Exact Linear Convergence Rate Analysis for Low-Rank Symmetric Matrix Completion via Gradient Descent

Trung Vu and Raviv Raich

School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA

vutru, raich@oregonstate.edu

Symmetric Matrix Completion Problem (SMCP)

1	?	?	1
?	?	6	?
?	6	?	2
1	?	2	1

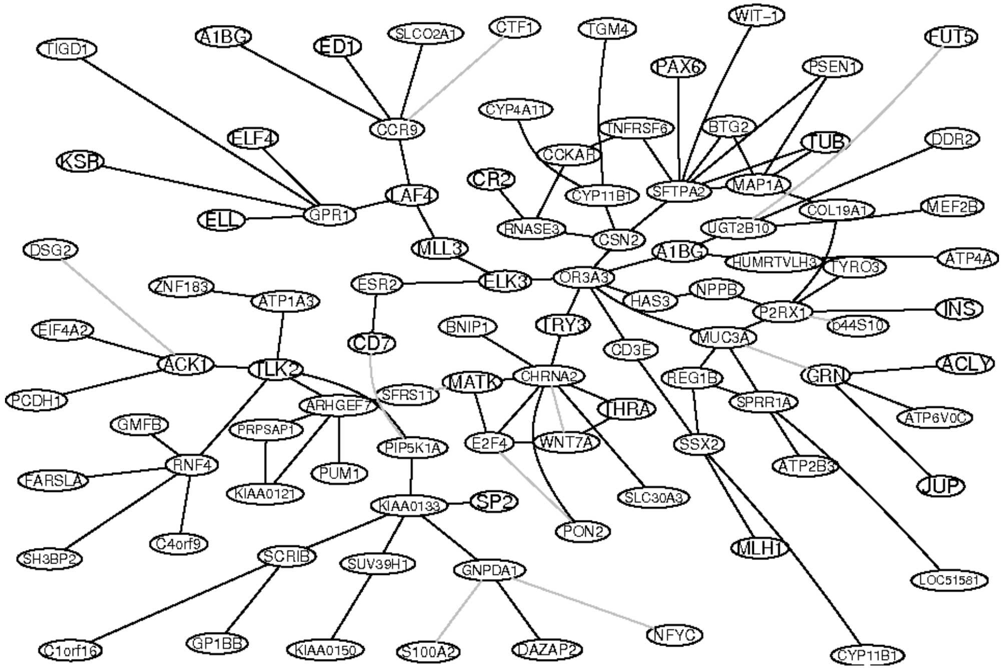
$n = 4, r = 1, s = 8$

- Positive semidefinite (PSD) rank-r matrix $M \in \mathbb{R}^{n\times n}$
  - $M = XX^T$, where $X \in \mathbb{R}^{n\times r}$
- Sampling set Ω with cardinality s
  - Ω is symmetric!

find	M_{ij} for $(i,j) \in \Omega^c$
given	$rank(M) \leq r$
and	M_{ij} for $(i,j) \in \Omega$

Applications

- Maximum likelihood estimation (MLE) of covariance matrices in Gaussian graphical models [1]
- Density matrix completion in quantum state tomography [2]
- Low-rank approximation of correlation matrices in finance and risk management [3]

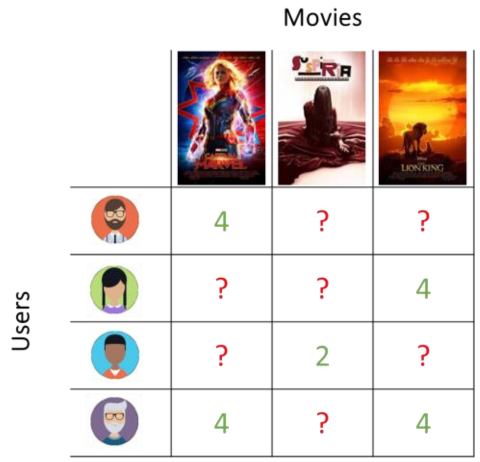


Schäfer, Juliane, and Korbinian Strimmer. "Learning Large-Scale Graphical Gaussian Models from Genomic Data." In *AIP Conference Proceedings*, vol. 776, no. 1, pp. 263-276. American Institute of Physics, 2005.

