

Global Optimality in Inductive Matrix Completion

Mohsen Ghassemi[†], Anand D. Sarwate[†], Naveen Goela[‡]

[†] Department of ECE, Rutgers University









[‡] Technicolor Research and Innovation Lab

April 17, 2018



Inductive Matrix Completion. What? Why?







Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix

						
	✓		✓			✓
	✓			✓		✓
	✓	✓			✓	✓
	✓		✓	✓	✓	✓
					✓	

Inductive Matrix Completion. What? Why?

Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix












- Recommender systems
- image inpainting

						
	✓		✓			✓
	✓			✓		✓
	✓	✓			✓	✓
	✓		✓	✓	✓	✓
					✓	

Inductive Matrix Completion. What? Why?

Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix

- Recommender systems
- image inpainting

						
	✓		✓			✓
	✓			✓		✓
	✓	✓			✓	✓
	✓		✓	✓	✓	✓
					✓	










Inductive Matrix Completion [Jain and Dhillon '13]

- In many applications, side information is available

Inductive Matrix Completion. What? Why?

Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix

- Recommender systems
- image inpainting

						
	✓		✓			✓
	✓			✓		✓
	✓	✓			✓	✓
	✓		✓	✓	✓	✓
					✓	

Inductive Matrix Completion [Jain and Dhillon '13]

- In many applications, side information is available
- Inductive matrix completion (IMC) incorporates side information in form of features of the row and column entities

Inductive Matrix Completion. What? Why?

Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix

- Recommender systems
- image inpainting

	Star Wars	Star Wars	Star Wars	Star Wars	Star Wars	Star Wars
User 1	✓		✓			✓
User 2	✓			✓		✓
User 3	✓	✓			✓	✓
User 4	✓		✓	✓	✓	✓
User 5						✓

Inductive Matrix Completion [Jain and Dhillon '13]

- In many applications, side information is available
- Inductive matrix completion (IMC) incorporates side information in form of features of the row and column entities
- Benefits:
 - Reduce sample complexity
 - Allow for inductive prediction on new users/items



Inductive Matrix Completion. What? Why?

Low-rank Matrix Completion:
given some entries, find a matching
low-rank matrix

- Recommender systems
- image inpainting

	✓		✓			✓
	✓			✓		✓
	✓	✓			✓	✓
	✓		✓	✓	✓	✓
					✓	

Inductive Matrix Completion [Jain and Dhillon '13]

- In many applications, side information is available
- Inductive matrix completion (IMC) incorporates side information in form of features of the row and column entities
- Benefits:
 - Reduce sample complexity
 - Allow for inductive prediction on new users/items

This talk: study the optimization landscape of the IMC Model



Low-Rank Matrix Completion

- Low-rank matrix completion problem

$$\min_{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{M}_{\Omega}^* - \mathbf{M}_{\Omega}\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{M}) \leq r.$$


└─ indices of m given entries



Low-Rank Matrix Completion

- Low-rank matrix completion problem

$$\min_{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}} \left\| \mathbf{M}_{\Omega}^* - \mathbf{M}_{\Omega} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{M}) \leq r.$$


 indices of m given entries

- This is too hard! Two tractable approaches:

Convex

$$\min_{\mathbf{M}} \left\| \mathbf{M}_{\Omega}^* - \mathbf{M}_{\Omega} \right\|_F^2 + \lambda \|\mathbf{M}\|_*$$

Nonconvex


$$\min_{\mathbf{U}, \mathbf{V}} \left\| \mathbf{M}_{\Omega}^* - [\mathbf{U}\mathbf{V}^T]_{\Omega} \right\|_F^2 + R(\mathbf{U}, \mathbf{V})$$



Low-Rank Matrix Completion

- Low-rank matrix completion problem

$$\min_{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}} \left\| \mathbf{M}_{\Omega}^* - \mathbf{M}_{\Omega} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{M}) \leq r.$$


 indices of m given entries

- This is too hard! Two tractable approaches:

Convex

$$\min_{\mathbf{M}} \left\| \mathbf{M}_{\Omega}^* - \mathbf{M}_{\Omega} \right\|_F^2 + \lambda \|\mathbf{M}\|_*$$

Nonconvex

$$\min_{\mathbf{U}, \mathbf{V}} \left\| \mathbf{M}_{\Omega}^* - [\mathbf{U}\mathbf{V}^T]_{\Omega} \right\|_F^2 + R(\mathbf{U}, \mathbf{V})$$

- Sample complexity $O(n \text{ polylog}(n) \text{ poly}(r))$, where n is the dimension of \mathbf{M} .



