Absence of spurious solutions far from ground truth: A low-rank analysis with high-order losses

Ziye Ma, Ying Chen, Javad Lavaei, Somayeh Sojoudi

UC Berkeley

May 02, 2024

Contents

Introduction

- We study the following low-rank matrix recovery problem:

$$\begin{aligned}\min_{X\in\mathbb{R}^{n\times r}} f(X) &:= \frac{1}{2}\|\mathcal{A}(XX^T)-b\|^2 \\ &= \frac{1}{2}\|\mathcal{A}(XX^T-ZZ^\top)\|^2\end{aligned} \qquad (1)$$

- $\mathcal{A}:\mathbb{R}^{n\times n}\to\mathbb{R}^m$:

$$\mathcal{A}(M)=[\langle A_1,M\rangle,\ldots,\langle A_m,M\rangle]^T,$$

- $A_1,\ldots,A_m\in\mathbb{R}^{n\times n}$ are called sensing matrices. $b=\mathcal{A}(M^*)$.

Introduction

Definition 1

The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is said to satisfy the δ-RIP_{2r} property for some constant $\delta \in [0, 1)$ if the inequality

$$(1-\delta)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1+\delta)\|M\|_F^2$$

holds for all $M \in \mathbb{R}^{n\times n}$ with $\mathrm{rank}(M) \leq 2r$.

Introduction

Definition 1

The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is said to satisfy the δ-RIP$_{2r}$ property for some constant $\delta \in [0, 1)$ if the inequality

$$(1-\delta)\|M\|_F^2 \le \|\mathcal{A}(M)\|^2 \le (1+\delta)\|M\|_F^2$$

holds for all $M \in \mathbb{R}^{n\times n}$ with $\mathrm{rank}(M) \le 2r$.

- [Recht et al., 2010, Candès and Tao, 2010] As along as $\delta_{5r^*} \le 1/10$, the SDP relaxation was tight and M^* could be recovered exactly.

Introduction

Definition 1

The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is said to satisfy the δ-RIP$_{2r}$ property for some constant $\delta \in [0, 1)$ if the inequality

$$(1-\delta)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1+\delta)\|M\|_F^2$$

holds for all $M \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(M) \leq 2r$.

- [Recht et al., 2010, Candès and Tao, 2010] As along as $\delta_{5r^*} \leq 1/10$, the SDP relaxation was tight and M^* could be recovered exactly.
- [Bhojanapalli et al., 2016] For factorized problem (1), as long as $\delta_{2r} \leq 1/5$, all second-order critical points (SOPs) of (1) are ground truth solutions.

Introduction

Definition 1

The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is said to satisfy the δ-RIP$_{2r}$ property for some constant $\delta \in [0, 1)$ if the inequality

$$(1-\delta)\|M\|_F^2 \le \|\mathcal{A}(M)\|^2 \le (1+\delta)\|M\|_F^2$$

holds for all $M \in \mathbb{R}^{n\times n}$ with $\mathrm{rank}(M) \le 2r$.

- [Recht et al., 2010, Candès and Tao, 2010] As along as $\delta_{5r^*} \le 1/10$, the SDP relaxation was tight and M^* could be recovered exactly.
- [Bhojanapalli et al., 2016] For factorized problem (1), as long as $\delta_{2r} \le 1/5$, all second-order critical points (SOPs) of (1) are ground truth solutions.
- [Zhang et al., 2019] $\delta_{2r} = 1/2$ was a sharp bound when $r = r^*$, meaning that as long as $\delta_{2r} < 1/2$, all problem instances of (1) are free of spurious solutions, and once $\delta_{2r} \ge 1/2$, it is possible to establish counter-examples with SOPs not corresponding to ground truth solutions.

