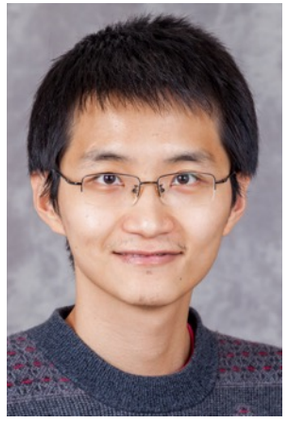
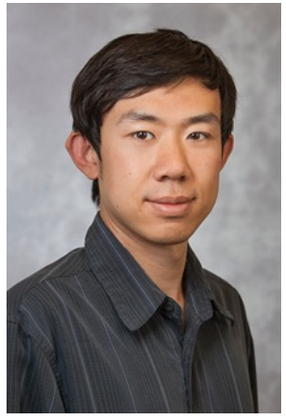
Non-Convex Matrix Completion Against a Semi-Random Adversary

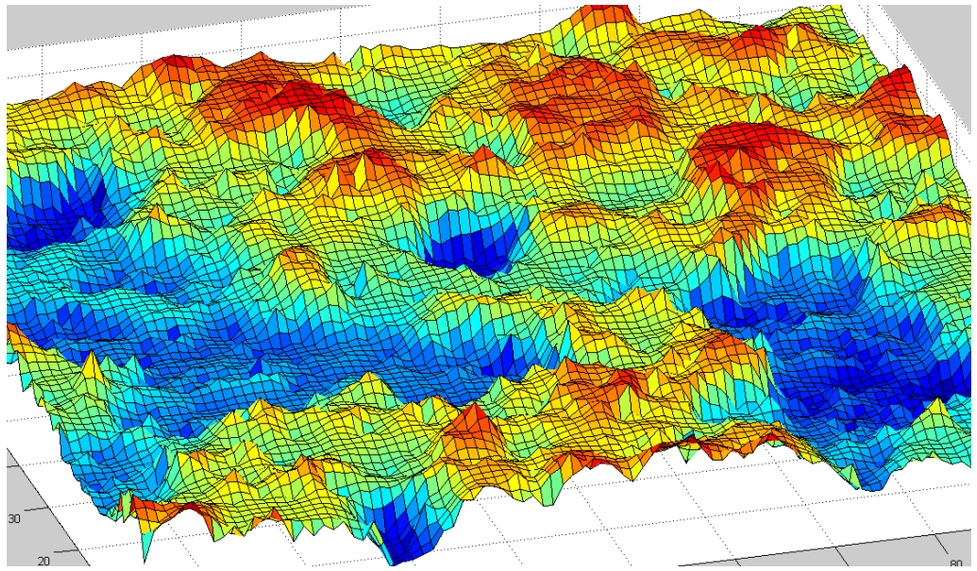




Yu Cheng Rong Ge

Duke University

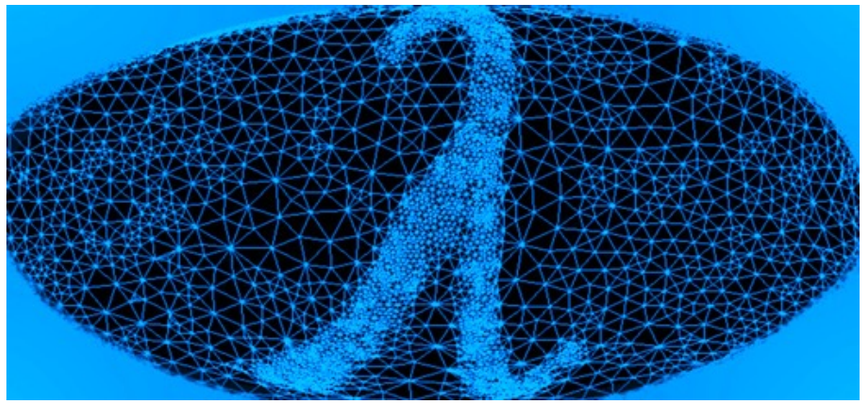
Non-Convex Optimization



Why non-convex optimization works well in practice?

- Are all local optima (approximately) globally optimal?

Spectral Graph Theory



Discrete / Combinatorial

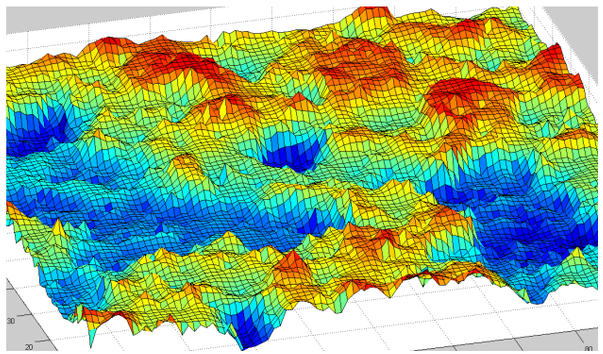
Graphs Paths Trees Flows Cuts …

Continuous / Numerical

Matrices Eigenvalues
Iterative Methods …

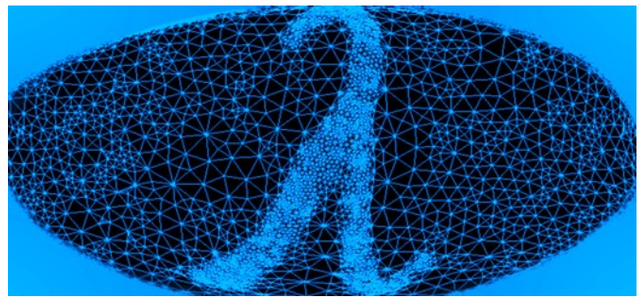
This Talk

Non-Convex



Matrix Completion

Spectral

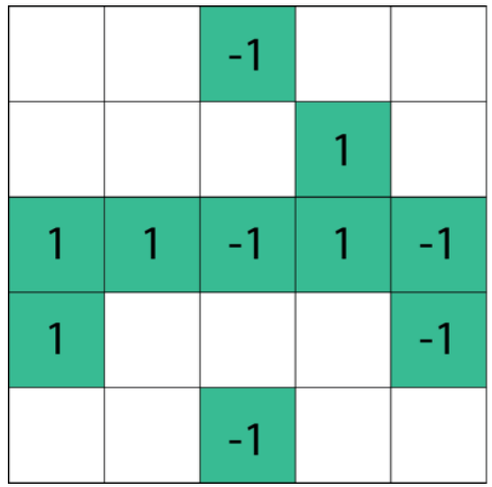


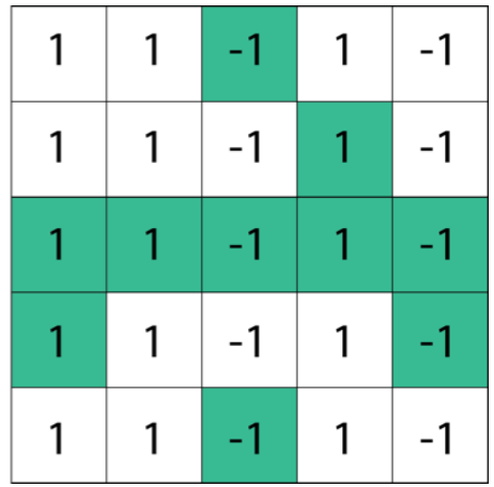
Matrix Completion

An unknown $n \times n$ rank r matrix $M^\star = ZZ^T$ $(Z \in \mathbb{R}^{n \times r})$.

