Provable Alternating Minimization Methods for Low-rank Matrix Estimation Problems

Prateek Jain
Microsoft Research India

Joint work with Praneeth Netrapalli, Sujay Sanghavi, Inderjit S. Dhillon

Talk Outline

- Low-rank Estimation Problems
  - recommendation systems (e.g. Netflix Challenge)
  - Multi-label Learning
  - ….
- Alternating Minimization methods
  - Most popular method
  - Little theoretical understanding
- Study Alternating Minimization method
  - General technique to analyze alternating minimization
  - Linear convergence to optima
- Guarantees for:
  - Matrix Completion
  - RIP-based Matrix Sensing
  - Rank-one Operator based Matrix Sensing
- Conclusions

Low-rank Matrix Completion

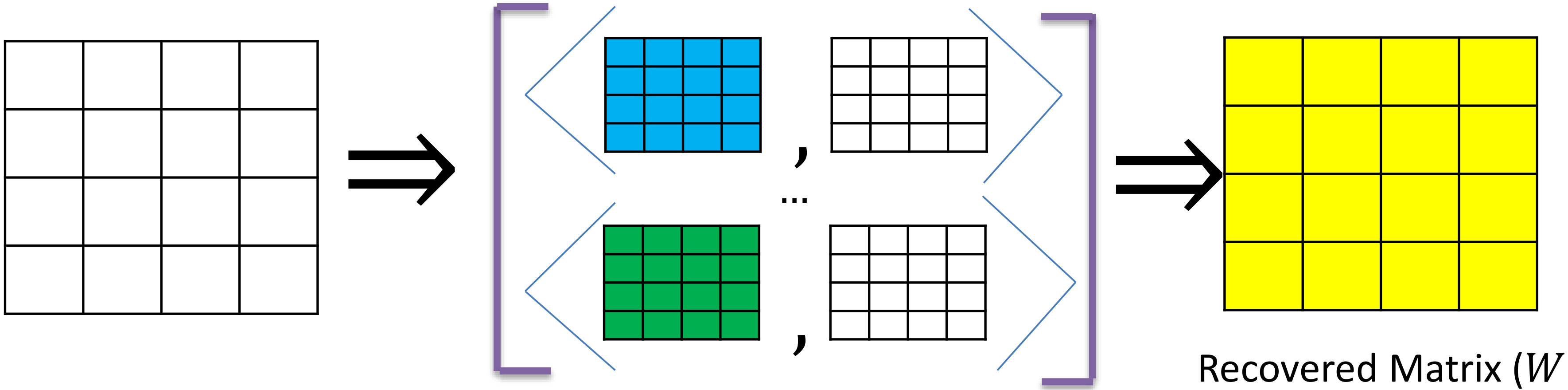
users

movies	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

□ - unknown rating 🟨 - rating between 1 to 5

- Task: Complete ratings matrix
- Applications: recommendation systems, PCA with missing entries

Low-rank Matrix Sensing



Hidden matrix (W_*)

Observed Measurements (A(W_*))

Recovered Matrix (W)

Low-rank Matrix Estimation—Linear Measurements

$$\mathbb{A}(W^*) = b$$

- $\mathbb{A}: \mathbf{R}^{\boldsymbol{d_1 \times d_2}} \to \boldsymbol{R^m}$
  - Linear operator
  - $\mathbb{A} = \{\mathbf{A_1}, \mathbf{A_2}, \dots, \boldsymbol{A_m}\}$

$$\mathbb{A}(W) = \begin{bmatrix} \langle A_1, W \rangle \\ \langle A_2, W \rangle \\ \vdots \\ \langle A_m, W \rangle \end{bmatrix}$$

- Optimization Version:

$$\min_{W} ||\mathbb{A}(W) - b||_2^2$$
$$s.t.\ \ rank(W) \leq k$$

Low-rank Matrix Estimation

$$\min_{W} ||\mathbb{A}(W) - b||_2^2$$
$$s.t.\ \ rank(W) \leq k$$

- NP-hard in general
  - Hard to even approximate within $\log(d_1 + d_2 + m)$ [MJCD'08]
- Tractable solutions for a variety of important problems
  - Matrix completion
  - RIP based matrix sensing

Existing method: Trace-norm minimization

$$\min_{W} ||\mathbb{A}(W) - b||_2^2$$
$$s.t.\ ||W||_* \leq \tau_k$$

- $||X||_*$: sum of singular values
- Several powerful results:
  - Matrix Completion: [CR08, CT08, Gross09, Recht11….]
  - RIP based Matrix Sensing: [RFP10]
- However, convex optimization methods for this problem don't scale well
  - Intermediate iterates can have rank much larger than "k"
  - SVD computation per step

Projected Gradient based Methods

- $W_0 = 0$
- For t=1:T

$$Z = W_t - \eta \mathbb{A}^{\mathrm{T}}(\mathbb{A}(W_{\mathrm{t}}) - \mathrm{b})$$

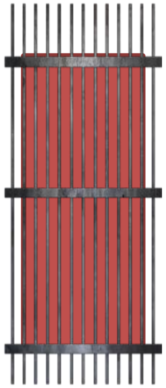
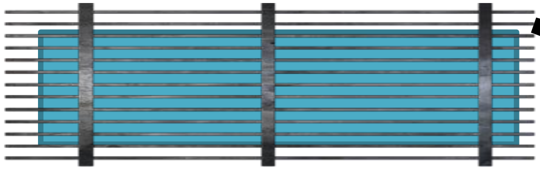
$$W_{t+1} = \arg\min_{W} ||Z - W||_F^2,$$

$$s.t., \quad rank(W) \leq k$$

- Known analysis for RIP based matrix sensing
- No guarantees for other problems like matrix completion

[JMD10] [GM, FOCM11] [MTT10]

