Guaranteed Matrix Completion via Non-convex Factorization

Zhi-Quan (Tom) Luo

The Chinese University of Hong Kong, Shenzhen, China
and
The University of Minnesota, Minneapolis, MN, USA

joint work with Ruoyu Sun

July 5, 2015
Outline

1. Introduction and Result
Outline

1. Introduction and Result

2. Technical Details
   - Formulation, Algorithm and Initial Point
   - Proof Ideas
   - Proof Sketch of Step 1
Netflix Challenge

- **Netflix prize** (2006): 500,000 customers, 17,000 movies, 100 million ratings (1.2% of all ratings).

- **Challenge**: predict missing ratings ($1,000,000 prize).

- One approach: exploring **low-rank structure**

- **Low-rank matrix completion**: complete a low-rank matrix from a few entries [Candes-Recht-09]
Idea of Matrix Completion (MC)

- **Low-rank**: $M \in \mathbb{R}^{n \times n}$, with rank $r \ll n$.
- **Partial observation**: $\Omega \subset [n] \times [n]$: set of sampled positions.
- If $M = E_{11}$, little hope to recover $M$.

$$E_{11} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

- Assumptions on $M$, $\Omega$ are needed.
Matrix Completion

- Result [Candes-Recht’09]: recover $M \in \mathbb{R}^{n \times n}$ from $O(nr \log n)$ entries via convex optimization, if
  - $M$ is “generic” (incoherent).
  - Set of observations $\Omega$ are “generic” (random).

$$M = \begin{bmatrix}
  x_1y_1 & x_1y_2 & x_1y_3 & \cdots & x_1y_n \\
  x_2y_1 & x_2y_2 & x_2y_3 & \cdots & x_2y_n \\
  \vdots & \vdots & \vdots & \cdots & \vdots \\
  x_ny_1 & x_ny_2 & x_ny_3 & \cdots & x_ny_n
\end{bmatrix}$$

- $O(nr \log n)$ v.s. $n^2$ ambient dim.
  - $O(nr)$ is necessary: $nr$ free variables
  - $\log n$ due to coupon collecting effect: need $n \log n$ samples to cover each row/column.
Why is Matrix Completion Interesting?

1) Natural extension of sparsity to matrix domain.
   - Related low-rank pursuing approaches:
     - Phase retrieval; Robust PCA; Tensor completion

2) Practice: lots of applications.
   - Recommendation systems, computer vision, network anomaly detection, etc.

3) Theory: A fundamental mathematical problem.
   - Related to: statistics, learning, optimization, linear algebra, information theory, etc.
Matrix Factorization v.s. Nuclear Norm

- **Two popular methods**: nuclear norm, matrix factorization.
  - **Method 1**: Nuclear norm minimization [Candes-Recht-09]
    \[
    \min_{Z \in \mathbb{R}^{n \times n}} \frac{1}{2} \| \mathcal{P}_\Omega(M - Z) \|_F^2 + \lambda \| Z \|_*.
    \]  
    where
    \[
    \mathcal{P}_\Omega(X) = \begin{cases} 
    X_{ij}, & \text{if } (i,j) \in \Omega \\
    0, & \text{if } (i,j) \notin \Omega
    \end{cases}
    \]
    - \(n^2\) variables; nonsmooth but convex
  - **Method 2**: matrix factorization (MF) based formulation [Koren09]:
    \[
    P_0 : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} \frac{1}{2} \| \mathcal{P}_\Omega(M - XY^T) \|_F^2 + \lambda (\| X \|_F^2 + \| Y \|_F^2). 
    \]  
    - \(nr\) variables; smooth but nonconvex
How can this be possible?

The matrix factorization (MF) formulation [Koren09]:

$$ P_0 : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} \frac{1}{2} \| \mathcal{P}_\Omega (M - XY^T) \|_F^2 $$

Even if we can solve \((P_0)\), how can we be sure \(M = X^* (Y^*)^T\)?

$$ \| \mathcal{P}_\Omega (M - XY^T) \|_F^2 = 0 \implies \| (M - XY^T) \|_F^2 = 0. $$

Observation: Let \(p = |\Omega|/n^2\). If \(\Omega\) and \(M, X, Y\) are independent, then

$$ E \left[ \| \mathcal{P}_\Omega (M - XY^T) \|_F^2 \right] = p \| (M - XY^T) \|_F^2. $$

Difficulty: the iterates \((X, Y)\) cannot be independent of \(\Omega\)!

