Convex and Nonconvex Optimization for Low Rank Matrix Completion

Zijian Zhao, Yichen Ji

February 29, 2024

THE UNIVERSITY OF CHICAGO

Introduction: Low Rank Matrix Completion

Goal: Estimate a low rank matrix by its partial and noisy entries
Technical Requirement: Statistically and computationally efficient algorithm
Valid Paradigms: Both convex relaxation and nonconvex optimization

Problem Formulation

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} F(X) \quad \text{s.t.} \quad \text{rank}(X) \leq r \tag{1}$$

where $F : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ is a given convex loss function.

Nonconvex Approach I: Formulation

Burer-Monteiro Factorization[BM03]

Let $X = LR^T$, where $L \in \mathbb{R}^{n_1 \times r}$ and $R \in \mathbb{R}^{n_2 \times r}$, we have

$$\min_{\substack{L \in \mathbb{R}^{n_1 \times r}, \\ R \in \mathbb{R}^{n_2 \times r}}} f(L, R) := F(LR^T) + \mathsf{reg}(L, R) \tag{2}$$

Here, $f(L, R)$ represents the objective function composed of the convex loss $F(LR^T)$ and a regularization term $\mathsf{reg}(L, R)$.

Key change: for low rank case $(r \ll \min\{n_1, n_2\})$, the size of all the variables is approximately linear (L, R) in $n_1 + n_2$, while originally the variables (X) are quadratic.

Nonconvex Approach II: Iterative Schemes I

Three major classes of iterative schemes to find global optimum [CC18]:

(Projected) Gradient Descent [BM05, KMO09, KMO10, CW15, WCCL16, SL16, ZL16, MWCC18]

Minimize a loss function $f(L, R)$ w.r.t (L, R):

$$L^{t+1} = \mathcal{P}_{\mathcal{L}} \left[L^t - \eta^t \nabla_L f(L^t, R^t) \right] \tag{3}$$

$$R^{t+1} = \mathcal{P}_{\mathcal{R}} \left[R^t - \eta^t \nabla_R f(L^t, R^t) \right] \tag{4}$$

where η^t is the step size and $\mathcal{P}_{\mathcal{L}}, \mathcal{P}_{\mathcal{R}}$ denote the Euclidean projection onto the sets $\mathcal{L}$ and $\mathcal{R}$.

(Projected) gradient descent is the best by its simple form, cheap iteration cost and efficiency [CC18].

Nonconvex Approach II: Iterative Schemes II

Alternating Minimization[JNS13, Har14]

Hold other factors constant, optimize one of the factors alternatively by a convex problem:

$$L^{t+1} = \arg\min_{L \in \mathbb{R}^{n_1 \times r}} f(L, R^t) \tag{5}$$

$$R^{t+1} = \arg\min_{R \in \mathbb{R}^{n_2 \times r}} f(L^{t+1}, R) \tag{6}$$

Singular Value Projection (SVP) [JMD10, NUNS⁺14, JN15]

GD on $F(LR^T)$ in the $n_1 \times n_2$ matrix space, then use SVD to project back to the factor space:

$$(L^{t+1}, R^{t+1}) = \mathsf{SVD}_r \left[L^t R^{t^T} - \eta^t \nabla F(L^t R^{t^T}) \right] \tag{7}$$

where $\mathsf{SVD}_r(Z)$ returns the top rank-r factors of Z.

Conclusion for Nonconvex and Inspirations for Convex Relaxation

Nonconvex Optimization:

- Efficient
- Theoretical guaranteed for estimation accuracy
- Properties such as local convergence, implicit regularization, global convergence (saddle-escaping algorithms with strict saddle property) would give satisfactory results.

Convex Relaxation:

- Worthwhile to solve a semidefinite and convex programs for large-scale or high-dimensional problems?
- Faithful in practice
- Not explained for long until [CCF$^+$19]

Convex Relaxation

Problem Reformulation from (1)

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} F(X) + \lambda \, \mathrm{rank}(X) \quad (8)$$

where F is a convex function and $\lambda > 0$ is a regularization parameter.

Convex Relaxation

Remove the nonconvex rank function by convex terms.

$$\min_{X \in \mathbb{R}^{n \times n}} g(X) := F(X) + \lambda \|X\|_* \quad (9)$$

where $\|\cdot\|_*$ is the nuclear norm. More specifically, we consider:

$$\min_{X \in \mathbb{R}^{n \times n}} \quad g(\boldsymbol{X}) := \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 + \lambda \|X\|_* \quad (10)$$

High Level Ideas for Convex Proofs

Main difficulty for convex relaxation solutions: it does not have closed-form solutions that would give reliable guarantees.

