STAT 37797: Mathematics of Data Science

Introduction

Cong Ma

University of Chicago, Winter 2024

An article in Harvard Data Science Review

4. Balance of Statistical and Computational Efficiencies

When we have limited data, the emphasis on statistical efficiency to make the best use of the available data has naturally become an important focus of statistics research. We do not think statistical efficiency will become irrelevant in the big data era; often inference is made locally and the relevant data that are available to infer around a specific subpopulation remain limited. On the other hand, useful statistical modeling and data analysis must take into account constraints on data storage, communication across sites, and the quality of numerical approximations in the computation. An 'optimally efficient' statistical approach is far from optimal in practice if it relies on optimization of a highly nonconvex and nonsmooth objective function, for instance. The need to work with streaming data for real-time actions also calls for a balanced approach. This is where statisticians and computer scientists, as well as experts from related domains (e.g., operation research, mathematics, and subject-matter science) can work together to address efficiency in a holistic way.

Challenges and Opportunities in Statistics and Data Science: Ten Research Areas

by Xuming He and Xihong Lin

Key messages

- Statistical efficiency is still relevant in big data era

Key messages

- Statistical efficiency is still relevant in big data era
  — *big data vs big parameters (high-dimensional statistics)*

Key messages

- Statistical efficiency is still relevant in big data era
  — *big data vs big parameters (high-dimensional statistics)*
- Computational efficiency cannot be ignored

Key messages

- Statistical efficiency is still relevant in big data era
  — *big data vs big parameters (high-dimensional statistics)*
- Computational efficiency cannot be ignored
  — *due to limited computation/memory*

Key messages

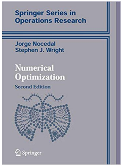
- Statistical efficiency is still relevant in big data era
  — *big data vs big parameters (high-dimensional statistics)*
- Computational efficiency cannot be ignored
  — *due to limited computation/memory*
- "A ... procedure is far from optimal in practice if it relies on optimization of a highly nonconvex and nonsmooth objective function"

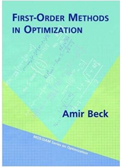
Key messages

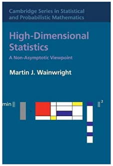
- Statistical efficiency is still relevant in big data era
  — *big data vs big parameters (high-dimensional statistics)*
- Computational efficiency cannot be ignored
  — *due to limited computation/memory*
- "A ... procedure is far from optimal in practice if it relies on optimization of a highly nonconvex and nonsmooth objective function"
  — *hmm...nonconvexity maybe our friend*

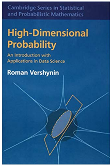
Main theme of this course

By blending statistical and computational theory, we can extract useful information from big data more efficiently









(nonconvex)optimization

(high-dimensional) statistics

Outline

- A motivating example: low-rank matrix completion
- Topics covered in this course
- Course logistics

A motivating example:
low-rank matrix completion

Noisy low-rank matrix completion

unknown rank-r matrix $\boldsymbol{\Theta}^{\star} \in \mathbb{R}^{d \times d}$

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

sampling set Ω

Noisy low-rank matrix completion

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

unknown rank-r matrix $\boldsymbol{\Theta}^\star \in \mathbb{R}^{d\times d}$

sampling set Ω

observations: $Y_{i,j} = \Theta^\star_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: estimate $\boldsymbol{\Theta}^\star$

Motivation 1: recommendation systems

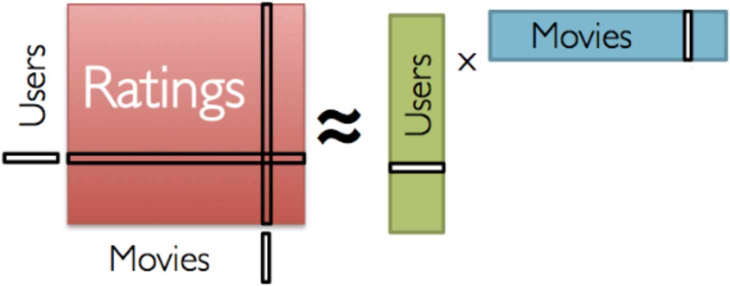
- Netflix challenge: Netflix provides highly incomplete ratings from nearly 0.5 million users & 20k movies
- How to predict unseen user ratings for movies?