- **Resampling** = use different samples at each iteration and discard.
  - Not practical: waste of resources; accuracy pre-determined
  - No exact recovery: infinite samples for exact recovery!
Nuclear Norm Formulation

- Nuc-norm formulation ($\in$ SDP) is convex (global convergence):
  - Standard SDP solvers (interior point method)
  - **Proximal gradient** method and variants [Toh-Yun10], [Ma-Goldfarb-Chen11]
    - Linear convergence under certain conditions [Agarwal-Negahban,Wainwright12], [Hou-Zhou-So-Luo13]
  
- Pros: convex; **guaranteed recovery** [Candes-Recht-09].

- Cons: slow for big data (requires SVD per-iteration); large memory requirement
Matrix Factorization Formulation

- Algorithms for **non-convex** MF model (converge to stationary points):
  - **Alt-Min** [Koren-09],[Wen-Yin-Zhang-12],
  - **SGD** (Stochastic Gradient Descent) [Koren-09], [Funk-06]
  - Other Alt-Min methods: multi-block [Yu-Hsieh-Si-Dhillon-12], block majorization [Hastie-Mazumder-Lee-Zadeh-14]

- **Pros:**
  - Fast in practice, little storage
  - **Flexibility:** can incorporate data aspects

- **Cons:** limited performance analysis (more later)

- **Our goal:** bridge the gap between theory and practice
Formulation

- Start from a **constrained version** (extra requirements on factors $X, Y$):

\[
P'_1 : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} \quad F(X, Y) \triangleq \frac{1}{2} \| \mathcal{P}_\Omega (M - XY^T) \|_F^2.
\]

\[
(X, Y) \in K_1 := \{ \|X\|_F \leq \beta_T, \quad \|Y\|_F \leq \beta_T, \}
\]

\[
(X, Y) \in K_2 := \{ \|X^{(i)}\| \leq \beta_1, \quad \|Y^{(i)}\| \leq \beta_1, \quad \forall \ i. \}
\]

- $K_1$: boundedness
- $K_2$: incoherence
Consider a penalized version of the above problem:

\[
P_1 : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} \tilde{F}(X, Y) \triangleq \frac{1}{2} \| \mathcal{P}_\Omega (M - XY^T) \|_F^2 + G(X, Y),
\]

where

\[
G(X, Y) \triangleq \rho G_1 \left( \frac{3\|X\|_F^2}{2\beta_T^2} \right) + \rho G_1 \left( \frac{3\|Y\|_F^2}{2\beta_T^2} \right) + \rho \sum_{i=1}^n G_1 \left( \frac{3\|X^{(i)}\|_2^2}{2\beta_i^2} \right) + \rho \sum_{j=1}^n G_1 \left( \frac{3\|Y^{(j)}\|_2^2}{2\beta_i^2} \right),
\]

and \( G_1(z) = \max(z - 1, 0)^2 \), and \( \rho \) is a large enough constant.
Our Contributions

Theorem [Sun-L.’14]

Suppose $M \in \mathbb{R}^{n \times n}$ has rank $r$ and

- is incoherent
- has a condition number $\kappa = \sigma_1 / \sigma_r$.

For i.i.d random observation set $\Omega \subseteq [n] \times [n]$ with size

$$|\Omega| \geq Cn \log n \cdot \text{poly}(r, \kappa, \mu),$$

with specific initialization, many standard algorithms for $(P_1)$ converge to global optima, AND recovering $M$ w.h.p.

Remark: initialization and formulation will be specified later.

Standard algorithms include GD, SGD, Alt-Min.
Proof Idea (1)

- Why global convergence possible for non-convex problems?
  Basin of attraction + good initial point.