Rectangular Matrix Completion as SMCP

- Rectangular (non-symmetric) matrix completion
  - $A \in \mathbb{R}^{n_1 \times n_2}$ has rank r
    - $A = YZ^T$, where $Y \in \mathbb{R}^{n_1 \times r}, Z \in \mathbb{R}^{n_2 \times r}$
  - $|\widetilde{\Omega}| = \tilde{s}$

- Semidefinite lifting
  - $X = \begin{bmatrix} Y \\ Z \end{bmatrix} \in \mathbb{R}^{n \times r}$, where $n = n_1 + n_2$
  - $M = XX^T = \begin{bmatrix} YY^T & YZ^T \\ ZY^T & ZZ^T \end{bmatrix} \in \mathbb{R}^{n \times n}$
    - M also has rank r



$n_1 = 4, n_2 = 3$

$r = 1, \tilde{s} = 5$

?	?	?	?	4	?	?
?	?	?	?	?	?	4
?	?	?	?	?	2	?
?	?	?	?	4	?	4
4	?	?	4	?	?	?
?	?	2	?	?	?	?
?	4	?	4	?	?	?

$n = 7$
$r = 1, s = 10$

Existing Approaches

Approach	Problem formulation	Existing algorithms
Linearly-constrained nuclear norm minimization	$\min_{Z \in \mathbb{R}^{n \times n}} \|Z\|_* \ \mathrm{s.t.} \sum_{(i,j)\in\Omega} \left|Z_{ij} - M_{ij}\right| = 0$	Iterative soft thresholding, i.e., SVT [4], APG [5], CGD [6]
Rank-constrained least squares	$\min_{Z \in \mathbb{R}^{n \times n}} \sum_{(i,j)\in\Omega} \left(Z_{ij} - M_{ij}\right)^2 \ \mathrm{s.t.}\ rank(Z) \le r$	Iterative hard thresholding, i.e., SVP [7], NIHT [8], Accelerated IHT [9,10]
Low-rank factorization	$\min_{X \in \mathbb{R}^{n \times r}} \sum_{(i,j)\in\Omega} \left((XX^T)_{ij} - M_{ij}\right)^2$	Gradient descent [11,12], projected gradient descent [13]

Gradient Descent (GD) for SMCP

- SMCP as *unconstrained* *non-convex* optimization

$$\underbrace{\min_{X\in\mathbb{R}^{n\times r}} \frac{1}{4}\|P_\Omega(XX^T - M)\|_F^2}_{f(X)}$$

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij} = \begin{cases} Z_{ij} & \text{if } (i,j)\in\Omega, \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 1 (Non-convex) Gradient Descent

Input: $\boldsymbol{X}^0, \mathcal{P}_\Omega(\boldsymbol{M}), \eta$

Output: $\{\boldsymbol{X}^k\}$

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: $\quad \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\underbrace{\mathcal{P}_\Omega(\boldsymbol{X}^k{\boldsymbol{X}^k}^\top - \boldsymbol{M})\boldsymbol{X}^k}_{\nabla f(X^k)}$

Convergence Analysis of GD for SMCP

- Most focus on **global** guarantees
  - Standard assumptions:
    - M is μ-incoherent
    - Ω is a uniform sampling

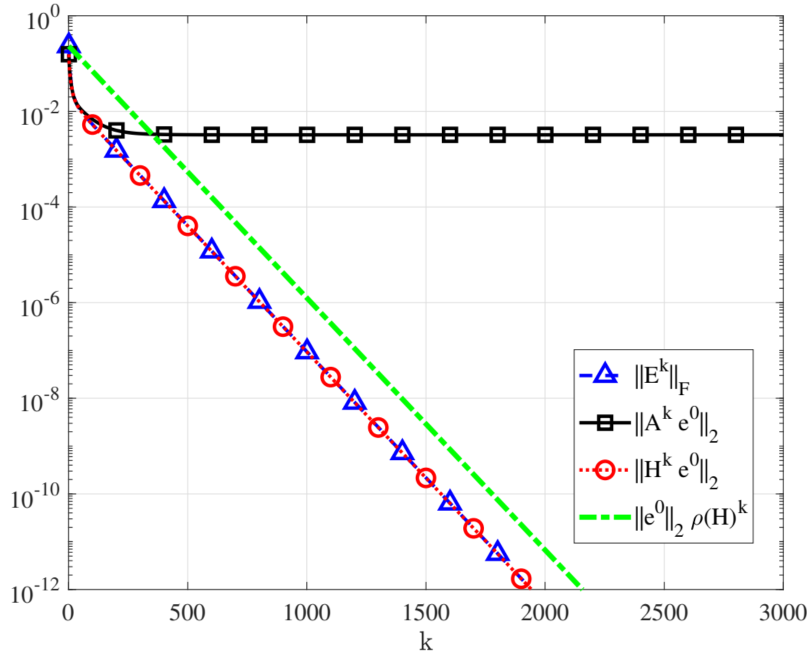
  - **(Ma *et. al.* 's result [12]**) If $s = O(\mu^3 r^3 n \, log^3 n)$, then *w.h.p* GD with spectral initialization converges globally at linear rate