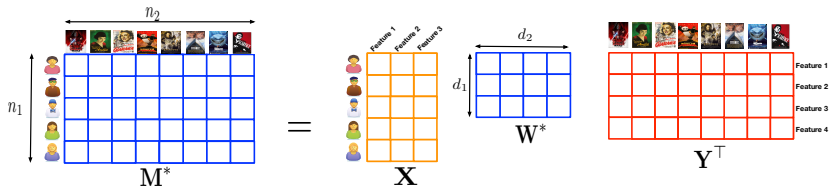
Side Information and IMC

- Examples of side information
 - Graph information (pairwise relationships)
 - Estimates of column/row spaces (e.g. in time varying applications)
- features of users and items



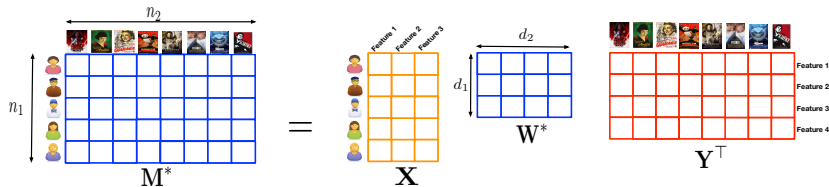
Side Information and IMC

- Examples of side information
 - Graph information (pairwise relationships)
 - Estimates of column/row spaces (e.g. in time varying applications)
 → features of users and items
- IMC models side information as knowledge of feature spaces



Side Information and IMC

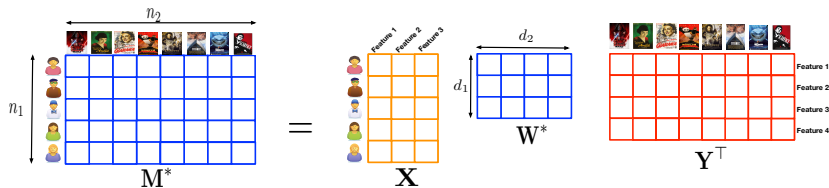
- Examples of side information
 - Graph information (pairwise relationships)
 - Estimates of column/row spaces (e.g. in time varying applications)
 → features of users and items
- IMC models side information as knowledge of feature spaces



$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \left\| \mathbf{M}_{\Omega}^* - [\mathbf{X}\mathbf{W}\mathbf{Y}^T]_{\Omega} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{W}) \leq r$$

Side Information and IMC

- Examples of side information
 - Graph information (pairwise relationships)
 - Estimates of column/row spaces (e.g. in time varying applications)
 → features of users and items
- IMC models side information as knowledge of feature spaces



$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \left\| \mathbf{M}_{\Omega}^* - [\mathbf{X}\mathbf{W}\mathbf{Y}^T]_{\Omega} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{W}) \leq r$$

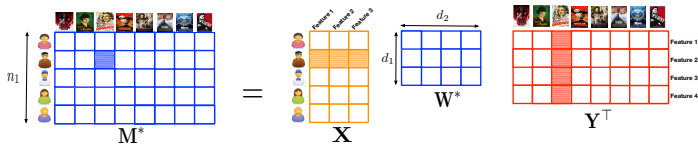
Completing $\mathbf{M}_{\Omega}^* \in \mathbb{R}^{n_1 \times n_2}$