(I) **Problem property**: basin of attraction (**hard**)

- a convex neighborhood in the space of $(X, Y)$
- nonconvex in the space of $M$
- every stationary point $(X^*, Y^*)$ of $(P_1)$ in the basin satisfies $X^*(Y^*)^T = M$
Proof of main result: (II) shows Algorithm 1-4 converge to a stationary point in the basin, which by (I) equals $M$ (global optimum).
Applicable Algorithms

- Define $x_t = (X_t, Y_t)$, $\Delta_t \triangleq x_{t+1} - x_t$.
- Our result applies to any algorithm with **two properties** (besides initialization):
  - (II.a) converges to stationary point;
  - (II.b) satisfies one of three mild conditions (in order to **keep the iterates in the “basin”**),

1. $\tilde{F}(x_t + \lambda \Delta_t) \leq 2\tilde{F}(x_0), \forall \lambda \in [0, 1] \ \forall t$;
2. $1 = \arg\min_{\lambda \in \mathbb{R}} \psi(x_t, \Delta_t; \lambda)$, where $\psi$ is a “convex upper bound”, $\forall t$;
3. $\tilde{F}(x_t) \leq 2\tilde{F}(x_0), \quad d(x_t, x_0) \leq \frac{5}{6}\delta, \forall t$.

where $\delta = O(\sigma_r)$. 
Applicable Algorithms

Three typical classes of applicable algorithms:

- **GD** with constant step-size and **SGD** satisfies 1);
- **Block coordinate descent** or more generally, BSUM, satisfies 2);
- Any descent algorithm with “bounded” update ($d(x_t, x_0) \leq \frac{5}{6} \delta$) at each iteration satisfies 3).
Related Works in Matrix Completion

- Result for MF model in Grassmann manifold [Keshavan-Montanari-Oh’09]

- Result for Alt-Min (variants)
  [Keshavan11],[Jain-Netrapalli-Sanhavi12],[Hardt13]

Table: Comparison with Recent Studies on Alt-Min

<table>
<thead>
<tr>
<th>Applicability</th>
<th>Studies on Alt-Min</th>
<th>Our work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>one algorithm (Alt-Min)</td>
<td>many algorithms</td>
</tr>
<tr>
<td>Form</td>
<td>require resampling</td>
<td>standard form</td>
</tr>
<tr>
<td>Technique</td>
<td>analysis of power method</td>
<td>random graph + perturbation</td>
</tr>
</tbody>
</table>
**Related Works for Other Problems**

- **Phase Retrieval (PR)**
  - Nuclear norm formulation (convex) [Candes-Eldar-Strohmer-Voroninski-11]
  - Non-convex **resampling** based Alt-Min [Netrapalli-Jain-Sanhavi-13]
  - Non-convex **gradient descent** (no resampling) [Candes-Li-Soltanolkotabi-14]

- **Non-convex sparse regression**
  [Zhang-Zhang-12], [Loh-Wainwright-13], [Fan-Xue-Zou-14]

- **EM algorithm** [Balakrishnan-Wainwright-14]

- **Power method** for computing eigenvectors
  - **Sparse PCA** [Wang-Lu-Liu-14], [Yuan-Zhang-13], [Deshpande-Montanari-14]
  - **Tensor decomposition**
    [Anandkumar-Ge-Hsu-Kadade-12], [Anandkumar-Ge-Janzamin-14]
Remarks on Non-convex Guarantee

**Remark 1:** geometry v.s. algorithm specific.
- Most works only analyze one algorithm, especially power-method
- Our work is about geometry: “basin of attraction”
- Few works on geometry: [Keshavan-Montanari-Oh’09], [Balakrishnan-Wainwright-14] (EM), [Candes-Li-Soltanolkotabi-14] (PR)

So many algorithms (variants) for MC and more are coming out, it is better to give a unified analysis. [KMO’09] does not cover SGD.

**Remark 2:** Good initialization is often necessary.
- Exception in sparse regression (perhaps due to convex loss?).
- Good initialization not found yet in [Balakrishnan-Wainwright-14] (EM), [Deshpande-Montanari-14] (PCA)
Formulation

- Start from a **constrained version** (extra requirements on factors $X, Y$):

$$P_1' : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} F(X, Y) \triangleq \frac{1}{2} \| \mathcal{P}_\Omega (M - XY^T) \|_F^2. \quad (8a)$$

$$\|X\|_F \leq \beta_T, \quad \|Y\|_F \leq \beta_T, \quad (8b)$$

$$\|X^{(i)}\| \leq \beta_1, \quad \|Y^{(i)}\| \leq \beta_1, \quad \forall i. \quad (8c)$$