Alternating Minimization

$\left\| b - \mathbb{A}\right($  $\times$  $\left)\right\|_F^2$

$$W^* \cong U \times V^T$$

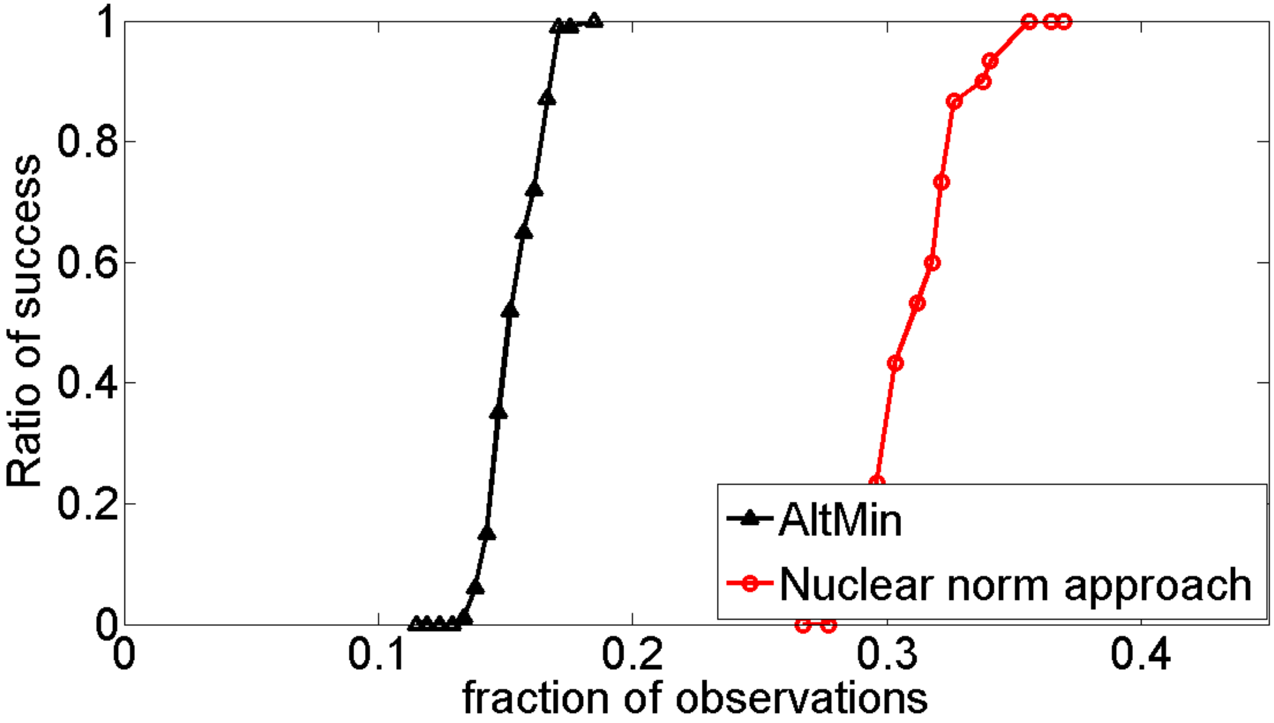
$$V^{t+1} = \min_V ||b - A(U^t V^T)\,||_2^2$$

$$U^{t+1} = \min_U ||b - A(U(V^{t+1})^T)||_2^2$$

Alternating Minimization

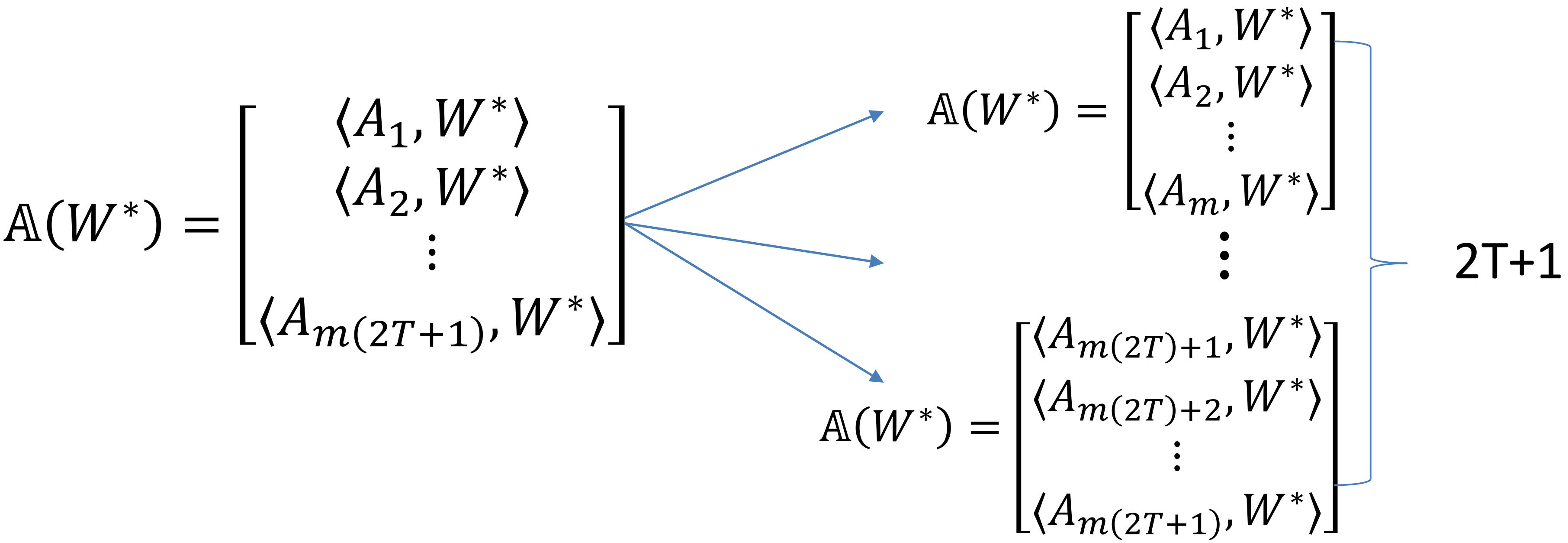
- Solving for U or V individually is "easy"
  - Only a least squares problem with $(m+n) \times k$ variables
- Several nice properties
  - Small storage requirement: store only U, V
  - Fast intermediate steps
    - no requirement of eigenvalue or singular value decomposition
  - Highly accurate in practice
    - forms an important component of the winning entry to Netflix Challenge

