergence
- Duality

Method

- Gradient Descent
- Quasi-Newton Method
- Conjugate Gradient

- Proximal Gradient Descent
- Stochastic Gradient Descent
- ADMM

The Julia Language

More on Wed