# 凸优化简介

文再文

wenzw@math.pku.edu.cn
北京大学— 北京国际数学研究中心

Acknowledgement: Prof. Lieven Vandenberghe and Prof. Wotao Yin

# 从解线性方程组谈起

Given matrix $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$

$$Ax = b$$

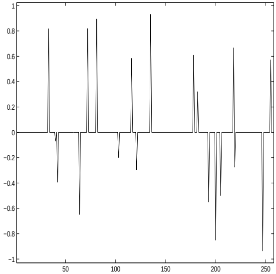
- structure of $A$: dense, banded, sparse . . . ?
- factorization: Cholesky, QR, eigenvalue

Suppose $m < n$, find a sparsest solution?

# Compressive Sensing

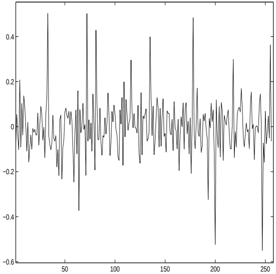
Find the sparest solution

- Given n=256, m=128.
- A = randn(m,n); u = sprandn(n, 1, 0.1); b = A*u;



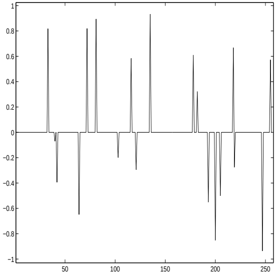

$$\begin{cases} \min\limits_{x} \|x\|_0 \\ \text{s.t. } Ax = b \end{cases}$$

(a) $\ell_0$-minimization




$$\begin{cases} \min\limits_{x} \|x\|_2 \\ \text{s.t. } Ax = b \end{cases}$$

(b) $\ell_2$-minimization

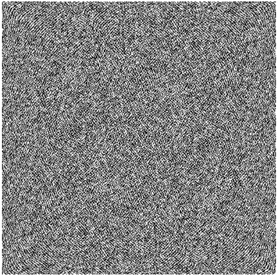



$$\begin{cases} \min\limits_{x} \|x\|_1 \\ \text{s.t. } Ax = b \end{cases}$$

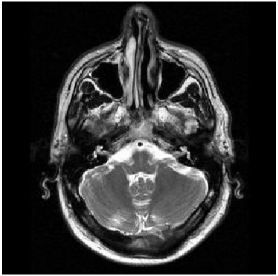
(c) $\ell_1$-minimization

# MRI: Magnetic Resonance Imaging


(a) MRI Scan



(b) Fourier Coefficients



(c) Image

Is it possible to cut the scan time into half?

# MRI: Magnetic Resonance Imaging

- MR images often have sparse sparse representations under some wavelet transform $\Phi$
- Solve
$$\min_u \|\Phi u\|_1 + \frac{\mu}{2}\|Ru - b\|^2$$
$R$: partial discrete Fourier transform
- The higher the SNR (signal-noise ratio) is, the better the image quality is.

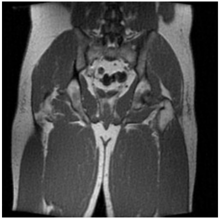
(a) full sampling

(b) 39% sampling, SNR=32.2

# MRI: Magnetic Resonance Imaging

(a) full sampling

(b) 39% sampling,
SNR=32.2



(c) 22% sampling,
SNR=21.4

(d) 14% sampling,
SNR=15.8

# Compressive sensing

Standard Acquisition: signal $x \in \mathbb{R}^n$

- Sample and compress: subject to the Nyquist rates
- Analog-to-digital converters may reach speed limit
- Time, power, speed, ... can become bottlenecks

Compressive Sensing: signal $x \in \mathbb{R}^n$

- Acquire less data $b_i = a_i^\top x^*$, $i = 1, \ldots, m \ll n$
- $A$ should be “random-like”
- Decoding is costly: recover $x^*$ from $Ax = b$

Difference: acquisition size reduced from $n$ to $m$

# Decoding in CS

Given $(A, b, \Psi)$, find the sparsest point:

$$x^* = \arg\min\{\|\Psi x\|_0 : Ax = b\}$$

From combinatorial to convex optimization:

$$\bar{x} = \arg\min\{\|\Psi x\|_1 : Ax = b\}$$

1-norm is sparsity promoting

- Basis pursuit (Donoho et al 98)
- Many variants: $\|Ax - b\|_2 \leq \sigma$ for noisy $b$
- Greedy algorithms
- Theoretical question: when is $\|\cdot\|_0 \leftrightarrow \|\cdot\|_1$ ?