Input: $M^\star_{ij}$ for a set of observed entries $(i, j) \in \Omega$.

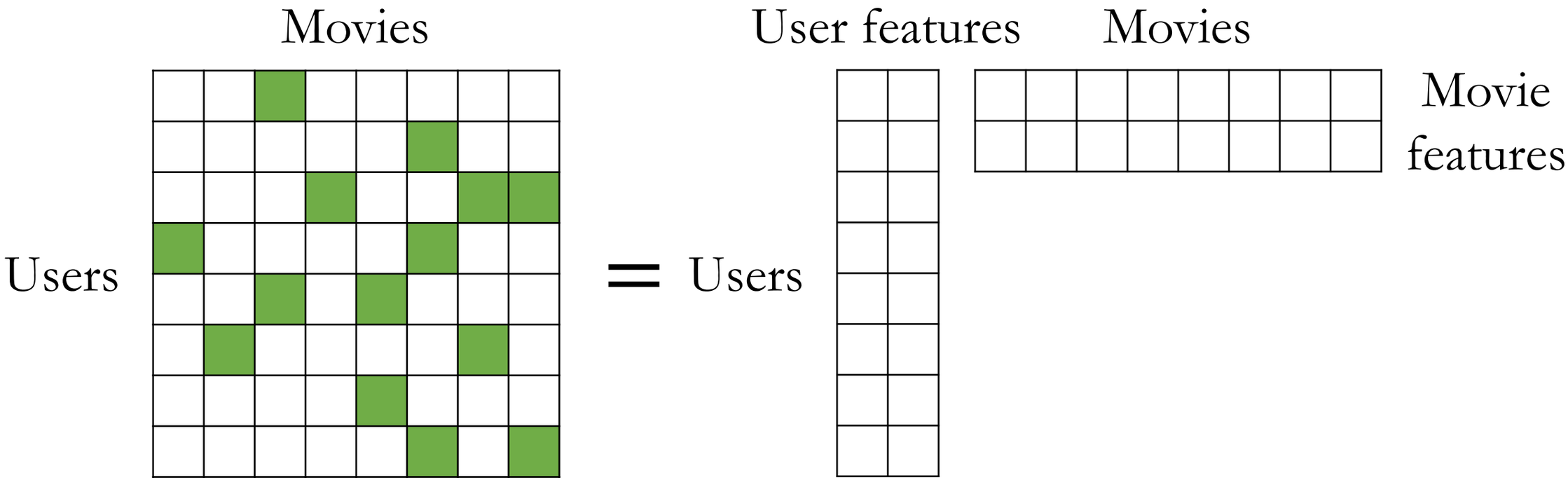
Goal: Recover $M^\star$.



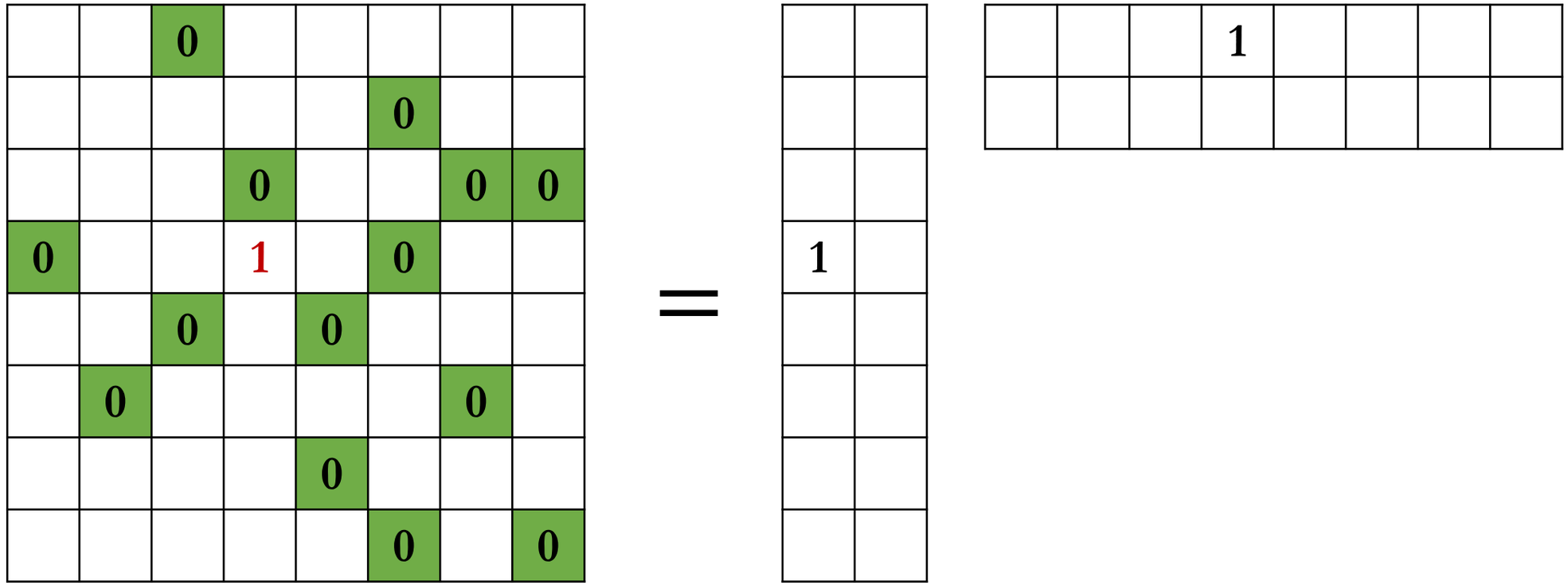


Matrix Completion

A basic machine learning problem with applications to recommendation systems and collaborative filtering:

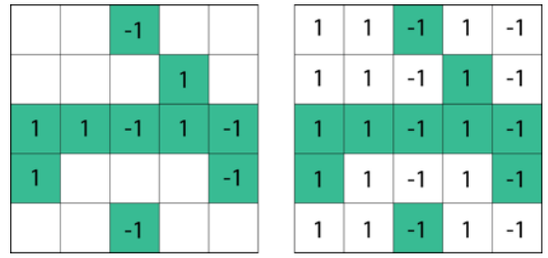


Incoherence



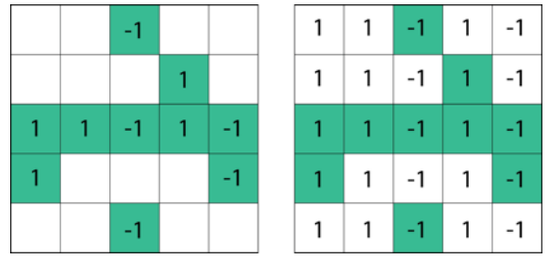
SVD: $M^{\star} = XDY^{T}$, the rows of X and Y have similar norms.

Previous Work



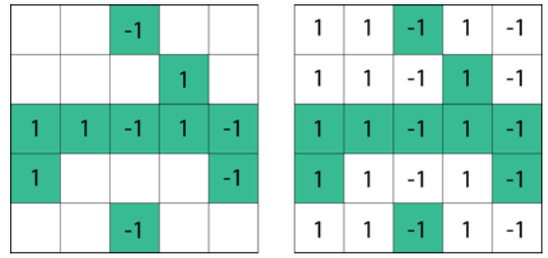
- Convex Relaxation (e.g., [Candès and Tao '10] [Recht '11])
  - min $\|M\|_\star$ s.t. $M_{ij} = M^\star_{ij}$ min $rank(M)$ s.t. $M_{ij} = M^\star_{ij}$ (minimizing sum of the singular values of M).
  - Can be solved using SDP in time $O(n^4)$.

Previous Work



- Convex Relaxation (e.g., [Candès and Tao '10] [Recht '11])
  - min $\|M\|_\star$ s.t. $M_{ij} = M^\star_{ij}$
- Non-Convex Approaches
  - min $f(X) = \sum_{(i,j)\in\Omega} \left(M^\star_{ij} - (XX^T)_{ij}\right)^2$ for $X \in \mathbb{R}^{n\times r}$.

Non-Convex Approaches



- With careful initialization (typically via SVD)

  [Jain et al. '13, Hardt and Wootters '14, Chen and Wainwright '15]

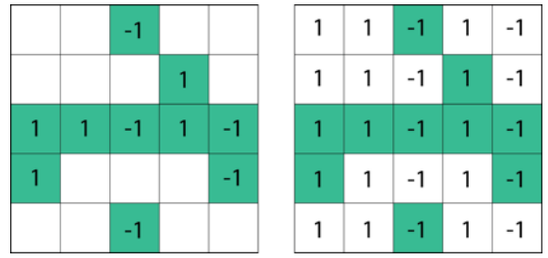
  - Strong convexity near ground-truth

  [Sun and Luo '15, Zhao et al. '15, Zheng and Lafferty '16, Tu et al. '15]

- Without requiring initialization: The non-convex objective (with regularization) has no bad local optima.

  [Sa et al. '15, Ge et al. '16, Park et al. '16, Ge et al. '17]

Previous Work



- Convex Relaxation (e.g., [Candès and Tao '10] [Recht '11])

- Non-Convex Approaches (e.g. [Jain et al. '13, Hardt and Wootters '14, Chen and Wainwright '15, Sun and Luo '15, Zhao et al. '15, Zheng and Lafferty '16, Tu et al. '15, Sa et al. '15, Ge et al. '16, Park et al. '16, Ge et al. '17])

All of these works require uniformly random observations!

$$p_{ij} = \Pr[(i,j) \in \Omega] = p \quad \text{(independently)}$$

Motivating Questions [Cheng Ge '18]

- What happens if $p_{ij} \geq p$?
  - Semi-random adversary.

Are the non-convex approaches robust in this semi-random model?

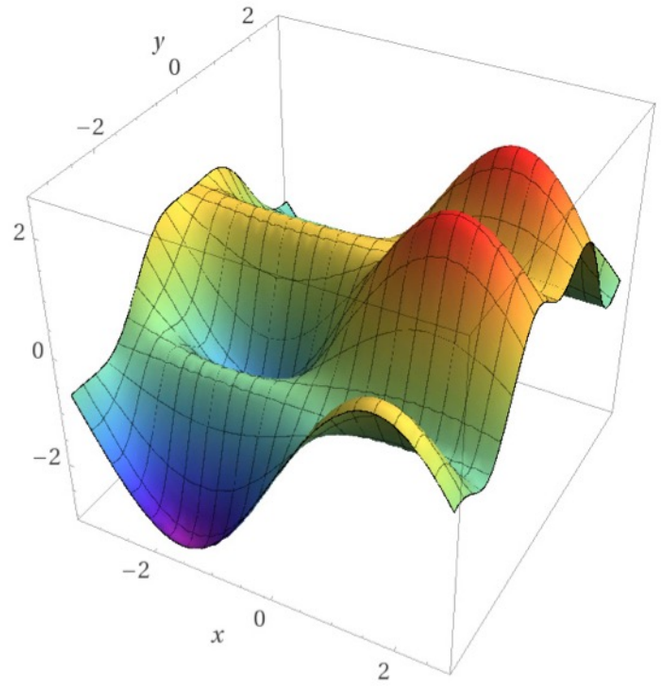
If not, is there a way to fix the non-convex algorithms while preserving their efficiency?

Our Results

- Counter-examples
- Preprocessing step
- No bad local optima

Our Results

- Counter-examples
- Preprocessing step
- No bad local optima



Counter Examples

$f(X) = \sum_{(i,j)\in\Omega} \left(M^{\star}_{ij} - (XX^T)_{ij}\right)^2$ has bad local optimum.

$$M^{\star} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, P = \begin{pmatrix} 9p & p \\ p & 9p \end{pmatrix}, \quad X = \begin{pmatrix} \sqrt{0.8} \approx 0.89 \\ -0.89 \end{pmatrix}$$

Counter Examples

$$M^\star = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, P = \begin{pmatrix} 9p & p \\ p & 9p \end{pmatrix}, \qquad x = \begin{pmatrix} 0.89 \\ -0.89 \end{pmatrix}$$

$$\frac{1}{p}\mathbb{E}[M^\star \mid \Omega] = \begin{pmatrix} 9 & 1 \\ 1 & 9 \end{pmatrix}$$

Can verify that $\nabla f(x) = 0, \nabla^2 f(x) \succcurlyeq 0$.