Empirical Performance of Alternative Minimization



- However, the overall problem is non-convex
  - No known analysis for recovery of exact M
  - Only convergence to local minima known

AltMin Algorithm



$$\mathbb{A}(W^*) = \begin{bmatrix} \langle A_1, W^*\rangle \\ \langle A_2, W^*\rangle \\ \vdots \\ \langle A_{m(2T+1)}, W^*\rangle \end{bmatrix}$$

$$\mathbb{A}(W^*) = \begin{bmatrix} \langle A_1, W^*\rangle \\ \langle A_2, W^*\rangle \\ \vdots \\ \langle A_m, W^*\rangle \end{bmatrix}$$

$$\vdots$$

$$\mathbb{A}(W^*) = \begin{bmatrix} \langle A_{m(2T)+1}, W^*\rangle \\ \langle A_{m(2T)+2}, W^*\rangle \\ \vdots \\ \langle A_{m(2T+1)}, W^*\rangle \end{bmatrix}$$

- Initialization:
  - $U_0 =$ Largest singular vector of $\sum_i A_i b_i$
- t+1-th Iteration:

$$V^{t+1} = \min_V ||b - A(U^t V^T)||_2^2$$

$$U^{t+1} = \min_U ||b - A(U(V^{t+1})^T)||_2^2$$

Conditions on $\mathbb{A}$

$$\mathbb{A}(W_* = U_*\Sigma_*V_*^T) = \begin{bmatrix} \langle A_1, W_* \rangle \\ \langle A_2, W_* \rangle \\ \vdots \\ \langle A_m, W_* \rangle \end{bmatrix}$$

- Assume that $\mathbb{A}$ satisfies:
  - Initialization: $||svd(\sum_i A_i b_i) - W_*||_2 \leq \delta\ ||W||_*$
  - Concentration:

$$|| \frac{1}{m} \sum_{i=1}^{m} A_i v_p v_q^T A_i^T - \langle v_p, v_q \rangle\, I||_2 \leq \delta\ ||v_p||_2 ||v_q||_2$$

$$|| \frac{1}{m} \sum_{i=1}^{m} A_i^T u_p u_q^T A_i \ - \langle u_p, u_q \rangle I||_2 \leq \delta\ ||u_p||_2 ||u_q||_2$$

  - $\delta \leq \frac{1}{100 \cdot k^{1.5} \cdot \beta}, \beta = \sigma_*^1 / \sigma_*^k$
- u_p, v_p, u_q, v_q independent of A_i's

[J., Dhillon, Arxiv'13]

Main General Result

- Assume $\mathbb{A}$ satisfies Property 1, 2
- For all $t \geq 1$,

$$dist(U_{t+1}, U_*) \leq \frac{1}{2} dist(U_t, U_*)$$

$$dist(V_{t+1}, V_*) \leq \frac{1}{2} dist(V_t, V_*)$$

- After $T = O\left(\log\left(\frac{||W_*||_F}{\epsilon}\right)\right)$:

$$||W_T - W_*||_2 \leq \epsilon$$

[J., Dhillon, Arxiv'13]

Distance Function

$$dist(U, U_*) = \; ||U_\perp^T U_*||_2$$

- $U_\perp$: basis of space orthogonal to $span(U)$
- Largest principal angle between U, U_*
- Commonly used distance function between subspaces
- For 1-d subspaces:

$$dist(u, u_*) = \sqrt{1 - \langle u, u_* \rangle^2}$$

Proof Sketch

$$v_{t+1} = \arg\min_{v} \sum_{i=1}^{m} (\langle A_i, u_t v^T \rangle - \langle A_i, u_* v_*^T \rangle)^2$$

$$\sum_{i=1}^{m} (\langle A_i, u_t v_{t+1}^T \rangle - \langle A_i, u_* v_*^T \rangle) A_i^T u_t = 0$$

$$\underbrace{\left(\sum_i A_i^T u_t u_t^T A_i \right)}_{B} v_{t+1} = \underbrace{\left(\sum_i A_i^T u_t u_*^T A_i \right)}_{C} v_*$$

$$v_{t+1} = \langle u_t, u_* \rangle v_* - B^{-1} \left(\sum_i A_i^T u_t u_* (I - u_t u_t^T) A_i \right) v_*$$

[J., Netrapalli, Sanghavi, STOC'13]

Proof Sketch

$$v_{t+1} = \underbrace{\langle u_t, u_* \rangle}_{\substack{\text{Power Method Term} \\ W_*^T u_t}} v_* \quad - \quad \underbrace{B^{-1} \left(\sum_i \overset{u_p}{A_i^T u_t} \overbrace{u_* (I - u_t u_t^T) A_i}^{u_q} \right) v_*}_{\text{Error Term}}$$

$$\langle u_p, u_q \rangle = 0, \qquad ||u_q||_2 = dist(u_t, u_*)$$

- Applying concentration inequality:
  - Error term $\leq 2\delta \, dist(u_t, u_*)$
- Error decay follows by using:
  - Bound on error term
  - Lower bound on $\langle u_t, u_* \rangle$ by initialization

