

# Northwestern University

## **Bayesian Statistics & Machine Learning Working Group**

.....

### **Matrix Optimization Problems in Statistics and Machine Learning**

October 27, 2021

**Tim Tsz-Kit Lau**

**4<sup>th</sup> Year Ph.D. Candidate**

**Department of Statistics**

**Northwestern University**

`timlautk@u.northwestern.edu`; `https://timlautk.github.io`

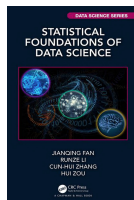
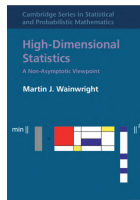
# Today's Roadmap

1. Preliminaries
2. Nonsmooth problems in machine learning
3. Matrix regression
4. Low-rank matrix completion
5. Low-rank noisy matrix completion—Convex versus nonconvex formulations
  - 5.1 Convex relaxation formulation
  - 5.2 Nonconvex Burer–Monteiro factorization formulation
  - 5.3 Convex versus nonconvex formulations? A brief discussion
6. Some low-rank matrix optimization algorithms
7. Covariance estimation
8. Graphical Lasso

# Materials

Content mainly taken from:

- High-Dimensional Statistics: A Non-Asymptotic Viewpoint (Wainwright, 2019)
- Statistical Foundations of Data Science (Fan et al., 2020)
- Nonsmoothness in machine learning..., *Set-Valued Var Anal.* (Iutzeler and Malick, 2020)
- Noisy matrix completion..., *SIAM J. Optim.* (Chen et al., 2020b)



## Additional Materials

- If you are interested in going deeper into matrix methods for data science, you might refer to [Spectral Methods for Data Science: A Statistical Perspective](#) (Chen et al., 2020a)
- For mathematical background in numerical linear algebra, matrix analysis and computations, see e.g.,
  - Matrix Analysis (Horn and Johnson, 2012)
  - Matrix Computations (Golub and Van Loan, 2013)
- See also the recent papers:
  - Implicit Regularization in Nonconvex Statistical Estimation..., *FoCM* (Ma et al., 2019)
  - An equivalence between critical points for rank constraints versus low-rank factorizations, *SIAM J. Optim.* (Ha et al., 2020)
  - On critical points of quadratic low-rank matrix optimization problems, *IMA J. Numer. Anal.* (Uschmajew and Vandereycken, 2020)
  - Low-rank matrix recovery with composite optimization..., *FoCM* (Charisopoulos et al., 2021)

# Preliminaries

## Linear Algebra and Matrix Analysis

- For a rectangular matrix  $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$  with  $d_1 \geq d_2$ , its ordered *singular values* are

$$\sigma_{\max}(\mathbf{X}) = \sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_{d_2}(\mathbf{X}) = \sigma_{\min}(\mathbf{X}) \geq 0$$

- For a rectangular matrix  $\mathbf{Y} \in \mathbb{R}^{d \times d}$ , its ordered *eigenvalues* are

$$\gamma_{\max}(\mathbf{Y}) = \gamma_1(\mathbf{Y}) \geq \gamma_2(\mathbf{Y}) \geq \cdots \geq \gamma_d(\mathbf{Y}) = \gamma_{\min}(\mathbf{Y})$$

- If  $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$  with  $d_1 \geq d_2$  and  $\mathbf{R} := \mathbf{X}^\top \mathbf{X}$ , then  $\gamma_j(\mathbf{R}) = (\sigma_j(\mathbf{X}))^2$  for  $j = 1, \dots, d_2$

- Write  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{d_2})^\top$  and  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)^\top$

- $\mathcal{S}_d$ : the set of symmetric matrices in  $\mathbb{R}^{d \times d}$

- $\mathcal{S}_d^+$ : the set of symmetric positive semidefinite (i.e.,  $\gamma_{\min} \geq 0$ ) matrices in  $\mathbb{R}^{d \times d}$

- $\mathcal{S}_d^{++}$  the set of symmetric positive definite (i.e.,  $\gamma_{\min} > 0$ ) matrices in  $\mathbb{R}^{d \times d}$

## Linear Algebra and Matrix Analysis

- **Singular value decomposition (SVD)** for  $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$  with  $d_1 \geq d_2$

$$\mathbf{X} = \sum_{i=1}^{d_2} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \mathbf{U} \mathbf{S} \mathbf{V}^\top,$$

where  $\mathbf{U} = (\mathbf{u}_1 \cdots \mathbf{u}_{d_2}) \in \mathbb{R}^{d_1 \times d_2}$  and  $\mathbf{V} = (\mathbf{v}_1 \cdots \mathbf{v}_{d_2}) \in \mathbb{R}^{d_2 \times d_2}$  are orthogonal matrices (i.e.,  $\mathbf{U} \mathbf{U}^\top = \mathbf{I}_{d_1}$ ,  $\mathbf{V} \mathbf{V}^\top = \mathbf{I}_{d_2}$ );  $\mathbf{S} = \text{Diag}(\sigma_1, \dots, \sigma_{d_2}) \in \mathbb{R}^{d_2 \times d_2}$