\equiv

Recovering $\mathbf{W}^* \in \mathbb{R}^{d_1 \times d_2}$

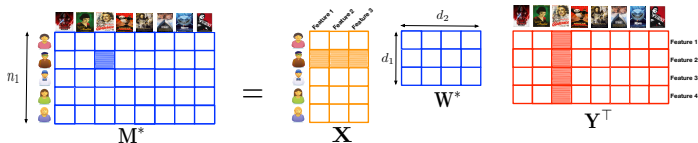


IMC as A Matrix Sensing Problem



$$M_{ij}^* = \mathbf{x}_j^T \mathbf{W}^* \mathbf{y}_j = \langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W}^* \rangle$$

IMC as A Matrix Sensing Problem

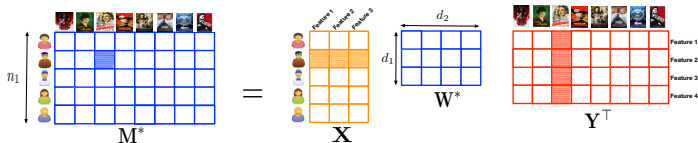


$$M_{ij}^* = \mathbf{x}_j^T \mathbf{W}^* \mathbf{y}_j = \langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W}^* \rangle$$

- Rewrite the IMC objective function

$$\|M_{\Omega}^* - [XWY^T]_{\Omega}\|_F^2 = \sum_{(i,j) \in \Omega} |M_{ij}^* - \langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W} \rangle|^2 = \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2$$

IMC as A Matrix Sensing Problem

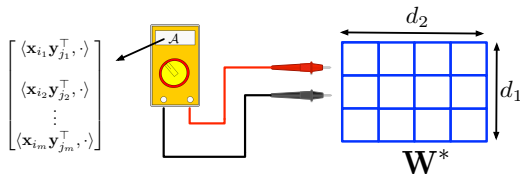


$$\mathbf{M}_{ij}^* = \mathbf{x}_j^T \mathbf{W}^* \mathbf{y}_j = \langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W}^* \rangle$$

- Rewrite the IMC objective function

$$\|\mathbf{M}_{\Omega}^* - [\mathbf{X}\mathbf{W}\mathbf{Y}^T]_{\Omega}\|_F^2 = \sum_{(i,j) \in \Omega} |\mathbf{M}_{ij}^* - \langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W} \rangle|^2 = \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2$$

- Matrix sensing problem operator \mathcal{A} :
randomly select
(row,col) of (\mathbf{X}, \mathbf{Y})



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$

Convex

Nonconvex



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$

Convex

- Sample complexity (exact):
 $O(rd \log d \log n)$

Nonconvex

- Sample complexity (inexact):
 $O(r^3 d \log d \max\{r, \log n\} \log(\frac{1}{\epsilon}))$



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$

Convex

- Sample complexity (exact):
 $O(rd \log d \log n)$
- ✗ Not scalable (SVD, SDP, ...)

Nonconvex

- Sample complexity (inexact):
 $O(r^3 d \log d \max\{r, \log n\} \log(\frac{1}{\epsilon}))$
- ✓ Computationally efficient



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$

Convex

- Sample complexity (exact):
 $O(rd \log d \log n)$
- ✗ Not scalable (SVD, SDP, ...)
- ✓ Theoretically understood
(convex program)

Nonconvex

- Sample complexity (inexact):
 $O(r^3 d \log d \max\{r, \log n\} \log(\frac{1}{\epsilon}))$
- ✓ Computationally efficient



Convex vs Nonconvex Formulation

Convex Formulation:

$$\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathcal{A}(\mathbf{W}^*) - \mathcal{A}(\mathbf{W})\|_2^2 + \lambda \|\mathbf{W}\|_*$$

Nonconvex Formulation:

$$\min_{\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}} \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T) - \mathcal{A}(\mathbf{UV}^T)\|_2^2 + R(\mathbf{U}, \mathbf{V})$$

Convex

- Sample complexity (exact):
 $O(rd \log d \log n)$
- ✗ Not scalable (SVD, SDP, ...)
- ✓ Theoretically understood
(convex program)

Nonconvex

- Sample complexity (inexact):
 $O(r^3 d \log d \max\{r, \log n\} \log(\frac{1}{\epsilon}))$
 - ✓ Computationally efficient
 - Not as well understood
- Our Goal: understand better**



Why Care about Nonconvex Optimization Theory?