- We consider a **penalized version** of the above problem:

$$P_1 : \min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{n \times r}} \tilde{F}(X, Y) \triangleq \frac{1}{2} \| \mathcal{P}_\Omega (M - XY^T) \|_F^2 + G(X, Y), \quad (9)$$

where

$$G(X, Y) \triangleq \rho G_1 \left( \frac{3\|X\|_F^2}{2 \beta_T^2} \right) + \rho G_1 \left( \frac{3\|Y\|_F^2}{2 \beta_T^2} \right) + \rho \sum_{i=1}^{n} G_1 \left( \frac{3\|X^{(i)}\|_F^2}{2 \beta_1^2} \right) + \rho \sum_{j=1}^{n} G_1 \left( \frac{3\|Y^{(j)}\|_F^2}{2 \beta_1^2} \right),$$

in which $G_1(z) = \max(z - 1, 0)^2$, and $\rho$ is a large enough constant.
Consider **four typical algorithms**, all using the previous initialization.

- **Algorithm 1**: GD (Gradient descent).
- **Algorithm 2**: two-block Alt-Min.
- **Algorithm 3**: row BSUM (block successive upper bound minimization).
  - Difference choice of blocks compared to Algorithm 2.
- **Algorithm 4**: SGD.

**Remark**: We cover 3 classes of **first order methods**:
GD, alternating method (BCD-type), SGD (incremental gradient method).
Choice of Initial Point

- Let \( p = |\Omega|/n^2 \). **Initialization** consists of two steps.
  - **Step 1:** SVD.
    \[
    \text{Compute } (\tilde{X}_0, D_0, \tilde{Y}_0) = \text{SVD}_r \left( \frac{1}{p} \mathcal{P}_\Omega(M) \right).
    \]
    Define \( \hat{X}_0 = \tilde{X}_0 D_0^{1/2}, \hat{Y}_0 = \tilde{Y}_0 D_0^{1/2} \).
  - **Step 2:** Scaling (to force incoherence)
    Define new matrices \( X_0, Y_0 \) to make
    \[
    \|X_0^{(i)}\|^2 \leq 2\beta_1^2/3, \|Y_0^{(j)}\|^2 \leq 2\beta_1^2/3.
    \]
    \[
    X_0^{(i)} = \frac{\hat{X}_0^{(i)}}{\|\hat{X}_0^{(i)}\|} \min \left\{ \|\hat{X}_0^{(i)}\|, \sqrt{\frac{2}{3}} \beta_1 \right\}, \forall i.
    \]
    \[
    Y_0^{(j)} = \frac{\hat{Y}_0^{(j)}}{\|\hat{Y}_0^{(j)}\|} \min \left\{ \|\hat{Y}_0^{(j)}\|, \sqrt{\frac{2}{3}} \beta_1 \right\}, \forall j.
    \]

- **Claim** of good initialization: \( (X_0, Y_0) \in K_1 \cap K_2 \cap K(\delta) \), where
  \[
  K(\delta) := \{(X, Y) \mid \|M - XY^T\| \leq \delta = O(\sigma_r)\}.
  \]
(I) Local Strong Convexity?

Local strong convexity:
\[ \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \geq c \| x - x^* \|^2, \forall x \text{ close to } x^* \]
(I) Local Convexity-Like Property

Lemma 1

Under the conditions of Theorem 1 (main result), w.h.p. the following holds: for any \((X, Y) \in K_1 \cap K_2 \cap K(\delta)\), there exists \(U, V \in \mathbb{R}^{n \times r}\) such that \(UV^T = M\) and

\[
\langle \nabla_X \tilde{F}(X, Y), X - U \rangle + \langle \nabla_Y \tilde{F}(X, Y), Y - V \rangle \geq \frac{p}{9} \beta_T \left( \|X - U\|_F^2 + \|Y - V\|_F^2 \right) \geq \frac{p}{9} \|M - XY^T\|_F^2
\]

Remark: \((??)\) approximately viewed as local (strong) convexity
Main Difficulties and Techniques

- **Difficulty 1**: Iterates depend on $\Omega$.
  - Hard to estimate $\mathcal{P}_\Omega(Z)$ if $Z$ depends on $\Omega$
  - **Re-sampling** avoids the difficulty, **artificially**

- **Solution**: Random graph lemma in [K-M-O-09] (due to [Feige-Ofek-03]).