$$\rho \geq 1 - \frac{2}{125\kappa^2}$$

Large condition number κ implies a loose bound on the linear rate!

Contribution

- We studies the **local** convergence of GD for SMCP
  - We establish a deterministic condition on M and Ω for linear convergence
    - Do not require standard assumptions
    - Do not require asymptotic regime
  - We provide the exact linear rate in *closed-form*
    - Tighter than the global bound in [12]
    - Match well the convergence behavior in practice



Preliminaries

- Rank-r eigendecomposition of M

$$M = U\Sigma U^T$$

→

- $U \in R^{n\times r}$ is a semi-orthogonal matrix
- $\Sigma \in R^{r\times r}$ is a diagonal matrix

- Projection onto the null space of M

$$P_{U_\perp} = I_n - UU^T$$

$$P_r(M+E) - M = \underbrace{E - P_{U_\perp} E P_{U_\perp}}_{\nabla P_r(M)\cdot E} + O\left(\|E\|_F^2\right)$$

- Vectorization of the projection onto the tangent plane of the set of rank-r matrices at M

$$P_1 = I_{n^2} - P_{U_\perp} \otimes P_{U_\perp}$$

→

$$\text{vec}\left(E - P_{U_\perp} E P_{U_\perp}\right) = P_1 \text{vec}(E)$$

- Vectorization of the projection onto the set of symmetric matrices

$$P_2 = \frac{I_{n^2} + T_{n^2}}{2}$$

→

$$\text{vec}\left(\frac{E + E^T}{2}\right) = P_2 \text{vec}(E)$$

A Recursion on the Error

- Recall the GD update

$$X^{k+1} = X^k - \eta P_\Omega\left(X^k {X^k}^T - M\right) X^k$$

- Let $E^k = X^k {X^k}^T - M$ be the error matrix

$$\Rightarrow E^{k+1} = E^k - \eta\left(P_\Omega(E^k)M + MP_\Omega(E^k)\right) + O\left(\|E^k\|_F^2\right)$$

- Let $e^k = \mathrm{vec}(E^k)$ be the error vector

$$\Rightarrow e^{k+1} = \underbrace{(I_{n^2} - \eta(M \oplus M)S^T S)}_{A} e^k + O\left(\|e^k\|_2^2\right)$$

$$M \oplus M = M \otimes I_n + I_n \otimes M$$

$$S \in \mathbb{R}^{s \times n^2}: \begin{cases} SS^T = I_s \\ vec\left(P_\Omega(E)\right) = S^T S vec(E) \end{cases}$$

Convergence of Nonlinear Difference Equations

$$e^{k+1} = Ae^k + O\left(\|e^k\|_2^2\right)$$

- Polyak's result [14]:
  - If $\|A^k\| \leq c(\epsilon)(\rho+\epsilon)^k$ for $\rho < 1$ and any $\epsilon > 0$, then for sufficiently small $\|e^0\|$:

$$\|e^k\| \leq C(\epsilon)\|e^0\|(\rho+\epsilon)^k$$

- Vu and Raich's result [15]:
  - Let $\rho = \rho(A)$ be the spectral radius of A. If $\rho < 1$, then for sufficiently small $\|e^0\|$:

$$\|e^k\| \leq K(\rho, \|e^0\|)\|e^0\|\rho^k$$

→ Can we apply the result directly to show the linear convergence of GD for SMCP?

Unfortunately, **NO**. Since $\rho(A) \geq 1$!