- Success of local search algorithm (e.g. AM, SGD, ...) in matrix factorization, deep learning, etc.



Why Care about Nonconvex Optimization Theory?

- Success of local search algorithm (e.g. AM, SGD, ...) in matrix factorization, deep learning, etc.



Why Care about Nonconvex Optimization Theory?

- Success of local search algorithm (e.g. AM, SGD, ...) in matrix factorization, deep learning, etc.
- Better understanding of behavior around stationary points
 - SGD (and other stochastic variants) escape strict saddle points [Ge et al. '15, Jin et al. '17]
 - SGD escapes sharp local minima [Kleinberg et al., 2018]



Why Care about Nonconvex Optimization Theory?

- Success of local search algorithm (e.g. AM, SGD, ...) in matrix factorization, deep learning, etc.
- Better understanding of behavior around stationary points
 - SGD (and other stochastic variants) escape strict saddle points [Ge et al. '15, Jin et al. '17]
 - SGD escapes sharp local minima [Kleinberg et al., 2018]
- Better understanding of *optimization landscape*



Why Care about Nonconvex Optimization Theory?

- Success of local search algorithm (e.g. AM, SGD, ...) in matrix factorization, deep learning, etc.
- Better understanding of behavior around stationary points
 - SGD (and other stochastic variants) escape strict saddle points [Ge et al. '15, Jin et al. '17]
 - SGD escapes sharp local minima [Kleinberg et al., 2018]
- Better understanding of *optimization landscape*

We study the optimization landscape of the IMC problem



Geometric Properties of Nonconvex IMC Objective

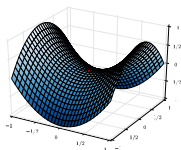
“Nice” properties of the IMC objective function make recovery using local algorithms possible:



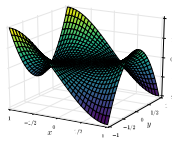
Geometric Properties of Nonconvex IMC Objective

“Nice” properties of the IMC objective function make recovery using local algorithms possible:

- Escapable saddles: there is a *descent direction* at saddle points



Strict Saddle Point*



Non-Strict Saddle*

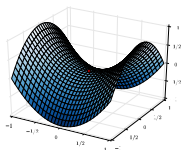
*Credit: Wikimedia Commons



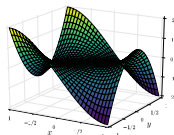
Geometric Properties of Nonconvex IMC Objective

“Nice” properties of the IMC objective function make recovery using local algorithms possible:

- Escapable saddles: there is a *descent direction* at saddle points



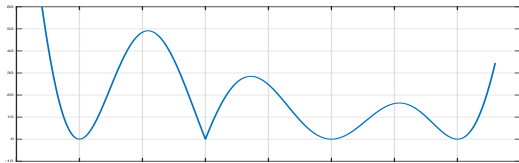
Strict Saddle Point*



Non-Strict Saddle*

*Credit: Wikimedia Commons

- No poor local minimum: all its local minima are globally optimum



Proof Strategy

We employ the framework by Ge et al. [2017] for nonconvex matrix recovery



Proof Strategy

We employ the framework by Ge et al. [2017] for nonconvex matrix recovery

- First we describe the strategy for symmetric matrices

$$\mathbf{M}^* = \mathbf{U}^*[\mathbf{U}^*]^T$$



Proof Strategy

We employ the framework by Ge et al. [2017] for nonconvex matrix recovery

- First we describe the strategy for symmetric matrices

$$\mathbf{M}^* = \mathbf{U}^*[\mathbf{U}^*]^T$$

- Show that around stationary points, the “difference” Δ between the current point and its *nearest true solution* (invariant to rotation) is a **descent direction**



Proof Strategy

We employ the framework by Ge et al. [2017] for nonconvex matrix recovery

- First we describe the strategy for symmetric matrices

$$\mathbf{M}^* = \mathbf{U}^*[\mathbf{U}^*]^T$$

- Show that around stationary points, the “difference” Δ between the current point and its *nearest true solution* (invariant to rotation) is a **descent direction**

