Provable Non-convex Optimization for ML

Prateek Jain

Microsoft Research India

Overview

$$\min_{X} f(X)$$
$$s.t.\ rank(X) \leq r$$

- Projected gradient descent
- Alternating Minimization

Our Results

- RIP/RSC based Linear Regression

$$\min_X ||A(X) - b||_2^2 \quad s.t. \quad rank(X) \leq r$$

  - $A(\cdot)$: RIP operator
  - $A(\cdot)$: RSC operator (statistical setting)

- Matrix Completion

$$\min_X ||P_\Omega(X - M)||_F^2 \quad s.t. \quad rank(X) \leq r$$

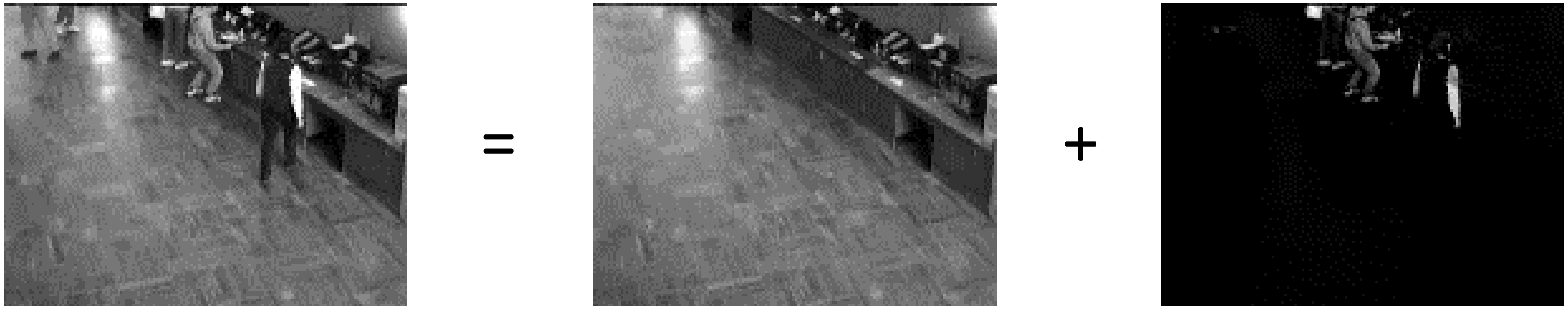
  - Ω: randomly sampled, M: incoherent matrix

- Non-convex Robust PCA

$$\min_X ||M - X||_0^2 \quad s.t. \quad rank(X) \leq r$$

  - $M = L + S$, L: low-rank incoherent matrix, S: sparse matrix

Foreground/Background Separation



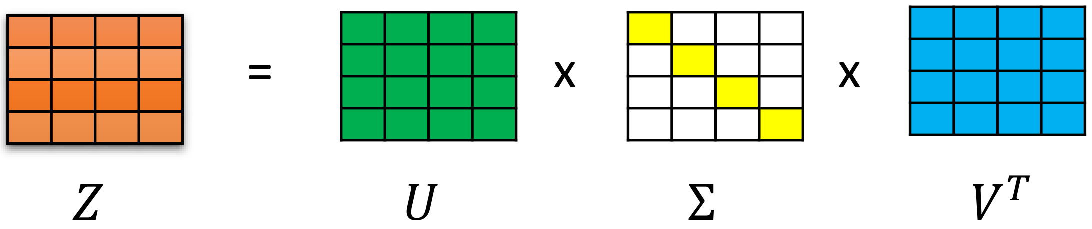
Non-convexity of Low-rank manifold

$$0.5\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} + 0.5\begin{bmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}0.5 & 0 & 0\\ 0 & 0.5 & 0\\ 0 & 0 & 0\end{bmatrix}$$

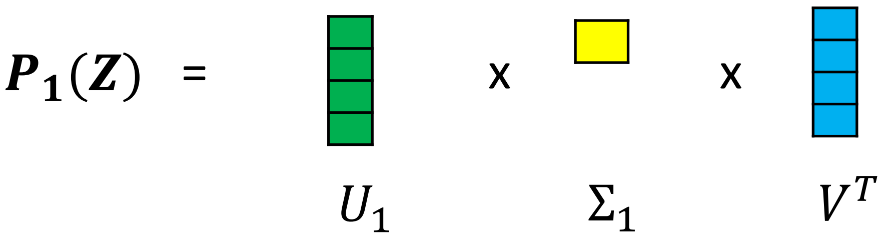
Projection onto set of Low-rank Matrices

- Non-convex projections: NP-hard in general
- But $P_r(Z)$ can be computed efficiently:

$$Z = U\Sigma V^T$$



- $P_r(Z) = U_r\Sigma_r V_r^T$



Convex-projections vs Non-convex Projections

- For non-convex sets, we only have:

$$\forall Y \in C, \qquad ||P_r(Z) - Z|| \leq ||Y - Z||$$

  - 0-th order condition

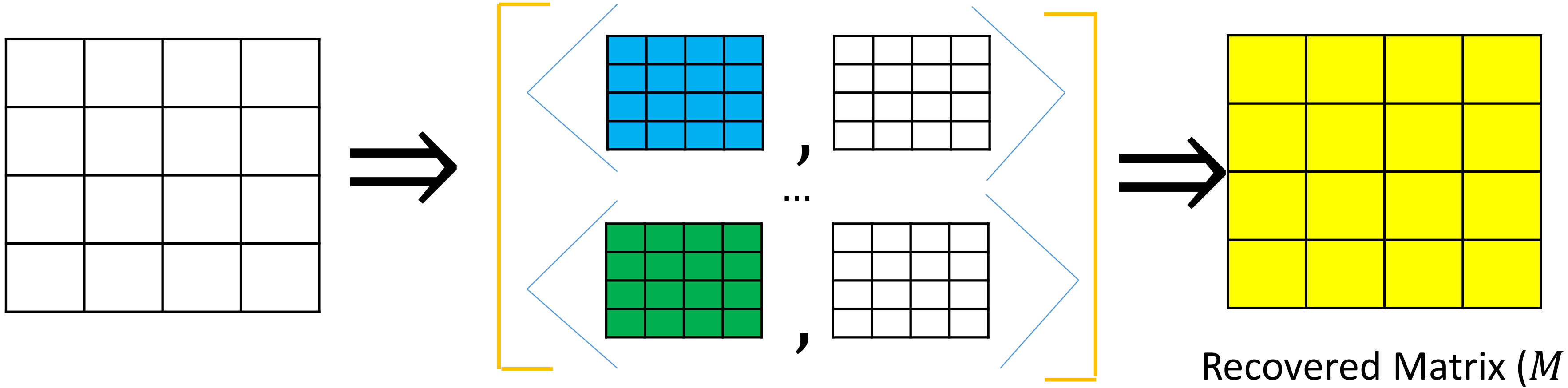
- But, for projection onto convex set C:

$$\forall Y \in C, \qquad ||Z - P_C(Z)||^2 \leq \langle Y - Z, P_C(Z) - Z \rangle$$

  - 1-st order condition

- 0 order condition sufficient for convergence of Proj. Grad. Descent?
  - In general, NO ☹
  - But, for certain *specially structured* problems, YES!!!

Low-rank Matrix Regression



Hidden matrix (M)

Observed Measurements (A(M))

Recovered Matrix (M)

Matrix Linear Regression

$$\mathbb{A}(M) = b$$

- $\mathbb{A}: \mathbf{R}^{\boldsymbol{n \times n}} \rightarrow \boldsymbol{R^d}$
  - Linear operator
  - $\mathbb{A} = \{\mathbf{A_1}, \mathbf{A_2}, \dots, \boldsymbol{A_d}\}$

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

- Optimization Version:

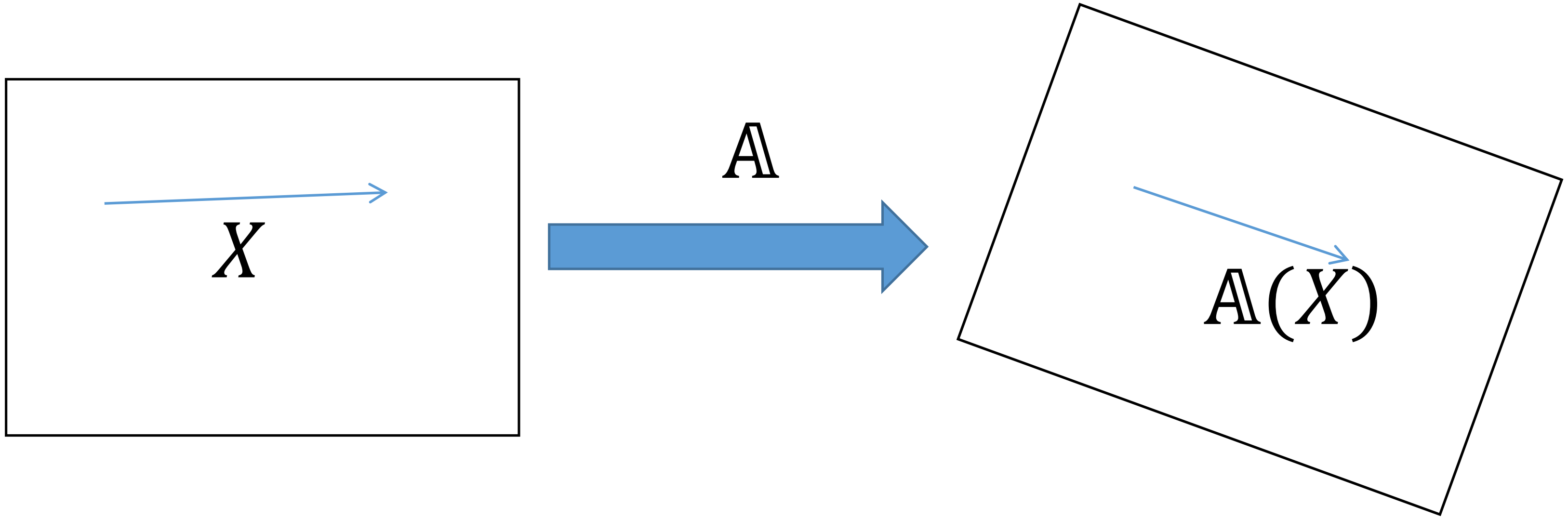
$$\min_X ||\mathbb{A}(X) - b||_2^2$$
$$s.t. \;\; rank(X) \leq r$$

Low-rank Matrix Estimation

$$\min_X ||\mathbb{A}(X) - b||_2^2$$
$$s.t.\ \ rank(X) \le r$$

- NP-hard in general
    - Hard to even approximate within $\log(n + d)$ [Meka, J., Caramanis, Dhillon'08]
- Tractable solutions under certain conditions
    - RIP conditions

Restricted Isometry Property



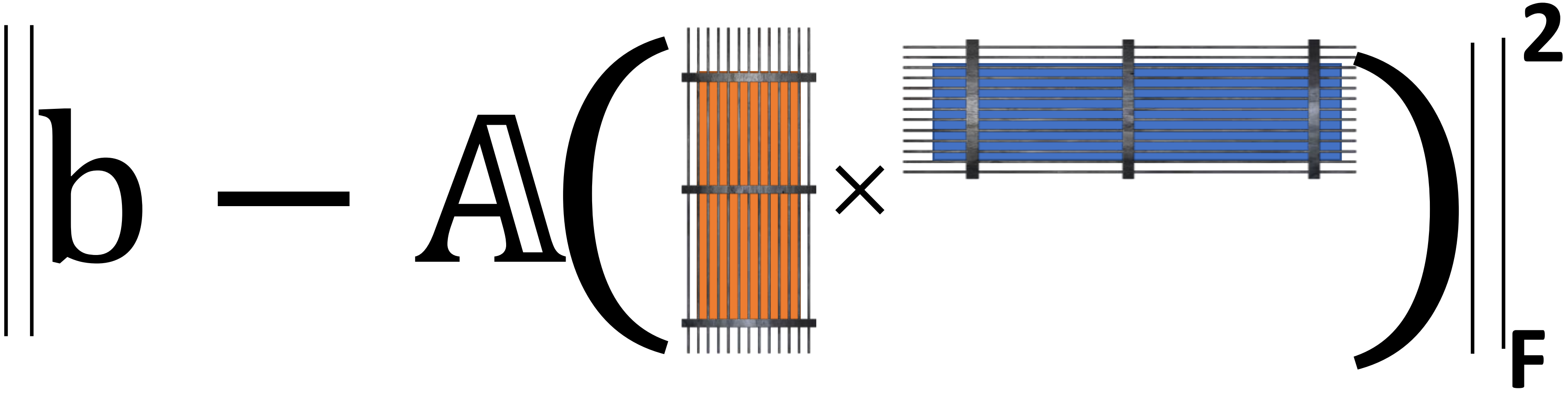
- For all rank-r matrix (X):
  $$(1-\delta_r)||X||_F^2 \leq ||\mathbb{A}(X)||_2^2 \leq (1+\delta_r)||X||_F^2$$
- Examples:
  - $\mathbb{A}$: sampled from multivariate normal distribution
  - $\mathrm{m} = O(\frac{r}{\delta_r^2} n \log n)$

Approach 1: Trace-norm minimization

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

- $||X||_*$: sum of singular values
- Provable recovery of M
  - RIP based Matrix Sensing: [Recht, Fazel, Parrilo'07]
  - For Gaussian distributed samples: $O(r\, n \log n)$
- However, convex optimization methods for this problem don't scale well
  - SVD computation per step
  - Intermediate iterates can have rank much larger than "r"

Approach 2: Alternating Minimization



$$M \qquad \cong \qquad U \qquad \times \qquad V^T$$

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

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

- Provable convergence to M [J., Netrapalli, Sanghavi'13]
  - RIP property satisfied
  - Gaussian distribution: $O(nr^3 \log n)$
    - Suboptimal bounds

Approach 3: Projected Gradient based Methods

- $X_0 = 0$
- For t=1:T

$$X_t = P_r\left(X_{t-1} - \eta \mathbb{A}^{\mathrm{T}}(\mathbb{A}(X_{\mathrm{t-1}}) - \mathrm{b})\right)$$

- $P_r(Z)$: projection onto set of rank-r projection
- Singular Value Projection
- Several other variants exist (ADMiRA [Lee, Bresler’09])

Guarantees

- SVP converges to global optima
  - $\delta_{2r} \leq 1/3$
  - For Gaussians: $O(r\, n \log n)$
  - Info. theoretically optimal
- Noisy case analysis also available
- Analysis: a simple extension of analysis of iterative hard thresholding [Garg, Khandekar'08]

[J., Meka, Dhillon 2009]

Extensions

- Optimize general f

$$\min_{X} f(X)$$
$$s.t.\ rank(X) \leq r$$

- Assume RSC-style condition: $\forall X, s.t. rank(X) \leq r$

$$(1+\delta_r)I \succcurlyeq \nabla^2 f(X) \succcurlyeq (1-\delta_r)I$$

- SVP converges to the optima for such a case as well [J., Kar, Tewari'14]
- Extensions to the "statistical setting" as well

Summary

$$\min_X f(X)$$
$$s.t.\ rank(X) \leq r$$

- Projected gradient descent converges to the global optima
  - Assuming certain RSC/RIP style conditions
- Standard matrix sensing:
  - Information theoretic optimal bounds
- Analysis:
  - Only requires 0-th order property

$$||Y - Z|| \geq ||P_r(Z) - Z||, \qquad \forall Y \in C$$

Low-rank Matrix Completion

Low-rank Matrix Completion

users

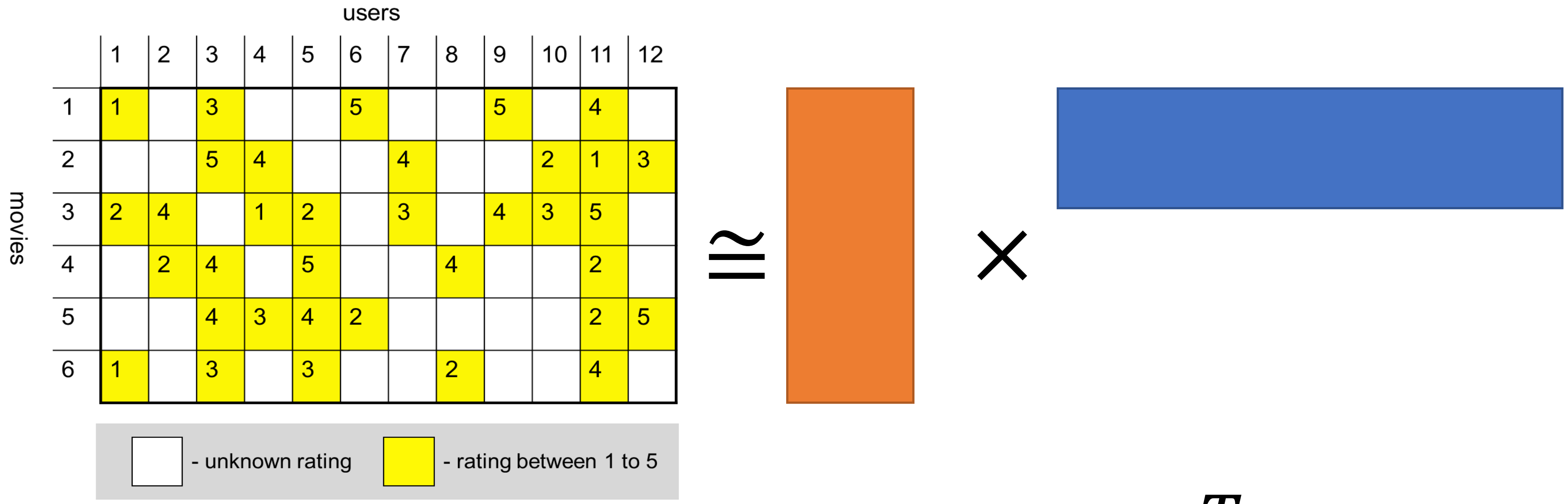
movies

	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



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

$$\cong \quad \times$$

$$M \qquad \cong U \quad \times \quad V^T$$

- M: characterized by U, V
- No. of variables:
  - U: $\mathrm{n} \times r = nr$
  - V: $n \times r = nr$