- **Eigendecomposition** or **spectral decomposition** for *diagonalizable matrices*, e.g.,  $\mathbf{Y} \in \mathcal{S}_d$  is diagonalizable via a *unitary transformation*

$$\mathbf{Y} = \sum_{i=1}^d \gamma_i \mathbf{q}_i \mathbf{q}_i^\top = \mathbf{Q} \mathbf{G} \mathbf{Q}^{-1},$$

where  $\mathbf{Q} = (\mathbf{q}_1 \cdots \mathbf{q}_d) \in \mathbb{R}^{d \times d}$ , with  $\mathbf{Q}^{-1} = \mathbf{Q}^\top$ ;  $\mathbf{G} = \text{Diag}(\gamma_1, \dots, \gamma_d) \in \mathbb{R}^{d \times d}$

# Linear Algebra and Matrix Analysis

**Norms** of  $\mathbf{X} = (x_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n} \in \mathbb{R}^{m \times n}$  with  $m \geq n$

- Elementwise  $\ell_1$ -norm

$$\|\mathbf{X}\|_1 := \sum_{i=1}^m \sum_{j=1}^n |x_{i,j}|$$

- Elementwise  $\ell_\infty$ -norm

$$\|\mathbf{X}\|_\infty := \max_{1 \leq i \leq m, 1 \leq j \leq n} |x_{i,j}|$$

- Nuclear norm

$$\|\mathbf{X}\|_{\text{nuc}} := \sum_{i=1}^n \sigma_i(\mathbf{X}) = \|\boldsymbol{\sigma}(\mathbf{X})\|_1 = \text{tr}(\sqrt{\mathbf{X}^\top \mathbf{X}})$$



# Linear Algebra and Matrix Analysis

- **Norms** of  $\mathbf{X} = (x_{i,j})_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \in \mathbb{R}^{d_1 \times d_2}$  with  $d_1 \geq d_2$ 
  - $\ell_2$ -operator/spectral norm

$$\|\mathbf{X}\|_2 \equiv \|\mathbf{X}\|_S := \sigma_{\max}(\mathbf{X})$$

- Frobenius norm

$$\|\mathbf{X}\|_F := \sqrt{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} x_{i,j}^2}$$

- **Frobenius/trace inner product** of  $\mathbf{X} = (x_{i,j})_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \in \mathbb{R}^{d_1 \times d_2}$  and  $\mathbf{Y} = (y_{i,j})_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \in \mathbb{R}^{d_1 \times d_2}$

$$\langle\langle \mathbf{X}, \mathbf{Y} \rangle\rangle_F := \text{tr}(\mathbf{X}^\top \mathbf{Y}) = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} x_{i,j} y_{i,j} \quad \text{with} \quad \langle\langle \mathbf{X}, \mathbf{X} \rangle\rangle_F = \|\mathbf{X}\|_F^2$$

# Linear Algebra and Matrix Analysis

- **Rank** of  $\mathbf{X} = (x_{i,j})_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \in \mathbb{R}^{d_1 \times d_2}$  with  $d_1 \geq d_2$ , denoted by  $\text{rank}(\mathbf{X})$
- Let  $r \leq \min\{d_1, d_2\}$ . The smooth manifold of fixed rank- $r$  matrices is

$$\mathcal{M}_r := \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\mathbf{X}) = r\}$$

- The closure of  $\mathcal{M}_r$  in  $\mathbb{R}^{d_1 \times d_2}$  is the **cone**

$$\mathcal{M}_{\leq r} := \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\mathbf{X}) \leq r\}$$

# Nonsmooth Problems in Machine Learning

# Nonsmooth Problems in Machine Learning

- Recall the **regularized ERM framework** for *matrix*-valued variables of interest

$$\underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \{L(\Theta) + \lambda h(\Theta)\}$$

- Purposes of *nonsmooth* **regularization**:
  - As  $d$  grows (or,  $d_1$  and  $d_2$ ), the problem becomes **ill-conditioned** and reduces the *interpretability* and *stability* of the model
  - Remedy: introduce a *prior* on the *structure* of the model  $\theta$  or  $\Theta$
  - The use of **regularization** to promote the prior structure
  - Use an additive *nonsmooth* function  $h$  enforcing this structure:  
**Nonsmoothness** of functions *traps* optimal solutions in *low-dimensional manifold*

## Nonsmooth Problems in Machine Learning

- Define the *proximity operator* of a function  $h: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  with a parameter  $\gamma > 0$  by

$$\text{prox}_{\gamma h}(\varphi) := \underset{\phi \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ h(\phi) + \frac{1}{2\gamma} \|\phi - \varphi\|_2^2 \right\}$$

- The *proximity operator* of a function  $h: \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}$  with a parameter  $\gamma > 0$  is similarly defined

$$\text{prox}_{\gamma h}(\Phi) := \underset{\Psi \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ h(\Psi) + \frac{1}{2\gamma} \|\Psi - \Phi\|_F^2 \right\}$$