Only direction: toward the minimum



Local minima = true solutions

There is a descent direction



Saddle points have to be strict



Proof Strategy

We employ the framework by Ge et al. [2017] for nonconvex matrix recovery

- First we describe the strategy for symmetric matrices

$$\mathbf{M}^* = \mathbf{U}^*[\mathbf{U}^*]^T$$

- Show that around stationary points, the “difference” Δ between the current point and its *nearest true solution* (invariant to rotation) is a **descent direction**

Only direction: toward the minimum



Local minima = true solutions

There is a descent direction



Saddle points have to be strict

- Later: the strategy for general case $\mathbf{M}^* = \mathbf{U}^*[\mathbf{V}^*]^T$



A Descent Direction Around Stationary Points

- $\nabla f(\mathbf{U}) \approx 0$ around stationary points, so $\delta^T \nabla^2 f(\mathbf{U}) \delta$ becomes dominant in Taylor series expansion



A Descent Direction Around Stationary Points

- $\nabla f(\mathbf{U}) \approx 0$ around stationary points, so $\delta^T \nabla^2 f(\mathbf{U}) \delta$ becomes dominant in Taylor series expansion
- $\Delta = \mathbf{U} - \mathbf{U}^*$ is a descent direction around a stationary point iff for $\mathbf{d} = \text{vec}(\Delta)$, we have $\mathbf{d}^T \nabla^2 f(\mathbf{U}) \mathbf{d} < 0$



A Descent Direction Around Stationary Points

- $\nabla f(\mathbf{U}) \approx 0$ around stationary points, so $\delta^T \nabla^2 f(\mathbf{U}) \delta$ becomes dominant in Taylor series expansion
- $\Delta = \mathbf{U} - \mathbf{U}^*$ is a descent direction around a stationary point iff for $\mathbf{d} = \text{vec}(\Delta)$, we have $\mathbf{d}^T \nabla^2 f(\mathbf{U}) \mathbf{d} < 0$
- We use RIP property of operator \mathcal{A} in the objective function

$$f(\mathbf{U}) = \|\underbrace{\mathcal{A}(\mathbf{U}^* [\mathbf{U}^*]^T - \mathbf{U} \mathbf{U}^T)}_{\text{At most rank-}2r}\|^2$$

\mathcal{A} is $(2r, \delta_{2r})$ -RIP

$\delta = \text{vec}(\Delta)$

$\delta^T \nabla^2 f(\mathbf{B}) \delta < 0$
around stationary points, unless
 $\Delta = \mathbf{B} - \mathbf{B}^* = 0$ (= recovery)



Operator \mathcal{A} satisfies RIP

\mathcal{A} is $(2r, \delta_{2r})$ -RIP

$\delta = \text{vec}(\Delta)$

$\delta^T \nabla^2 f(\mathbf{B}) \delta < 0$
around stationary points, unless
 $\Delta = \mathbf{B} - \mathbf{B}^* = 0$ (= recovery)



Operator \mathcal{A} satisfies RIP

\mathcal{A} is $(2r, \delta_{2r})$ -RIP

$\delta = \text{vec}(\Delta)$

$\delta^T \nabla^2 f(\mathbf{B}) \delta < 0$
around stationary points, unless
 $\Delta = \mathbf{B} - \mathbf{B}^* = 0$ (= recovery)

Theorem (Operator \mathcal{A} is $(2r, \delta_{2r})$ -RIP)

If $m = O(\mu^2 dr \max\{r^2, \log^2 n\} \log(36\sqrt{2}/\delta)/\delta^2)$, then there exists $h > 0$ such that with probability at least $1 - 2e^{-hm}$, the linear operator \mathcal{A} is $(2r, 2\delta)$ -RIP.