Intuition: x is the eigenvector of $M^\star \circ P$.

Counter Examples

$f(X) = \sum_{(i,j)\in\Omega} \left(M^{\star}_{ij} - (XX^T)_{ij}\right)^2$ has bad local optimum.

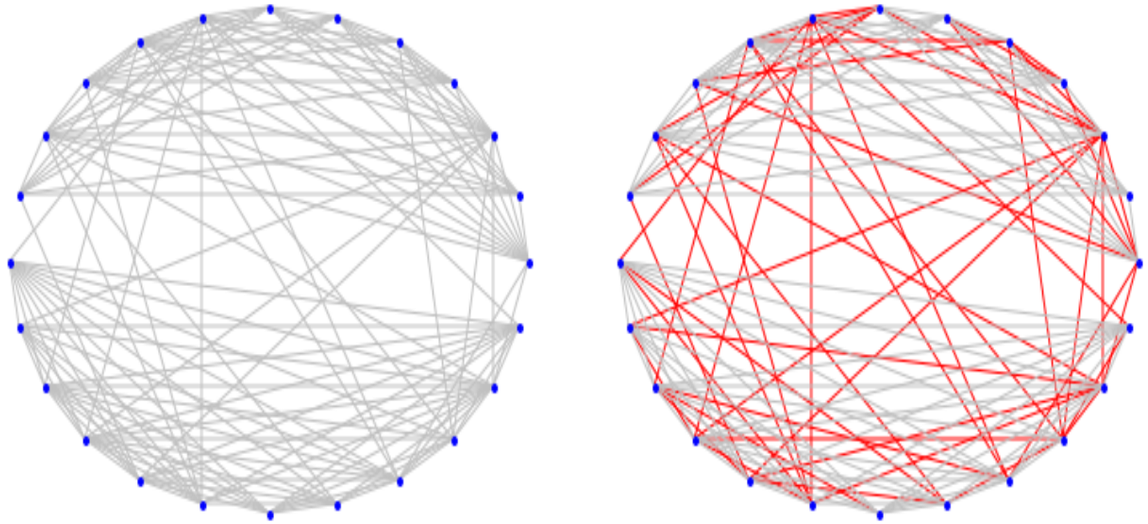
Another example: SVD cannot find the right subspace:

- $\text{span}(M^{\star}) \perp \text{span}(M^{\star} \circ P)$.

There are bad local optima and SVD does not give good initialization!

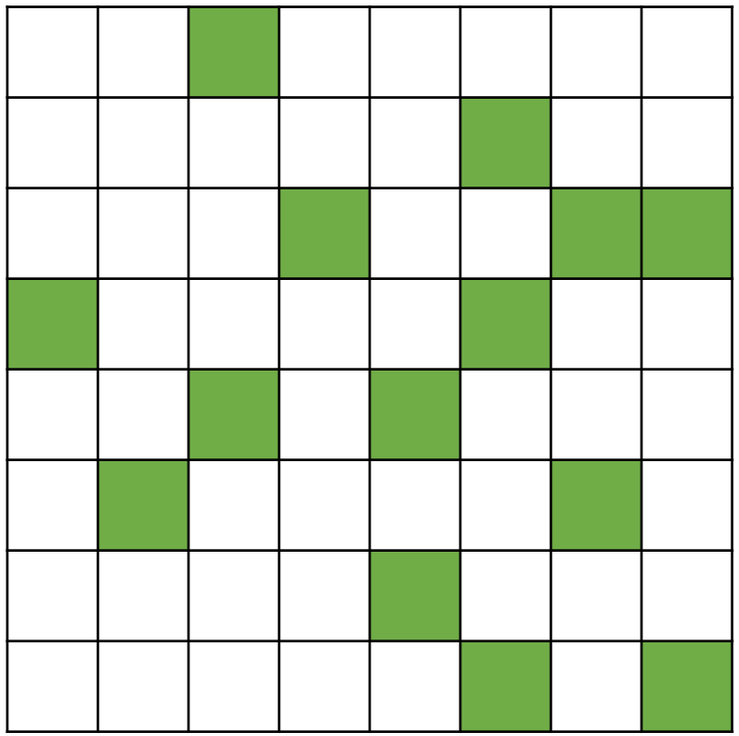
Our Results

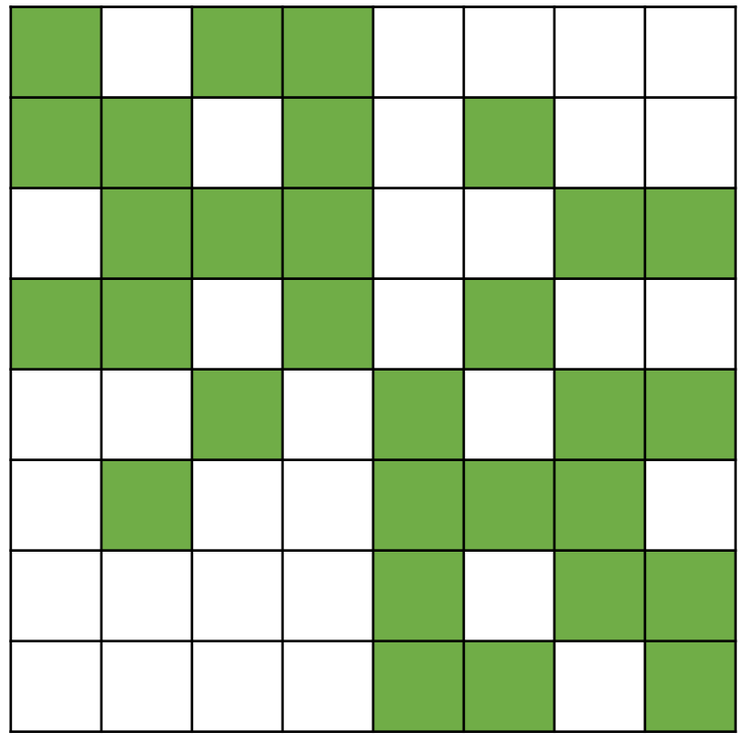
- Counter-examples
- Preprocessing step
- No bad local optima