In general, we cannot infer missing ratings

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

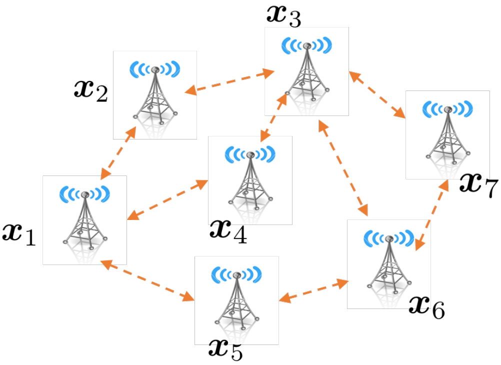
Underdetermined system (more unknowns than equations)

... unless rating matrix has some structure



low-rank approximation $\longrightarrow$ a few factors explain most of data

Motivation 2: sensor localization



- Observe partial pairwise distances
- Goal: infer distance between every pair of nodes

Motivation 2: sensor localization

Introduce location matrix

$$\boldsymbol{X} = \begin{bmatrix} - & \boldsymbol{x}_1^\top & - \\ - & \boldsymbol{x}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{x}_d^\top & - \end{bmatrix} \in \mathbb{R}^{d\times 3}$$

then distance matrix $\boldsymbol{D} = [D_{i,j}]_{1\le i,j\le d}$ can be written as

$$\boldsymbol{D} = \underbrace{\underbrace{\begin{bmatrix} \|\boldsymbol{x}_1\|_2^2 \\ \vdots \\ \|\boldsymbol{x}_d\|_2^2 \end{bmatrix} \boldsymbol{1}^\top}_{\text{rank } 1} + \underbrace{\boldsymbol{1} \cdot \left[\|\boldsymbol{x}_1\|_2^2, \cdots, \|\boldsymbol{x}_d\|_2^2\right]}_{\text{rank } 1} - \underbrace{2\boldsymbol{X}\boldsymbol{X}^\top}_{\text{rank } 3}}_{\text{low rank}}$$

$\text{rank}(\boldsymbol{D}) \ll d \quad \longrightarrow \quad$ low-rank matrix completion

Least-squares estimator

$$\begin{aligned} \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}}{\text{minimize}} \quad & f(\boldsymbol{\Theta}) = \sum_{(i,j) \in \Omega} (\Theta_{i,j} - Y_{i,j})^2 \\ \text{subject to} \quad & \text{rank}(\boldsymbol{\Theta}) = r \end{aligned}$$

Least-squares estimator

$$\begin{array}{ll} \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}}{\text{minimize}} & f(\boldsymbol{\Theta}) = \sum_{(i,j) \in \Omega} (\Theta_{i,j} - Y_{i,j})^2 \\ \text{subject to} & \text{rank}(\boldsymbol{\Theta}) = r \end{array}$$

— *This is also MLE when noise follows Gaussian*

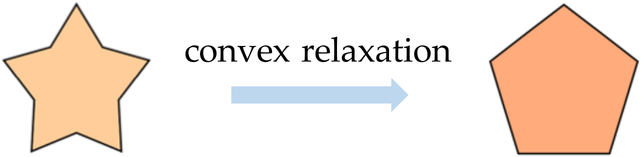
Least-squares estimator

$$\begin{aligned} \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}}{\text{minimize}} \quad & f(\boldsymbol{\Theta}) = \sum_{(i,j) \in \Omega} \left(\Theta_{i,j} - Y_{i,j}\right)^2 \\ \text{subject to} \quad & \text{rank}(\boldsymbol{\Theta}) = r \end{aligned}$$

— *This is also MLE when noise follows Gaussian*

Challenge: nonconvexity $\Longrightarrow$ computational hardness

Popular workaround: convex relaxation



Relax nonconvex problems into convex ones by finding convex surrogates

Convex relaxation for matrix completion

Replace rank constraint by nuclear norm constraint

$$\begin{aligned} \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad & f(\boldsymbol{\Theta}) = \sum_{(i,j)\in\Omega} (\Theta_{i,j} - Y_{i,j})^2 \\ \text{subject to} \quad & \cancel{\text{rank}(\boldsymbol{\Theta}) = r} \quad \|\boldsymbol{\Theta}\|_* \leq t \end{aligned}$$

— $\|\boldsymbol{\Theta}\|_* = \sum_{i=1}^{d} \sigma_i(\boldsymbol{\Theta})$

Convex relaxation for matrix completion

Replace rank constraint by nuclear norm constraint

$$\begin{array}{ll} \underset{\mathbf{\Theta}\in\mathbb{R}^{d\times d}}{\text{minimize}} & f(\mathbf{\Theta}) = \sum_{(i,j)\in\Omega} \left(\Theta_{i,j} - Y_{i,j}\right)^2 \\ \text{subject to} & \cancel{\text{rank}(\mathbf{\Theta}) = r} \quad \|\mathbf{\Theta}\|_* \le t \end{array}$$