Operator \mathcal{A} satisfies RIP

\mathcal{A} is $(2r, \delta_{2r})$ -RIP

$\delta = \text{vec}(\Delta)$

$\delta^T \nabla^2 f(\mathbf{B}) \delta < 0$
around stationary points, unless
 $\Delta = \mathbf{B} - \mathbf{B}^* = 0$ (= recovery)

Theorem (Operator \mathcal{A} is $(2r, \delta_{2r})$ -RIP)

If $m = O(\mu^2 dr \max\{r^2, \log^2 n\} \log(36\sqrt{2}/\delta)/\delta^2)$, then there exists $h > 0$ such that with probability at least $1 - 2e^{-hm}$, the linear operator \mathcal{A} is $(2r, 2\delta)$ -RIP.

Proof steps:

- 1 **given** a rank- $2r$ matrix, $\|\mathcal{A}(\mathbf{W})\|_2^2 \approx \|\mathbf{W}\|_F^2$ w.h.p.
- 2 **for all** rank- $2r$ matrices, $\|\mathcal{A}(\mathbf{W})\|_2^2 \approx \|\mathbf{W}\|_F^2$ w.h.p.



Operator \mathcal{A} satisfies RIP: Step 1

Lemma

If the number of measurements $m = O(4\mu^2\bar{r}^2 \log(2/\rho))$ for a constant $\rho > 0$, then for a **given** matrix \mathbf{W} of rank $2r$, with probability at least $1 - \rho$, for some positive constants C and c , we have

$$(1 - \delta_{2r}) \|\mathbf{W}\|_F^2 \leq \|\mathcal{A}(\mathbf{W})\|_2^2 \leq (1 + \delta_{2r}) \|\mathbf{W}\|_F^2$$



Operator \mathcal{A} satisfies RIP: Step 1

Lemma

If the number of measurements $m = O(4\mu^2\bar{r}^2 \log(2/\rho))$ for a constant $\rho > 0$, then for a *given* matrix \mathbf{W} of rank $2r$, with probability at least $1 - \rho$, for some positive constants C and c , we have

$$(1 - \delta_{2r}) \|\mathbf{W}\|_F^2 \leq \|\mathcal{A}(\mathbf{W})\|_2^2 \leq (1 + \delta_{2r}) \|\mathbf{W}\|_F^2$$

Proof Idea:

- Employ *Bernstein Inequality* to show concentration of $\|\mathcal{A}(\mathbf{W})\|_2^2$ around its mean $\|\mathbf{W}\|_F^2$



Operator \mathcal{A} satisfies RIP: Step 2

We want to show concentration of $\|\mathcal{A}(\mathbf{W})\|_2^2$ around $\|\mathbf{W}\|_F^2$ for all $\mathbf{W} \in \mathbb{R}^{d \times d} : \text{rank}(\mathbf{W}) \leq 2r$.



Operator \mathcal{A} satisfies RIP: Step 2

We want to show concentration of $\|\mathcal{A}(\mathbf{W})\|_2^2$ around $\|\mathbf{W}\|_F^2$ for all $\mathbf{W} \in \mathbb{R}^{d \times d} : \text{rank}(\mathbf{W}) \leq 2r$.

- Define $\mathbb{S}_{2r}^d = \{\bar{\mathbf{W}} \in \mathbb{R}^{d \times d} : \text{rank}(\bar{\mathbf{W}}) \leq 2r, \|\bar{\mathbf{W}}\|_F = 1\}$ and its ϵ -net $\bar{\mathbb{S}}_{2r}^d$



Operator \mathcal{A} satisfies RIP: Step 2

We want to show concentration of $\|\mathcal{A}(\mathbf{W})\|_2^2$ around $\|\mathbf{W}\|_F^2$ for all $\mathbf{W} \in \mathbb{R}^{d \times d}$: $\text{rank}(\mathbf{W}) \leq 2r$.

- Define $\mathbb{S}_{2r}^d = \{\bar{\mathbf{W}} \in \mathbb{R}^{d \times d} : \text{rank}(\bar{\mathbf{W}}) \leq 2r, \|\bar{\mathbf{W}}\|_F = 1\}$ and its ϵ -net $\bar{\mathbb{S}}_{2r}^d$
- Step 1: for a given $\bar{\mathbf{W}} \in \bar{\mathbb{S}}_{2r}^d$

$$\mathbb{P}\left(\left|\|\mathcal{A}(\bar{\mathbf{W}})\|_2^2 - 1\right| > \delta_{2r}\right) \leq \rho$$



Operator \mathcal{A} satisfies RIP: Step 2

We want to show concentration of $\|\mathcal{A}(\mathbf{W})\|_2^2$ around $\|\mathbf{W}\|_F^2$ for **all** $\mathbf{W} \in \mathbb{R}^{d \times d}$: $\text{rank}(\mathbf{W}) \leq 2r$.