[J., Netrapalli, Sanghavi, STOC’13] [J., Dhillon, Arxiv’13]

Summary

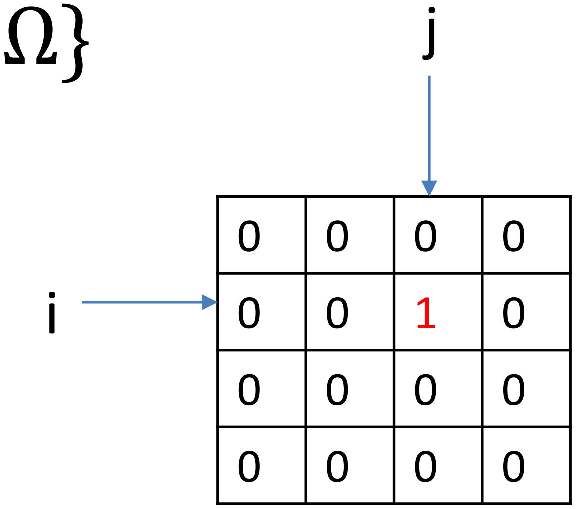
- AltMin: Power method with Error Term
- Error term bounded using concentration assumption
- Lower bound on the “correct” term by initialization assumption
- Geometric convergence

Low-rank Matrix Completion

$$\min_{W} \; Error_{\Omega}(W) = \sum_{(i,j)\in\Omega} \left(W_{ij} - W_{ij}^{*}\right)^{2}$$

$$s.t \qquad \mathbf{rank}(W) \leq k$$

- Ω: set of known entries
- $\mathbb{A} = \{\mathrm{A}_{\mathrm{ij}}, \mathrm{ij} \in \Omega\}$
  - $A_{ij} = e_i e_j^T$



$$\begin{array}{c|cccc} & & & j & \\ \hline & 0 & 0 & 0 & 0 \\ i & 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 \end{array}$$

Our Results

- Show Property 1, 2 of General Theorem
- Assumptions:
  - Ω is sampled uniformly, i.e.,
    $$|\Omega| = O(k^7 \beta^6 (d_1 + d_2) \log(d_1 + d_2)\ )$$
    - $\beta = \sigma_1/\sigma_k$
  - W_*: rank-k “incoherent” matrix
    - Most of the entries are similar in magnitude
- Initialization property follows by [KMO’09]
- Decay property follows by using incoherence of U_*, V_*, U_t (recall that $W_* = U_* \Sigma_* V_*^T$)
  - Challenge: Show that V_{t+1} is incoherent

[J., Netrapalli, Sanghavi, STOC’13]

Alternating Minimization	Trace-Norm Minimization
$\|\|W_* - UV^T\|\|_F \leq \epsilon \|\|W\|\|_F$ after $O(\log\left(\frac{1}{\epsilon}\right))$ steps	Requires $O(\log\left(\frac{1}{\epsilon}\right))$ steps
Each step require solving 2 least squares problems	Require Singular value decomposition
Intermediate iterate always have rank-k	Intermediate iterates can have rank larger than k
Assumptions: random sampling and incoherence	Similar assumption
$m = O(k^7 \beta^6 d \log(d)\ )$ $d = d_1 + d_2$	$m = O(k\ d \log(d)\ )$ $d = d_1 + d_2$

Comparison to Keshavan'12

- Independent of our work
- Show results for Matrix Completion
  - Alternating minimization method
  - Similar linear convergence

$$|\Omega| = O(k\beta^8(d_1 + d_2)\log(d_1 + d_2))$$

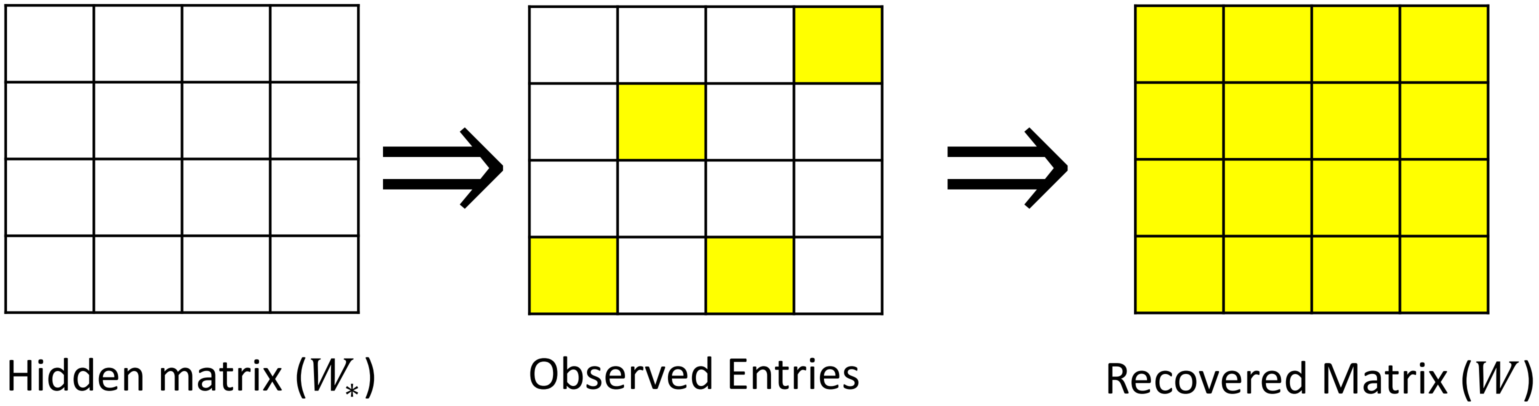
  - Ours:

$$|\Omega| = O(k^7\beta^6(d_1 + d_2)\log(d_1 + d_2))$$

[J., Netrapalli, Sanghavi, STOC'13]