Introduction

Nice guarantees exist when $\delta < 1/2$, but things become more complicated if δ exceeds that range

Introduction

Nice guarantees exist when $\delta < 1/2$, but things become more complicated if δ exceeds that range

- **Benign landscape near** M^*. [Zhang et al., 2019] proved that when $\delta_{2r} \geq 1/2$ for $r = 1$, we can ensure the absence of spurious solutions in a local region that is close to M^*.

Introduction

Nice guarantees exist when $\delta < 1/2$, but things become more complicated if δ exceeds that range

- **Benign landscape near M^*.** [Zhang et al., 2019] proved that when $\delta_{2r} \geq 1/2$ for $r = 1$, we can ensure the absence of spurious solutions in a local region that is close to M^*.
- **Over-parametrization with $r \geq r^*$.** [Zhang, 2022] proved that if $r > r^*[(1+\delta_n)/(1-\delta_n) - 1]^2/4$, with $r^* \leq r < n$, then every SOP $\hat{X}$ satisfies that $\hat{X}\hat{X}^\top = M^*$.

Introduction

Nice guarantees exist when $\delta < 1/2$, but things become more complicated if δ exceeds that range

- **Benign landscape near** M^*. [Zhang et al., 2019] proved that when $\delta_{2r} \geq 1/2$ for $r = 1$, we can ensure the absence of spurious solutions in a local region that is close to M^*.
- **Over-parametrization with** $r \geq r^*$**.** [Zhang, 2022] proved that if $r > r^*[(1 + \delta_n)/(1 - \delta_n) - 1]^2/4$, with $r^* \leq r < n$, then every SOP $\hat{X}$ satisfies that $\hat{X}\hat{X}^\top = M^*$.
- **The SDP approach.** When using SDP, it was recently proven in [Yalcin et al., 2023] that as long as the RIP constant δ_{2r^*} is lower than the maximum of $1/2$ and $2r^*/(n + (n - 2r^*)(2l - 5))$, the global solution of the SDP relaxation corresponds to M^*.

Introduction

- In summary, various studies have been conducted to address the optimization landscape of (1) when the RIP constant is larger than $1/2$.

Introduction

- In summary, various studies have been conducted to address the optimization landscape of (1) when the RIP constant is larger than $1/2$.

- These method require either
  1. increase the complexity of the algorithm by a large margin (via over-parametrization $r \gg r^*$, SDP relaxation, or tensor optimization).
  2. initialize the algorithm close to M^*.

Introduction

- In summary, various studies have been conducted to address the optimization landscape of (1) when the RIP constant is larger than $1/2$.

- These method require either
  1. increase the complexity of the algorithm by a large margin (via over-parametrization $r \gg r^*$, SDP relaxation, or tensor optimization).
  2. initialize the algorithm close to M^*.

- *Does there exist meaningful global guarantees for* (1) *in the case of* $\delta \geq 1/2$ *without increasing the computational complexity of the problem drastically?* .

Contents

Disappearance of Spurious Solutions

- We study the landscape far away from M^* in the problematic case $\delta_{2r} \geq 1/2$.

Lemma 2

A point X is a first-order critical point of (1) if

$$\nabla f(X) = \left(\sum_{i=1}^{m} \langle A_i, XX^\top - M^* \rangle A_i \right) X = 0 \tag{2}$$

and it is a second-order critical point if it satisfies the above condition together with

$$\nabla^2 f(X)[U, U] = \sum_{i=1}^{m} \langle A_i, UX^\top + XU^\top \rangle^2 + \langle A_i, XX^\top - M^* \rangle \langle A_i, 2UU^\top \rangle \geq 0 \quad \forall U \in \mathbb{R}^n \tag{3}$$