Preprocessing

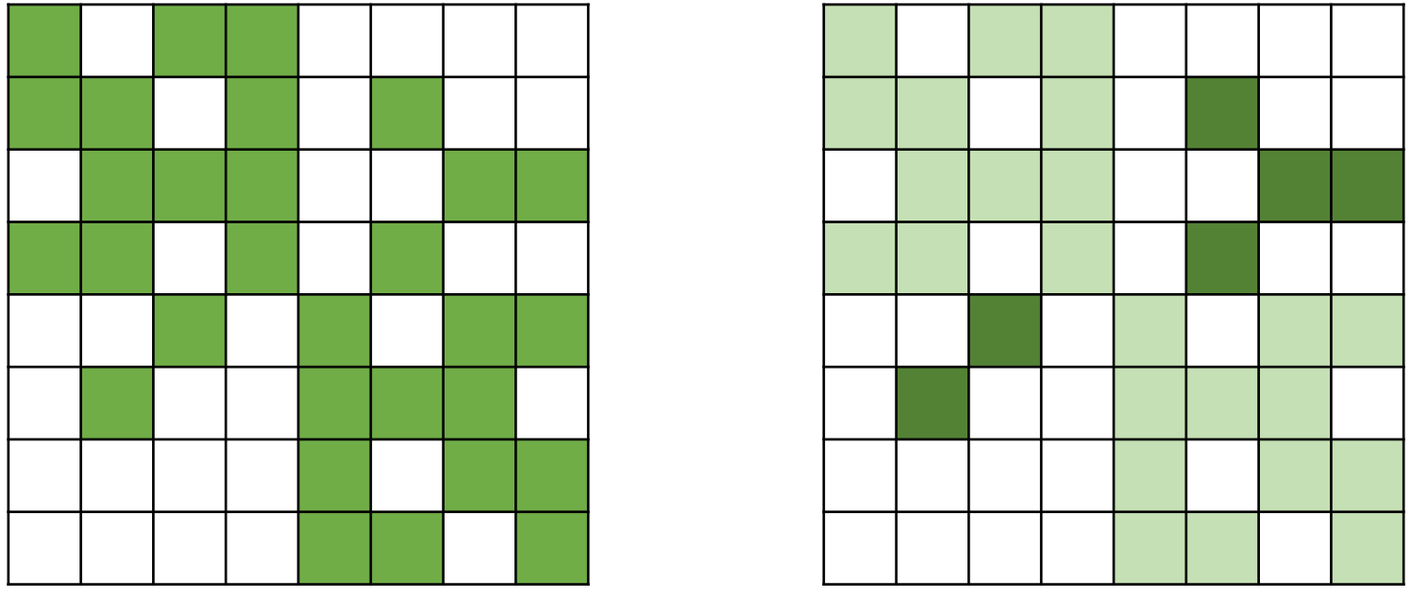
What went wrong in the counter-example?





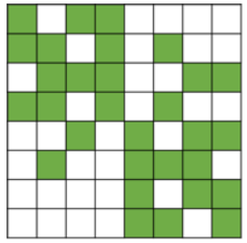
Preprocessing

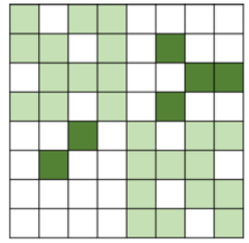
If we know the adversary is changing the probabilities this way…



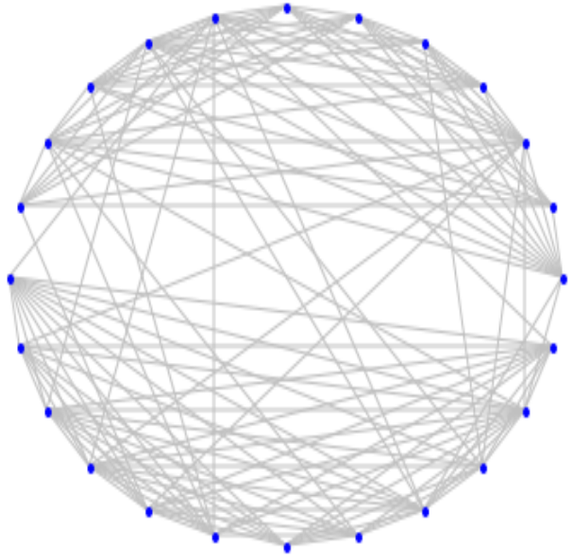
Key Idea: Re-weight the samples so that the input is "similar" to a random input.

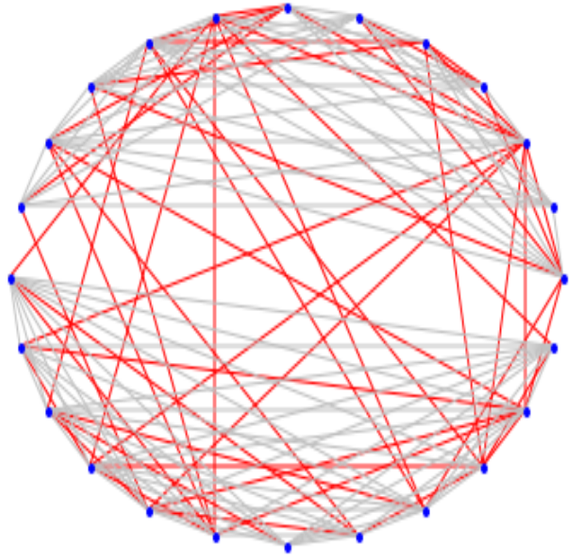
Preprocessing



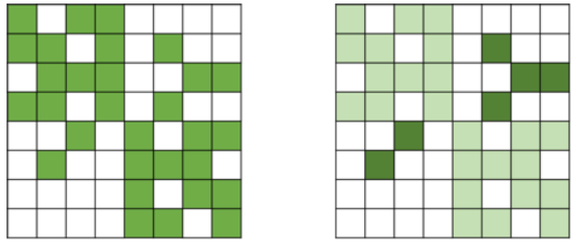


$$f(X) = \sum_{(i,j)\in\Omega} \boldsymbol{W_{ij}} \left(M^{\star}_{ij} - (XX^T)_{ij}\right)^2$$





Preprocessing

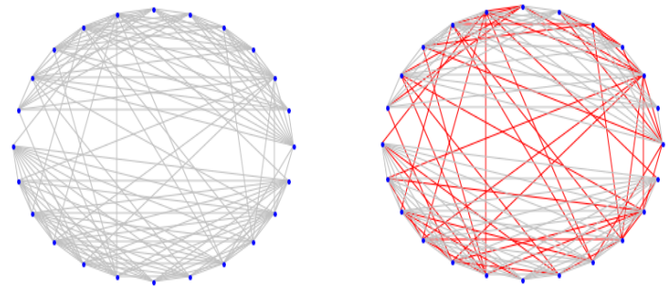


The graph formulation:

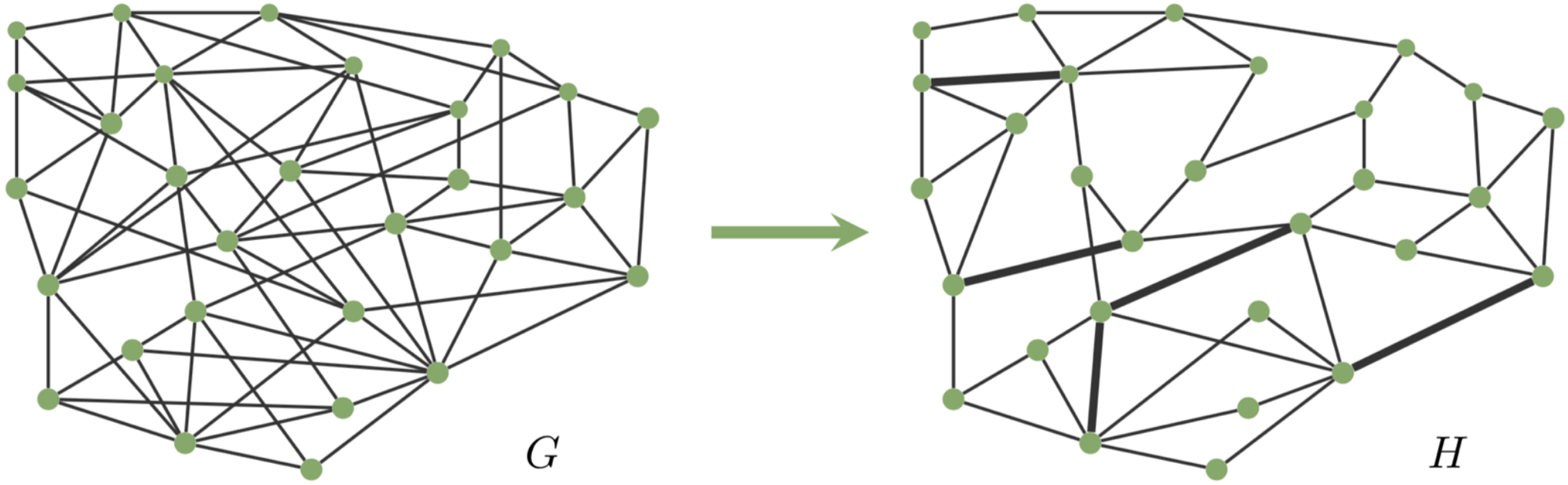
- Input: a graph G that contains a subgraph similar to H.
- Goal: Reweight G so that $G \approx H$.

For matrix completion, $H =$ complete graph / expander.

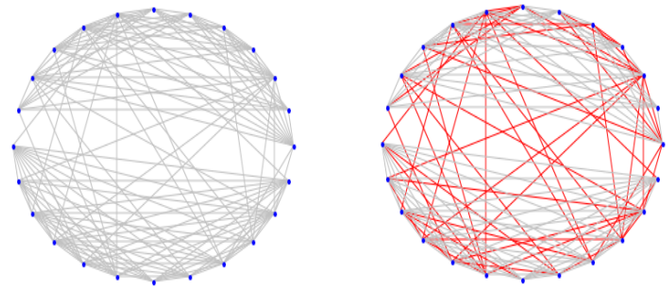
Preprocessing



Graph Sparsification [Spielman and Teng '11]



Preprocessing



Graph Sparsification:

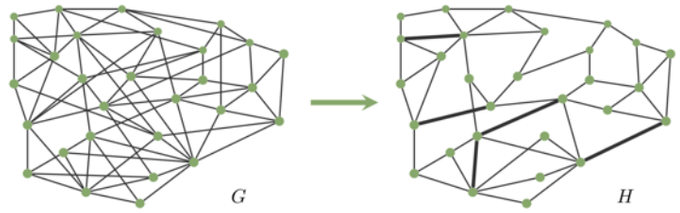
- ϵ-spectral sparsifier with $\tilde{O}(n/\epsilon^2)$ edges. [Spielman and Teng '11]
- ϵ-spectral sparsifier with $O(n/\epsilon^2)$ edges.
  - Runtime: $O(n^3 m)$ [Batson Spielman Srivastava '12].
  - Runtime: $\tilde{O}(m\, poly(1/\epsilon))$ [Lee and Sun '17].

$(n = |V|, m = |E|)$.

Preprocessing

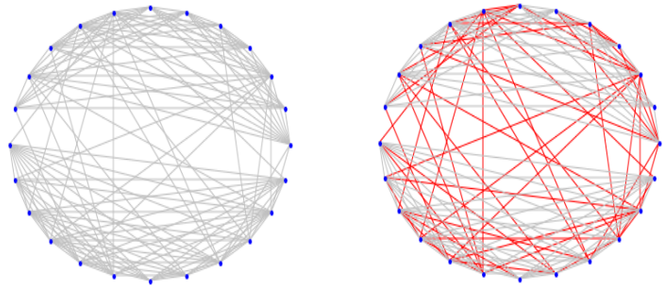
Problem 1. *Given a set S of m PSD matrices $M_1, \cdots, M_m$ with $\sum_{i=1}^m M_i = I$ and $0 < \varepsilon < 1$, find non-negative coefficients $\{c_i\}_{i=1}^m$ such that $|\{c_i | c_i \neq 0\}| = O\left(n/\varepsilon^2\right)$, and*

$$(1-\varepsilon)\cdot I \preceq \sum_{i=1}^m c_i M_i \preceq (1+\varepsilon)\cdot I.$$



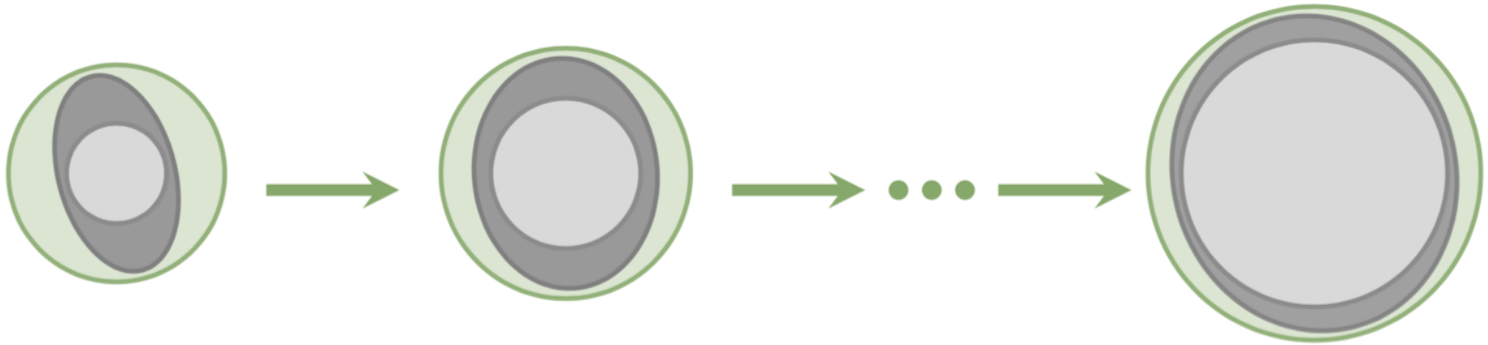
Problem 2. *For a set of m vectors $\{v_i\}_{i=1}^m$, assume there exists a subset S of vectors such that $(1-\beta)I \preceq \sum_{i\in S} v_i v_i^\top \preceq (1+\beta)I$. Compute a set of weights $w_i \geq 0$ such that*