- DoF: nr

Low-rank Matrix Completion

$$\min_{X} \; Error_{\Omega}(X) = \sum_{(i,j)\in\Omega} \left(X_{ij} - M_{ij}\right)^2 \quad = ||P_{\Omega}(X - M)||_F^2$$

$$s.t \qquad \mathbf{rank}(X) \le r$$

- Ω: set of known entries
- $P_{\Omega}(X)_{ij} = X_{ij}, (i,j) \in \Omega$
    - 0 otherwise

1			
		2	
		1	
	4		

M

$\Rightarrow$

1	0	0	0
0	0	2	0
0	0	1	0
0	4	0	0

$P_{\Omega}(M)$

Approach 1

- Convex relaxation: Replace $rank(X)$ with $||X||_*$
- Provably recovers M if:
  - M:rank-r incoherent matrix (non-spiky matrix)
    - $M = U\Sigma V^T, ||U^i||_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}}$
  - Ω: sampled uniformly at random and $|\Omega| \geq O(r\, n \log^2 n)$
- Worst Computation time: $O(n^3)$
- Refs: [Candes, Recht 2008], [Candes, Tao 2008], [Recht 2010]

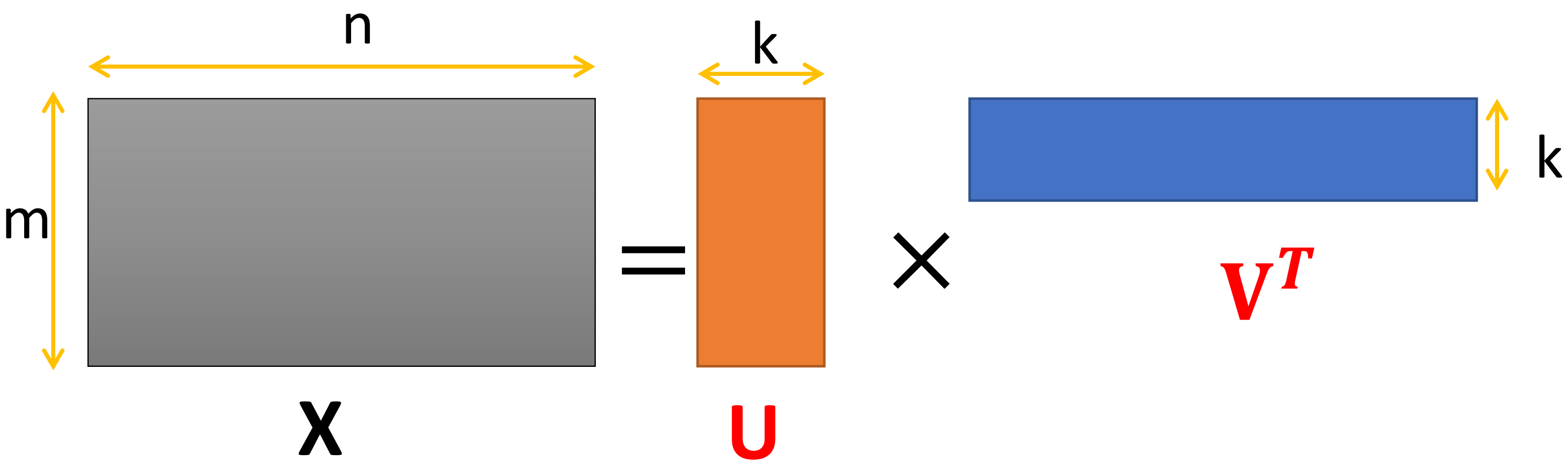
Incoherence?

Alternating Minimization

$$\min_{X} \; Error_{\Omega}(X) = \sum_{(i,j)\in\Omega} \left(X_{ij} - M_{ij}\right)^2$$

$$s.t \qquad \mathbf{rank}(X) \le r$$

- If X has rank-k:



$$V^{t+1} = \min_{V} Error_{\Omega}(U^{t}, V)$$

$$U^{t+1} = \min_{U} Error_{\Omega}(U, V^{t+1})$$

Initialization [JNS’13]

- Initialization:
  - SVD($P_\Omega(M)$, r)

0	3	0
2	5	0
0	0	2

$P_\Omega(M)$

Results [JNS'13]

- Assumptions: Ω: set of known entries
    - Ω is sampled uniformly s.t. $|\Omega| = O(k^7 n \log n \ \beta^6)$
        - $\beta = \sigma_1 / \sigma_r$
    - M: rank-k "incoherent" matrix
        - Most of the entries are similar in magnitude
- Then, $||M - UV^T||_F \leq \epsilon$ after only $O(\log\left(\frac{1}{\epsilon}\right))$ steps
- Improved analysis by Hardt-Wooters'14

Proof Sketch

- Assume Rank-1 case, i.e., $M = u^* {v^*}^T$
- Fixing u, update for v is given by:

$$v = \arg\min_v \sum_{(i,j)\in\Omega} \left(u_i v_j - u_i^* v_j^*\right)^2$$

$$v_j = \frac{\sum_{(i,j)\in\Omega} u_i u_i^*}{\sum_{(i,j)\in\Omega} u_i^2} \cdot v_j^*$$

- If $\Omega = [\mathrm{m}]\mathrm{x}[\mathrm{n}]$,

$$v_j = \langle u, u^* \rangle v_j^*$$

- Power method update!

Proof Sketch

$$v = \underbrace{M^T u}_{\text{Power Method Term}} - \underbrace{B^{-1}(B < u, u^* > -C)v^*}_{\text{Error Term}}$$

Problems:

1. Show error term decreases with iterations
2. Also, need to show “incoherence” of each v

Tools:

1. Spectral gap of random graphs
2. Bernstein-type concentration bounds

Bernstein?

Power Method?

Approach 3: Singular Value Projection

$$\text{Sample } \Omega$$

$$X_t = P_r(X_t - P_\Omega(X_t - M))$$

- Previous analysis applies only if $P_\Omega(\cdot)$ satisfies RIP
  - RIP holds but *only* for incoherent matrices
  - $X_t - M$: need not be incoherent

1	1	1
1	1	1
1	1	1

-

1	1	1
1	1	1
.5	.5	.5

=

0	0	0
0	0	0
.5	.5	.5

- Require: $X_t \to M$ in L_∞ norm

Guarantees

- Our approach:
  - Analyze $||X_t - M||_\infty$ instead!
  - At first seems tricky: $P_r(\cdot)$ optimal only w.r.t. spectral norm or Frobenius norm
- Three key tricks:
  - Use a Taylor series expansion technique by [Erdos et al’ 2013]
  - Convert L_∞-norm error bounds into $|| \cdot ||_2$ error bounds
  - Analyze $||H^a u||_\infty$

Setting up the proof (Rank-one Case)

$$\begin{aligned} X_t &= P_1\big(X_{t-1} - P_\Omega(X_{t-1} - M)\big) \\ &= P_1\big(M + X_{t-1} - M - P_\Omega(X_{t-1} - M)\big) \\ &= P_1\big(M + E_t - P_\Omega(E_t)\big) \\ &= P_1(M + H_t) \end{aligned}$$

- $H_t = E_t - P_\Omega(E_t)$
- $E[H_t] = 0$: assuming Ω is independent of E_t
- $E[H_t(i,j)^2] \leq \frac{||M - X_{t-1}||_\infty^2}{p}$
- $||H_t||_2 \leq \delta n ||M - X_{t-1}||_\infty$ (assuming $p \geq \log n / \delta^2$)
- $||M - X_t||_2 \leq 2||H_t||_2$ (but only spectral norm bound)

Matrix Bernstein?

Matrix Perturbation?

Davis-Kahan?