Low-rank Matrix Sensing

Matrix Completion:

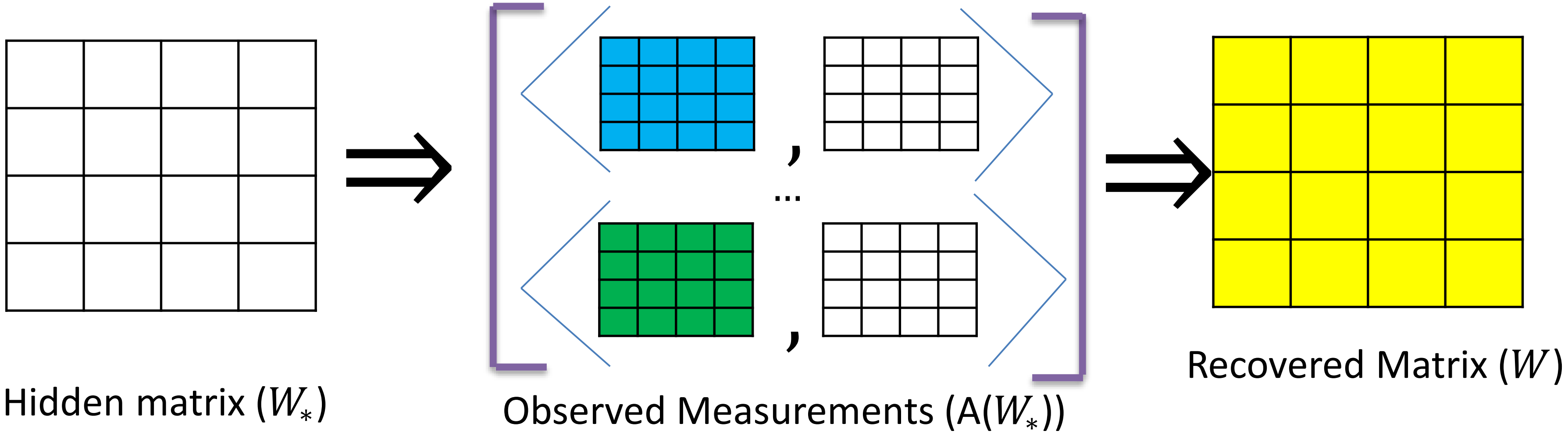


Hidden matrix (W_*)

Observed Entries

Recovered Matrix (W)

Matrix Sensing:

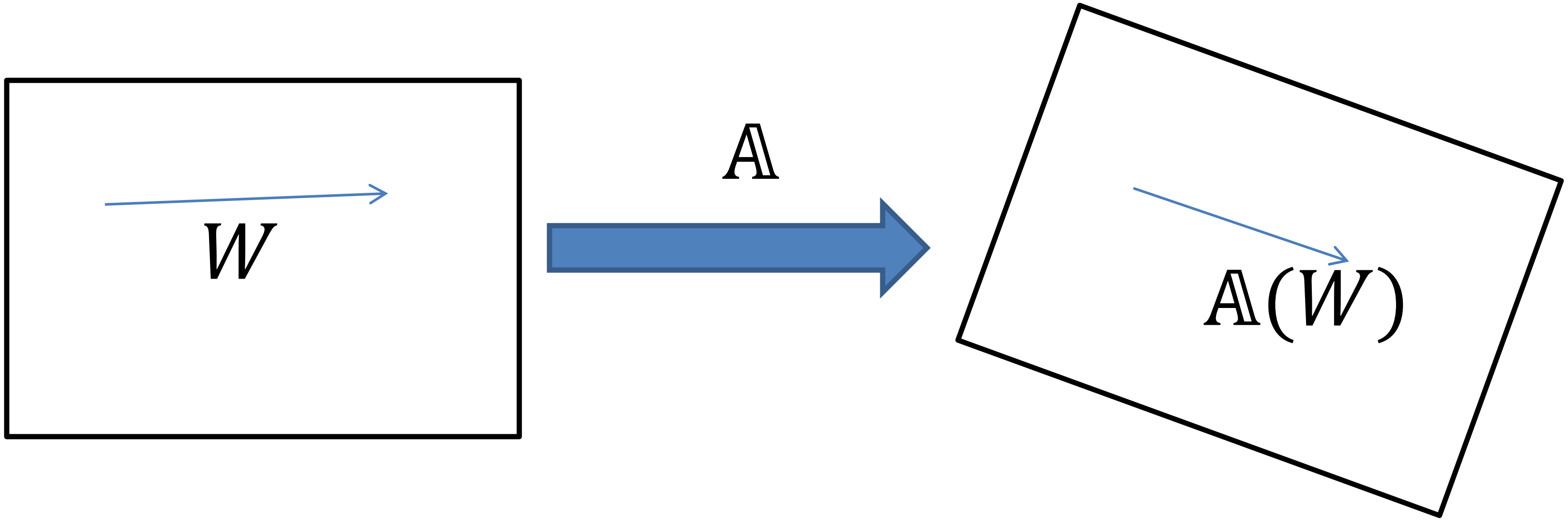


Hidden matrix (W_*)

Observed Measurements (A(W_*))

Recovered Matrix (W)

Restricted Isometry Property



- For all rank-k matrix (W):

$$(1-\delta_k)||W||_F^2 \le ||\mathbb{A}(W)||_2^2 \le (1+\delta_k)||W||_F^2$$

- Examples:
  - $\mathbb{A}$: sampled from multivariate normal distribution
  - $\mathrm{m} = O(\frac{k}{\delta_k^2}(d_1+d_2)\log(d_1+d_2))$