$$(1-O(\beta)-\epsilon)I \preceq \sum_i w_i v_i v_i^\top \preceq I.$$



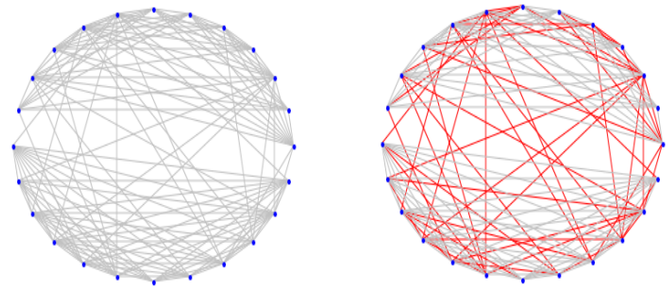
Linear-Sized Graph Sparsification

- Maintain current sum of rank-one matrices $A^t = \sum_i w_i^t v_i v_i^T$, and upper/lower barrier values u_t, l_t such that $l_t\, I \prec A \prec u_t\, I$.



- Add edges iteratively and deterministically.
  - A good edge always exists because on average the edges are good.

Preprocessing

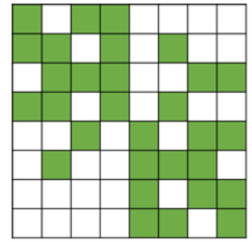


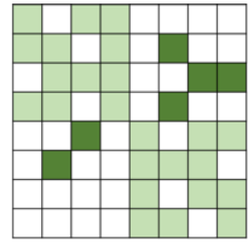
- Conceptual contributions:
  - BBS is powerful because they pick edge deterministically.
  - Useful for semi-random settings.

- Technical contributions:
  - Hidden good subset S. (Random sampling no longer works!)
  - $\sum_{i \in S} v_i v_i^T \approx I$ instead of $\sum_i v_i v_i^T = I$.

    (Upper/lower barriers move at different rates.)

Preprocessing





Implication to matrix completion:

- If $p \geq \cdots$, then there exists a set of good weights.
  - $W_{ij} = 1$ for the entries revealed by nature.
  - $W_{ij} = 0$ for the entries added by adversary.
- Preprocessing can recover $\widetilde{W}$ supported on Ω such that $\|W - J\|_2 \leq O(\epsilon n)$ in time $\tilde{O}(m\,\mathrm{poly}(1/\epsilon))$.

We will show this is enough to remove bad local optima.

Our Results

- Counter-examples
- Preprocessing step
- No bad local optima

Non-Convex Matrix Completion

$$f(X) = \sum_{(i,j)\in\Omega} W_{ij}\left(M^{\star}_{ij} - (XX^T)_{ij}\right)^2$$

$$f(X) = \langle XX^T - ZZ^T, XX^T - ZZ^T\rangle_W$$

[Ge Jin Zheng '17]: Sufficient to focus on the direction of $\Delta = X - Z$.

$$\langle \nabla f(X), \Delta\rangle = 2\langle X\Delta^T + \Delta^T X, XX^T - ZZ^T\rangle_W = 0$$

$$\frac{1}{2}\left\langle \nabla^2 f(X), \Delta\Delta^{\mathrm{T}}\right\rangle = \langle X\Delta^T + \Delta^T X, X\Delta^T + \Delta^T X\rangle_W + 2\langle \Delta\Delta^T, XX^T - ZZ^T\rangle_W \geq 0$$

Non-Convex Matrix Completion

$$\langle X\Delta^T + \Delta^T X, XX^T - ZZ^T \rangle_W = 0$$

$$\langle X\Delta^T + \Delta^T X, X\Delta^T + \Delta^T X \rangle_W + 2\langle \Delta\Delta^T, XX^T - ZZ^T \rangle_W \geq 0$$

$$\Delta = X - Z \qquad \text{Let } A = X\Delta^T + \Delta^T X$$

$$B = XX^\top - ZZ^T = X\Delta^T + \Delta^T X - \Delta\Delta^T$$

$$\langle A, B \rangle_W = 0$$

$$\begin{aligned}\langle A, A \rangle_W + 2\langle A - B, B \rangle_W &= \langle A, A \rangle_W - 2\langle B, B \rangle_W \\ &= \langle A - B, A - B \rangle_W - 3\langle B, B \rangle_W \\ &= \langle \Delta\Delta^T, \Delta\Delta^T \rangle_W - 3\langle B, B \rangle_W \geq 0.\end{aligned}$$

Non-Convex Matrix Completion

$$\langle \Delta\Delta^T, \Delta\Delta^T \rangle_W - 3\langle M^\star - XX^T, M^\star - XX^T \rangle_W \geq 0$$

[Ge Lee Ma '16, Ge Jin Zheng '17]:

$$\langle \Delta\Delta^T, \Delta\Delta^T \rangle \leq 2\langle M^\star - XX^T, M^\star - XX^T \rangle$$

- Need to show: for specific low rank matrices the weighted norm is similar to the Frobenius norm.

Non-Convex Matrix Completion

- Need to show: for specific low rank matrices the weighted norm is similar to the Frobenius norm.

This is clear under (uniformly) random sampling as we can rely on concentration bounds.

Now we need to replace matrix concentration bounds with deterministic conditions.