- A key tool to deal with *explicit nonsmoothness* in optimization
- For various nonsmooth functions, their proximity operators admit **closed forms**

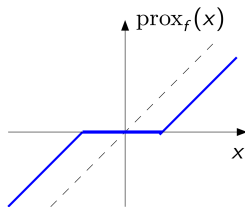
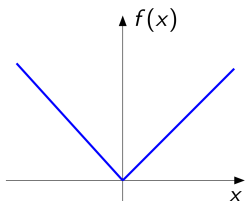
# Nonsmooth Problems in Machine Learning

## Sparse structure with $\ell_1$ -norm $\|\cdot\|_1$

- To look at the *sparsity* pattern of a vector, we look at whether each coordinate is null or not
- The proximity operator of the  $\ell_1$ -norm is the *soft-thresholding* operator

$$\theta_i = \text{prox}_{\gamma|\cdot|}(\varphi_i) = \text{sign}(\varphi_i) \cdot \max\{|\varphi_i| - \gamma, 0\},$$

where  $\theta = (\theta_i)_{1 \leq i \leq d}$  and  $\varphi = (\varphi_i)_{1 \leq i \leq d}$



# Nonsmooth Problems in Machine Learning

## Low-rank with nuclear norm $\|\cdot\|_{\text{nuc}}$

- For matrix-valued problems, the notion of being low-rank is closely related to the sparsity of the vector of singular values
- Recall that the nuclear norm is the  $\ell_1$ -norm of the vector of singular values
- The proximity operator of the nuclear norm is, by the SVD  $\mathbf{U} \text{Diag}(\boldsymbol{\sigma}) \mathbf{V}^\top$ ,

$$\boldsymbol{\Theta} = \mathbf{U} \text{Diag}(\boldsymbol{\varphi}) \mathbf{V}^\top \text{ with } \varphi_i = \text{prox}_{\gamma|\cdot|}(\sigma_i),$$

where  $\boldsymbol{\sigma} = (\sigma_i)_{1 \leq i \leq d}$  and  $\boldsymbol{\varphi} = (\varphi_i)_{1 \leq i \leq d}$

# Matrix Regression



## Matrix Regression

- **Linear regression:** observations  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , with a vector of covariates  $\mathbf{x}_i \in \mathbb{R}^d$ , a response variable  $y_i \in \mathbb{R}$  and some type of noise variable  $\varepsilon_i$

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle + \varepsilon_i$$

- **Linear matrix regression:** observations  $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$ , with a matrix of covariates  $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ , a response variable  $y_i \in \mathbb{R}$  and some type of noise variable  $\varepsilon_i$

$$y_i = \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle_{\text{F}} + \varepsilon_i$$

- Define the *observation operator*  $\mathcal{X}_n: \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$  with elements  $[\mathcal{X}_n(\boldsymbol{\Theta})]_i = \langle \mathbf{X}_i, \boldsymbol{\Theta} \rangle_{\text{F}}$ , then we obtain the vector equation

$$\mathbf{y} = \mathcal{X}_n(\boldsymbol{\Theta}^*) + \boldsymbol{\varepsilon},$$

where  $\mathbf{y} \in \mathbb{R}^n$  and  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$

## Matrix Regression

- Many applications in which the regression matrix  $\Theta^*$  is either low-rank, or well approximated by a low-rank matrix
- An appropriate estimator would be a rank-penalized form of least squares

$$\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathcal{X}_n(\Theta)\|_2^2 + \lambda \cdot \operatorname{rank}(\Theta) \right\},$$

which is a **nonconvex** form of least squares (i.e., *computationally hard to solve*)

- Replacing the **nonconvex** rank penalty by the **convex** nuclear norm, we instead solve the **convex** program

$$\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathcal{X}_n(\Theta)\|_2^2 + \lambda \|\Theta\|_{\text{nuc}} \right\}$$

# Low-Rank Matrix Completion

## Motivating Example—The Netflix Problem





				...	...	
	4	*	3	...	...	*
	3	5	*	...	...	2
	5	4	3	...	...	3
	2	*	*	...	...	1

Figure: The Netflix problem ([Wainwright, 2019](#), Figure 10.1(a))

## Low-Rank Matrix Completion

- Estimating an unknown matrix  $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$  based on (noisy) observation of a subset of its entries
- Ill-posed problem unless further structure is imposed
- E.g., low-rank or can be well approximated by a low-rank matrix
- Consider the noisy observation model

$$\tilde{y}_i = \Theta_{a(i), b(i)} + \frac{\varepsilon_i}{\sqrt{d_1 d_2}},$$

where  $\varepsilon_i$  is some form of observation noise, and  $(a(i), b(i))$  are the row and column indices of the  $i$ th observation

- **Noiseless** case if  $\varepsilon_i = 0$  for all  $i$

## Low-Rank Matrix Completion as Matrix Regression

- Define the *mask matrix*  $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$

$$\mathbf{X}_i(j, k) = \begin{cases} \sqrt{d_1 d_2} & \text{if } (j, k) = (a(i), b(i)) \\ 0 & \text{otherwise} \end{cases}$$