# Sufficient condition for recovery

- $\{x : Ax = b\} \equiv \{\bar{x} + v : v \in Null(A)\}$
- $\bar{x}$ uniquely solves $\ell_1$-problem iff

$$\|\bar{x} + v\|_1 > \|\bar{x}\|_1, \forall v \in Null(A)$$

- Let $S = \{i : \bar{x}_i \neq 0\}$ and $Z = \{i : \bar{x}_i = 0\}$, we have

$$\begin{aligned} \|\bar{x} + v\|_1 &= \|\bar{x}_S + v_S\|_1 + \|0 + v_Z\|_1 \\ &= \|\bar{x}\|_1 + (\|v_Z\|_1 - \|v_S\|_1) + \\ &\quad (\|\bar{x}_S + v_S\|_1 - \|\bar{x}_S\|_1 + \|v_S\|_1) \\ &\geq \|\bar{x}\|_1 + (\|v_Z\|_1 - \|v_S\|_1) \end{aligned}$$

- Hence, $\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1$ if $\|v_Z\|_1 - \|v_S\|_1 \geq 0$
- $\|v_S\|_1 \leq \sqrt{|S|}\|v_S\|_2 \leq \sqrt{\|\bar{x}\|_0}\|v\|_2$,
- Sufficient condition: $\sqrt{\|\bar{x}\|_0} < \frac{1}{2}\frac{\|v\|_1}{\|v\|_2}, \quad \forall v \in \mathrm{Null}(A) \setminus \{0\}$

# Sufficient condition for recovery

- $1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}, \quad \forall v \in \mathbb{R}^n \setminus \{0\}$
- Garnaev and Gluskin established that for any natural number $p < n$, there exist $p$-dimensional subspaces $V_p \subset \mathbb{R}^n$ in which

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{C\sqrt{n-p}}{\sqrt{\log(n/(n-p))}}, \forall v \in V_p \setminus \{0\},$$

- vectors in the null space of $A$ will satisfy, with high probability, the Garnaev and Gluskin inequality for $V_p = \mathrm{Null}(A)$ and $p = n - m$.
- for a random Gaussian matrix $A$, $\bar{x}$ will uniquely solve $\ell_1$-min with high probability whenever

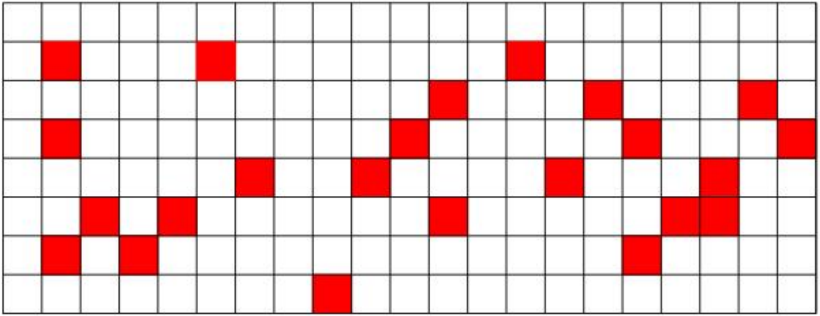
$$\|\bar{x}\|_0 < \frac{C^2}{4} \frac{m}{\log(n/m)}.$$

# Algorithmic challenges

- Has large-scale and/or dense data in practice
- Has a nonsmooth objective function
- CS: small errors in A or b can cause large errors in the solution (when A does not obey the RIP)
- Linear algebra in matrix problems are much more expensive
- Standard (simplex, interior-point) methods not suitable

# Netflix Problem: 1 million dollar award

- Given $m$ movies $x \in \mathcal{X}$ and $n$ customers $y \in \mathcal{Y}$
- predict the "rating" $W(x, y)$ of customer $y$ for movie $x$
- training data: known ratings of some customers for some movies
- Goal: complete the matrix
- other applications: collaborative filtering, system identification, etc.



# Matrix Rank Minimization

Given $X \in \mathbb{R}^{m\times n}$, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, $b \in \mathbb{R}^p$, we consider

- matrix completion problem:

$$\min \ \mathrm{rank}(X), \ \text{s.t. } X_{ij} = M_{ij}, (i,j) \in \Omega$$

- the matrix rank minimization problem:

$$\min \ \mathrm{rank}(X), \ \text{s.t. } \mathcal{A}(X) = b$$

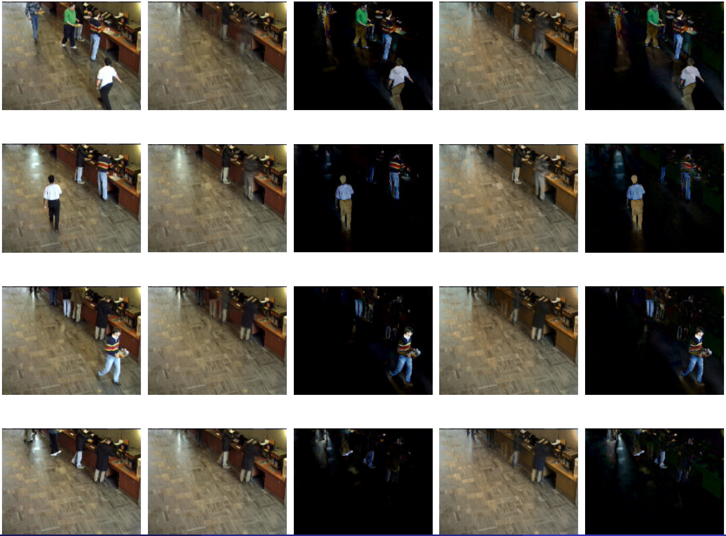
- nuclear norm minimization:

$$\min \ \|X\|_* \ \text{ s.t. } \mathcal{A}(X) = b$$

where $\|X\|_* = \sum_i \sigma_i$ and $\sigma_i = i$th singular value of matrix $X$.