Alternating Minimization	Trace-Norm Minimization
$\Vert W_* - UV^T \Vert_F \leq \epsilon \, \Vert W \Vert_F$ after $O(\log\left(\frac{1}{\epsilon}\right))$ steps	Requires O$(\log(\frac{1}{\epsilon}))$ steps
Each step require solving 2 least squares problems	Require Singular value decomposition
Intermediate iterate always have rank-k	Intermediate iterates can have rank much higher than k
Assumptions: "random" measurement matrix A	Similar assumption
$m = O(k^3 \beta^2 d \log(d)\ )$ $d = d_1 + d_2$	$m = O(k\ d \log(d)\ )$ $d = d_1 + d_2$

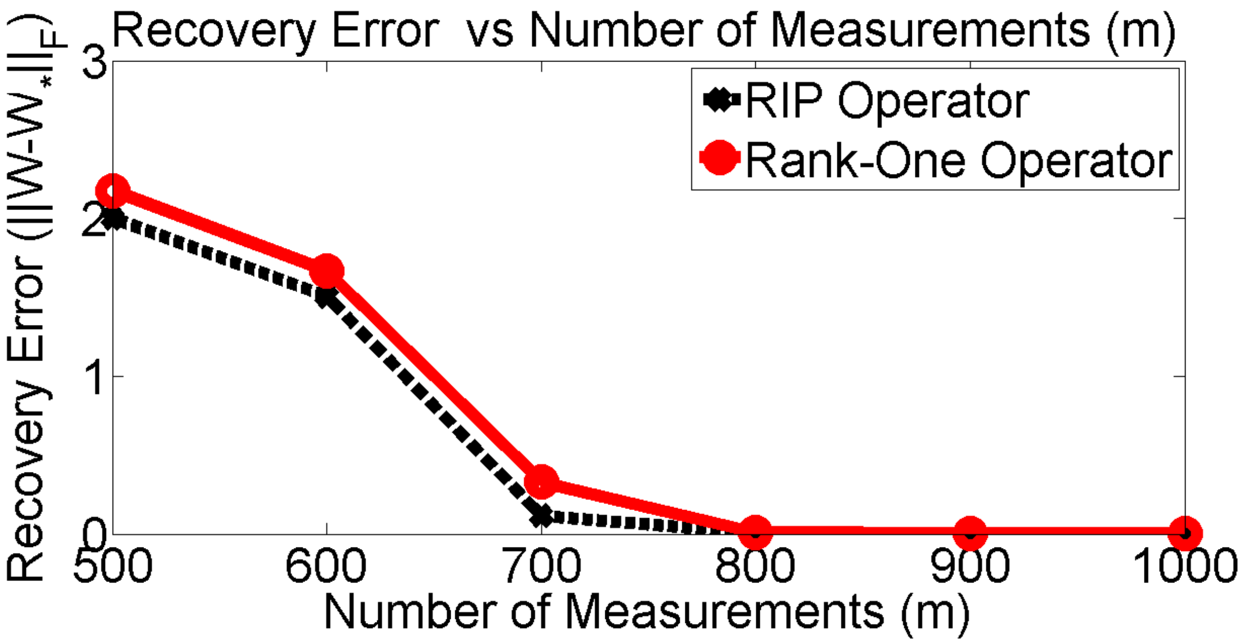
[J., Netrapalli, Sanghavi, STOC'13]

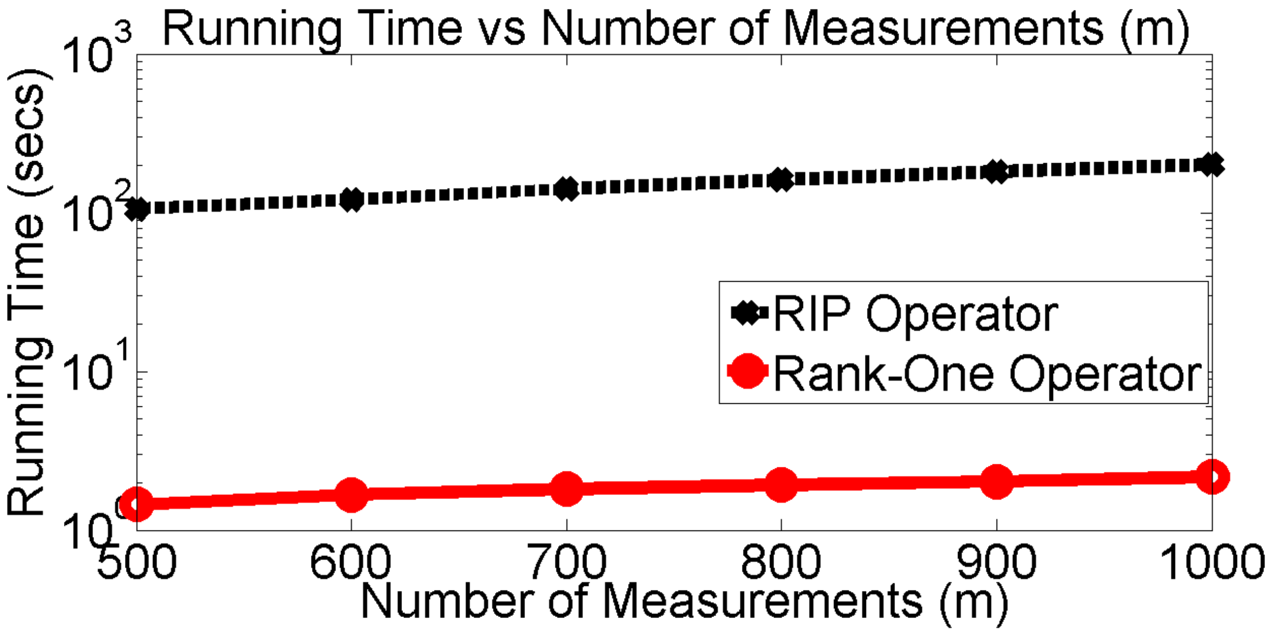
Rank-One Measurements for Matrix Sensing

$$\mathbb{A}(W_*) = \begin{bmatrix} \langle A_1, W \rangle \\ \langle A_2, W \rangle \\ \vdots \\ \langle A_m, W \rangle \end{bmatrix} = \begin{bmatrix} x_1^T W_* y_1 \\ x_2^T W_* y_2 \\ \vdots \\ x_m^T W_* y_m \end{bmatrix}$$