Key Step 1

- Let v, λ be the largest eigenvector/value of $M + H_t$

$$(M + H_t)v = \lambda v$$

$$\left(I - \frac{H_t}{\lambda}\right)v = \frac{Mv}{\lambda}$$

$$v = \left(I - \frac{H_t}{\lambda}\right)^{-1} \frac{Mv}{\lambda} = \frac{Mv}{\lambda} + \sum_{a=1}^{\infty} \left(\frac{H_t}{\lambda}\right)^a \frac{Mv}{\lambda}$$

- $X_t = \lambda v v^T$

$$M - X_t = M - \lambda v v^T$$

$$= M - M\ \frac{vv^T}{\lambda} M - \sum_{a\geq 0, b\geq 0, a+b\geq 1}^{\infty} \left(\frac{H_t}{\lambda}\right)^a \frac{Mvv^T M^T}{\lambda} \left(\frac{H_t}{\lambda}\right)^b$$

Key Step 2

$M = u^* {u^*}^T$

$$||M - X_t||_\infty$$
$$\leq ||M - M\,\frac{vv^T}{\lambda}M||_\infty + \sum_{a\geq 0, b\geq 0, a+b\geq 1}^{\infty} \left|\left(\frac{H_t}{\lambda}\right)^a \frac{Mvv^TM^T}{\lambda}\left(\frac{H_t}{\lambda}\right)^b\right|_\infty$$

- $M = u^* {u^*}^T$

$$||M - M\,\frac{vv^T}{\lambda}M||_\infty \leq \max_{i,j} e_i^T u^* \left(1 - {u^*}^T \frac{vv^T}{\lambda} u^*\right) {u^*}^T e_j$$
$$\leq \max_{i,j} |e_i^T u^*||e_j^T u^*|\,|1 - \ \left({u^*}^T v\right)^2/\lambda|$$

$$\leq \frac{\mu^2}{n}4||H_t||_2 \leq 8\mu^2\delta||M - X_{t-1}||_\infty$$

Key Step 3

- Need to bound
$$||(H_t)^a u^*||_\infty$$
- $H_t = M - X_{t-1} - P_\Omega(M - X_{t-1})$
- $(H_t)^a$ has several correlated entries
  - Use technique of [Erdos et al'2013]
  - Intuitively, counts the total no. of paths between any pair of nodes
- Bound: $||(H_t)^a u^*||_\infty \leq \frac{\mu}{\sqrt{n}}(\delta||M - X_{t-1}||_\infty c \log n)^a$
- Sum up terms to bound $||M - X_t||_2$

Guarantee for SVP

- At t-th step :

$$||M - X_t||_\infty \leq .5\ ||M - X_{t-1}||_\infty$$

- After $\log(\frac{\mu}{\epsilon})$ steps: $||M - X_t||_\infty \leq \epsilon$
- Sample complexity: $|\Omega| \geq nr^2\mu^2 \left(\frac{\sigma_1}{\sigma_r}\right)^2 \log^2 n \log\frac{1}{\epsilon}$
  - Dependence on condition number!!!

[Netrapalli, J.'14]

Stagewise-SVP

```
• $X_0 = 0$
• For k=1…r
    • For t=1:T
        • $X_t = P\_r(X_{t-1} - P_\Omega(X_{t-1} - M))$
    • End For
    • $X_0 = X_T$
• End For
```

Guarantees

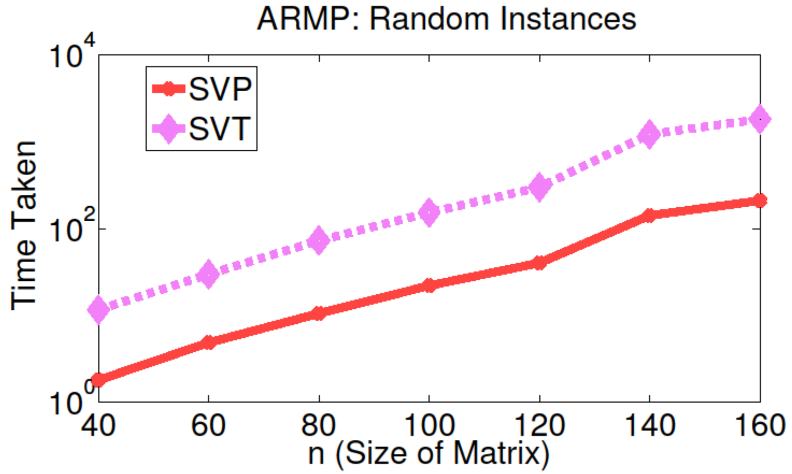
- After t-th step of k-th stage:

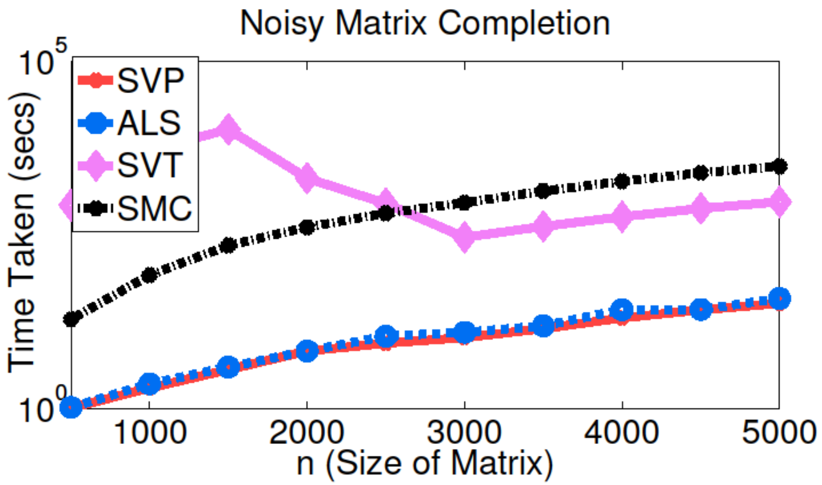
$$||M - X_t||_\infty \leq \frac{2\mu^2 r}{n}(\sigma_{k+1} + \left(\frac{1}{2}\right)^t \sigma_k)$$

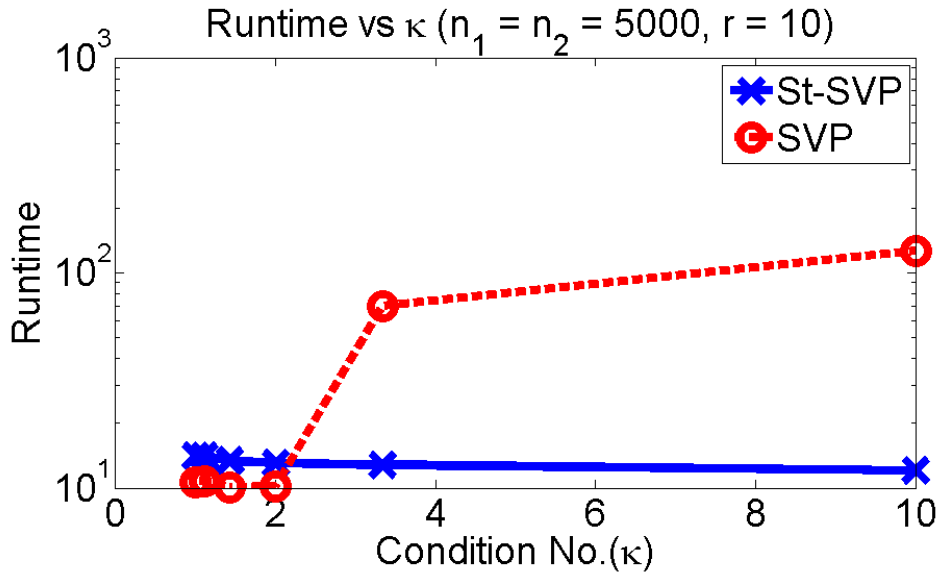
- M: rank-r i.e. $\sigma_{r+1} = 0$
- After $\mathrm{T} = \log(\frac{1}{\epsilon})$ steps of r-th stage: $||M - X_T||_\infty \leq \epsilon$
- Sample complexity: $|\Omega| \geq nr^4\mu^2 \log n \log 1/\epsilon$
- Computation complexity: $O(nr^6\mu^2 \log n\ \log\frac{1}{\epsilon})$
  - Linear in n
  - No explicit dependence on σ_1/σ_r

[Netrapalli, J.'14]

Simulations



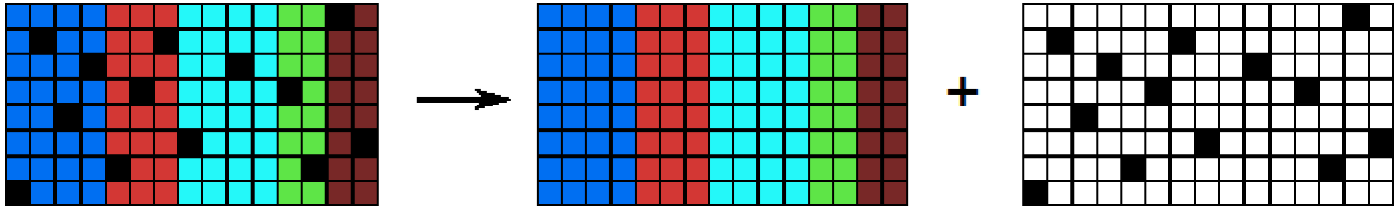




Summary

- Study matrix completion problem
- Projected gradient descent works!
- With some tweaks, obtain a nearly linear time algorithm for matrix completion
    - No explicit dependence on condition number
- Future work:
    - Remove dependence on ϵ for sample complexity
    - AltMin: remove condition no. dependence using similar techniques?