# Video separation

- Partition the video into moving and static parts



# Sparse and low-rank matrix separation

- Given a matrix $M$, we want to find a low rank matrix $W$ and a sparse matrix $E$, so that $W + E = M$.
- Convex approximation:

$$\min_{W,E} \; \|W\|_* + \mu\|E\|_1, \text{ s.t. } W + E = M$$

- Robust PCA

# Extension of sparsity

- Sparse inverse covariance estimation for a given empirical covariance matrix $S \in S^n$

$$\max_{X \succ 0} \ \log\det X - \mathrm{Tr}(SX) - \lambda \|X\|_1$$

- Sparse principal component analysis (PCA)
  - Variations:

$$\max \ x^\top \Sigma x \ \text{ s.t. } \mathrm{Card}(x) \le k, \ \|x\| = 1$$

$$\max \ x^\top \Sigma x - \rho \, \mathrm{Card}(x), \ \text{ s.t. } \|x\| = 1$$

  - SDP relaxations:

$$\max \ \mathrm{Tr}(\Sigma X) - \rho \|X\|_1, \ \text{ s.t. } \mathrm{Tr}(X) = 1, \ X \succeq 0$$

  - Other formulations:

$$\max \ \mathrm{Tr}(V^\top \Sigma V) - \rho \|V\|_1, \ \text{s.t. } V^\top \Sigma V \text{ is diagonal}, \ V^\top V = I$$

# Portfolio optimization

- $r_i$, random variable, the rate of return for stock $i$
- $x_i$, the relative amount invested in stock $i$
- Return: $r = r_1 x_1 + r_2 x_2 + \ldots + r_n x_n$
- expected return: $R = E(r) = \sum E(r_i) x_i = \sum \mu_i x_i$
- Risk: $V = Var(r) = \sum_{i,j} \sigma_{ij} x_i x_j = x^\top \Sigma x$

$$
\begin{aligned}
\min \ & \frac{1}{2} x^\top \Sigma x, \\
\text{s.t. } & \sum \mu_i x_i \geq r_0 \\
& \sum x_i = 1, \\
& x_i \geq 0
\end{aligned}
$$

# Correlation Matrices

A correlation matrix satisfies

$$X = X^\top,\ X_{ii} = 1,\ i = 1, \ldots, n,\ X \succeq 0.$$

Example: (low-rank) nearest correlation matrix estimation

$$\begin{aligned} &\min \frac{1}{2} \|X - C\|_F^2, \\ &\text{s.t. } X = X^\top,\ X_{ii} = 1,\ i = 1, \ldots, n,\ X \succeq 0 \end{aligned}$$

- objective fun.: $\|W \odot (X - C)\|_F^2$
- lower and upper bounds
- rank constraints $\mathrm{rank}(X) \le r$

# Optimization Formulation

Mathmatical optimization problem

$$
\begin{aligned}
&\min f(x) \\
\text{s.t. } & c_i(x) = 0, i \in \mathcal{E} \\
& c_i(x) \geq 0, i \in \mathcal{I}
\end{aligned}
$$

- $x = (x_1, \dots, x_n)^\top$: variable
- $f(x) : \mathbb{R}^n \to \mathbb{R}$: objective function
- $c_i(x) : \mathbb{R}^n \to \mathbb{R}$: constraints
- **optimal solution** $x^*$: a feasible point with the smallest value of $f$

# Classification

- Continuous versus discrete optimization
- Unconstrained versus constrained optimization
- Global and local optimization
- Stochastic and deterministic optimization
- Linear/nonlinear/quadratic programming, Convex/nonconvex optimization
- Least square problem, equation solving
- sparse optimization, PDE-constrained optimization, robust optimization

# Nonlinear optimization

- **local optimization methods** (nonlinear programming)
  - find a point that minimizes $f$ among feasible points near it
  - fast, can handle large problems
  - require initial guess
  - provide no information about distance to (global) optimum
- **global optimization methods**
  - find the (global) solution
  - worst-case complexity grows exponentially with problem size

# Brief History of Convex Optimization

- Theory (convex analysis):
- Algorithms
  - 1947: simplex algorithm for linear programming (Dantzig)
  - 1960s: early interior-point methods (Fiacco & McCormick, Dikin)
  - 1970s: ellipsoid method and other subgradient methods
  - 1980s: polynomial-time interior-point methods for linear programming (Karmarkar 1984)
  - late 1980s2000s: polynomial-time interior-point methods for nonlinear convex optimization (Nesterov & Nemirovski 1994)
  - 2010s: first-order methods
- Application
  - before 1990: mostly in operations research; few in engineering
  - since 1990: many new applications in engineering (control, signal processing, communications, circuit design, . . . ); new problem classes (semidefinite and second-order cone programming, robust optimization)

# Course Goals

- recognize/formulate problems as convex optimization problems
- understand the basic knowlege of convex optimization
- familar with the basic algorithms and develop code for problems of moderate size