Tuning-Free Convex Methods for Noisy Matrix Completion

Yuepeng Yang

joint work with Cong Ma

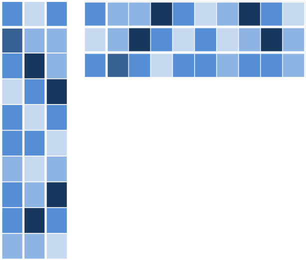
Noisy low-rank matrix completion

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

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

sampling set Ω

Noisy low-rank matrix completion



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

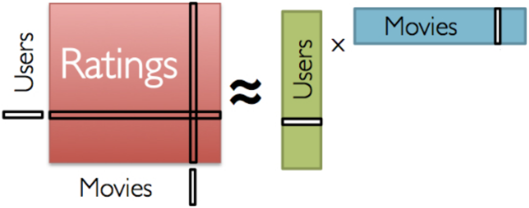
$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark &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

sampling set Ω

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

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

One application: Netflix challenge



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

Convex relaxation for matrix completion

observations: $M_{i,j} = L^{\star}_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$

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

Convex relaxation for matrix completion

observations: $M_{i,j} = L^{\star}_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$

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

convex relaxation:

$$\boldsymbol{L}_{\mathsf{cvx}} := \underset{\boldsymbol{L} \in \mathbb{R}^{n \times n}}{\text{argmin}} \underbrace{\sum_{(i,j) \in \Omega} \left(L_{i,j} - M^{\star}_{i,j}\right)^2}_{\text{squared loss}} + \lambda \|\boldsymbol{L}\|_*$$

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

Known statistical guarantees

- **random sampling:** each $(i, j) \in \Omega$ indep. with prob. p
- **random noise:** i.i.d. sub-Gaussian noise with variance proxy σ^2
- true matrix $\boldsymbol{L}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, well-conditioned, incoherent

Setting $\lambda \asymp \sigma\sqrt{np}$ yields minimax optimal estimation rate (Negahban and Wainwright '12, Chen et al '20)

Issue: tuning parameter λ requires knowledge of both σ and p

A solution: square root MC

— borrowing ideas from square root Lasso (Belloni et al. '11)

$$\boldsymbol{L}_{\mathsf{cvx}} \quad := \quad \underset{\boldsymbol{L}\in\mathbb{R}^{n\times n}}{\mathsf{argmin}} \quad \underbrace{\sum_{(i,j)\in\Omega} \left(L_{i,j} - M_{i,j}\right)^2}_{\text{squared loss}} + \lambda \|\boldsymbol{L}\|_*$$

$$\downarrow$$

$$\boldsymbol{L}_{\mathsf{cvx}} \quad := \quad \underset{\boldsymbol{L}\in\mathbb{R}^{n\times n}}{\mathsf{argmin}} \quad \underbrace{\sqrt{\sum_{(i,j)\in\Omega} \left(L_{i,j} - M_{i,j}\right)^2}}_{\text{sqaure root squared loss}} + \lambda \|\boldsymbol{L}\|_*$$

The intuition

λ is often chosen based on size of sub-gradient:

$$\frac{\partial}{\partial \boldsymbol{L}} \left(\sum_{(i,j)\in\Omega} \left(L_{i,j} - M_{i,j}\right)^2 \right) = 2\mathcal{P}_\Omega(\boldsymbol{L} - \boldsymbol{M}) \tag{1}$$

$$\frac{\partial}{\partial \boldsymbol{L}} \sqrt{\sum_{(i,j)\in\Omega} \left(L_{i,j} - M_{i,j}\right)^2} = \frac{\mathcal{P}_\Omega(\boldsymbol{L} - \boldsymbol{M})}{\|\mathcal{P}_\Omega(\boldsymbol{L} - \boldsymbol{M})\|_{\mathrm{F}}} \tag{2}$$

When $\boldsymbol{L} = \boldsymbol{L}^\star$, (1) is σ-dependent, (2) is not!