- $A_i = x_i y_i^T, \qquad x_i, y_i \sim N(0, I), \forall i$
- Property 1, 2 of General Theorem satisfies:
  - $m \geq C\, k^4 \beta^2 (d_1 + d_2) \log(d_1 + d_2)$
- Significantly more efficient signal acquisition
- Drawback: Not universal
  - Requires new $\mathbb{A}$ for each signal W_*

[J., Dhillon, Arxiv'13]


Recovery Error vs Number of Measurements (m)
Recovery Error ($||W-W_*||_F$)
3
2
1
0
500
600
700
800
900
1000
Number of Measurements (m)
RIP Operator
Rank-One Operator


Running Time vs Number of Measurements (m)
Running Time (secs)
10^3
10^2
10^1
10^0
500
600
700
800
900
1000
Number of Measurements (m)
RIP Operator
Rank-One Operator

Removing Condition Number Dependence

- Main challenge:
  - Require “good” initialization
    - Necessary, even power method requires it
  - The largest subspace dominates initialization step
  - Solution: “remove” one subspace at a time

Stagewise Altmin

- Stage r = 1 to k
  - Initialize by projected gradient $P_r(W - \eta \mathbb{A}^{\mathrm{T}}(\mathbb{A}(\mathrm{W}) - \mathrm{b}))$
  - $W = U_0 \Sigma_0 V_0^T$
  - For t=1 to T

$$V^{t+1} = \min_V ||b - \mathbb{A}(U^t V^T)||_2^2$$

$$U^{t+1} = \min_U ||b - \mathbb{A}(U(V^{t+1})^T)||_2^2$$

  - $W = U_T V_T^T$
  - End-Stage

[J., Netrapalli, Sanghavi, STOC'13]

General Result

- Let $\mathbb{A}$ satisfy Property 1,2
  - Concentration:
    $||A_i v_p v_q^T A_i^T - \langle v_p, v_q \rangle \, I||_2 \le \delta \, ||v_p||_2 ||v_q||_2$
    $||A_i^T u_p u_q^T A_i - \langle u_p, u_q \rangle I||_2 \le \delta \, ||u_p||_2 ||u_q||_2$
  - $\delta \le \frac{1}{10k^2}$
- After $T = O\left(\log\left(\frac{||W_*||_F}{\epsilon}\right)\right)$:

$$||W_T - W_*||_2 \le \epsilon$$

Results

- Results for other problems
  - RIP based Matrix Sensing:

    $$m = O(k^4(d_1 + d_2)) \log(d_1 + d_2))$$

  - Rank-one Operator based Matrix Sensing

    $$m = O(k^5(d_1 + d_2)) \log(d_1 + d_2))$$

- Not applied to Matrix Completion
  - Challenge: Incoherence for projected gradient step

[J., Netrapalli, Sanghavi, STOC’13] [J., Dhillon, Arxiv’13]

Phase Retrieval

$$y_i = |\langle a_i, x_* \rangle|, \qquad 1 \leq i \leq m,$$
$$x_* \in C^n$$

- Only magnitudes of measurements available
- Applications in several areas
- Recent theoretical results
  - – Assume $a_i \sim N(0, I)$
  - – PhaseLift: trace norm based relaxation

$$y_i^2 = a_i^T x_* x_*^T a_i$$

  - – Relax $xx^T \rightarrow X$

PhaseLift

$$\begin{aligned} &\min ||X||_* \\ &s.t.\;\; y_i^2 = \langle X, a_i a_i^T \rangle \\ &\qquad\quad X \succcurlyeq 0 \end{aligned}$$

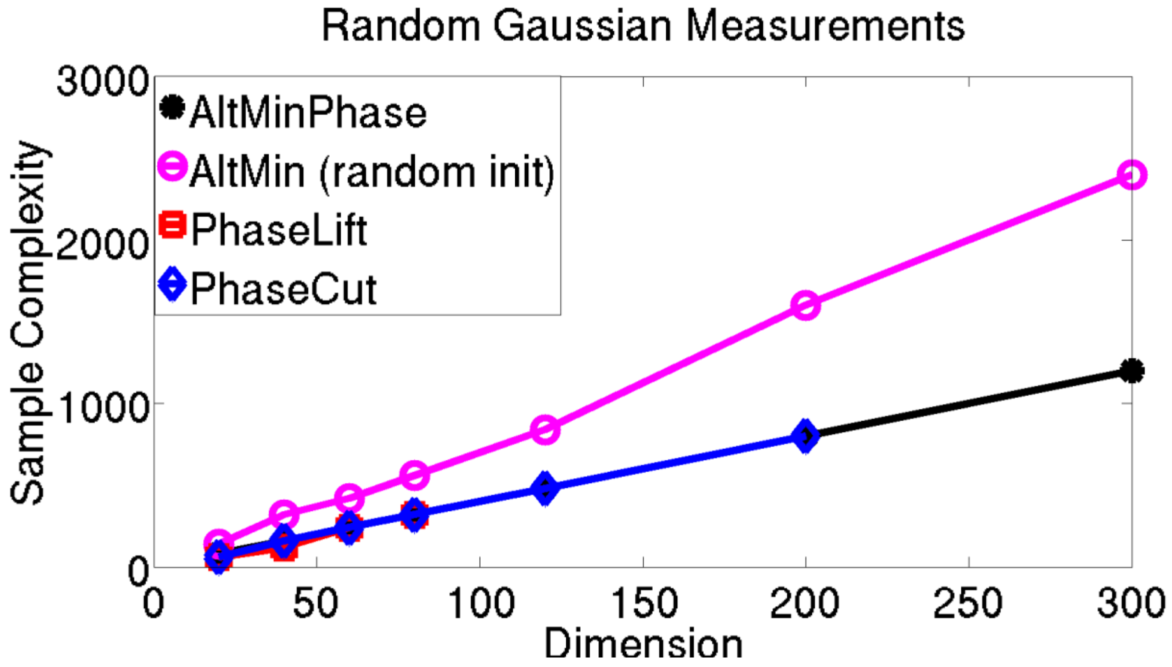
- Exact recovery if $m = O(n \log n)$ [CTV11]
- Later improved to $m = O(n)$ [CL12]
- Optimization procedure is computationally expensive