Non-Convex Matrix Completion

Lemma 4.3 (Preserving the Norm via Spectral Properties). *For any matrices* $X \in \mathbb{R}^{n_1 \times r}$, $Y \in \mathbb{R}^{n_2 \times r}$, *and* $W \in \mathbb{R}^{n_1 \times n_2}$, *we have*

$$|\|XY^\top\|_W^2 - \|XY^\top\|_F^2| \leq \|W - J\|\|X\|_F\|Y\|_F \max_i \|X_i\| \max_i \|Y_i\|.$$

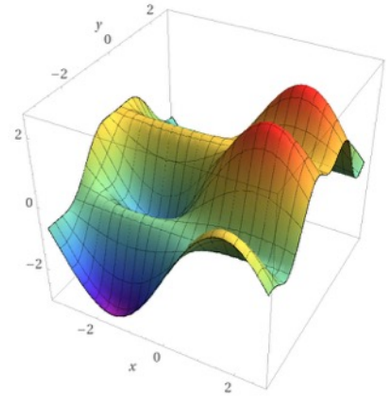
After preprocessing,

any local optima of $f(X)$ satisfies that

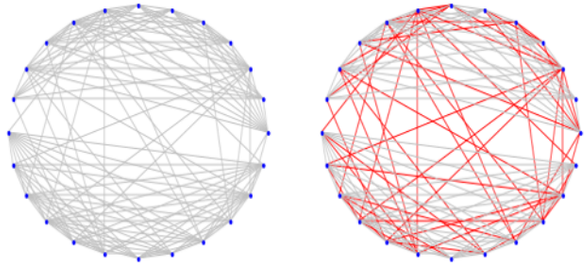
$$\|M^\star - XX^T\|_F^2 \leq \epsilon \|M^\star\|_F^2$$

Our Contributions

- Counter-examples



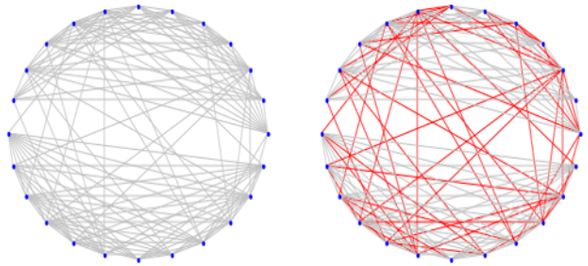
- Preprocessing step



- No bad local optima

Our Contributions

- Study non-convex matrix completion when $p_{ij} \geq p$. Existing non-convex approaches do not work.
- Preprocessing step



- No bad local optima

Our Contributions

- Study non-convex matrix completion when $p_{ij} \geq p$. Existing non-convex approaches do not work.

- Efficient preprocessing step that reweights the samples.

- No bad local optima

Our Contributions

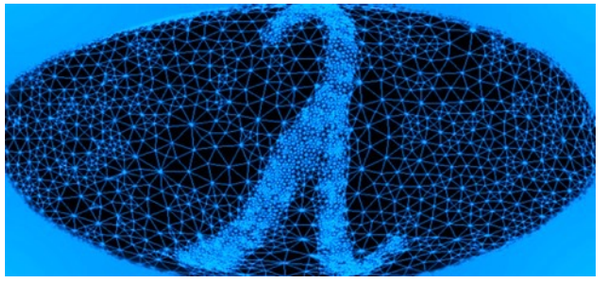
- Study non-convex matrix completion when $p_{ij} \geq p$. Existing non-convex approaches do not work.

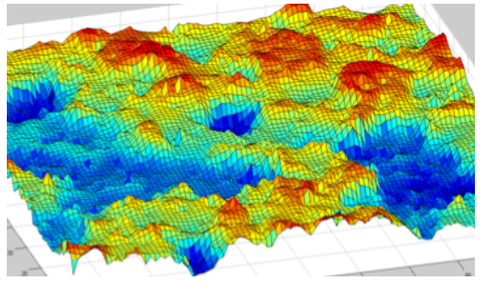
- Efficient preprocessing step that reweights the samples.

- After the reweighting, the non-convex objective has no bad local optima.

Our Contributions

Lightweight convex optimization + Non-convex

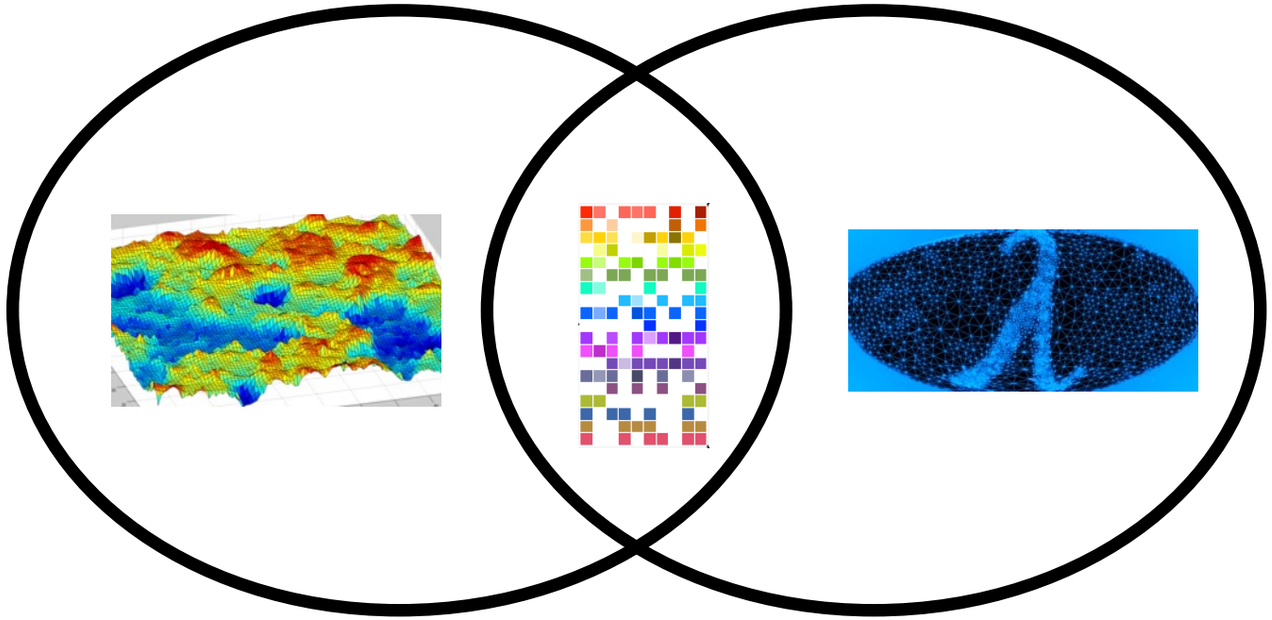




Open Problems

- Technical open problem: Exact Recovery?
  - Can the adversary create bad local optima by revealing just a few entries?

- More applications of BSS (or packing SDP solvers in general) in semi-random settings.



Thanks!
Questions?