Disappearance of Spurious Solutions

Theorem 3

Assume that (1) *satisfies the RIP*$_{r+r^*}$ *property with constant* $\delta \in [0, 1)$. *Given a first-order critical point* $\hat{X} \in \mathbb{R}^{n \times r}$ *of* (1), *if it satisfies the inequality*

$$\|\hat{X}\hat{X}^\top - M^*\|_F^2 > 2\frac{1+\delta}{1-\delta}\operatorname{tr}(M^*)\sigma_r(\hat{X})^2, \tag{4}$$

then $\hat{X}$ *is not a second-order critical point and is a strict saddle point with* $\nabla^2 f(\hat{X})$ *having a strictly negative eigenvalue not larger than*

$$2(1+\delta)\sigma_r(\hat{X})^2 - \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2(1-\delta)}{\operatorname{tr}(M^*)} \tag{5}$$

Disappearance of Spurious Solutions

Theorem 4 ([Zhang and Zhang, 2020])

Assume that (1) *satisfies the RIP property with constant* $\delta \in [0, 1)$. *Given an arbitrary constant* $\tau \in (0, 1 - \delta^2)$, *if a second-order critical point* $\hat{X} \in \mathbb{R}^{n \times r}$ *of* (1) *satisfies*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau \lambda_{r^*}(M^*) \tag{6}$$

then $\hat{X}$ *corresponds to the ground truth solution.*

Theorem 5

Consider the problem (1) *under the RIP*$_{r+r^*}$ *property with a constant* $\delta \in [0, 1)$. *Assume that its ground truth solution* M^* *satisfies the following inequality*

$$\|M^*\|_F \frac{\operatorname{tr}(M^*)}{\lambda_{r*}^2(M^*)} \leq \frac{\sqrt{r}}{2\sqrt{2}}(1+\delta)^{1/2}(1-\delta)^{7/2}, \tag{7}$$

Then, every second-order critical point $\hat{X}$ *of* (1) *satisfies*

$$\hat{X}\hat{X}^\top = M^*$$

Contents

Higher-order Loss Functions

- Although Theorem 3 proves that critical points far away from the ground truth are strict saddle points, the time needed to escape such points depends on the local curvature of the function [Ge et al., 2017, Jin et al., 2021].
- Therefore, it is essential to understand whether the curvatures at saddle points could be enhanced to reshape the landscape favorably.
- In this work, we address this problem by the use of a modified loss function

$$\min_{X \in \mathbb{R}^{n \times r}} f^l_\lambda(X) := f(X) + \lambda f^l(X) \tag{8}$$

where

$$f^l(X) := \frac{1}{l} \|\mathcal{A}(XX^\top) - b\|_l^l \tag{9a}$$

$$h^l(M) := \frac{1}{l} \|\mathcal{A}(M) - b\|_l^l \tag{9b}$$

Higher-order Loss Functions

Theorem 6

Assume that the operator $\mathcal{A}(\cdot)$ satisfies the RIP$_{r+r^*}$ property with constant $\delta \in [0,1)$. Consider the high-order optimization problem (8) *such that $l \geq 2$ is even. Given a first-order critical point $\hat{X} \in \mathbb{R}^{n\times r}$ of* (8)*, if*

$$D^2 \geq \mathrm{tr}(M^*)\sigma_r^2(\hat{X})\frac{(1+\delta)+\lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2+\lambda C(l)(1-\delta)^{l/2}D^{l-2}}, \tag{10}$$

then $\hat{X}$ is a strict saddle point with $\nabla^2 f(\hat{X})$ having a strictly negative eigenvalue not larger than

$$\begin{aligned}&\left[2(1+\delta)\sigma_r(\hat{X})^2 - \frac{D^2(1-\delta)}{\mathrm{tr}(M^*)}\right] + \\ \lambda D^{l-2}&\left[2(1+\delta)^{l/2}(l-1)\sigma_r(\hat{X})^2 - 2\frac{(1-\delta)^{l/2}C(l)D^2}{\mathrm{tr}(M^*)}\right]\end{aligned} \tag{11}$$

where $D := \|\hat{X}\hat{X}^\top - M^*\|_F$, $C(l) := m^{(2-l)/2}\left(\frac{2^l-1}{l}-1\right)$