Robust Principal Component Analysis



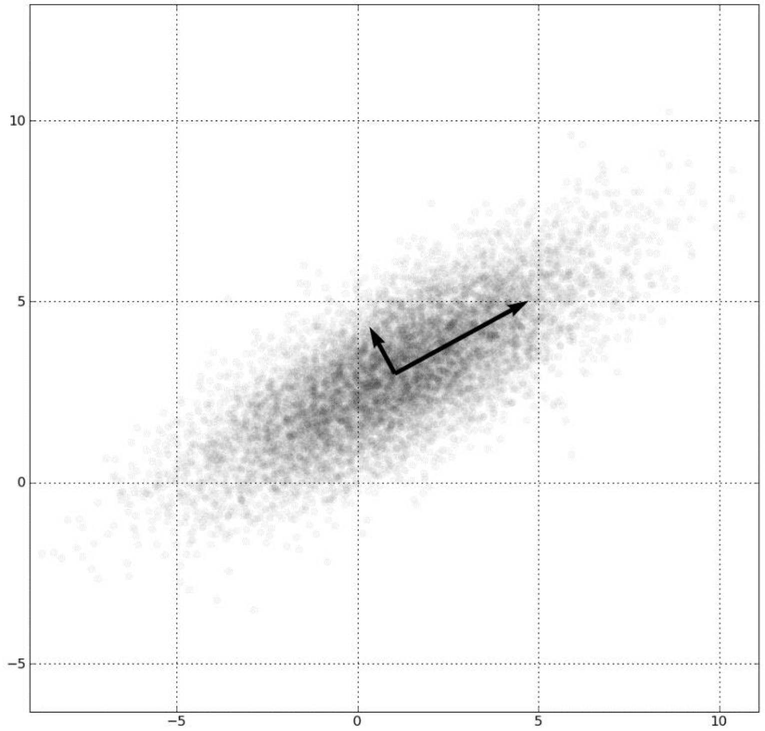
$$M \quad = \quad L \quad + \quad S$$

Principal Component Analysis

- $X = [x_1 \; x_2 \ldots x_n]$

- PCA: find best rank-r approx. of X
    - Top r −singular components of X
    - $X_r = P_r(X)$

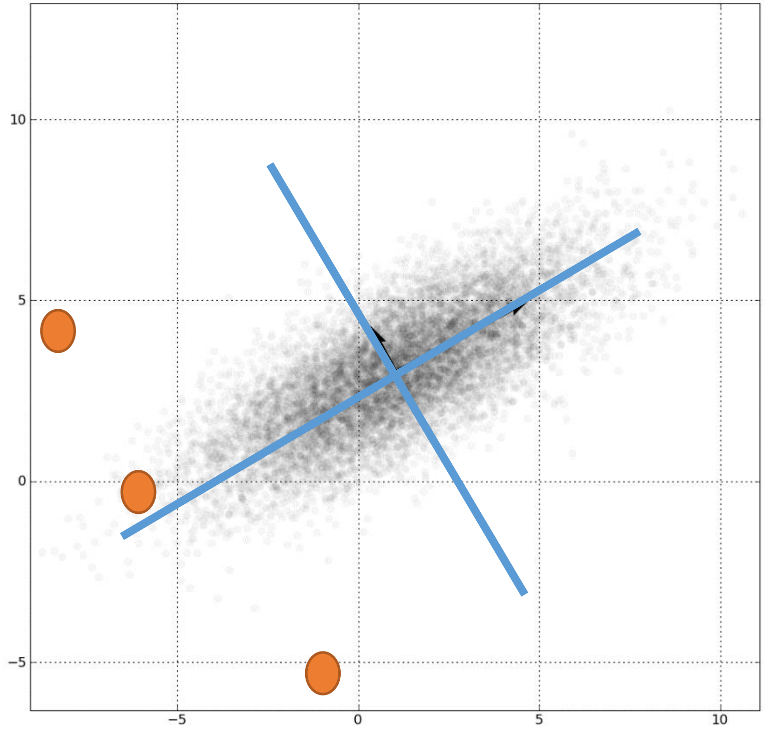
- $||X - X_r||_2 = \sigma_{r+1}$
    - Frobenius norm guarantees



PCA with Corruption?

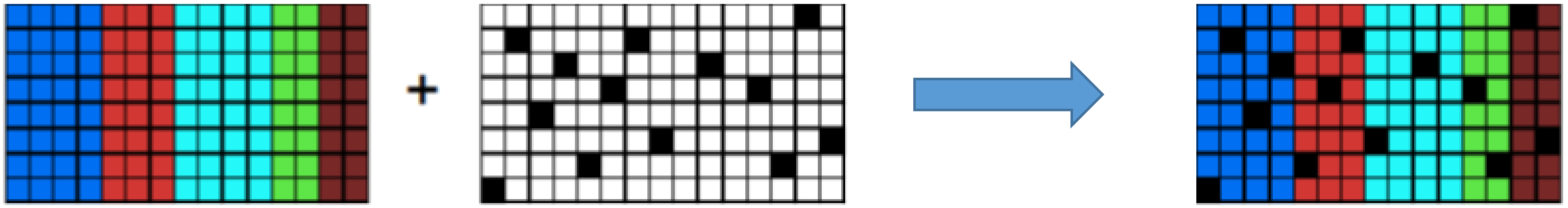
- $X = [x_1 \;\; x_2 \ldots x_n] + E$

- $||X - P_r(X + E)||_2 \leq \sigma_{r+1} + 2||E||_2$



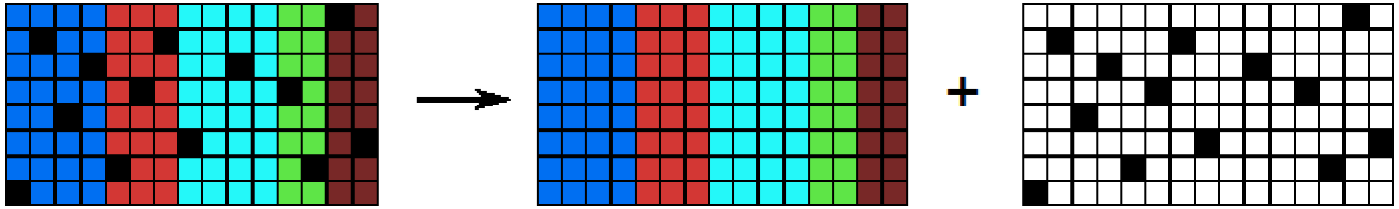
Sparse Corruptions?

- Can we do better?
  - If E is sparse?



- E.g.
  - Each point can be corrupted in a few random co-ordinates

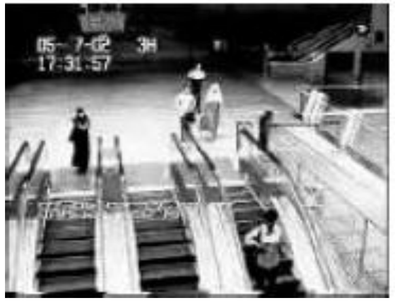
Robust PCA



$$M = L + S$$

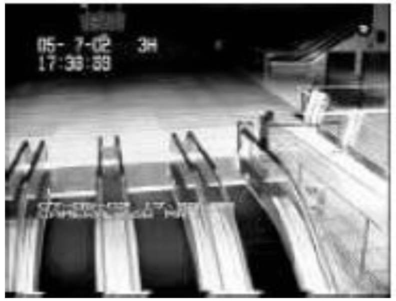
- M: given matrix
  - L: low-rank matrix
  - S: sparse matrix