High-level idea: Prove that the nonconvex solutions are close to convex solutions (i.e. a tight approximation to convex solutions) [CCF+19].

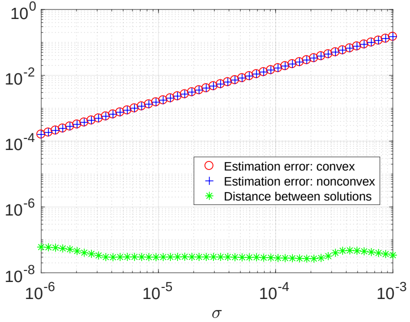

Figure: Empirical Evidence for Closeness

Main Results - Model Assumptions

Random Sampling (Assumption 1)

Each index (i, j) belongs to the index set Ω independently with probability p.

Random Noise (Assumption 1)

The noise matrix $E = [E_{ij}]_{1 \leq i,j \leq n}$ is composed of i.i.d. zero-mean sub-Gaussian random variables with sub-Gaussian norm at most $\sigma > 0$, i.e. $\|E_{ij}\|_{\psi_2} \leq \sigma$

Incoherence condition

A rank-r matrix $M^\star \in \mathbb{R}^{n \times n}$ with SVD $M^\star = U^\star \Sigma^\star V^{\star\top}$ is said to be μ-incoherent if $\|U^\star\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \|U^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}$ and $\|V^\star\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \|V^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}$

When Rank and condition number are constants O(1)

Theorem 1 Let $M^\star$ be rank-r and μ-incoherent with a condition number κ, where the rank and the condition number satisfy $r, \kappa = O(1)$. Suppose that Assumption 1 holds and take $\lambda = C_\lambda \sigma \sqrt{np}$ in (10) for some large enough constant $C_\lambda > 0$. Assume the sample size obeys $n^2 p \geq C\mu^2 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma \lesssim \sqrt{\frac{np}{\mu^3 \log n}} \|M^\star\|_\infty$ for some sufficiently small constant $c > 0$. Then with probability exceeding $1 - O\left(n^{-3}\right)$:

1. Any minimizer $\boldsymbol{Z}_{\mathrm{cvx}}$ of (10) obeys

$$\|Z_{\mathrm{cvx}} - M^\star\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|_{\mathrm{F}}\,; \quad \|Z_{\mathrm{cvx}} - M^\star\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|$$

$$\|Z_{\mathrm{cvx}} - M^\star\|_\infty \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\mu n \log n}{p}} \|M^\star\|_\infty\,.$$

2. Letting $Z_{\mathrm{cvx},r} \triangleq \arg\min_{Z:\mathrm{rank}(\boldsymbol{Z}) \leq r} \|\boldsymbol{Z} - \boldsymbol{Z}_{\mathrm{cvx}}\|_{\mathrm{F}}$ be the best rank-r approximation of $\boldsymbol{Z}_{\mathrm{cvx}}$, we have

$$\|Z_{\mathrm{cvx},r} - Z_{\mathrm{cvx}}\|_{\mathrm{F}} \leq \frac{1}{n^3} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|$$

Breakdown of required conditions and implications

- **Sample complexity**: the sample size needs to exceed the order of $npolylogn$, which is information-theoretically optimal up to some logarithmic term.
- **Noise size**: the size of the noise in each entry is allowed to be substantially larger than the maximum entry in the matrix i.e. can have a very small signal-to-noise ratio w.r.t. each observed entry.
- **Nearly low-rank structure of the convex solution**: the optimizer of the convex program is almost if not exact, rank-r.
- **Implicit regularization**: the convex approach implicitly controls the spikiness of its entries, without resorting to explicit regularization.
- **Entry-wise and spectral norm error control**: the estimation errors of the convex optimizer are fairly spread out across all entries, thus implying near-optimal entry-wise error control.
- **Statistical guarantees for fast iterative optimization methods**: when these convex optimization algorithms converge w.r.t. the objective value, they are guaranteed to return a statistically reliable estimate e.g. SVT, FPC, SOFT-IMPUTE, etc.