- Let  $y_i := \sqrt{d_1 d_2} \tilde{y}_i$ , then

$$y_i = \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle_F + \varepsilon_i$$

- Then we can treat low-rank matrix completion as a low-rank matrix regression problem
- Matrices might take discrete values such as yes/no votes coded in  $y_i \in \{\pm 1\}$ , then we can apply the **matrix logistic regression** model

$$\mathbb{P}(y_i | \mathbf{X}_i, \boldsymbol{\Theta}^*) = \frac{\exp(y_i \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle_F)}{1 + \exp(y_i \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle_F)}$$

# Low-Rank Noisy Matrix Completion

## Convex versus Nonconvex Formulations

## Low-Rank Noisy Matrix Completion

- Consider the task of estimating a rank- $r$  data matrix

$$\mathbf{M}^* = (M_{i,j}^*)_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \in \mathbb{R}^{d_1 \times d_2}$$

- Performed on a subset of noisy entries (*alternative formulation*)

$$M_{i,j} = M_{i,j}^* + \varepsilon_{i,j}, \quad \text{where } (i,j) \in \Omega$$

- $\Omega \subseteq \{1, \dots, d_1\} \times \{1, \dots, d_2\}$  denotes a set of indices
- $\varepsilon_{i,j}$ : additive noise at  $(i,j)$
- Solving noisy matrix completion via convex relaxation (e.g., nuclear norm regularization) practically exhibits excellent stability
- Far less understood **theoretically** compared to the noiseless setting



# Convex Relaxation Formulation

## Convex Relaxation

- Regularized least-squares formulation (regularization parameter  $\lambda > 0$ ):

$$\underset{\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \left\{ \frac{1}{2} \sum_{(i,j) \in \Omega} (Z_{i,j} - M_{i,j})^2 + \lambda \cdot \text{rank}(\mathbf{Z}) \right\}$$

- Computational intractability of rank minimization  $\Rightarrow$  convex relaxation, e.g., nuclear norm regularization:

$$\underset{\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} f(\mathbf{Z}) := \frac{1}{2} \sum_{(i,j) \in \Omega} (Z_{i,j} - M_{i,j})^2 + \lambda \|\mathbf{Z}\|_{\text{nuc}} \quad (1)$$

- The nuclear norm is a **convex surrogate** for the rank function

## Convex Relaxation Formulation

- Denote the solution to the convex relaxation formulation (1) by  $\mathbf{Z}_{\text{cvx}}$
- In the noiseless setting (i.e.,  $\varepsilon_{i,j} = 0$  for all  $(i,j) \in \Omega$ ), the solution to (1) is known to be **faithful** (i.e., zero estimation error;  $\|\mathbf{Z}_{\text{cvx}} - \mathbf{M}^*\|_F = 0$ ) even under near-minimal sample complexity
- In the noisy setting, the performance of convex relaxation remains largely unclear
- Candès and Plan (2010) derived that the estimation error  $\|\mathbf{Z}_{\text{cvx}} - \mathbf{M}^*\|_F$  is significantly larger than the oracle lower bound
- The effectiveness of convex relaxation in practice is not explained well
- Numerical experiments in Candès and Plan (2010) indicated that the performance of convex relaxation is far better than their theoretical bounds

# The Natural yet Challenging Questions

1. Where does the convex program (1) stand in terms of its **stability** vis-à-vis additive noise?
2. Can we establish **statistical performance guarantees** that match its practical effectiveness?

# Nonconvex Burer–Monteiro Factorization Formulation

## Nonconvex Burer–Monteiro Factorization Formulation

- The *nonconvex* **Burer–Monteiro** factorization approach: Represent  $\mathbf{Z}$  as

$$\mathbf{Z} = \mathbf{X}\mathbf{Y}^\top \text{ with low-rank factors } \mathbf{X} \in \mathbb{R}^{d_1 \times r}, \mathbf{Y} \in \mathbb{R}^{d_2 \times r}$$

- Solve the nonconvex regularized least-squares problem:

$$\underset{\mathbf{X} \in \mathbb{R}^{d_1 \times r}, \mathbf{Y} \in \mathbb{R}^{d_2 \times r}}{\text{minimize}} \left\{ \frac{1}{2} \sum_{(i,j) \in \Omega} [(\mathbf{X}\mathbf{Y}^\top)_{i,j} - M_{i,j}]^2 + \text{reg}(\mathbf{X}, \mathbf{Y}) \right\}$$

- Intimate connection with the convex program (1)*: if the solution to (1) has rank  $r$ , then it must coincide with the solution to

$$\underset{\mathbf{X} \in \mathbb{R}^{d_1 \times r}, \mathbf{Y} \in \mathbb{R}^{d_2 \times r}}{\text{minimize}} \frac{1}{2} \sum_{(i,j) \in \Omega} [(\mathbf{X}\mathbf{Y}^\top)_{i,j} - M_{i,j}]^2 + \underbrace{\frac{\lambda}{2} [\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2]}_{\text{reg}(\mathbf{X}, \mathbf{Y})} \quad (2)$$