Prior works

— for $\|\boldsymbol{L} - \boldsymbol{L}^{\star}\|_{\mathrm{F}}$, ignoring log factors

Minimax limit	$O(\sigma\sqrt{n/p})$
Gaïffas and Klopp '17	$O(\max\{\sigma, \|\boldsymbol{L}\|_{\infty}\}\sqrt{n/p})$
Zhang, Yan, and Wright '21	$O(\sigma n^2)$

Main results for $r, \kappa = O(1)$

- **Random sampling**: Each (i,j) observed with prob $p \gtrsim \frac{\log^3 n}{n}$
- **Random noise**: Sub-Gaussian noise with sd $\sigma \lesssim \sqrt{\frac{np}{\log n}} \|\boldsymbol{L}^\star\|_\infty$
- **Regularity condition**: $\boldsymbol{L}^\star$ is incoherent and well-conditioned

$$\boldsymbol{L}_{\mathsf{cvx}} := \operatorname*{argmin}_{\boldsymbol{L} \in \mathbb{R}^{n \times n}} \quad \sqrt{\sum_{(i,j)\in\Omega} \big(L_{i,j} - M_{i,j}\big)^2} + \lambda \|\boldsymbol{L}\|_*,$$

Main results for $r, \kappa = O(1)$

- **Random sampling**: Each (i, j) observed with prob $p \gtrsim \frac{\log^3 n}{n}$
- **Random noise**: Sub-Gaussian noise with sd $\sigma \lesssim \sqrt{\frac{np}{\log n}} \|\boldsymbol{L}^\star\|_\infty$
- **Regularity condition**: $\boldsymbol{L}^\star$ is incoherent and well-conditioned

$$\boldsymbol{L}_{\mathsf{cvx}} := \operatorname*{argmin}_{\boldsymbol{L} \in \mathbb{R}^{n \times n}} \quad \sqrt{\sum_{(i,j) \in \Omega} \left(L_{i,j} - M_{i,j}\right)^2} + \lambda \|\boldsymbol{L}\|_*,$$

Theorem 1 (Yang and Ma, 2022)

Set $\lambda = 32/\sqrt{n}$. With high probability, $\boldsymbol{L}_{\mathrm{cvx}}$ achieves

$$\|\boldsymbol{L}_{\mathrm{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma \sqrt{\frac{n}{p}}$$

Implications

- Minimax optimal for a wide range of noise sizes $\sigma \lesssim \sqrt{\frac{np}{\log n}} \|\boldsymbol{L}^{\star}\|_{\infty}$

- Improves the error bound in (Zhang, Yan, and Wright '21) from $O(\sigma n^2)$ to $O(\sigma \sqrt{n/p})$

- A byproduct of our analysis: $\boldsymbol{L}_{\mathrm{cvx}}$ is nearly rank-r

A peek at analysis

—follows general roadmap of bridging convex and nonconvex analysis, e.g. Chen et al '20

- Sqrt-MC $\boldsymbol{L}_{\mathsf{cvx}} := \underset{\boldsymbol{L}\in\mathbb{R}^{n\times n}}{\text{argmin}} \quad \sqrt{\sum_{(i,j)\in\Omega}\left(L_{i,j} - M_{i,j}\right)^2} + \lambda\|\boldsymbol{L}\|_*$ is hard to analyze due to non-smoothness.

- We know more about nonconvex schemes than the convex ones

A peek at analysis

- "Translate" Sqrt-MC into smooth, nonconvex counterpart

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{argmin}} \ \min_{\theta} \frac{\sum_{(i,j)\in\Omega}\left([\boldsymbol{X}\boldsymbol{Y}^\top]_{i,j}-M_{i,j}\right)^2}{\theta}+\theta+\lambda(\|\boldsymbol{X}\|_{\mathrm{F}}^2+\|\boldsymbol{Y}\|_{\mathrm{F}}^2)$$

- Suffices to show:

  (1) Nonconvex solution $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}})$ is close to $\boldsymbol{L}_{\mathsf{cvx}}$,

  (2) $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}})$ is optimal

Conclusion and discussion

- We sharpen analysis of Sqrt-MC, a tuning-free convex scheme for noisy matrix completion
- Analysis is based on a nonconvex proxy that is close to both convex solution and ground truth
- Future directions: inference on $\boldsymbol{L}^\star$

Reference

Belloni, A., Chernozhukov, V., and Wang, L. (2011).
Square-root lasso: pivotal recovery of sparse signals via conic programming.
Biometrika, 98(4):791–806.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020).
Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization.
SIAM journal on optimization, 30(4):3098–3121.

Gaïffas, S. and Klopp, O. (2017).
High dimensional matrix estimation with unknown variance of the noise.
Statistica Sinica, pages 115–145.

Negahban, S. and Wainwright, M. J. (2012).
Restricted strong convexity and weighted matrix completion: Optimal bounds with noise.
The Journal of Machine Learning Research, 13(1):1665–1697.

Zhang, J., Yan, J., and Wright, J. (2021).
Square root principal component pursuit: Tuning-free noisy robust matrix recovery.
Advances in Neural Information Processing Systems, 34.