Foreground + Background Separation



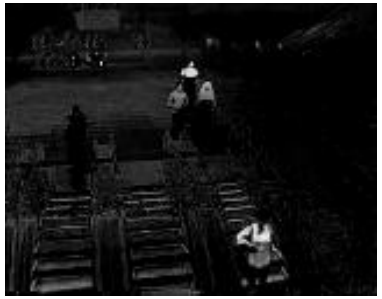
Original Video

=

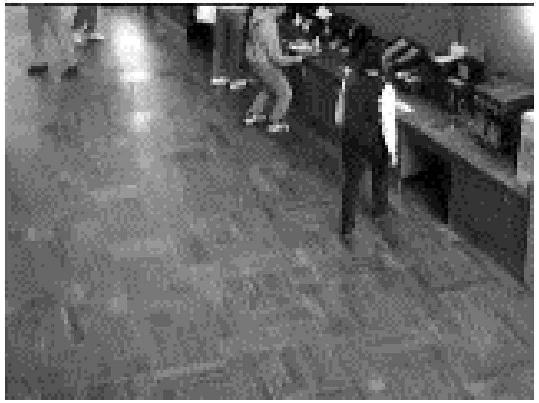


Background

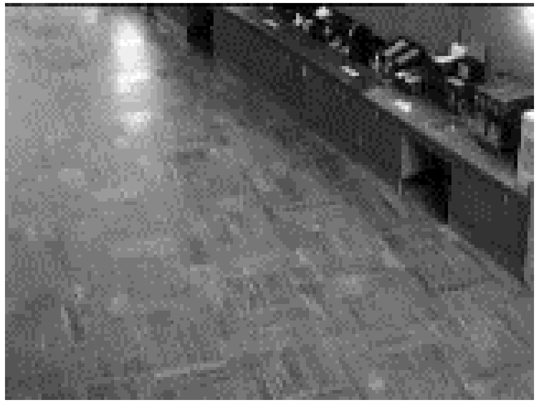
+



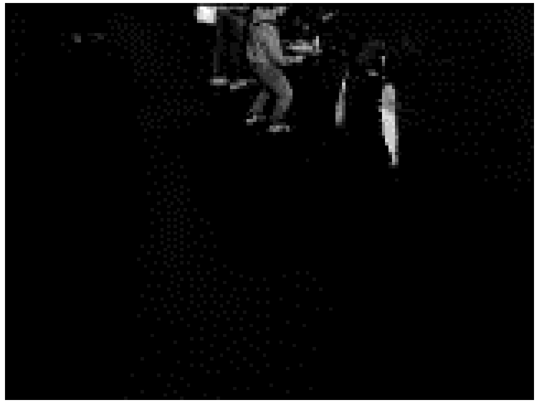
Foreground



=

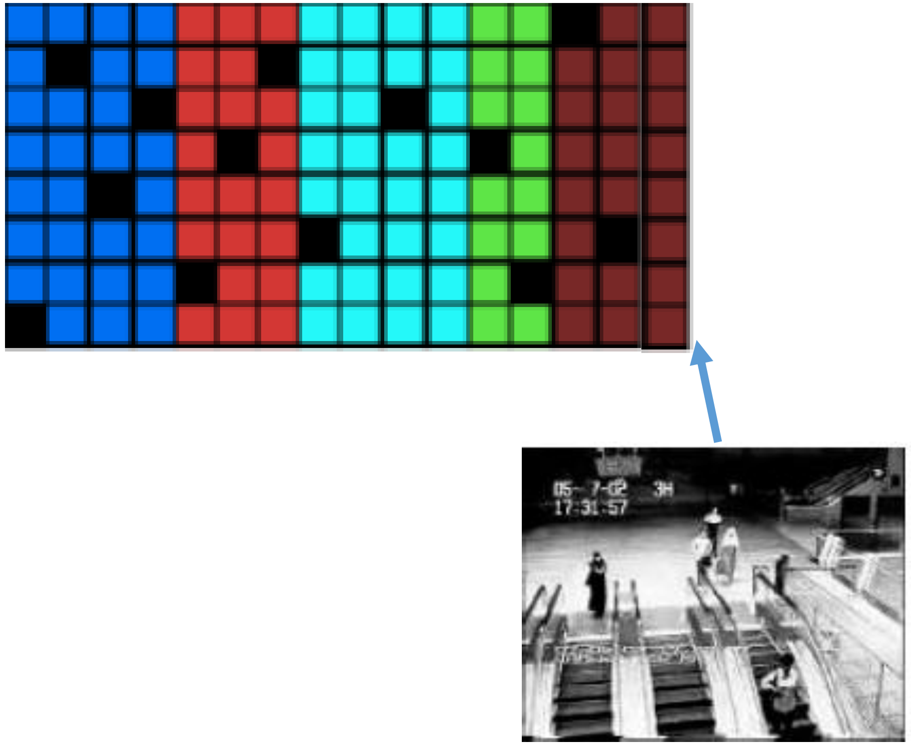


+

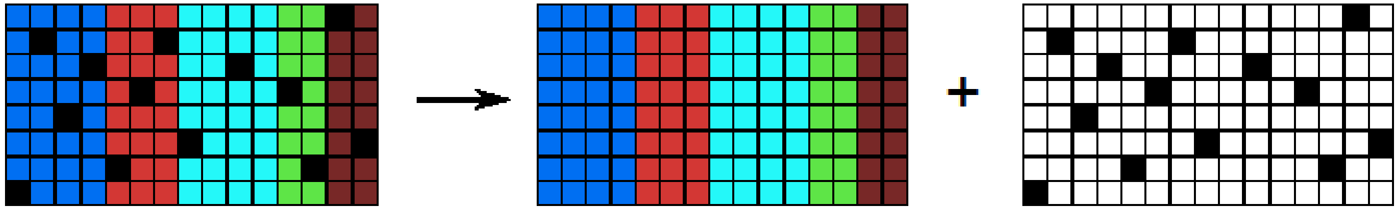


Foreground + Background Separation

- Each 64×64 frame: 4096-dimensional vector



Robust PCA



$$M \quad = \quad L \quad + \quad S$$

- $M \in R^{n \times n}$: given matrix
  - L: low-rank matrix
  - S: sparse matrix
- NP-hard problem in general

Identifiability?

$$\begin{bmatrix}1&0&0\\0&0&1\\0&0&0\end{bmatrix} = \begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix} + \begin{bmatrix}0&0&0\\0&0&1\\0&0&0\end{bmatrix}$$

$$M = L + S$$

- Assumptions:
  - L is incoherent--- $L_{ij} \leq \mu||L||_F/n$
  - S is row and column sparse

Existing Method

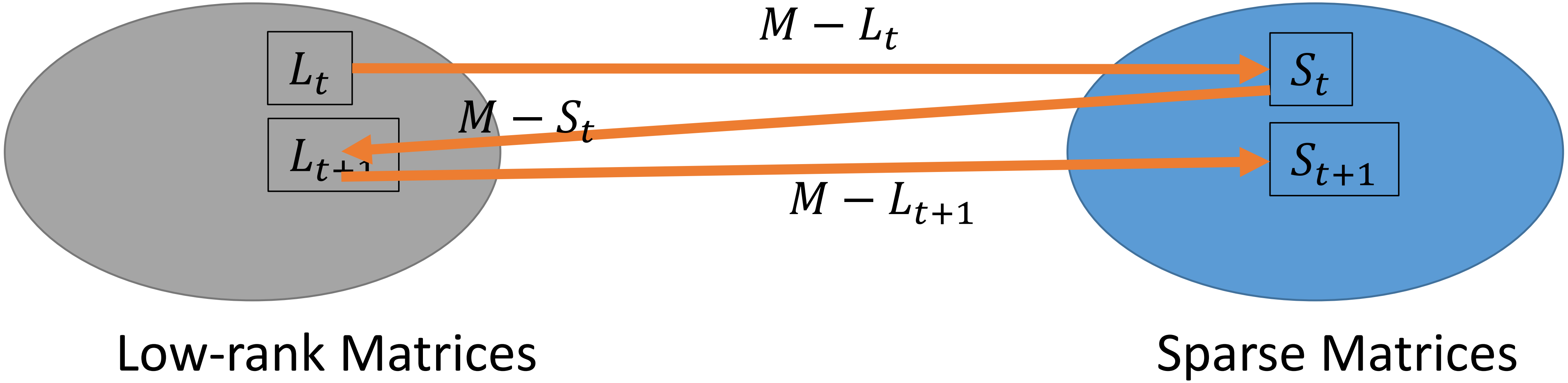
$$\min_{\hat{L},\hat{S}} rank(\hat{L}) + \lambda||\hat{S}||_0$$
$$s.t.\;\; M = \hat{L} + \hat{S}$$

- $||\hat{L}||_* = \sum_i \sigma_i(\hat{L})$
- Convex program,
- Assumptio
    - M
    - $s \leq \frac{n}{\mu^2 r}$
- Recover L, S [Cha asekharan et al’2009, Candes et al’2009]

Question: PCA time complexity for Robust PCA?
That is, $O(n^2 r)$ algorithm?