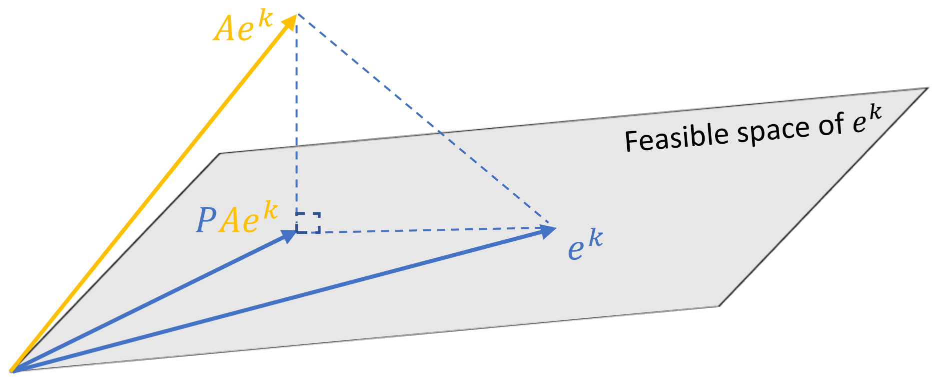
$A = I_{n^2} - \eta(M \oplus M)S^T S$

Structural Constraints on the Error

Structural properties of the error matrix

$$E = XX^T - M$$

1. $E = E^T$
2. $P_r(M + E) = M + E,$
   - where P_r is the projection onto the set of rank-r matrices
3. $M + E$ is PSD