- Define $\mathbb{S}_{2r}^d = \{\bar{\mathbf{W}} \in \mathbb{R}^{d \times d} : \text{rank}(\bar{\mathbf{W}}) \leq 2r, \|\bar{\mathbf{W}}\|_F = 1\}$ and its ϵ -net $\bar{\mathbb{S}}_{2r}^d$
- Step 1: for a given $\bar{\mathbf{W}} \in \bar{\mathbb{S}}_{2r}^d$

$$\mathbb{P}\left(\left|\|\mathcal{A}(\bar{\mathbf{W}})\|_2^2 - 1\right| > \delta_{2r}\right) \leq \rho$$

- Union bound:

$$\mathbb{P}\left(\max_{\bar{\mathbf{W}} \in \bar{\mathbb{S}}_{2r}^d} \left|\|\mathcal{A}(\bar{\mathbf{W}})\|_2^2 - 1\right| > \delta_{2r}\right) \leq |\bar{\mathbb{S}}_{2r}^d| \rho$$

- Some algebra and setting $\bar{\mathbf{W}} = \frac{\mathbf{W}}{\|\mathbf{W}\|_F}$ conclude the proof.



Proof Strategy for Asymmetric Matrices

- Symmetric matrices:

\mathcal{A} is $(2r, \delta_{2r})$ -RIP



Δ is a descent direction
around stationary points of $f(\mathbf{B})$

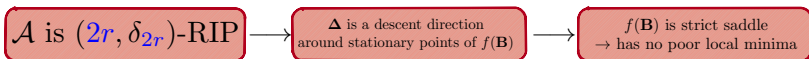


$f(\mathbf{B})$ is strict saddle
→ has no poor local minima

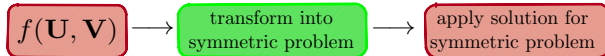


Proof Strategy for Asymmetric Matrices

- Symmetric matrices:



- What about general asymmetric case?



Reformulation into a Symmetric Problem

- The asymmetric objective function

$$f(\mathbf{U}, \mathbf{V}) = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{U}\mathbf{V}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$



Reformulation into a Symmetric Problem

- The asymmetric objective function

$$f(\mathbf{U}, \mathbf{V}) = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{UV}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$

- Construct the symmetric matrix \mathbf{N}

$$\mathbf{M} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V}^T \end{bmatrix} \longrightarrow \mathbf{N} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \\ \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{U}^T & \mathbf{V}^T & \mathbf{B}^T \end{bmatrix}$$



Reformulation into a Symmetric Problem

- The asymmetric objective function

$$f(\mathbf{U}, \mathbf{V}) = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{UV}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$

- Construct the symmetric matrix \mathbf{N}

$$\mathbf{M} = \begin{bmatrix} \mathbf{U} & \mathbf{V}^T \end{bmatrix} \longrightarrow \mathbf{N} = \begin{bmatrix} \mathbf{U} & \mathbf{U}^T & \mathbf{V}^T \\ \mathbf{V} & & \mathbf{B}^T \end{bmatrix}$$

\mathbf{B}

- Instead of $\mathbf{M} = \mathbf{UV}^T$, we work with $\mathbf{N} = \begin{bmatrix} \mathbf{UU}^T & \mathbf{UV}^T \\ \mathbf{VU}^T & \mathbf{VV}^T \end{bmatrix}$



Reformulation into a Symmetric Problem

- One can define linear operator \mathcal{T} such that

$$\|\mathcal{T}(\mathbf{B}^*[\mathbf{B}^*]^T - \mathbf{B}\mathbf{B}^T)\|^2 = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{U}\mathbf{V}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$