Our Approach: Alternating Projections

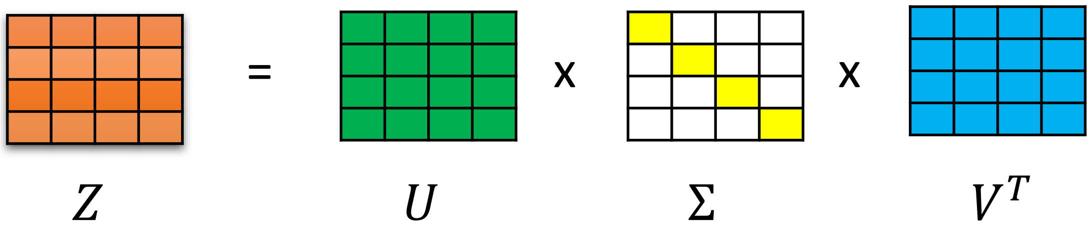
- Goal: $M = L + S$
  - L: low-rank matrix
  - S: sparse matrix
- M=$L_t + S_t$



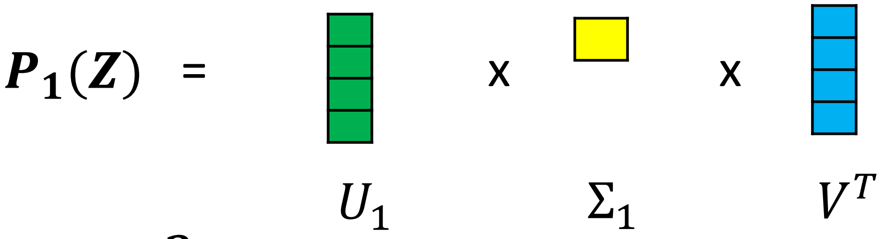
Projection onto Low-rank Matrices

- Non-convex projections: NP-hard in general
- But $P_r(Z)$ can be computed efficiently:

$$Z = U\Sigma V^T$$



- $P_r(Z) = U_r \Sigma_r V_r^T$



- Time complexity: $O(n^2 r)$

Projection onto Sparse Matrices

- Non-convex projection
- $HT_{\zeta}(Z)$: removes all elements with magnitude smaller than ζ

1	0.1	0.22
0.1	0.01	.9
0.11	0.02	0.12

$$\xrightarrow{HT_{0.5}}$$

1	0	0
0	0	.9
0	0	0

Non-convex RPCA

- $L_0 \to 0$
- $\zeta = \mu^2 r/n$
- For t=1, 2, … T
    - $\zeta = \frac{1}{4} \cdot \zeta$
    - $S_t = HT_\zeta(M - L_t)$
    - $L_{t+1} = P_r(M - S_t)$
- Output, L_T, S_T

Computation Time

- Each round: 1 SVD + 1 Hard Thresholding

- Time complexity per round: $O(n^2r)$

- No. of rounds?

Results

- After t-th step:

$$||L - L_{t+1}||_\infty \leq \frac{1}{2}||L - L_t||_\infty$$

- $T = \log\left(\frac{||L||_\infty}{\epsilon}\right)$, $\quad ||L_T - L||_\infty \leq \epsilon$
- Computation complexity: $O(n^2 r \log\frac{1}{\epsilon})$
  - $O(\log\frac{1}{\epsilon})$ more expensive than PCA
- Assumption: $M = L + S$
  - $s \leq \frac{n}{\mu^2 r} \cdot \boxed{\frac{\sigma_r^2}{\sigma_1^2}}$
  - Worse requirement than Hsu et al'2011

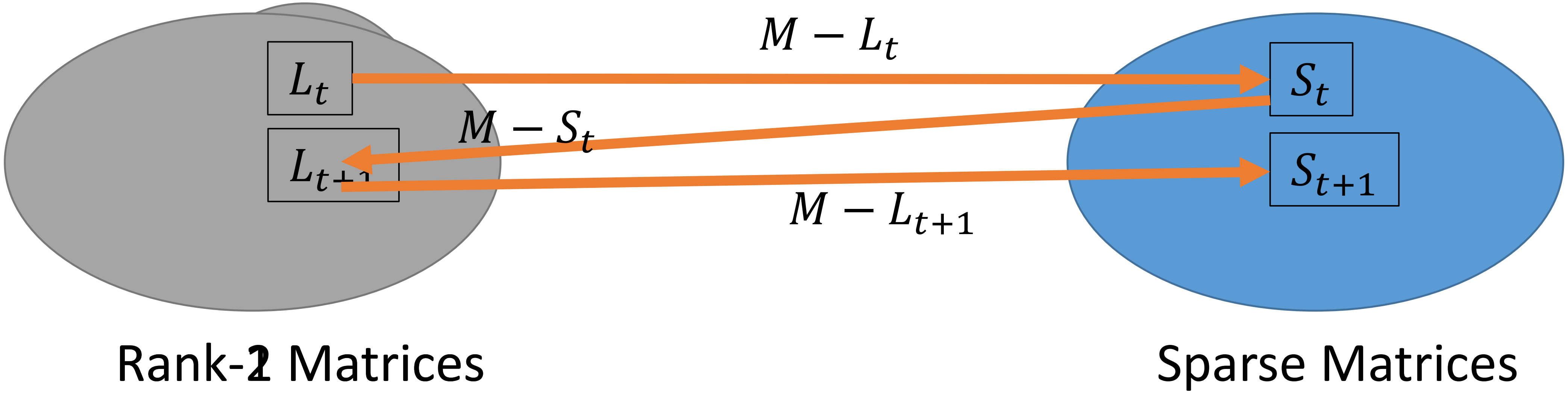
[NUSAJ'14]

Remove Condition No. Dependence?

- Stagewise procedure
  - k-th stage projects onto rank-k matrices
  - $1 \leq k \leq r$

2nd Stage

1st Stage



Result

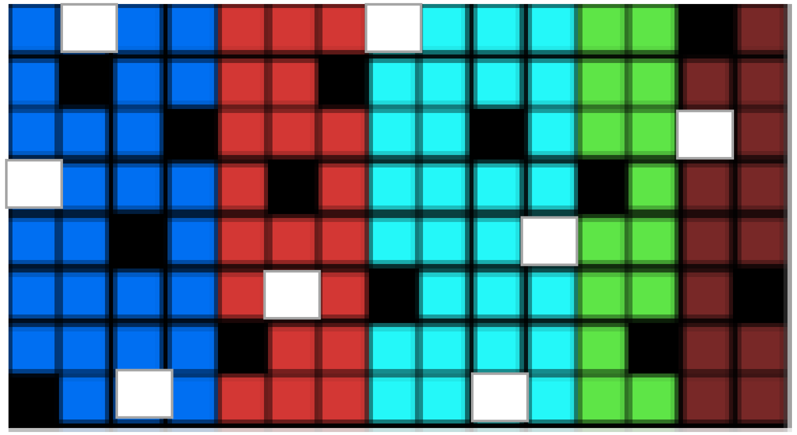
- $T = \log(\frac{1}{\epsilon})$

$$||L_T - L||_2 \leq \epsilon$$

- Assumption: $s \leq \frac{n}{\mu^2 r}$
  - s: number of corrupted entries in any row or column
  - Same as convex relaxation approach (Hsu et al'2011)
- Running time: $O(n^2 r^2 \log\frac{1}{\epsilon})$

[NUSAJ'14]

Missing Entries?



- Assuming missing entries are corrupted entries
- Allows for $O(\frac{n^2}{r})$ missing entries

Proof Technique

- $L_{t+1} = P_r(M - S_t) = P_r(L + S - S_t) = P_r(L + E_t)$

- Standard SVD guarantees:
  - $||L_{t+1} - L||_2 \leq ||E_t||_2 \sim O(1)$
  - $supp(S_{t+1}) \neq supp(S)$
  - Hence, $E_{t+1} = S - S_{t+1}$ can be dense

- Goal: ensure
  - $\mathrm{supp}(S_{t+1}) \subseteq supp(S)$
  - $||S - S_{t+1}||_\infty \leq .5\ ||S - S_t||_\infty$

- But for this, we need $||L_{t+1} - L||_\infty \leq .5\ ||E_t||_\infty$

A Novel Perturbation Lemma

$$||P_r(L + E_t) - L||_\infty \leq .5\ ||E_t||_\infty$$

- If:
  - E_t: sparse
  - L: incoherent

- Much tighter than the standard matrix perturbation results
  - $||P_r(L + E_t) - L||_2 \leq 2||E_t||_2$

Proof Sketch (Rank-1 case)

- $L = uu^T$
- $L_{t+1} = P_1(L + E_t), \quad L_{t+1} = vv^T$

$$(L + E_t)v = v$$

$$(I - E_t)v = Lv$$

$$v = (I - E_t)^{-1}\, Lv = Lv + \sum_{a=1}^{\infty} (E_t)^a\, Lv$$

$$\begin{aligned} L - L_{t+1} &= L - vv^T \\ &= L - Lvv^T L - \sum_{a\geq 0, b\geq 0, a+b\geq 1}^{\infty} (E_t)^a\, Lvv^T L^T (E_t)^b \end{aligned}$$