# Nonconvex Burer–Monteiro Factorization Formulation

- This can be verified by the fact that

$$\|Z\|_{\text{nuc}} = \inf_{\mathbf{X} \in \mathbb{R}^{d_1 \times r}, \mathbf{Y} \in \mathbb{R}^{d_2 \times r}: \mathbf{X}\mathbf{Y}^T = \mathbf{Z}} \left\{ \frac{1}{2} \|\mathbf{X}\|_F^2 + \frac{1}{2} \|\mathbf{Y}\|_F^2 \right\}$$

for any rank- $r$  matrix  $\mathbf{Z}$

- **Challenging** to predict when the *rank- $r$  assumption* of the solution to (1) can possibly hold
- Simple **first-order optimization methods** (gradient descent and its variants) with proper initialization are often effective in solving (2) despite its nonconvexity

## Nonconvex Burer–Monteiro Factorization Formulation

- Theoretically, algorithms tailored to the nonconvex formulation (2) often enable *exact recovery* in the *noiseless* setting
- In a wide range of noisy settings, the nonconvex approach achieves *appealing estimation accuracy*
- Could be **significantly better** than those bounds derived for convex relaxation



# Convex versus Nonconvex Formulations?

## A Brief Discussion

## Empirical Evidence

- Fix  $d_1 = d_2 = 1000 =: d, r = 5$
- Generate  $\mathbf{M}^* = \mathbf{X}^* \mathbf{Y}^{*\top}$ , where  $\mathbf{X}^*, \mathbf{Y}^* \in \mathbb{R}^{d \times r}$  are random orthonormal matrix
- Each entry  $M_{i,j}^*$  of  $\mathbf{M}^*$  is observed with probability  $p = 0.2$  independently
- Then corrupted by an independent Gaussian noise  $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$
- Solve the convex program (1) by the **proximal gradient** method (cf. **iterative soft-thresholding algorithm** for the Lasso)
- Solve the nonconvex program (2) by **gradient descent** with spectral initialization (see [Chi et al., 2019](#), for details)
- Denote the solution to the convex program (1) by  $\mathbf{Z}_{\text{cvx}}$  (resp. the nonconvex program (2) by  $\mathbf{Z}_{\text{ncvx}}$ )

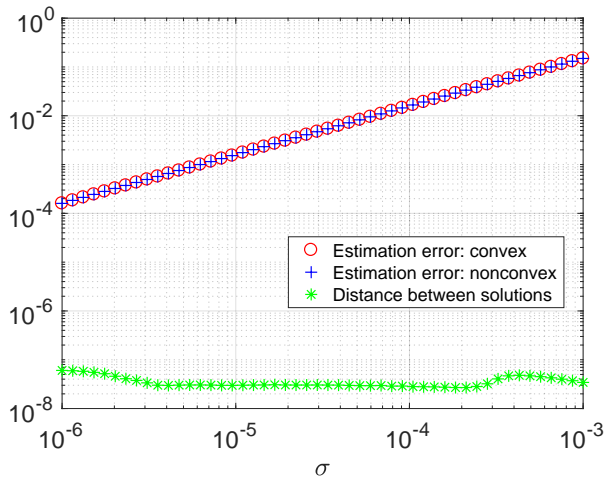


Figure: The relative estimation error  $\|\mathbf{Z}_{\text{cvx/ncvx}} - \mathbf{M}^*\|_F / \|\mathbf{M}^*\|_F$  of (1) and (2), and the relative distance  $\|\mathbf{Z}_{\text{cvx}} - \mathbf{Z}_{\text{ncvx}}\|_F / \|\mathbf{M}^*\|_F$  vs. the standard deviation  $\sigma$  of the noise, averaged over 20 independent trials (Chen et al., 2020b, Fig. 1).

## Empirical Evidence

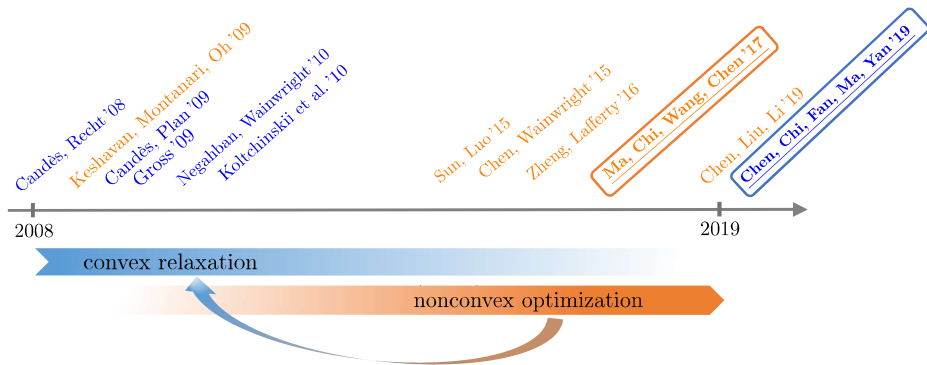
- The distance between the convex and the nonconvex solutions seems extremely small
- The relative estimation errors of both  $\mathbf{Z}_{\text{cvx}}$  and  $\mathbf{Z}_{\text{ncvx}}$  are substantially larger
- The estimate returned by the nonconvex approach serves as a **remarkably accurate approximation** of the convex solution
- Recall that the nonconvex approach is often guaranteed to achieve **intriguing statistical guarantees** vis-à-vis random noise (Ma et al., 2019)
- Suggest that **the convex program is equally stable**
- The difficulty in rigorously justifying the above numerical observations has been noted in the literature (see e.g., Keshavan et al., 2010)