Reformulation into a Symmetric Problem

- One can define linear operator \mathcal{T} such that

$$\|\mathcal{T}(\mathbf{B}^*[\mathbf{B}^*]^T - \mathbf{B}\mathbf{B}^T)\|^2 = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{U}\mathbf{V}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$

- Reformulated Symmetric problem

$$\min_{\mathbf{B}} \|\mathcal{T}(\mathbf{B}^*[\mathbf{B}^*]^T - \mathbf{B}\mathbf{B}^T)\|^2$$



Reformulation into a Symmetric Problem

- One can define linear operator \mathcal{T} such that

$$\|\mathcal{T}(\mathbf{B}^*[\mathbf{B}^*]^T - \mathbf{B}\mathbf{B}^T)\|^2 = \|\mathcal{A}(\mathbf{U}^*[\mathbf{V}^*]^T - \mathbf{U}\mathbf{V}^T)\|^2 + R(\mathbf{U}, \mathbf{V})$$

- Reformulated Symmetric problem

$$\min_{\mathbf{B}} \|\mathcal{T}(\mathbf{B}^*[\mathbf{B}^*]^T - \mathbf{B}\mathbf{B}^T)\|^2$$

- Now we can use the proof strategy of the symmetric problem

\mathcal{T} is $(2r, \delta_{2r})$ -RIP

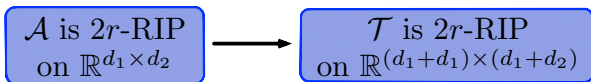
$$\delta = \text{vec}(\Delta)$$

$\delta^T \nabla^2 f(\mathbf{B}) \delta < 0$
around stationary points, unless
 $\Delta = \mathbf{B} - \mathbf{B}^* = 0$ (= recovery)



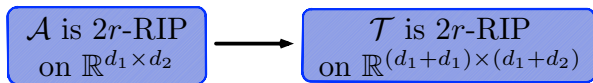
Operator \mathcal{T} Satisfies RIP

- If we use the common regularizer $R(\mathbf{U}, \mathbf{V}) = \frac{1}{4} \|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2$



Operator \mathcal{T} Satisfies RIP

- If we use the common regularizer $R(\mathbf{U}, \mathbf{V}) = \frac{1}{4} \|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2$

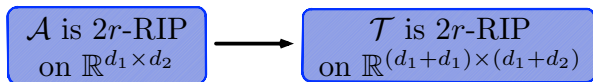


- Showing \mathcal{A} is $2r$ -RIP is similar to the symmetric case

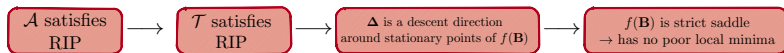


Operator \mathcal{T} Satisfies RIP

- If we use the common regularizer $R(\mathbf{U}, \mathbf{V}) = \frac{1}{4} \|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2$



- Showing \mathcal{A} is $2r$ -RIP is similar to the symmetric case
- Therefore, in general case it is also sufficient to show \mathcal{A} satisfies RIP



Conclusion

We studied the optimization landscape of the IMC problem. Given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the *factored IMC objective function*,

- All saddle point are escapable
- There is no poor local minimum
- Global optimization results in exact recovery



Conclusion

We studied the optimization landscape of the IMC problem. Given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the *factored IMC objective function*,

- All saddle point are escapable
- There is no poor local minimum
- Global optimization results in exact recovery

Implication: local search algorithms can escape saddle points.



Conclusion

We studied the optimization landscape of the IMC problem. Given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the *factored IMC objective function*,

- All saddle point are escapable
- There is no poor local minimum
- Global optimization results in exact recovery

Implication: local search algorithms can escape saddle points.

→ **SGD will efficiently solve the IMC problem**



Conclusion

We studied the optimization landscape of the IMC problem. Given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the *factored IMC objective function*,

- All saddle point are escapable
- There is no poor local minimum
- Global optimization results in exact recovery

Implication: local search algorithms can escape saddle points.

→ **SGD will efficiently solve the IMC problem**

Next steps:

- Experiments to understand non-asymptotic behavior.
- Extension to other side-information models.