Proof Sketch

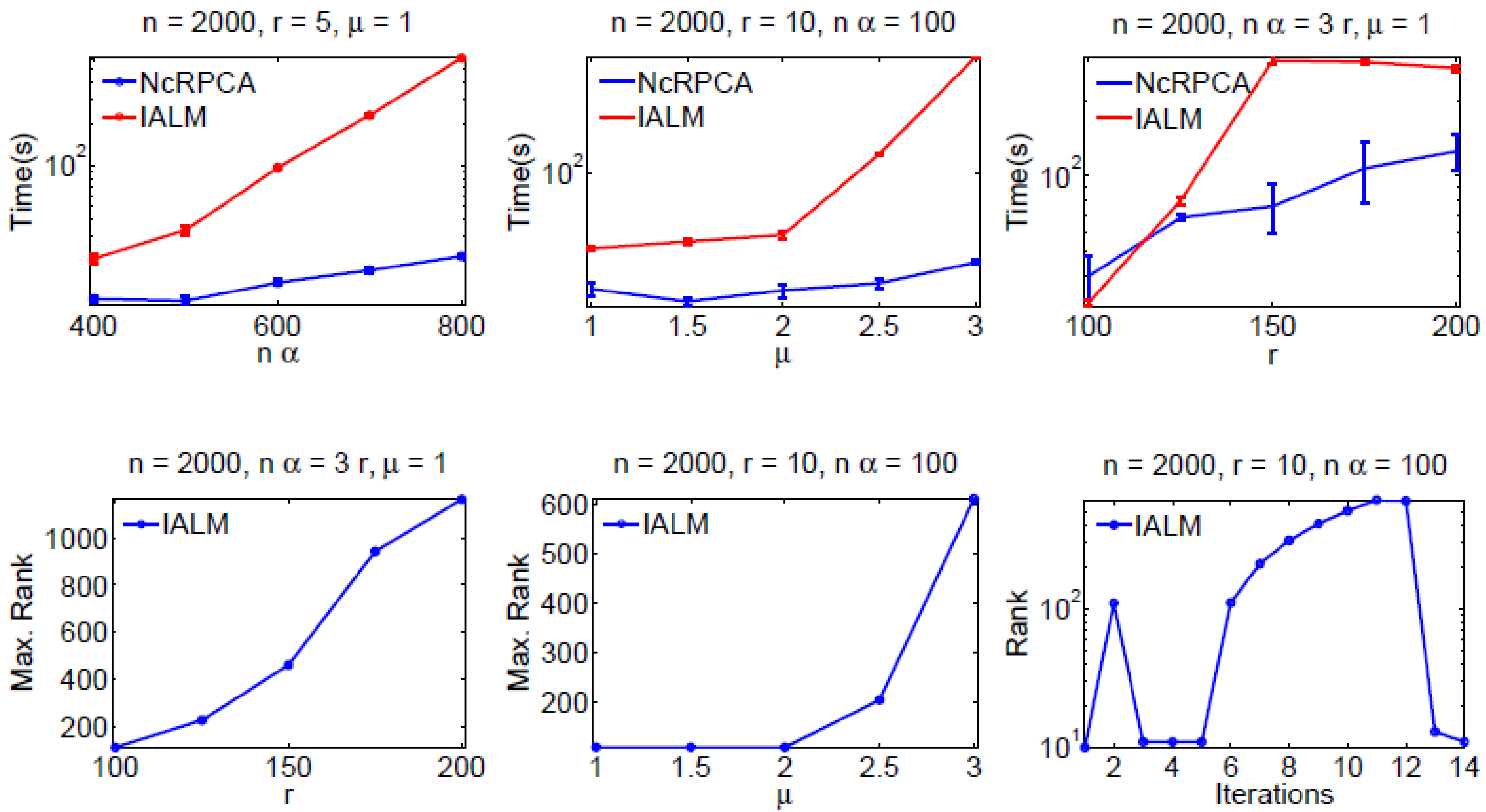
- Using $L = uu^T$

$$L - L_t = (1 - \langle u, v\rangle^2)L + \langle u, v\rangle^2 \sum_{a\geq 0, b\geq 0, a+b\geq 1}^{\infty} (E_t)^a \, uu^T (E_t)^b$$

- $||(E_t)^a u||_\infty$: small
  - E_t: sparse
  - u: incoherent ($||u||_\infty \leq \mu/\sqrt{n}$)

- Bound $|| \cdot ||_\infty$ of each term

Empirical Results (Synthetic Datasets)



n = 2000, r = 5, μ = 1

n = 2000, r = 10, n α = 100

n = 2000, n α = 3 r, μ = 1

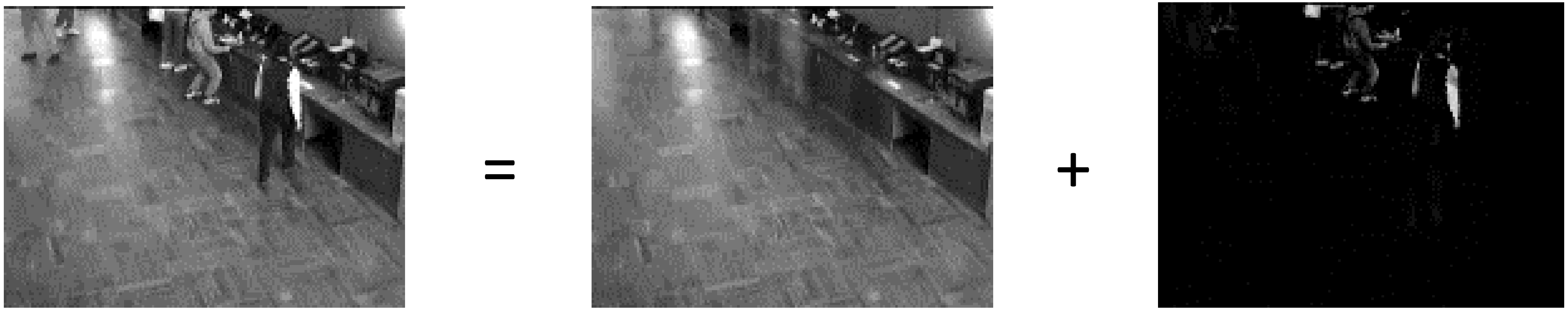
n = 2000, n α = 3 r, μ = 1

n = 2000, r = 10, n α = 100

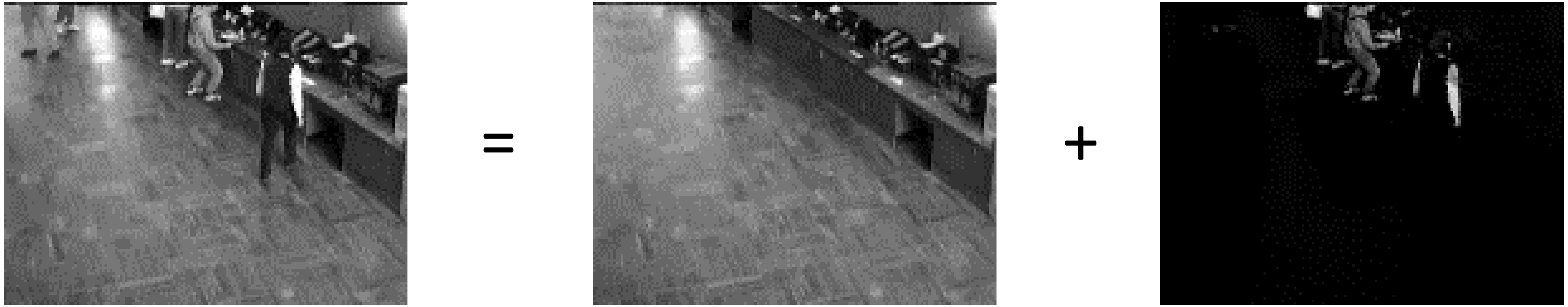
n = 2000, r = 10, n α = 100

Empirical Results

Convex Method. Runtime: 1700 sec



Non-Convex Method. Runtime: 70 sec



Summary

- Robust PCA
  - Low-rank+Sparse Decomposition
- Alternating Projection Method
- Under standard assumptions
  - Linear rate of convergence
  - Computation time: Recovery in O(PCA), for constant rank matrices
- Key analysis tool: a strong perturbation bound for SVD

Future Work

- RIP/RSC based Matrix sensing:
  - Necessity of the required RIP/RSC conditions?
- Matrix completion:
  - Remove dependence of $|\Omega|$ on error ϵ
  - Optimal dependence of $|\Omega|$ on r
- Robust PCA:
  - Extension to [Candes et al'09] style conditions
  - Can handle $O(\frac{n}{\mu^2})$ corruptions per row (currently, $O(\frac{n}{\mu^2 r})$)
- Develop a more generic framework to jointly analyze these problems
  - Similar to unified M-estimator technique of [Negahban et al'09]

Thanks!