- **Difficulty 2**: Distance to a factor space
  - Recall (11) bounds $\sup_{UV^T=M} \langle (U, V) - (X, Y), \nabla \tilde{F} \rangle$.
  - $U, V$ are **coupled**. Estimate $\text{dist}((X, Y), S)$, where $S = \{(U, V) \mid UV^T = M\}$.
  - In [K-M-O-09], estimate $d(U, X), d(V, Y)$ independently (in Grassman manifold).

- Related to **perturbation analysis** [Wedin-70], but much more difficult.
A Decomposition Result

- If \( M \) is close to \( XY^T \), then \( \exists \ M = UV^T \) s.t. \( U, V \) are close to \( X, Y \), resp.

Proposition 1

Suppose
(1) \( \|M - XY^T\|_F = d \leq \frac{\Sigma_{min}}{10} \);
(2) \( \|X\|_F \leq \beta_T, \|Y\|_F \leq \beta_T \);

Then there exist \( U, V \in \mathbb{R}^{n \times r} \) such that
(a) \( UV^T = M \);
(b) \( \|U - X\|_F \leq \frac{2\beta_T}{\Sigma_{min}} d, \|V - Y\|_F \leq \frac{4\beta_T}{\Sigma_{min}} d \);

- Related to perturbation analysis [Wedin70]: if \( M \) and \( Z \) are close, then their singular vector spaces are also close.
- Prop 1 is not enough! Lemma 1 requires an advanced version of Prop 1.
Proof of Lemma 1: Outline

- Since $\tilde{F} = F + G$, only need to find factorization $M = UV^T$ s.t.:

- Step 1:
  \[ \phi_F \triangleq \langle \nabla_X F, X - U \rangle + \langle \nabla_Y F, Y - V \rangle \geq \frac{p}{9}d^2, \quad (12) \]

  where $d \triangleq \| M - XY^T \|_F \leq O(\Sigma_{\min})$.

- Step 2:
  \[ \phi_G \triangleq \langle \nabla_X G, X - U \rangle + \langle \nabla_Y G, Y - V \rangle \geq 0. \quad (13) \]

- In the rest, we only prove Step 1. (Omit Step 2, which is much more complicated.)
Difficulty of Step 1

- Need to prove $\phi_F \geq \frac{p}{9}d^2$, where

$$
\phi_F = \langle \nabla_X F, X - U \rangle + \langle \nabla_Y F, Y - V \rangle \\
= \langle \mathcal{P}_\Omega (XY^T - M)Y, X - U \rangle + \langle \mathcal{P}_\Omega (XY^T - M)^TX, Y - V \rangle \\
= \langle \mathcal{P}_\Omega (XY^T - M), (X - U)Y^T + X(Y - V)^T \rangle.
$$

- **Question**: how to bound $\|\mathcal{P}_\Omega (XY^T - M)\|_F$ by $d = \|M - XY^T\|_F$?

- **Guess**: $E(\mathcal{P}_\Omega (S)) = pS$, thus: w.h.p. $\|\mathcal{P}_\Omega (S)\|_F^2 \geq pd^2/2, \forall S$?

  - Game: pick $\Omega$ first, adversary picks $S$ depending on $\Omega$
  - If $(1, 1) \notin \Omega$, picks $S = dE_{11}$, then $\mathcal{P}_\Omega (S) = 0 < pd^2/2$.

- **Remark**:
  - In our problem, $X_k, Y_k$ depend on $\Omega$.
  - By resampling, this technical difficulty is (artificially) avoided.
Solution of Step 1

- **Lemma 2** [C-R-09]: \( \| \mathcal{P}_\Omega A \|_F^2 \geq \frac{p}{2} \| A \|_F^2 \), where

\[
A \in \mathcal{T} \triangleq \{ \hat{U}W_2^T + W_1 \hat{V}^T \mid W_1, W_2 \in \mathbb{R}^{n \times r} \}.
\]  

**Intuition:** \( \mathcal{T} \) is the “tangent space” of a fixed incoherent matrix \( M \), thus independent of \( \Omega \).