## Another Question

- Can we leverage *existing theory* for the **nonconvex** scheme to improve the *statistical analysis* of the **convex relaxation** approach?
- If one can formally justify the **proximity** between the convex and the nonconvex solutions, then it is possible to *propagate* the appealing **stability guarantees** from the **nonconvex** scheme to the **convex** approach

## Convex versus Nonconvex Formulations? A Brief Discussion

- An **approximate critical point** of the *nonconvex formulation* serves as an **extremely tight approximation** of the *convex solution*
- Allowing to transfer the **desired statistical guarantees** of the nonconvex approach to its convex counterpart
- Some important directions for future exploration:
  - Approximate low-rank structure  
(Current theory based upon assuming exactly low-rank ground-truth matrix  $\mathbf{M}^*$ )
  - Extension to deterministic noise (currently i.i.d. sub-Gaussian noise)
  - Extension to structured matrix completion
  - Extension to robust PCA and blind deconvolution
- For more comprehensive results, see [Chen et al. \(2020b\)](#)



“Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization”, Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, 2019

# Some Low-Rank Matrix Optimization Algorithms



## Algorithm for Convex Formulation of Low-Rank Optimization Problems

- Recall the low-rank optimization problem (i.e., rank-constrained) with  $r < \min\{d_1, d_2\}$

$$\underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \{f(\Theta) : \text{rank}(\Theta) \leq r\} \equiv \underset{\Theta \in \mathcal{M}_{\leq r}}{\text{minimize}} f(\Theta)$$

- Projected gradient descent (PGD); a.k.a. *iterative hard thresholding*

$$\Theta_{k+1} = \text{proj}_{\mathcal{M}_{\leq r}}(\Theta_k - \gamma \nabla f(\Theta_k))$$

- $\text{proj}_{\mathcal{M}_{\leq r}}$  is performed by taking the top  $r$  components of a SVD
- The **Eckart–Young–Mirsky theorem** (a *very important* theorem in numerical linear algebra): the **best rank- $r$  approximation** of a matrix  $\Theta$  (in the  $\ell_2$ -operator norm) is

$$\Theta_r = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad \text{where} \quad \sum_{i=1}^{d_2} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \text{ is the SVD of } \Theta$$

# Proximal Algorithms for Regularized Matrix Estimation

- Recall the regularized matrix estimation problem

$$\underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \{L(\Theta) + \lambda h(\Theta)\}$$

- Proximal gradient algorithm

$$\Theta_{k+1} = \text{prox}_{\gamma \lambda h}(\Theta_k - \gamma \nabla L(\Theta_k))$$

- Need the expression of  $\text{prox}_{\gamma h}$  with  $h = \|\cdot\|_{\text{nuc}}$

# Algorithm for Nonconvex Factorization Formulation of Low-Rank Optimization Problems

- Recall the nonconvex Burer–Monteiro factorization formulation of the low-rank optimization problem

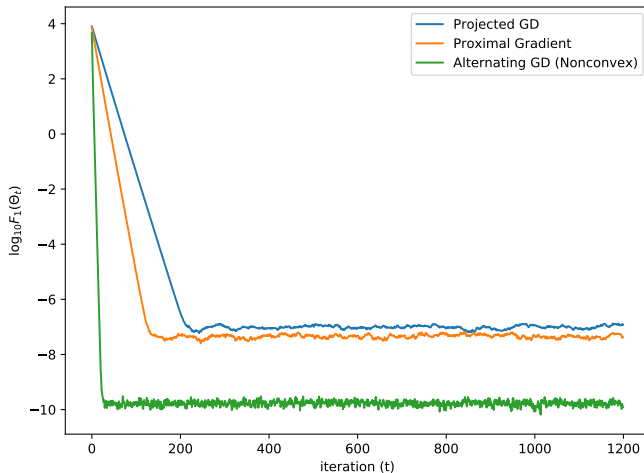
$$\underset{\Theta_1 \in \mathbb{R}^{d_1 \times r}, \Theta_2 \in \mathbb{R}^{d_2 \times r}}{\text{minimize}} \quad f(\Theta_1 \Theta_2^\top)$$