When rank and condition number can grow with n

Theorem 2 Let $M^\star$ be rank-r and μ-incoherent with a condition number κ. Suppose Assumption 1 holds and take $\lambda = C_\lambda \sigma \sqrt{np}$ in (10) for some large enough constant $C_\lambda > 0$. Assume the sample size obeys $n^2 p \geq C \kappa^4 \mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma \sqrt{\frac{n}{p}} \leq c \frac{\sigma_{\min}}{\sqrt{\kappa^4 \mu r \log n}}$ for some sufficiently small constant $c > 0$. Then with probability exceeding $1 - O\left(n^{-3}\right)$,

1. Any minimizer Z_{cvx} of (10) obeys

$$\|Z_{\text{cvx}} - M^\star\|_{\text{F}} \lesssim \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|_{\text{F}} ; \|Z_{\text{cvx}} - M^\star\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|$$

$$\|Z_{\text{cvx}} - M^\star\|_\infty \lesssim \sqrt{\kappa^3 \mu r} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|M^\star\|_\infty$$

2. Letting $Z_{\text{cvx},r} \triangleq \arg\min_{Z:\text{rank}(\boldsymbol{Z}) \leq r} \|\boldsymbol{Z} - Z_{\text{cvx}}\|_{\text{F}}$ be the best rank-r approximation of $\boldsymbol{Z}_{\text{cvx}}$, we have

$$\|Z_{\text{cvx},r} - Z_{\text{cvx}}\|_{\text{F}} \leq \frac{1}{n^3} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|M^\star\|$$

Approximate nonconvex optimizer

Idea: Lack of closed-form primal solution to (10) $\rightarrow$ Invoke an iterative nonconvex algorithm to approximate such a primal solution.
Algorithm: Construction of an approximate primal solution

- Initialization: $X^0 = X^\star; Y^0 = Y^\star$
- Gradient updates: for $t = 0, 1, \ldots, t_0 - 1$ do

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right) = \boldsymbol{X}^t - \frac{\eta}{p}\left(\mathcal{P}_\Omega\left(\boldsymbol{X}^t \boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\boldsymbol{Y}^t + \lambda \boldsymbol{X}^t\right)$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right) = \boldsymbol{Y}^t - \frac{\eta}{p}\left(\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t \boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\right]^\top \boldsymbol{X}^t + \lambda \boldsymbol{Y}^t\right)$$

where $\eta > 0$ is the step size, $\mathcal{P}_\Omega$ represents the projection on the the subspace matrices supported on Ω.
Note: This algorithm is not practical since it starts from the ground truth, so it's mainly used to simplify theoretical analysis. One can apply spectral initialization to make it practical.

Thank You

References I

Samuel Burer and Renato D.C. Monteiro.
A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization.
Mathematical Programming, 95(2):329–357, February 2003.

Samuel Burer and Renato D.C. Monteiro.
Local Minima and Convergence in Low-Rank Semidefinite Programming.
Mathematical Programming, 103(3):427–444, July 2005.

Yudong Chen and Yuejie Chi.
Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, May 2018.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan.
Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization, October 2019.

References II

Yudong Chen and Martin J. Wainwright.
Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees, September 2015.

Moritz Hardt.
Understanding Alternating Minimization for Matrix Completion.
In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660, October 2014.

Prateek Jain, Raghu Meka, and Inderjit Dhillon.
Guaranteed Rank Minimization via Singular Value Projection.
In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

References III

Prateek Jain and Praneeth Netrapalli.
Fast Exact Matrix Completion with Finite Samples.
In *Proceedings of The 28th Conference on Learning Theory*, pages 1007–1034. PMLR, June 2015.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi.
Low-rank matrix completion using alternating minimization.
In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, June 2013. Association for Computing Machinery.

Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh.
Matrix Completion from Noisy Entries.
In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

References IV

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh.
Matrix Completion From a Few Entries.
IEEE Transactions on Information Theory, 56(6):2980–2998, June 2010.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen.
Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval and Matrix Completion.
In *Proceedings of the 35th International Conference on Machine Learning*, pages 3345–3354. PMLR, July 2018.

Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain.
Non-convex Robust PCA.
In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

References V

Ruoyu Sun and Zhi-Quan Luo.

Guaranteed Matrix Completion via Non-Convex Factorization.

IEEE Transactions on Information Theory, 62(11):6535–6579, November 2016.

Ke Wei, Jian-Feng Cai, Tony F. Chan, and Shingyu Leung.

Guarantees of Riemannian Optimization for Low Rank Matrix Recovery.

SIAM Journal on Matrix Analysis and Applications, 37(3):1198–1222, January 2016.

Qinqing Zheng and John Lafferty.

Convergence Analysis for Rectangular Matrix Completion Using Burer-Monteiro Factorization and Gradient Descent, November 2016.