— $\|\mathbf{\Theta}\|_* = \sum_{i=1}^{d} \sigma_i(\mathbf{\Theta})$

convex relaxation (regularized version):

$$\underset{\mathbf{\Theta}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(\Theta_{i,j} - Y_{i,j}\right)^2 + \lambda\|\mathbf{\Theta}\|_*$$

Convex relaxation: pros and cons

convex relaxation (regularized version):

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(\Theta_{i,j} - Y_{i,j}\big)^2 + \lambda\|\boldsymbol{\Theta}\|_*$$

Pro: often achieve statistical optimality

Convex relaxation: pros and cons

convex relaxation (regularized version):

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(\Theta_{i,j} - Y_{i,j}\right)^2 + \lambda\|\boldsymbol{\Theta}\|_*$$

Pro: often achieve statistical optimality

Issue: expensive in computation/memory

Can we solve matrix completion with lower computational cost?

Spectral methods

- Assumption: each entry is observed indep. with probability p

Spectral methods

- Assumption: each entry is observed indep. with probability p
- Key observation: let

$$\hat{Y}_{i,j} = \begin{cases} \frac{1}{p} Y_{i,j}, & \text{if } (i,j) \text{ is observed,} \\ 0, & \text{otherwise} \end{cases}$$

  we have $\mathbb{E}[\hat{\boldsymbol{Y}}] = \boldsymbol{\Theta}^\star$

Spectral methods

- Assumption: each entry is observed indep. with probability p
- Key observation: let

$$\hat{Y}_{i,j} = \begin{cases} \frac{1}{p} Y_{i,j}, & \text{if } (i,j) \text{ is observed,} \\ 0, & \text{otherwise} \end{cases}$$

we have $\mathbb{E}[\hat{\boldsymbol{Y}}] = \boldsymbol{\Theta}^\star$

spectral method:

use best rank-r approximation to $\hat{\boldsymbol{Y}}$ as estimator of $\boldsymbol{\Theta}^\star$

Spectral methods

- Assumption: each entry is observed indep. with probability p
- Key observation: let

$$\hat{Y}_{i,j} = \begin{cases} \frac{1}{p} Y_{i,j}, & \text{if } (i,j) \text{ is observed,} \\ 0, & \text{otherwise} \end{cases}$$

we have $\mathbb{E}[\hat{\boldsymbol{Y}}] = \boldsymbol{\Theta}^\star$

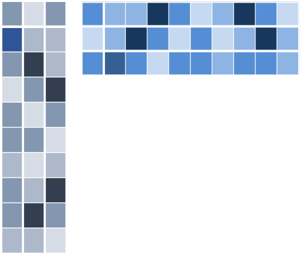
spectral method:

use best rank-r approximation to $\hat{\boldsymbol{Y}}$ as estimator of $\boldsymbol{\Theta}^\star$

— simple, but sometimes statistically inefficient

Nonconvex optimization

Represent low-rank matrix by $\boldsymbol{L}\boldsymbol{R}^\top$ with $\underbrace{\boldsymbol{L}, \boldsymbol{R} \in \mathbb{R}^{d\times r}}_{\text{low-rank factors}}$



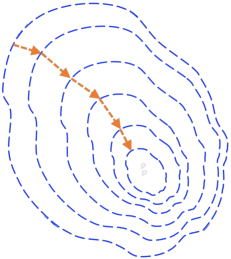
$$\underset{\boldsymbol{L},\boldsymbol{R}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{L},\boldsymbol{R}) = \sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{L}\boldsymbol{R}^\top\right)_{i,j} - Y_{i,j} \right]^2$$

Two-stage algorithm

$$\underset{\boldsymbol{L},\boldsymbol{R}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{L},\boldsymbol{R}) = \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{L}\boldsymbol{R}^\top\big)_{i,j} - Y_{i,j} \Big]^2$$

Two-stage algorithm

$$\underset{\boldsymbol{L},\boldsymbol{R}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{L},\boldsymbol{R}) = \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{L}\boldsymbol{R}^\top\big)_{i,j} - Y_{i,j}\Big]^2$$

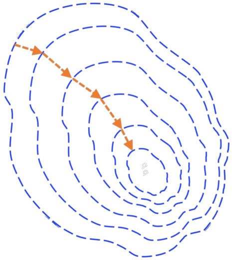


- **spectral initialization:** $(\boldsymbol{L}^0, \boldsymbol{R}^0)$
  — top singular vectors of $\hat{\boldsymbol{Y}}$
- **gradient descent:** for $t = 0, 1, \ldots$

$$\boldsymbol{L}^{t+1} = \boldsymbol{L}^t - \eta_t \nabla_{\boldsymbol{L}} f(\boldsymbol{L}^t, \boldsymbol{R}^t)$$
$$\boldsymbol{R}^{t+1} = \boldsymbol{R}^t - \eta_t \nabla_{\boldsymbol{R}} f(\boldsymbol{L}^t, \boldsymbol{R}^t)$$