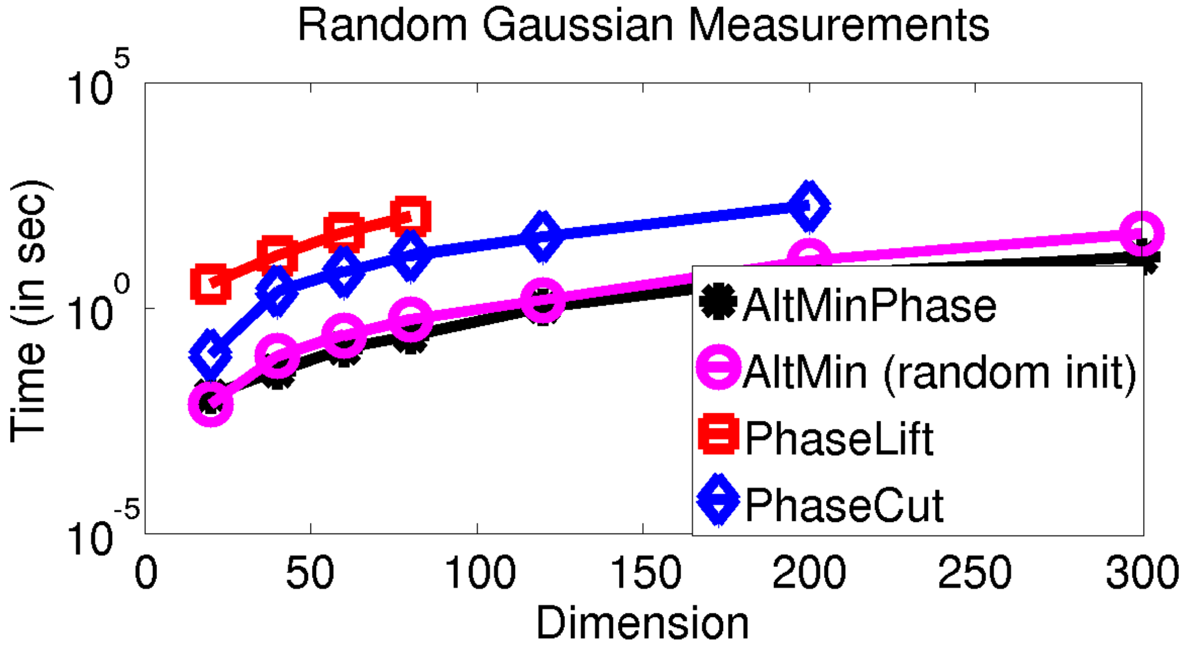
Alternating Minimization

$$\min_{P,x} ||Py - Ax||_2^2$$

- P: phase of Ax_*
- Alternating minimization:
  - $P_t = Phase(Ax_t)$
  - $x_{t+1} = (A^T A)^{-1} A^T P_t y$
- Initialization: largest singular vector of $\sum_i y_i^2 a_i a_i^T$
- Exact recovery if $m = \Omega(n \log^3 n)$

[Netrapalli, J., Sanghavi, Arxiv’13]





[Netrapalli, J., Sanghavi, Arxiv'13]

Summary

$$\begin{aligned} \min_{W} \quad & ||\mathbb{A}(W) - b||_2^2 \\ s.t. \quad & \mathbf{rank}(W) \leq k \end{aligned}$$

$k \ll$ dimensions(W)

- Popular approach: trace-norm relaxation

$$\begin{aligned} \min_{W} \quad & ||\mathbb{A}(W) - b||_2^2 \\ s.t. \quad & ||\boldsymbol{W}||_* \leq \lambda(k) \end{aligned}$$

  - $||W||_*$: sum of singular values
  - Convex formulation
  - Proven to solve rank problem
    - assumptions on error function
  - Non-smooth optimization problem: doesn’t scale well

Summary

$$\min_{W} \quad Error(W) = ||\mathbb{A}(W) - b||_2^2$$
$$s.t. \quad \mathbf{rank}(W) \leq k$$

- Alternating minimization: empirically successful
  - $W = UV^T$

$$V^{t+1} = \min_{V} \quad Error(U^t, V), U^{t+1} = \min_{U} Error(U, V^{t+1})$$

  - Computationally efficient
  - Prone to local minima
    - Little work on convergence guarantees

Summary

- Provide generic conditions for which AltMin works well
  - Provide an enhanced stage-wise AltMin procedure to remove condition number dependence
- Provide results for:
  - Low-rank Matrix Completion
  - Low-rank Matrix Sensing
- Provide convergence to the global optima guarantees
  - Use similar assumptions as existing methods
  - But slightly worse no. of measurements (or entries)

Future Work

- Optimal scaling for k in the sample complexity bounds
- Matrix completion: remove dependence on β: condition no. of W_*
- Application of our technique to other problems:
  - Robust PCA
  - Non-negative Matrix Approximation

Thank You!!!

Questions?