Higher-order Loss Functions

- This can be compared to the lifted technique proposed in [Ma et al., 2023]. The presented method can amplify the negative curvature of those points X that satisfy

$$\|XX^\top - M^*\|_F^2 \geq \frac{1+\delta}{1-\delta}\,\mathrm{tr}(M^*)\sigma_r^2(\hat{X})$$

Higher-order Loss Functions

- This can be compared to the lifted technique proposed in [Ma et al., 2023]. The presented method can amplify the negative curvature of those points X that satisfy

$$\|XX^\top - M^*\|_F^2 \geq \frac{1+\delta}{1-\delta}\operatorname{tr}(M^*)\sigma_r^2(\hat{X})$$

- Where in comparison to (10) the multiplicative factor to $\operatorname{tr}(M^*)\sigma_r^2(\hat{X})$ becomes

$$\frac{(1+\delta)+\lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2+\lambda C(l)(1-\delta)^{l/2}D^{l-2}},$$

which is on the order of magnitude of

$$\mathcal{O}\left(l\left(\frac{\sqrt{m}}{2}\right)^l\left(\frac{1+\delta}{1-\delta}\right)^{l/2}\right),$$

Higher-order Loss Functions

- This can be compared to the lifted technique proposed in [Ma et al., 2023]. The presented method can amplify the negative curvature of those points X that satisfy

$$\|XX^\top - M^*\|_F^2 \geq \frac{1+\delta}{1-\delta}\,\mathrm{tr}(M^*)\sigma_r^2(\hat{X})$$

- Where in comparison to (10) the multiplicative factor to $\mathrm{tr}(M^*)\sigma_r^2(\hat{X})$ becomes

$$\frac{(1+\delta)+\lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2+\lambda C(l)(1-\delta)^{l/2}D^{l-2}},$$

which is on the order of magnitude of

$$\mathcal{O}\left(l\left(\frac{\sqrt{m}}{2}\right)^l\left(\frac{1+\delta}{1-\delta}\right)^{l/2}\right),$$

- This means that by utilizing a high-order loss, we can recover some of the desirable properties of an over-parametrized technique.

Contents

Simulation Results

n	λ	$\lambda_{\min}(\nabla^2 f^l(\hat{X}))$	$\lambda_{\max}(\nabla^2 f^l(\hat{X}))$	$\lambda_{\min}(\nabla^2 f^l(X^*))$	$\lambda_{\max}(\nabla^2 f^l(X^*))$
3	0	1.821	3.642	2.18	4.36
3	0.5	1.779	3.855	2.18	4.36
3	5	1.594	7.422	2.18	4.36
3	50	1.470	55.028	2.18	4.36
5	0	0.429	3.898	0.54	4.72
5	0.5	0.421	4.106	0.54	4.72
5	5	0.385	9.117	0.54	4.72
5	50	0.354	69.816	0.54	4.72
7	0	0.516	3.642	0.72	5.08
7	0.5	0.502	4.122	0.72	5.08
7	5	0.456	10.006	0.72	5.08
7	50	0.433	75.786	0.72	5.08

Simulation Results

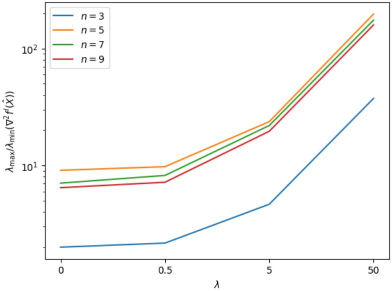
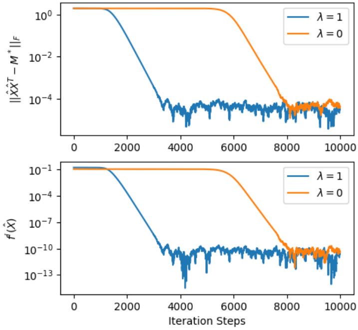
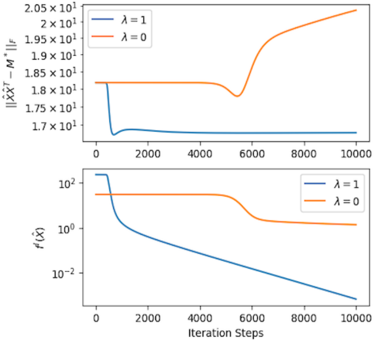