- Alternating gradient descent**

$$\Theta_{1,k+1} = \Theta_{1,k} - \gamma \nabla_{\Theta_1} f(\Theta_{1,k} \Theta_{2,k}^\top)$$

$$\Theta_{2,k+1} = \Theta_{2,k} - \gamma \nabla_{\Theta_2} f(\Theta_{1,k+1} \Theta_{2,k}^\top)$$

# Experiments (Google Colab)



# Experiments

1. Matrix regression ([Google Colab](#))
  - 1.1 Rank-constrained formulation  
(solved by **projected gradient descent**)
  - 1.2 Nuclear norm regularization formulation  
(solved by **proximal gradient algorithm**)
  - 1.3 Nonconvex Burer–Monteiro factorization formulation  
(solved by **alternating gradient descent**)
2. Sparse covariance estimation (see [documentations](#) of `scikit-learn`)
3. Graphical Lasso (see [documentations](#) of `scikit-learn`)

# Covariance Estimation

## Covariance Estimation

- Let  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  be a collection of  $n$  independent and identically distributed (i.i.d.) samples from a distribution in  $\mathbb{R}^d$  with zero mean and covariance matrix  $\mathbf{\Sigma} = \text{Cov}(\mathbf{x}_1) \in \mathcal{S}_d^+$
- A natural estimator is the *sample covariance matrix*

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n} \mathbf{X}^\top \mathbf{X},$$

where  $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}$

- Note that  $\mathbb{E} \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{\Sigma}$  for all  $i \in \{1, \dots, n\}$ , so  $\hat{\mathbf{\Sigma}}$  is an unbiased estimator of  $\mathbf{\Sigma}$
- Usually, the goal is to bound the error measure in the  $\ell_2$ -operator norm

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$$

## Regularized Covariance Estimator

- If  $\Sigma$  is assumed to be *sparse*, the  $\ell_1$ -regularized covariance estimator is found by solving the convex program

$$\underset{\Sigma \in \mathcal{S}_d^+}{\text{minimize}} \left\{ \frac{1}{2} \|\Sigma - \hat{\Sigma}\|_F^2 + \lambda \|\Sigma\|_1 \right\}$$

- Could be hard to solve since it might involves **computationally prohibitive** *iterative* procedures
- Instead, we consider estimators based on **thresholding** (i.e., *non-iterative*)
- (Hard) thresholding operator:

$$T_\lambda(u) = u \cdot \mathbb{1}\{|u| > \lambda\} = \begin{cases} u & \text{if } |u| > \lambda, \\ 0 & \text{otherwise} \end{cases}$$



# Thresholding-based Covariance Estimation

- The thresholding estimator is  $T_{\lambda_n}(\hat{\Sigma})$  (elementwise thresholding)
- The parameter  $\lambda_n > 0$  is suitably chosen as a function of  $n$  and  $d$

## Theorem

*Let  $\{\mathbf{x}_i\}_{i=1}^n$  be an i.i.d. sequence of zero-mean random vectors with covariance matrix  $\Sigma$ , and suppose that each component of  $\mathbf{x}_{i,j}$  is sub-Gaussian with parameter at most  $\sigma$ . If  $n > \log d$ , then for any  $\delta > 0$ , the thresholded sample covariance matrix  $T_{\lambda_n}(\hat{\Sigma})$  with  $\lambda_n/\sigma^2 = 8\sqrt{(\log d)/n} + \delta$  satisfies*

$$\mathbb{P}(\|\| T_{\lambda_n}(\hat{\Sigma}) - \Sigma \| \|_2 \geq 2\lambda_n \|\| \mathbf{A} \| \|_2) \leq 8 \exp \left\{ -\frac{n}{16} \min\{\delta, \delta^2\} \right\}.$$

# Graphical Lasso

## Gaussian Graphical Model I

- **Undirected graphical model:**  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , set of vertices  $\mathcal{V} = \{1, \dots, d\}$
- Each vertex  $j \in \mathcal{V}$  is associated with a random variable  $x_j \in \mathcal{X}_j$
- Suppose that the random variables are jointly Gaussian with mean zero and covariance matrix  $\mathbf{\Sigma}$ , i.e.,  $\mathbf{x} = (x_1, \dots, x_d)^\top \sim \mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma})$
- A zero-mean Gaussian random vector  $\mathbf{x} = (x_i)_{1 \leq i \leq d}^\top \in \mathbb{R}^d$  with covariance matrix  $\mathbf{\Sigma} \in \mathcal{S}_d^{++}$  admits a density of the form

$$\mathbb{P}(x_1, \dots, x_d; \mathbf{\Theta}^*) \propto \sqrt{\det \mathbf{\Theta}^*} \cdot e^{-\frac{1}{2} \langle \mathbf{x}, \mathbf{\Theta}^* \mathbf{x} \rangle} = \sqrt{\det \mathbf{\Theta}^*} \cdot \prod_{(j,k) \in \mathcal{E}} \underbrace{e^{-\frac{1}{2} \mathbf{\Theta}_{j,k}^* x_j x_k}}_{\psi_{j,k}(x_j, x_k)},$$

where  $\mathbf{\Theta}^* = \mathbf{\Sigma}^{-1}$  is the inverse covariance (a.k.a. precision) matrix,  $\mathcal{E}$  is the set of edges of the graph,  $\psi_{j,k}$  is a function of a pair of edges  $(j, k)$