Two-stage algorithm

$$\underset{\boldsymbol{L},\boldsymbol{R}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{L},\boldsymbol{R}) = \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{L}\boldsymbol{R}^{\top}\big)_{i,j} - Y_{i,j} \Big]^2$$



- **spectral initialization:** $(\boldsymbol{L}^0, \boldsymbol{R}^0)$
  — top singular vectors of $\hat{\boldsymbol{Y}}$
- **gradient descent:** for $t = 0, 1, \ldots$

$$\boldsymbol{L}^{t+1} = \boldsymbol{L}^{t} - \eta_t \nabla_{\boldsymbol{L}} f(\boldsymbol{L}^t, \boldsymbol{R}^t)$$
$$\boldsymbol{R}^{t+1} = \boldsymbol{R}^{t} - \eta_t \nabla_{\boldsymbol{R}} f(\boldsymbol{L}^t, \boldsymbol{R}^t)$$

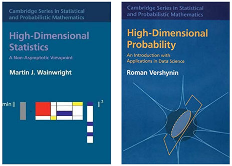
nonconvex estimator achieves optimal estimation error

— *Ma, Wang, Chi, Chen '17*

Main theme of this course

By blending statistical and computational theory, we can extract useful information from big data more efficiently

(nonconvex)optimization



(high-dimensional) statistics

Tentative topics

- Spectral methods
  - Classic ℓ_2 matrix perturbation theory
  - Matrix concentration inequalities
  - Applications of spectral methods (ℓ_2 theory)
  - ℓ_∞ matrix perturbation theory
  - Applications of spectral methods (ℓ_∞ theory)
- Nonconvex optimization
  - Basic optimization theory
  - Generic local analysis for regularized gradient descent (GD)
  - Refined local analysis for vanilla GD
  - Global landscape analysis
  - Gradient descent with random initialization
- Convex relaxation (maybe)
  - Compressed sensing and sparse recovery
  - Phase transition and convex geometry
  - Low-rank matrix recovery
  - Robust principal component analysis
- Minimax lower bounds (maybe)

Logistics

Why you should not take this course

Why you **should not** take this course

- There will be quite a few THEOREMS and PROOFS ...

Why you **should not** take this course

- There will be quite a few THEOREMS and PROOFS ...
- Nonrigorous/heuristic arguments from time to time

Why you should consider taking this course

Why you should consider taking this course

- There will be quite a few THEOREMS and PROOFS ...

Why you **should** consider taking this course

- There will be quite a few THEOREMS and PROOFS ...
  - promote deeper understanding of scientific results

Why you **should** consider taking this course

- There will be quite a few THEOREMS and PROOFS ...
    - promote deeper understanding of scientific results
- Nonrigorous/heuristic arguments from time to time

Why you should consider taking this course

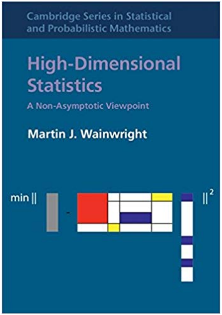
- There will be quite a few THEOREMS and PROOFS ...
  - promote deeper understanding of scientific results
- Nonrigorous/heuristic arguments from time to time
  - "nonrigorous" but grounded in rigorous theory
  - help develop intuition

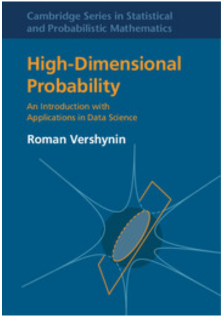
Prerequisites

- linear algebra
- probability theory
- a programming language (e.g., Matlab, Python, Julia, ...)
- *knowledge in convex optimization*

Textbooks

We recommend these books, but will not follow them closely





Useful references

- *Spectral Methods for Data Science: A Statistical Perspective*, Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma
- *Nonconvex optimization meets low-rank matrix factorization: An overview*, Yuejie Chi, Yue M. Lu, and Yuxin Chen
- *Convex optimization*, Stephen Boyd, and Lieven Vandenberghe

Course project

Two forms

- literature review
- original research
  - *You are strongly encouraged to combine it with your own research*

Course project

Two forms

- literature review
- original research
  - *You are strongly encouraged to combine it with your own research*

Three milestones

- proposal (due Feb. 1st): up to 1 page
- in-class presentation: last week (week 9) of class
- report (due Mar. 6th): up to 4 pages with unlimited appendix