Structural properties of the error vector

$$e = \mathrm{vec}(E) = \mathrm{vec}(XX^T - M)$$

1. $e = P_2 e$
2. $e = P_1 e + O(\|e\|_2^2)$
3. Negligible effect

$$\Rightarrow e = \underbrace{P_1 P_2}_{P} e + O(\|e\|_2^2)$$

$$P_1 = I_{n^2} - P_{U_\perp} \otimes P_{U_\perp}$$

$$P_2 = \frac{I_{n^2} + T_{n^2}}{2}$$

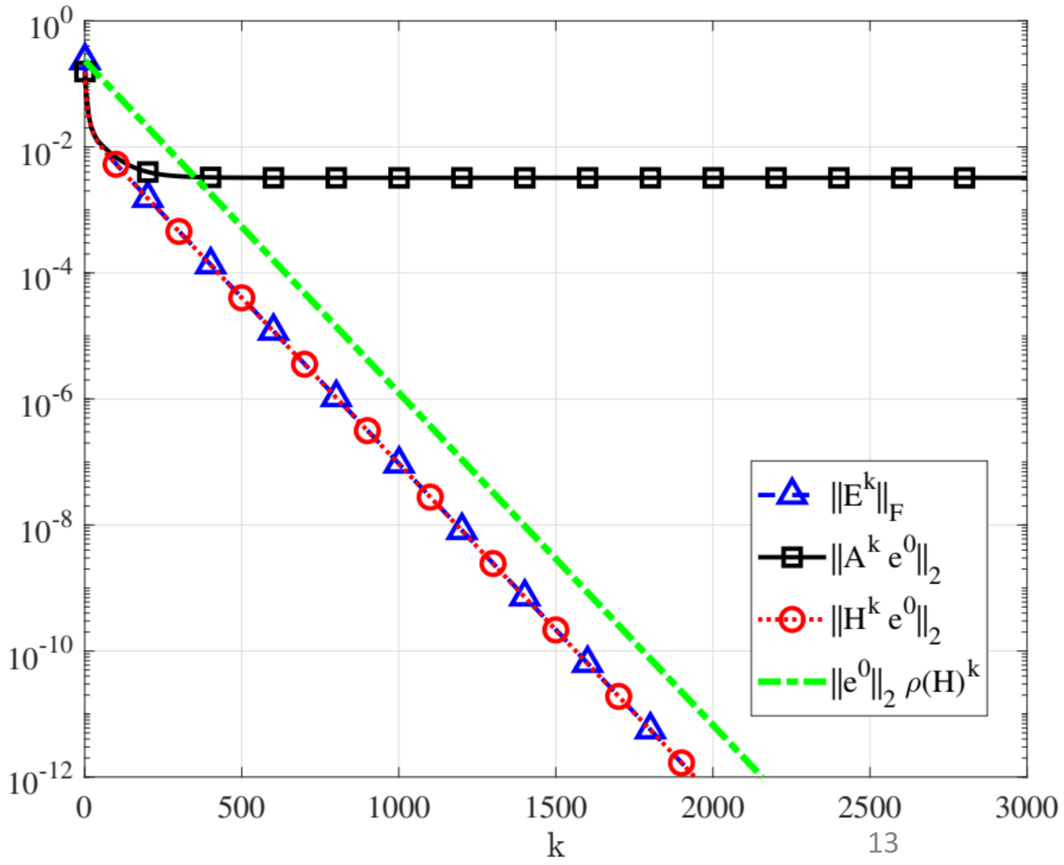
Determining the Linear Rate

- Integrating structural constraints

$$P = P_1 P_2 = \left(I_{n^2} - P_{U_\perp} \otimes P_{U_\perp}\right) \frac{I_{n^2} + T_{n^2}}{2}$$

$$H = P(I_{n^2} - \eta(M \oplus M) S^T S) P$$

$$e^{k+1} = \underbrace{PAP}_{H} e^k + O\left(\|e^k\|_2^2\right)$$



Theorem

If $\rho(H) < 1$, then there exists a neighborhood $\mathcal{N}(M)$ such that for any $X^0 {X^0}^T \in \mathcal{N}(M)$:

$$\|e^k\| \le C \|e^0\| \rho(H)^k$$

for some numerical constant $C > 0$.

Conclusions

- Local convergence analysis recovers the **exact** rate of linear convergence of GD for SMCP.

- Integrating **structural constraints** is the key to obtain the convergence rate for the nonlinear difference equation on the error.

- It is interesting to extend the analysis to the **non-symmetric** case and make connection to existing works on the **global** convergence.

References

1. Caroline Uhler. "Geometry of maximum likelihood estimation in Gaussian graphical models." Annals of Statistics 40, no. 1, pp. 238-261, 2012.
2. David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert. "Quantum state tomography via compressed sensing." Physical review letters 105, no. 15, 2010.
3. Zhenyue Zhang, and Lixin Wu. "Optimal low-rank approximation to a correlation matrix." Linear algebra and its applications 364, pp. 161-187, 2003.
4. Jian-Feng Cai, Emmanuel J Candes, and Zuowei Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.
5. Kim-Chuan Toh and Sangwoon Yun, “An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems,” Pacific Journal of Optimization, vol. 6, no. 615-640, pp. 15, 2010.
6. Martin Jaggi, “Revisiting Frank-Wolfe: Projection-free sparse convex optimization,” in Proceedings of the International Conference on Machine Learning. PMLR, 2013, pp. 427–435.
7. Prateek Jain, Raghu Meka, and Inderjit S Dhillon, “Guaranteed rank minimization via singular value projection,” in Advances in Neural Information Processing Systems, 2010, pp. 937–945.
8. Jared Tanner and Ke Wei, “Normalized iterative hard thresholding for matrix completion,” SIAM Journal on Scientific Computing, vol. 35, no.5, pp. S104–S125, 2013.
9. Trung Vu and Raviv Raich, “Accelerating iterative hard thresholding for low-rank matrix completion via adaptive restart,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp.2917–2921.
10. Trung Vu and Raviv Raich, “Local convergence of the Heavy Ball method in iterative hard thresholding for low-rank matrix completion,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 3417–3421.
11. Ruoyu Sun and Zhi-Quan Luo, “Guaranteed matrix completion via non-convex factorization,” IEEE Transactions on Information Theory, vol.62, no. 11, pp. 6535–6579, 2016.
12. 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 International Conference on Machine Learning. PMLR, 2018, pp. 3345–3354.
13. Yudong Chen and Martin J Wainwright, “Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees,” arXiv preprint arXiv:1509.03025, 2015.
14. Boris T. Polyak, “Some methods of speeding up the convergence of iteration methods,” USSR Computational Mathematics and Mathematical Physics, vol. 4, no. 5, pp. 1–17, 1964.
15. Trung Vu and Raviv Raich, “Adaptive step size momentum method for deconvolution,” in IEEE Statistical Signal Processing Workshop, 2018, pp. 438–442.

Thank you!

Check our full paper at https://arxiv.org/abs/2102.02396

Further results on local convergence analysis at https://trungvietvu.github.io/