## Gaussian Graphical Model II

- The components  $\mathbf{x} = (x_i)_{1 \leq i \leq d}^\top$  might satisfy various types of **conditional independence** relationships
  - $x_j$  is *conditionally independent* of  $x_k$  given the other variables  $x_{\setminus\{j,k\}}$

$$x_j \mid x_{\setminus\{j,k\}} \perp\!\!\!\perp x_k$$

- In the **Gaussian** case, this conditional independence holds if and only if  $\Theta^*$  has a zero in position  $(j, k)$
- *Conditional independence* is directly captured by the **sparsity** of  $\Theta^*$

# Graphical Lasso

- The rescaled negative log-likelihood of the multivariate Gaussian based on samples  $\{\mathbf{x}_i\}_{i=1}^n$  takes the form

$$\mathcal{L}_n(\Theta) = \left\langle \hat{\Sigma}, \Theta \right\rangle_{\text{F}} - \log \det \Theta,$$

with  $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\top}$

- If  $\hat{\Sigma}$  is invertible, then  $\hat{\Theta}_{\text{MLE}} = \hat{\Sigma}^{-1}$
- When  $d > n$ ,  $\hat{\Sigma}$  is always rank-deficient  $\Rightarrow$  **NOT** invertible
- $\ell_1$ -constraint on  $\Theta$  for the graph  $G$  having *relatively few* edges

# Graphical Lasso

- The graphical Lasso objective is

$$\hat{\Theta} = \operatorname{argmin}_{\Theta \in \mathcal{S}_d^{++}} \left\{ \left\langle \hat{\Sigma}, \Theta \right\rangle_F - \log \det \Theta + \lambda_n \|\Theta\|_{1,\text{off}} \right\},$$

where  $\|\cdot\|_{1,\text{off}}$  is the elementwise  $\ell_1$ -norm without diagonal entries defined by

$$\|\Theta\|_{1,\text{off}} := \sum_{\substack{1 \leq i, j \leq d \\ i \neq j}} |\theta_{i,j}| = \|\Theta\|_1 - \|\text{diag}(\Theta)\|_1$$

- Do not penalize the diagonal for preserving positive definiteness and preventing from introducing further **bias** to the estimator

# Graphical Lasso

- Bounds on the Frobenius norm error  $\|\hat{\Theta} - \Theta\|_F$

## Proposition

*Suppose that the inverse covariance matrix  $\Theta^*$  has at most  $m$  nonzero entries per row, and we solve the graphical Lasso with regularization parameter  $\lambda_n = 8\sigma^2 \left( \sqrt{(\log d)/n} + \delta \right)$  for some  $\delta \in (0, 1]$ . Then as long as  $6(\|\Theta^*\|_2 + 1)^2 \lambda_n \sqrt{md} < 1$ , the graphical Lasso estimate  $\hat{\Theta}$  satisfies*

$$\|\hat{\Theta} - \Theta\|_F^2 \leq \frac{9}{(\|\Theta^*\|_2 + 1)^4} m d \lambda_n^2$$

*with probability at least  $1 - 8e^{-n\delta^2/16}$ .*

## What We DID NOT Cover Today (Yet Important) I

1. *Statistical* properties (e.g., **inference** and **uncertainty quantification**) of matrix optimization problems, e.g., *noisy matrix completion* (e.g., construction of optimal confidence intervals for each missing entry, see Chen et al., 2019)
2. Other matrix optimization problems, e.g., (robust/sparse) principal component analysis, phase retrieval, blind deconvolution, factor models
3. Other more sophisticated algorithms for solving matrix optimization problems, e.g., coordinate descent, ADMM, other proximal algorithms and their stochastic (and variance-reduced) variants
4. Convergence analysis of matrix optimization algorithms
5. More delicate/specific structures in covariance or precision matrix estimation problems, e.g., simultaneously sparse and low-rank, sparse + low-rank, bandable, Toeplitz, Toeplitz + low-rank



## What We DID NOT Cover Today (Yet Important) II

6. **Proximal identification** of proximal algorithms (Iutzeler and Malick, 2020):  
Can the iterates of the proximal algorithms *identify* the desired structure eventually after some finite time?

# References I

- Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.
- Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.
- Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*, 2020a.
- Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020b.
- Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. Chapman and Hall/CRC, 2020.
- Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 4th edition, 2013.

## References II

- Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020.
- Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.
- Franck Iutzeler and Jérôme Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28(4):661–678, 2020.
- Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.
- Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2019.
- André Uschmajew and Bart Vandereycken. On critical points of quadratic low-rank matrix optimization problems. *IMA Journal of Numerical Analysis*, 40(4):2626–2651, 2020.
- Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

The End  
Thank you!