- Define \( A = U(Y - V)^T + (X - U)V^T \in \mathcal{T} \), \( B = (X - U)(Y - V)^T \), then

\[
XY^T - M = A + B.
\]

\[
\phi_F = \langle \mathcal{P}_\Omega (XY^T - M), (X - U)Y^T + X(Y - V)^T \rangle
= \langle \mathcal{P}_\Omega (A + B), A + 2B \rangle
= \| \mathcal{P}_\Omega (A) \|_F^2 + 2\| \mathcal{P}_\Omega (B) \|_F^2 + 3\langle \mathcal{P}_\Omega (A), \mathcal{P}_\Omega (B) \rangle.
\]

**Guess:** \( \approx pd^2 + 2pd^4 + 3pd^3 \)?

**We prove:** \( \approx pd^2 + 2p \frac{1}{10^2} d^2 + 3p \frac{1}{10} d^2 \). \( \Leftarrow \) “Worse than expected” bound!  

(16)
Solution of Step 1 (cont’d)

- **Step 1.1:** $\|\mathcal{P}_\Omega(A)\|_F \geq \frac{\sqrt{2p}}{3}d$, implied by Lemma 2 and $\|A\|_F \geq \frac{2}{3}d$ (to prove).
  - Note $\|A\|_F = \|(XY^T - M) - B\|_F \geq d - \|B\|_F$, so need to prove $\|B\|_F \leq \frac{1}{3}d$.

- Will prove stronger result: $\|B\|_F \leq O(d^2)$.

- **Step 1.2:** $\|\mathcal{P}_\Omega(B)\|_F^2 = \|\mathcal{P}_\Omega((U - X)(V - Y)^T)\|_F^2 \leq \frac{p}{100}d^2$.
  - Expectation of LHS = $pd^4$, so allow to lose a factor of $d^2$. 
Technical result for Step 1.2: A Random Graph Lemma

- Step 1.2 requires a random graph lemma in [Feige-Ofek-03,KMO’09]: Related to 2nd largest eigenvalue of adjacency matrix of a random graph.

Random Graph Lemma

\[ \exists \text{ constants } C_0, C_1, C_1 \text{ s.t. if } |\Omega| \geq C_0 n \log n, \text{ then w.h.p.,} \]

\[ \sum_{(i,j) \in \Omega} x_i y_j \leq C_1 p \|x\| \|y\| + C_2 \sqrt{np} \left( \sum_i x_i^2 \right) \left( \sum_i y_i^2 \right), \forall x, y \in \mathbb{R}^n. \quad (17) \]

- Does not require \( \Omega \) to be independent of \( (x, y) \)!
Examples

\[ \sum_{(i,j) \in \Omega} x_i y_j \leq C_1 p \|x\|\|y\| + C_2 \sqrt{n} p \left( \sum_i x_i^2 \right) \left( \sum_i y_i^2 \right), \ \forall x, y \in \mathbb{R}^n. \]

- E.g. 1: \( x = y = (d^2, \ldots, d^2)/\sqrt{n} \); note \( \|x\| = \|y\| = d^2 \).
  LHS = \( pd^4 \), RHS = \( O(pd^4 + \sqrt{n}pd^4/n^2) \approx O(pd^4) \). The first term is expected.

- E.g. 2: \( (i_0, j_0) \in \Omega, x = d^2 e_{i_0}, y = d^2 e_{j_0} \).
  LHS = \( d^4 \), RHS = \( O(pd^2 + \sqrt{n}pd^4) \approx O(pd^2 + d^4) \).
  - compare to \( pd^4 \) (expectation), lose \( d^2 \) (acceptable) or \( p \)
    (\( \approx 1/n \), un-acceptable).

- Need “incoherence” to control the second term \( \implies \) not to lose \( p \); can lose \( d^2 \).
Main contribution: recovery guarantee for non-convex factorization based matrix completion

- Key idea: “local strongly-convex basin”
- Apply to many first-order algorithms
- No resampling!

Math tools:
- Perturbation analysis;
- A random graph lemma (stronger than concentration bounds)

Future directions
- Noisy matrix completion
- Remove penalty on row-norms
- Robust PCA, tensor completion, etc.
Thank You!