Figure: The ratio between the largest and smallest eigenvalue of Hessian at the spurious local minimum $\lambda_{\max}/\lambda_{\min}(\nabla^2 f^l(\hat{X}))$ with respect to λ under different size n.

Simulation Results



(a) $\lambda = 0$ converges to ground truth



(b) $\lambda = 0$ converges to a spurious solution around the ground truth

Figure: The evolution of the objective function and the error between the obtained solution $\widehat{X}\widehat{X}^T$ and the ground truth M^* during the iterations of the perturbed gradient descent method, with a constant step-size. In both cases, high-order loss functions accelerate the convergence.

Simulation Results

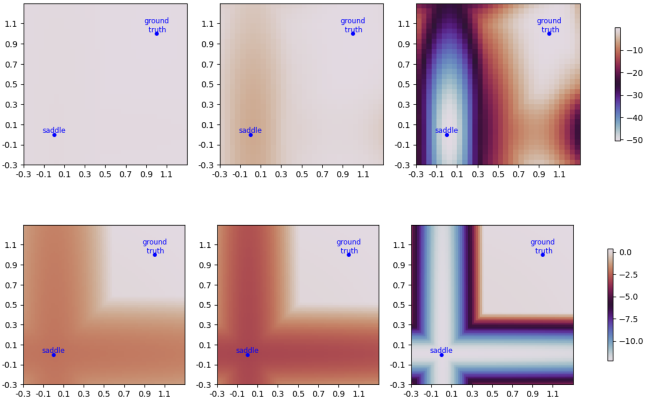


Figure: Different rows represents different problems. $\lambda = 0$ (left column), $\lambda = 0.5$ (middle column), $\lambda = 5$ (right column), with x-axis and y-axis as two orthogonal directions from the critical point to the ground truth.

References I

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016).
Global optimality of local search for low rank matrix recovery.
In *Advances in Neural Information Processing Systems*, volume 29.

Candès, E. J. and Tao, T. (2010).
The power of convex relaxation: Near-optimal matrix completion.
IEEE Transactions on Information Theory, 56(5):2053–2080.

Ge, R., Jin, C., and Zheng, Y. (2017).
No spurious local minima in nonconvex low rank problems: A unified geometric analysis.
In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1233–1242.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2021).
On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points.
Journal of the ACM (JACM), 68(2):1–29.

Ma, Z., Molybog, I., Lavaei, J., and Sojoudi, S. (2023).
Over-parametrization via lifting for low-rank matrix sensing: Conversion of spurious solutions to strict saddle points.
In *International Conference on Machine Learning*. PMLR.

Recht, B., Fazel, M., and Parrilo, P. A. (2010).
Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.
SIAM Review, 52(3):471–501.

References II

Yalcin, B., Ma, Z., Lavaei, J., and Sojoudi, S. (2023).
Semidefinite programming versus burer-monteiro factorization for matrix sensing.
In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, G. and Zhang, R. Y. (2020).
How many samples is a good initial point worth in low-rank matrix recovery?
In *Advances in Neural Information Processing Systems*, volume 33, pages 12583–12592.

Zhang, R. Y. (2022).
Improved global guarantees for the nonconvex burer–monteiro factorization via rank overparameterization.
arXiv preprint arXiv:2207.01789.

Zhang, R. Y., Sojoudi, S., and Lavaei, J. (2019).
Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery.
Journal of Machine Learning Research, 20(114):1–34.